Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising

Randall A. Lewis
Yahoo! Research
ralewis@yahoo-inc.com

Justin M.Rao
Yahoo! Research
jmrao@yahoo-inc.com

David H. Reiley
Yahoo! Research
reiley@yahoo-inc.com

ABSTRACT

Measuring the causal effects of online advertising (adfx) on user behavior is important to the health of the WWW publishing industry. In this paper, using three controlled experiments, we show that observational data frequently lead to incorrect estimates of adfx. The reason, which we label "activity bias," comes from the surprising amount of time-based correlation between the myriad activities that users undertake online. In Experiment 1, users who are exposed to an ad on a given day are much more likely to engage in brand-relevant search queries as compared to their recent history for reasons that had nothing do with the advertisement. In Experiment 2, we show that activity bias occurs for page views across diverse websites. In Experiment 3, we track account sign-ups at a competitor's (of the advertiser) website and find that many more people sign-up on the day they saw an advertisement than on other days, but that the true "competitive effect" was minimal. In all three experiments, exposure to a campaign signals doing "more of everything" in given period of time, making it difficult to find a suitable "matched control" using prior behavior. In such cases, the "match" is fundamentally different from the exposed group, and we show how and why observational methods lead to a massive overestimate of adfx in such circumstances.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Economics*; J.1 [**Computer Applications**]: Administrative Data Processing—*Business*

General Terms

Economics

Keywords

advertising effectiveness, field experiments, browsing behavior, causal inference, selection bias

1. INTRODUCTION

The largest publishers on the WWW monetize their businesses through the sale of sponsored search and display advertising. The health of the industry is dependent upon integrating advertisements into the user experience in a way that does not detract too much from user enjoyment but also provides the advertiser with a positive return on investment. Measuring the effectiveness of advertising (adfx) is a crucial step in this process; it allows for better design of user interfaces, provides essential feedback to advertisers about which strategies/creatives work and which do not, and allows a publisher to accurately gauge the value of various segments of its inventory. Despite these economically important reasons to "get it right," the industry has not settled on accepted standards for measuring adfx. The lack of standards has led to the use of a host of techniques, some of which incorporate large positive biases. In this paper we evaluate observational methods [14, 15, 5] by comparing them to the gold standard of controlled experiments [10, 13], which can distinguish causal effects from mere correlation. We identify an empirical regularity, "activity bias," that is a source of overestimation in observational estimates of the effects of online advertising.

Because clicks on advertisements are easily measured, online advertising automatically provides much more information on advertising effectiveness than most media can. However, advertisers would prefer to go beyond the click to measure causal effects on user behavior beyond the page on which the ad appears. For example, display advertising for a resort destination may stimulate users to visit travel websites, perform search queries for hotels or make online purchases of airline tickets. We will call these events of interest "dependent measures." Studies of these dependent measures are compelling in the online environment, especially compared with adfx studies in traditional media, because of the richness of individual-level data on ad exposure and outcomes. Yet, in our research, we have discovered that it is quite easy to overestimate such effects using observational data.

There are several different strategies used with observational data.[1] One is to compare users exposed to an advertising campaign to users not exposed to the same advertising campaign. Though frequently used, this method is rife with problems. In particular, the correlations observed between the dependent measure and advertising are often due to selection effects: the exposed and unexposed populations are different in outcomes for reasons having nothing to do with the advertising. A second technique attempts to correct for potential selection problems by attempting to

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

[1]For an example of a press release by a major adfx firm on observational methods, see [1]. Additional discussion of internet advertising effectiveness methods can be found in [8].

match exposed to unexposed users based on observed attributes of the user (regression with covariates, difference in differences, nearest-neighbor matching, propensity-score matching). The goal is to come as close as possible to the ideal of comparing two populations that are identical in all respects except for the advertising exposure. A third technique looks at users before and after ad exposure, and asks to what extent the dependent measure increases for these users after exposure. A weakness of the third technique is that other changes in product demand might be correlated with the timing of the campaign (news stories about the product, TV advertising, holiday increases in purchases), and thus the before-after changes might include changes unrelated to the advertising.

By contrast with observational methods, the experimental approach actually achieves the ideal of comparing apples to apples, so that observed differences can be attributed as true causal effects of the advertising campaign. An experiment takes a set of users and randomly assigns each either to the treatment group (eligible for exposure) or the control group (deliberately barred from exposure), just as in a drug trial using a placebo for the control group. By deliberately withholding ads from a randomly chosen control group of customers, we know that the control group has the same browsing behaviors as the treatment group who were exposed to the ad.

In an observational study, ad exposure is determined by user browsing activity: did this user satisfy the targeting requirements of the ad (which can include behavioral targeting), and did the user browse pages on which the ad was being served during the relevant time period? For example, a user who performs a search for "car insurance" on Google will see a search ad for Geico, while a user who does not perform that search will not see the Geico ad. Similarly, a user who recently searched for airfares on Orbitz, but did not purchase one, could be shown with a "retargeting" campaign featuring similar airfares on Orbitz. Finally, a user who browses many pages on the Yahoo! website today is much more likely to see a Honda display advertisement than a user who browses few or no pages on Yahoo! All of these represent examples of endogenous exposure to advertising: the user's behavior determines which ads they see.

This endogenous exposure can lead to overestimates of the effects of advertising. Suppose an observational study of adfx for the Orbitz campaign compares those who saw the ad with those who did not see the ad. Then the exposed group will contain only users who have recently searched for airfares, while the unexposed group will contain a number of users who have not. If the former are more likely to purchase, even in the absence of advertising, then the observational study will overestimate adfx, mistaking correlation for causation. Similarly, suppose an observational study attempts to estimate the increases in searches for the keyword "Honda" due to the Honda display advertising campaign on Yahoo!, using data from Yahoo! Search. Those actively browsing on Yahoo! during the time period of the campaign are both more likely to see the display ad and more likely to perform a search. Browsing behavior causes both advertising and searches, and this spurious correlation will lead to an overestimate of the true causal adfx.

Both of the hypothetical observational studies above are a bit of a straw man. They lack sophistication, in the sense that they fail to attempt to match the samples using observable characteristics. A more sophisticated study of the Orbitz campaign would restrict attention only to those users who were qualified to see the ad, but didn't happen to browse Yahoo that day. Similarly, a more sophisticated study of the Honda campaign would use the previous month's searches as a control variable, to control for heterogeneity in search behavior having nothing to do with the campaign. Both of these matching strategies are designed to eliminate the obvious sources of spurious correlation. What we show in this paper, using experiments to determine the "ground truth," is that even these matching strategies cannot fully eliminate the estimation bias: the overestimation remains economically substantial.

In the estimation of causal effects in the social sciences, good observational studies rely on several types of key identifying assumptions. One such assumption is the following: if people A and B looked exactly the same yesterday, and only A gets exposed to a treatment today, then B is a good control for what would have happened to the exposed individual. This assumption has intuitive appeal as something that might be approximately true, which explains the popularity of the technique. Because the matching technique can only be applied to observable characteristics, the key assumption is that no unobservable characteristics are correlated both with the treatment (advertising) and with the outcome of interest (purchases or searches). Unfortunately, the validity of this assumption cannot be tested in the absence of the experiment, and such absence is precisely the reason for making the assumption in the first place.

Another well-accepted identifying assumption is that a person's behavior yesterday is a good predictor of what she would likely do today. If an intervention (such as exposure to advertising) coincides with a change in behavior, we could then conclude that the advertising was the cause of that behavior. In the context of online advertising, we shall see below that this assumption surprisingly fails to hold in key examples.

Why do the above innocuous-sounding assumptions fail to hold in the advertising setting? First, browsing activity drives exposure to an advertising campaign, so users who saw a given campaign were more active on the publisher's site during that time period. Second, users' browsing behavior shows large variance over time: people browse very different numbers of pages from one day to the next. Third, users' browsing behavior across various websites appears to be positively correlated, at least in our examples. That is, someone who browses a website more than usual on a given day is also likely to be to be browsing other websites more than usual as well. Our key finding is that these three features combine to create what we call "activity bias," a tendency to overestimate the causal effects of advertising using online behavioral data.

Let's consider an example in more detail. When online purchases are the outcome of interest, then matching on past online purchases seems like a sensible way to clean up an observational study. For example, a difference-in-differences estimator would compare the before-after difference in exposed users' sign-up rate at Geico to the before-after difference in unexposed users' sign-up rate. If users were heterogeneous in their browsing and purchasing behavior, but users' browsing types remained relatively constant over time, this matching method would work well to estimate causal effects. However, we find that this is not the case. We demonstrate,

in three different settings, that users do not browse the web in a consistent manner over time; "lumpy" usage is quite common, and this creates non-causal correlation between ad viewing and purchases (or other online activities of interest). This is the source of activity bias.

Moreover, browsing is correlated across diverse Internet properties. As such, exposed users are more likely than the matched group to exhibit a host of browsing behaviors, some of which are the dependent measure of interest. That is, there is a surprising amount of positive correlation between the myriad activities that users undertake online. Because the exposed group does more of everything online during the relevant time period, estimates of adfx are likely to overstate the truth by a significant margin.

In the three experiments presented in this paper, we show that the assumption of common trends in usage between exposed and unexposed users that underlie observational methods fail to hold for a variety of online dependent measures (brand-relevant keyword searches, page views, account sign-ups) and that ignoring this failed assumption leads to massive overestimates of advertising causal effects.

In our first experiment, we measured the increase in keyword searches caused by a display-advertising campaign for a major American firm on the Yahoo! Front Page (yahoo.com). Our treatment-control comparison gave us an estimated increase of 5.4% in the number of users performing searches on a set of keywords related to that brand. We then compared this experimental estimate to the estimate we would have obtained, in the absence of a control group, using observational techniques. Depending on which set of covariates (various measures of user behavior in the week prior to the campaign) we used to match exposed to unexposed users, our estimates of search lift ranged from 871% to 1198%, all of which differ wildly from the truth. The reason is that individuals who actively visit Yahoo! on a given day are much more likely both to see the display ad and to do a Yahoo! search, by comparison with those who do not actively visit. Thus, in the observational data, the ad exposure and the search behavior are highly positively correlated, but not because one causes the other. Rather, both are caused by the level of Yahoo! browsing behavior. Note that even controlling for past Yahoo! search activity does not come close to providing a correct causal estimate, because it turns out that searching and browsing behavior varies quite a bit from one day to the next, though both are correlated with each other. This is the source of what we call "activity bias." In the absence of a randomized control group, the exposed users are guaranteed to be more active than the unexposed users, and this sample selection yields spurious results, even after controlling for past levels of activity.

One might easily imagine this first experiment to be unrepresentative. The results could depend heavily on the fact that both the stimulus (display advertising on Yahoo!) and the outcome measure (searches on relevant keywords on Yahoo!) relied on behaviors taking place in the same neighborhood of the Web, namely Yahoo! One might therefore hope that when measuring adfx across different websites, the correlation might be weak or zero, and therefore observational measures of adfx might be fairly accurate. Surprisingly, however, in our second and third experiments, we find evidence to the contrary.

In Experiment 2, we recruited users through Amazon Mechanical Turk (AMT) and exposed half to a 30-second video advertisement promoting Yahoo.com services, and half to a political video advertisement, which served as a control. We see roughly three times more activity, as compared to the week prior, on Yahoo.com for the treatment group on the day of exposure. Absent a control group, we would be tempted to conclude that the advertisement had a large, positive effect. However, the pattern of activity in the control group turns out to be almost identical, so that the experimental estimates of the adfx are very small and statistically insignificant. Looking at the treatment group before and after exposure, without the use of an experiment, would lead to overstating the effects of the ad by 350%. The reason is that being active on Amazon Mechanical Turk on a given day increased the chance of participating in the experiment, but this Amazon activity also correlated with activity on Yahoo!, independent of the ad exposure. Thus, we find activity bias even when measuring user outcomes on a very different location from the site where the ads were delivered, once again producing large overestimates of adfx relative to the gold standard of a controlled experiment.

The second experiment was not a fluke: in our third experiment, we find something very similar. As economists, we were very excited to discover an opportunity to measure the competitive effects of an ad campaign: an example where we could measure how one firm's advertising affects a rival firm's sales. We conducted an experiment with a major American online-services firm running a campaign of about 200M impressions on Yahoo!, and though we were not able to track conversions for that advertiser, we were lucky enough to have a conversion beacon installed on the new-account-sign-up page of one of its rivals. Exposed users were much more likely to sign up at the competitor's website on the day they saw the advertisement as compared to the week prior to exposure. This observational evidence might lead one to conclude that the ad had large, positive spillovers on the sales of a close competitor. However, once again the measured adfx turn out to be overestimated. The experiment shows that users in the control group exhibited a nearly identical lift in competitor sign-ups on the day they came to the Yahoo! site but did not see ads for this particular advertiser. Thus, the causal effect was very close to zero. Activity bias was what caused the treatment group to exhibit increased sign-ups on the day of ad exposure. Observational methods would significantly overstate the positive spillovers of advertising in this case.

The three experiments highlight the severe difficulty with causal inference using observational data in online environments. The experiments focused on three widely-used dependent measures in the study of adfx: searches, page views, and account sign-ups. In all three cases, data and plausible identifying assumptions were available to eliminate bias in the estimates of causal effects. Yet observational methods still led to massive overestimates of the true effects of the ad campaigns because a diversity of online activities are highly correlated with each other, and highly variable over time. Without the controlled experiments we conducted, we never would have been able to know how large the bias might be; instead, we would have had to rely on untestable assumptions. Given these demonstrated difficulties, we strongly advocate the use of randomized experiments to obtain valid measurements of causal effects.

Other work on estimating adfx in the online world has found that randomized experiments are more reliable than observational methods. Lewis and Reiley [10] perform a randomized experiment to examine the effect on sales of an online campaign for a major American department store. The paper finds that the ad had a significantly positive effect on sales. Interestingly, most of the online ad campaign's effect comes through offline sales. The authors show that industry-standard observational methods would lead to underestimates, rather than overestimates, of adfx. Although this is the opposite bias we find for the online dependent measures in this paper, the explanation is similar. It turns out that online browsing activity and offline shopping activity were *negatively* correlated in that setting (highly active browsers shopped less in stores), so the activity bias was negative rather than positive. Without the use of a control group, the authors would not have been able to uncover this important driver of bias and would have incorrectly concluded that the campaign had a negative effect. This paper is similar in tone to the work of Lalonde who showed that observational methods had great difficulty estimating the true effect, estimated through multiple randomized experiments, of labor training programs [7].

Other researchers argue in favor of observational studies of adfx and argue that although experimental studies minimize bias in estimation, their cost in terms of sacrificing reach or paying for charity ads can be high.[4] The authors propose using the observational technique of doubly robust propensity score matching [3] as a general tool when experiments are cost prohibitive. In Section 5, we argue that the cost of experiments can be reduced or eliminated with a basic modification to ad-serving infrastructure. More importantly, our results show that there are features of the data that could not have been uncovered in the absence of a controlled experiment and these features can strongly violate the identification assumptions required by propensity score estimation. We address this point in more detail at the end of Section 3.

2. EXPERIMENT 1: EFFECTS ON SEARCHES

In this experiment the goal is to estimate the causal impact of a display-advertising campaign on searches for the advertiser's brand name and related keywords. This outcome measure is particularly appealing for advertisers who are interested in generating offline transactions that cannot be matched easily to online ad exposure (as is typically the case for most advertising in most media). We worked with a major American advertiser to estimate the "buzz" generated by their advertising campaign by conducting a randomized experiment. In this section we first give the estimated impact of the ad based on comparison of treatment and control. We then show how the commonly used observational-data strategy of comparing exposed users to a "matched" pseudo-control group significantly overstates the ad's impact.

2.1 Design

The experiment was conducted for the large rectangular ad unit on the Yahoo! Front Page (www.yahoo.com). The campaign ran for a single day and was delivered as an "exclusive," meaning that it was shown on every US-originated visit to the Front Page that day. To create a control group, we randomly chose five percent of browser cookies to be excluded from this campaign. Control-group users instead saw a public-service announcement (PSA) unrelated to the advertisement used in the treatment group. Control-group users saw the PSA on each and every visit to the Front Page on the day in question, just as treatment-group users saw the advertiser's ad.

2.2 Results

Table 1 gives the overview and key results of Experiment 1. The row in italics gives the impact on searches during the day of the campaign. The campaign led to a 5.4% increase in propensity to search for relevant keywords, which is significant at the 0.05 level. For the remainder of this section, 5.4% serves as the comparison "ground truth" for the observational estimators.

Table 1: Experiment 1 Overview and Results

Group	Treatment	Control
Ad Creative	**Advertiser**	**PSA**
Location	**Y! FP**	**Y! FP**
Media	**LREC**	**LREC**
Impressions	218,509,781	11,515,109
Viewers	35,300,548	1,857,748
Avg. Impressions	6.20	6.20
Std. Dev. of Avg.	8.0	8.1
Clicks	63,014	5,278
Clickers	56,869	4,712
Searchers	55,170	2,755
CTR	0.029%	0.046%
Clicker Rate	0.161%	0.254%
Search Rate	*0.156%*	*0.148%*

Now suppose that we were in the usual adfx measurement environment, in the absence of an experiment. We will ignore the control group from which we withheld ads and instead compute results using observational methods on the remainder of users. For a sample of endogenously unexposed users, we used a machine-learning algorithm designed to find Yahoo! users as similar as possible to the users exposed to the ad. This algorithm produced a set of 15 million users to compare to the 35 million exposed users.[2]

The first observational technique one might try is to compare the number of searchers for this advertiser's brand-relevant terms between exposed and unexposed users. This frequently used technique [2] results in a poor measurement of the true lift: 1198% more searchers per person in the exposed group than in then unexposed group, compared with the ground truth of 5.4% from the experiment. An obvious problem is that users who are exposed to the Yahoo! Front-

[2]The model put a large amount of weight on eligibility to see the ad. This meant that all of the unexposed users in our sample ended up being people who viewed the Yahoo! Front Page on the day of the campaign, but did not view any ads from the campaign. As noted above, control-group members were excluded from the analysis. At first, it puzzled us that there could be any users who saw the Yahoo! Front Page but did not see the exclusive campaign designed to be shown to every user on that page. We subsequently realized that our matched sample proved to consisted of international users who visited the www.yahoo.com, who did not see the ad because it was targeted only to American viewers.

Table 2: Estimated Causal Effects Using Observational Methods

Model	(0)	(1)	(2)	(3)
Estimated search lift	1198%	894%	871%	872%
Day dummies	No	Yes	Yes	Yes
Session dummies	No	No	Yes	Yes
Page views	No	No	No	Yes
Minutes spent	No	No	No	Yes

Page campaign are relatively more active on the US Yahoo! pages, which also means that they are relatively more likely to be doing searches on Yahoo! on the day of the campaign, independent of the content of the display-advertising campaign they viewed.

For a more sophisticated attempt to account for the differences between exposed and unexposed users in the observational data, we employ regression analysis with "control variables" included. Since we are concerned about differences in search behavior between the type of user who manages to see the Y! Front-Page ad on a given day and the type of user who does not, our control variables describe search activity by each user in the recent past. For a seven-day period ending in the week before the campaign took place, these variables are are the number of days on which the user performed searches on Yahoo!, the number of Yahoo! Search sessions, the number of Yahoo! Search page views, and the number of minutes spent viewing pages in Yahoo! Search. Because the number of days and the number of sessions have a limited number of possible values, we express them as a full set of categorical (or "dummy") variables, while the number of page views and the number of minutes enter the model linearly. If these control variables fully account for the differences between the two groups, the regression approach should yield similar estimates to the experimental results.

We estimate variations on the following baseline model:

$$1\left(S_i > 0\right) = \alpha + \gamma * A_i + \beta X_i + \epsilon \qquad (1)$$

Here, S_i is the number of brand-relevant searches performed by individual i, and $1\left(\cdot\right)$ is the indicator function. A_i is a dummy variable equal to 1 if the individual was exposed to the ad campaign, and 0 if not. X_i is a vector of control variables, γ gives the exposure effect and X_i is a vector of control variables. The control variables we use are dummy variables for number of search days in the sample, number of sessions and total page views. γ represents the effect of interest, the increase in probability that an individual performed a relevant search as a result of the ad exposure. Each regression is estimated on data from 50 million users (35 million exposed plus 15 million unexposed).

Table 2 gives the estimated search lift for regression models with increasing numbers of control variables that might be used to control for observables in an observational study. Note that a special case of the model is the one with no control variables, which is equivalent to the simple exposed-unexposed difference of 1198% we computed earlier.

All of these models, whose 95% confidence intervals bound their estimates by roughly ±10%, drastically overstate the impact of the campaign as compared to the truth of 5.4%. Including more and more control variables does reduce the bias slightly, but not enough to get anywhere near the truth.

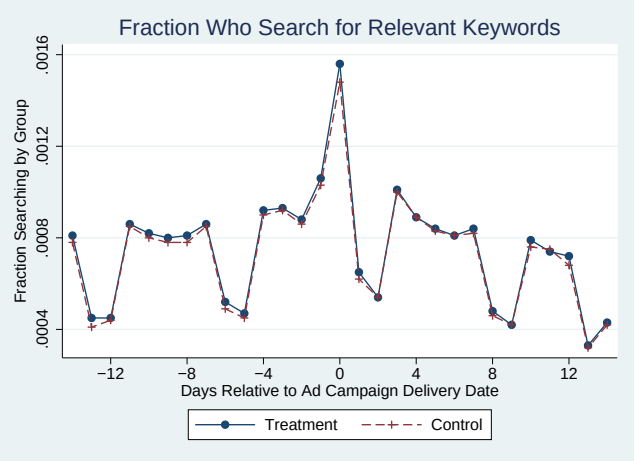

Figure 1: Brand keyword search patterns over time.

Note that without the use of an experimental control group, we would never know how far from the truth our observational estimates really were. Clearly, there are unobserved factors that affected both ad exposure and search behavior by users on the day of the campaign. These unobserved factors cannot be controlled for, even when using the observable data we have, and generate a significant positive bias in the adfx estimate.

Yet another observational technique we could use to estimate adfx is to look at the exposed users' level of searches for the advertiser's brand-related keywords over time. If the fraction of users performing relevant searches increases significantly on the day of the campaign, we might attribute that increase to the display advertising. Figure 1 displays the daily time series of the percentage of exposed users who searched for brand-relevant keywords, from two weeks before the day of the campaign to two weeks after the day of the campaign.

Compared with the rest of the month's worth of data, the day of the campaign shows approximately double the number of searcher on the advertiser's brand keywords. From this, we could easily conclude a 100% lift in searcher due to the advertising campaign. This is much better than the estimates of more than 800% obtained from an exposed-unexposed comparison, but it is still nearly twenty times the ground truth of 5.4% obtained in the experiment. A look at the data for the control group shows that the control group exhibits exactly the same spike in behavior, despite viewing a PSA instead of the advertiser's Front-Page ad! How could this be?

The reason is activity bias. On the day of the campaign, users who visited the Yahoo! Front Page (and thus saw either a treatment ad or a control ad) were much more likely to be active across the Yahoo! network than they were on other days. This means that they did more searches on Yahoo!, not just on this advertiser's brand keywords, but on all keywords.

An amusing observation comes from examining the day before the campaign ran. Overall, we see a very clear weekly pattern in the data, with more users searching on weekdays than on weekends. But on the day before the Friday campaign, we see a 20% greater level of search on these key-

words than on the other Thursdays of the weeks before and after. The search behavior on that Thursday appears to be somehow predicting the future, anticipating the campaign on Friday! In truth, what this shows us is a small amount of positive autocorrelation in the online behavioral data between adjacent days. People who view the Yahoo! home page on that Friday are twice as likely to search on Yahoo! on Friday as on the average day and 20% more likely to search on Yahoo! on that Thursday as well.

To summarize the results of this experiment, two kinds of observational techniques produce wildly incorrect estimates of adfx. Comparing exposed versus unexposed users produces overestimates of search lift on the order of 200 times the correct value as obtained in an experiment. These overestimates occur even when we control for a rich set of observable characteristics of the users, such as their past intensity of search behavior. A user's past search behavior turns out not to be a very good predictor of their search behavior on a given day, as we see in our second observational technique. For users exposed to the ad, comparing search behavior before versus during the campaign produces overestimates nearly 20 times the truth. In both cases, the culprit is activity bias: individuals' online activity varies quite a bit over time but is correlated across different activities, for reasons unobservable to the researcher. This correlation can easily mislead a researcher to infer large positive effects of advertising under plausible-sounding assumptions, but with an experiment we can see that the reasonable-sounding assumptions are incorrect and the observed correlations wildly misstate the true causal effects.

3. EXPERIMENT 2: EFFECTS ON CONSUMPTION OF YAHOO! PAGES

Experiment 2 offers additional evidence of activity bias. In this case, we see that browsing behavior is not just correlated within a single website (Yahoo!), but across diverse websites as well. An implication is that when ads and outcomes are measured in very different parts of the Web, activity bias can still produce overestimates of adfx.

3.1 Design

We conducted an experiment on a third-party site, Amazon Mechanical Turk (AMT), to measure the impact of a video creative promoting Yahoo! products. The goal of the advertisement was to increase visits to the Yahoo! Front Page and other Yahoo! properties, such as Mail and Sports. So in this case, the dependent measure was usage of Yahoo! properties, and subjects were exposed based on their activity on a third-party site.

We used AMT to recruit subjects for the study. Half ($n = 806$) of the $N = 1600$ subjects were shown a 30-second video ad promoting Yahoo! (treatment group), the remainder ($n = 794$) were shown a political campaign advertisement of similar length (control group). Subjects then completed a 3-minute survey about attitudes towards Yahoo! and the political candidate. Subjects were paid $0.40 for participation.[3] Using a third-party site to recruit subjects

[3]Typically online surveys to measure adfx use volunteers solicited through pop-up ads or offer a chance at a gift certificate for completion. Hulu.com "pays" customers in the form of running 250 ads for a charity of their choice. Based on click-through rates and the cost of pop-up ads, we found it cheaper to pay subjects directly using an amount that was quite generous by Mechanical Turk standards.

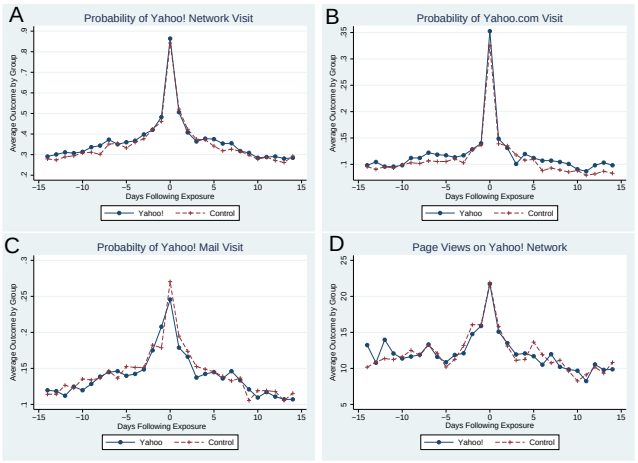

Figure 2: The effect on various Yahoo! usage metric of exposure to treatment/control ads. Panels A-C: Probability of at least 1 visit to the Yahoo! network, Yahoo.com and Mail respectively. Panel D: Total page views on the Yahoo! network.

to determine adfx through surveys is a common strategy used by measurement firms such as Dynamic Logic. The thought is that one will get a more representative sample and will avoid "preaching to the converted." A survey often measures the "effects" of a campaign that did not have a control group, so survey respondents are split into exposed/unexposed groups using the observational methodology we described in our discussion of Experiment 1. Our study is relatively unique not only because we use a controlled experiment, but also because we are able to go beyond typical survey measurements to measure effects on actual consumption of the advertised product (Yahoo! pages).

3.2 Results

Figure 2 gives the key results. Examining the group who saw the ad promoting Yahoo!, we see a large spike in activity on the day of exposure and, to a lesser extent, the days following across all metrics of Yahoo! usage. Imagine, for a moment, that we did not possess the control data, that the graphs only presented the solid lines, which correspond to the treatment group. Panels A-C show that users exposed to the Yahoo! video creative were approximately 3 times more likely to visit the Yahoo! network, Front Page and Mail on the day of ad exposure.[4] Panel D shows an increase of 10 network page views on the day of exposure as compared to users' recent history; some of this increase persists up to 3 days out. (One observation that might make us suspicious is that Panel D also shows evidence of a slight "build up effect," usage increasing prior to exposure, similar to what we saw for the exposed group in Experiment 1.) Without the aid of a control one might be tempted to conclude, based on this

[4]The same patterns are true for other Yahoo! properties, such as Sports and Finance, but the results are noisier due to lower overall usage. These additional graphs are available from the authors.

evidence, that the ad had a large but relatively short-lived effect.

Now let's bring the control data back into the picture. User behavior in the control group exhibits *the exact same patterns*. The plots are strikingly similar; it was not exposure to the Yahoo! video advertisement that drove an increase in Yahoo! usage, it was pure activity bias. The "conclusion" we derived from examining the treatment data alone was quite misleading indeed. A treatment-versus-control comparison convincingly demonstrates that the ad did not have a substantive impact on user behavior — this particular ad was completely ineffective. Regression analysis available from the authors confirms that the ocular assessment that the treatment and control plots do not differ in a statistically significant way.

From the two lines plotted together, it is clear there is something fundamentally different about browsing behavior on the day of exposure as compared to the recent past. This is important for statistical inference because the recent past is often used to match users. But if I am fundamentally different from my recent past self, then its very likely I will be different from a user matched to me based on past activity. This fact makes observational inference very difficult in this environment. Commonly employed methods to match based on observables such as propensity score [14], nearest neighbor, difference-in-difference [12] and simple regression with controls require the assumption that the time trend of the exposed and unexposed (pseudocontrol) groups are the same. Specifically one has to assume that one group does not experience an idiosyncratic shock to the dependent measure. Activity bias is simply an idiosyncratic shock that is correlated with exposure. Put more forcefully, *selecting comparison groups based on exposure to an ad automatically generates different time trends in online behaviors, independent of the actual content of the ad campaign.*

4. EXPERIMENT 3: COMPETITIVE EFFECTS OF ADVERTISING

In Experiment 3 we studied the advertising impact of a major American firm's campaign of approximately 200M impressions run on Yahoo![5] Firms are generally interested in understanding the effects of their advertising on competing firms in the industry. Does their campaign reduce (negative externality) or increase (positive externality) sign-ups for competitor firms? Positive externalities arise when the ad attracts people to the activity in question in a general way and some of that attracted activity goes to competing firms. For instance, a highway billboard for a restaurant at an upcoming exit might increase sales for all the restaurants on that exit. Negative externalities arise when potential customers of the competitor are swayed towards the advertiser. An example of this could be a ad in supermarket for a particular brand of beer. Perhaps it does not convince non-beer drinkers to buy beer, but increases sales for the advertising brand by reducing sales for other brands. The competitive impact of advertising is an important factor for the firm to consider when purchasing media. It can help determine the rate of return on investment and in the design of creatives.

[5]For confidentiality reasons, we cannot publish the name of the firm. We reveal only that this is a service industry, for which customers may apply online to open new accounts.

4.1 Design

We tracked new-account sign-ups at a major competitor's website using a beacon installed by the competitor on their sign-up page. Competitor sign-ups constitute the dependent measure. We used a 10% browser-cookie-based hold-out for the campaign in order to create two control groups. The control groups were not eligble for delivery of the advertiser's creative. The first control (5%) saw a Yahoo! branded news-unit instead. The news-unit provided real-time information on stock prices and trading conditions. The second control saw ads that normally run in that spot, but not from this advertiser's campaign. This means that the ad-server fired off the next ad in the queue for these users. The competitor firm we are studying was not running an ad in that location during that time period. We do note that a weakness of the second control is that we do not know exactly what ads were served in absence of the treatment creative, so in particular we cannot identify exactly which users would have been served ads for Firm A had they been in the treatment group. However, we show that behavior in the news-unit and no-ad control did not end up differing from each other.

4.2 Results

In Figure 3 we plot the probability of seeing a campaign ad for the 7 days prior to signing up, day of and 7 days after signing up for the three groups. For the treatment group, the graph shows a large spike in visits to the page where ads were shown on the day when those users signed up. Elevated levels of visitations are also evident for a few days prior and following the spike. Simply put, *people were much more likely to sign up on a day they were exposed to the ad.* This appears to be strong evidence that ad imparted a significant, positive externality on the competing firm. And based on this evidence alone, one might be tempted to reach this conclusion.

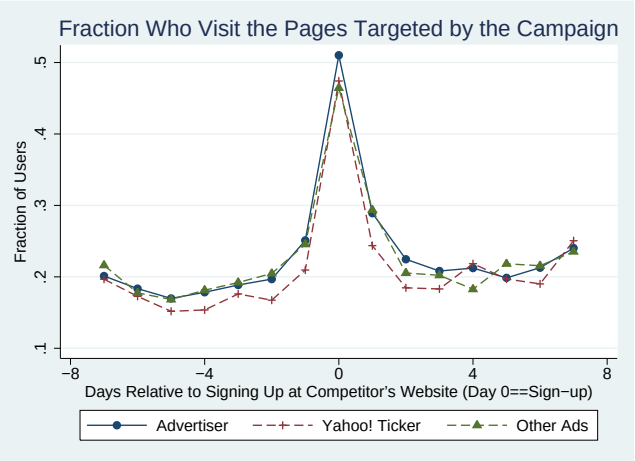

Figure 3: Fraction of users who saw the page targeted by the campaign relative to their sign-up at the competitor's website.

However, in Figure 3 we also see a very similar trend in the control group that were delivered ads (news-unit). People were much more likely to sign up on a day they were exposed to *an irrelevant ad* as well. Almost none of the spike

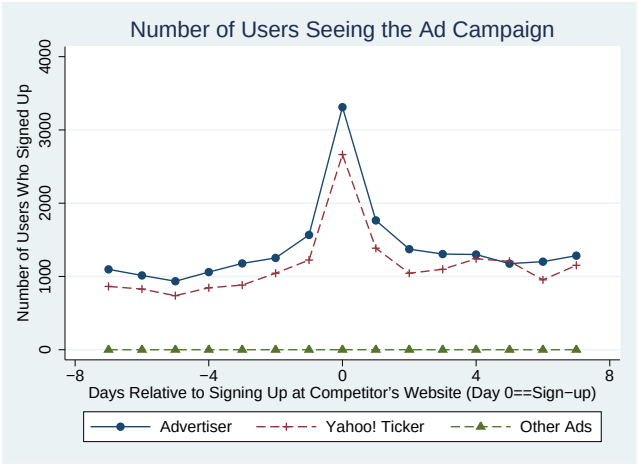

Figure 4: Number of users who saw the page targeted by the campaign relative to their sign-up at the competitor's website.

Table 3: Experiment 3 Treatment Effects

	Simple Avg	Efficient Avg
Treatment effect	369 (3.5%)	441 (4.3%)
Std. Error (Poisson)	293	288
t-statistic	1.26	1.53

in sign-ups had to do with the causal impact of the competing firm's ad. The spike is a mechanical artifact of activity bias. The fact that people were much more likely to sign up on a day of exposure simply does not mean that ad had a causal effect on behavior. This initially may seem counterintuitive, but becomes clear once we look at controlled experiments. Statistical assumptions can be self-sustaining when past precedents justify future research methods. In this case, experimental evidence torpedoes these relatively unquestioned assumptions.

We confirm this point in Figure 4, which plots the raw number of users exposed to the campaign relative to their sign-up date. We see the same patterns in Figure 4 as in Figure 3. Without the use of a control, we might be tempted to conclude from Figure 4 that ad led to an increase of 2000 sign-ups at the competitor's website. Table 3 establishes the true treatment effect. Overall, there was a slight lift in competitor sign-ups as a result of the campaign, although the result is not statistically significant. The treatment vs. control comparison shows an increase of between 369 and 441 sign-ups, far lower than the 2000 we might have deduced from Figure 4. Prior experimental work has found that display advertising in this industry has a significantly positive effect on an advertiser's own new account sign-ups [11]. The experimental results presented here indicate that the cross-effects of advertising in this industry are potentially smaller than the own-effects.

As an additional check of the key finding of our paper, we examine the "effect" of sign-ups at the third-party site on Yahoo! usage. In this case, there is no plausible reason why sign-ups would *cause* Yahoo! usage (the competitor was not a web publisher). The two behaviors might tend to occur together due to correlation alone. Figure 5 shows that this is indeed the case. Sign-ups are predictive of an increase in Yahoo! usage by about 100%, or 3 page views per user per day. Now of course it would be a serious mistake to conclude from these results that Yahoo! should subsidize advertisements for this firm. In this case the lack of causality is obvious. Both ad views and online sign-ups are actually caused by a third variable: spending time online, which varies from one day to the next, but is clearly correlated across websites.

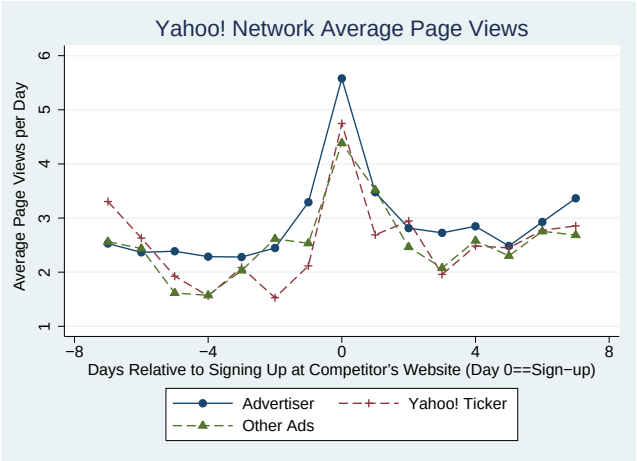

Figure 5: Sign-ups at third party site "cause" Yahoo! network page views.

Once again, standard observational methods overstate the true causal effect of the advertisement. In this case, the company might have falsely concluded that advertising in this venue provided too large a benefit to a major competitor, and might have revamped the media plan to account for this "finding."

5. DISCUSSION

To show that activity bias is not unique to online advertising, we consider a hypothetical example from the offline world. Suppose a supermarket wanted to estimate the effect of a highway billboard. The firm knows that it simply cannot compare drivers who saw the billboard to drivers who did not because there may be differences between the two populations (the drivers may come from neighborhoods of different income levels, for example). To combat this problem, the firm uses a database match to get a list of drivers who use the highway for regular commuting purposes. The billboard runs for a week and after the firm surveys drivers to determine exposure (0-5 days) and product usage. They find that exposed drivers were much more likely to shop at the supermarket in that week and the week following as compared to matched unexposed drivers and drivers who only saw the ad once. Activity bias says that exposed drivers were probably more active shopping and participating in offline commerce in general as compared to their unexposed counterparts. In this example, given that the unexposed group had been identified as commuters on that route but showed up rarely, it is likely they were either sick or out-of-town. Now the ad might have had a large positive effect, but there is no way to reach a clear conclusion based on the

data in hand. In fact, if such data were readily available, the company might also find that the billboard "caused" more speeding tickets; hopefully this would sound an alarm bell.

Our hypothetical example has been constructed to make very obvious how mistaken firms and researchers can be when they infer causality from observational studies. In the online-advertising world, the inferential error is less obvious, but no less important, and our experiments drive this point home. In the billboard example, it is not obvious how the company would run a controlled experiment. One option would be to use a billboard that displays two ads, one on even minutes, one on odd. Then in the survey, only use respondents who drove all week. This method is statistically valid, because conditional on driving all week, exposure is determined via a random draw (for an example using this technique for online display advertising, see [9]). We admit, however, that this could be prohibitively costly.

Some authors have asserted that the cost of experiments in the online world can also be prohibitive [4]. But as ad-serving technology improves, the cost of experiments declines. Currently, a careful experiment requires the advertiser (or the publisher) to sacrifice part of their advertising budget to fund impressions for a control campaign, perhaps a public-service announcement. One way around this problem is for publishers to pair campaigns of unrelated advertisers. For small campaigns, this method is satisfactory, but for large campaigns the sacrifice in reach might be unacceptable to the advertiser. A second option is to improve ad-serving technology to reduce cost. In this method, a small hold-out group is created, say 5% of maximum reach, and the ad server flags deliberate non-deliveries to this control group (i.e., when the ad would have been shown but is not and a replacement ad is shown instead). This reduces cost by eliminating the need to purchase PSAs for the control group, because those users will merely see the next ad opportunity in the stack.

Experimental studies of adfx represent a major advantage for online advertising versus offline media. Unlike print, television, radio, billboards, and event sponsorship, online ads can deliberately be "un-delivered" to a control group, and subsequent user behavior can be easily measured. We are optimistic that the industry will move towards experimental methods. In this paper we show another reason why they are necessary: the insidious effects of activity bias. Indeed, without the use of controlled experiments it would have been difficult to identify activity bias in the first place. There may well be additional biases lurking in observational studies, yet to be uncovered.

6. CONCLUSION

In this paper we presented three experiments on the causal effects of advertising. Accurately measuring the effect of advertising is essential to the online publishing industry because it can inform what placements and positions work well within the user interface, find which advertisers and brands get the most value from a given piece of inventory, and accurately report return-on-investment to advertisers. Online advertising has a distinct measurement advantage over other forms of marketing, due to the availability of individual-level data on exposures and outcomes. However, having lots of data is not sufficient for generating accurate estimates of adfx. Another key advantage of online advertising is the ability to control individual exposure to an ad campaign in a controlled experiment.

While observational studies attempt to control for observable characteristics of customers, experiments (through randomization) are the only way to control for their unobservable characteristics. The problem, as we show in this paper, is that unobservable characteristics of a consumer (such as whether they are likely to be using the Internet during a given time period) are highly correlated with both advertising exposure and with online outcome measures of interest to advertisers. Somewhat surprisingly, we find that positive correlation of activity even across very different websites, such as Amazon and Yahoo! This is the opposite of what we would observe if users had a fixed daily budget of online time, so that time on Amazon would substitute for time on Yahoo! Instead, in our examples, we see that on some days, a user does more of everything online, and other days, she does less of everything online.

This overall phenomenon, which we call activity bias, leads to severe overestimates of adfx when using observational or quasi-experimental methods. In Experiment 1, exposure to an ad campaign on Yahoo! was associated with a huge increase in brand-relevant keyword searches, but the actual causal effect was quite minimal. In Experiment 2, activity on a third-party site on a given day was highly predictive of using Yahoo! properties, but not relative to an experimental control group. In Experiment 3, exposure to an ad campaign on Yahoo! correlated with account sign-ups for a competitor to the advertiser, but our experiment revealed that the true causal effect was approximately zero. We believe that these three applications are quite representative of effects that we might expect to occur in studies of adfx; we did not cherry-pick these examples from a large set of experiments.

In all three cases, activity bias causes observational methods to overstate causal effects, sometimes by orders of magnitude. If a user is active enough to see a given ad or sign-up at a given site, she is likely more active than usual and will thus be doing more of everything here, there and everywhere. For online dependent measures, such as searches, page views and account sign-ups, observational methods will typically suffer from activity bias and overstate causal effects.

7. REFERENCES

[1] comScore Bids to Shake Up Online Ad Assessment. In *Market Research Institute Online, http://www.mrweb.com/drno/news11750.htm*, 2010.

[2] M. Abraham. The off-line impact of online ads. *Harvard Business Review*, 86(4):28, 2008.

[3] H. Bang and J. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[4] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–16. ACM, 2010.

[5] G. Fulgoni and M. Morn. How online advertising works: Whither the click. *Comscore.com Whitepaper*, 2008.

[6] J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

[7] R. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

[8] P. J. Lavrakas. An Evaluation of Methods Used to Assess the Effectiveness of Advertising on the Internet. In *Interactive Advertising Bureau Research Papers*, 2010.

[9] R. Lewis. *Where's the "Wear-Out?": Online Display Ads and the Impact of Frequency*. PhD thesis, MIT Dept of Economics, 2010.

[10] R. Lewis and D. Reiley. Does retail advertising work: Measuring the effects of advertising on sales via a controlled experiment on Yahoo! In *Working paper*, 2010.

[11] R. Lewis and T. Schreiner. *Can Online Display Advertising Attract New Customers?* PhD thesis, MIT Dept of Economics, 2010.

[12] K. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13, 1986.

[13] D. Reiley, S. Li, and R. Lewis. Northern exposure: A field experiment measuring externalities between search advertisements. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 297–304. ACM, 2010.

[14] P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41, 1983.

[15] D. Rubin and R. Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, 21(2):206–222, 2006.