Measuring Causal Impact of Online Actions Via Natural Experiments: Application to Display Advertising

Daniel N. Hill
Amazon.com, Inc.
Palo Alto, CA
daniehil@amazon.com

Robert Moakler
Integral Ad Science
& NYU Stern
New York, NY
rmoakler@stern.nyu.edu

Alan E. Hubbard
Division of Biostatistics
University of California
Berkeley, CA
hubbard@berkeley.edu

Vadim Tsemekhman
OpenDSP
Seattle, WA
vadimt@opendsp.com

Foster Provost
Integral Ad Science
& NYU Stern
New York, NY
fprovost@stern.nyu.edu

Kiril Tsemekhman*
Integral Ad Science
New York, NY
kiril@integralads.com

ABSTRACT

Predictive models are often employed to decide actions in interactive online systems. For example, ads are selectively served to users who are modeled as being inclined to purchase the product being advertised. News feed items are populated based on a model of the user's interests. A common consequence of these predictive models is the creation of a spurious correlation, or confounding, between the action and its desired outcome. In the above examples, the targeted users are likely to buy the product or find the news item regardless of the intervention. This presents a challenge for measuring the true impact of these systems.

Here we present a novel framework for estimating causal effects that relies on neither randomized experiments nor adjusting for the potentially explosive number of variables used in predictive models. We propose the identification and instrumentation of events that mediate the effect of the action. When the effect of an action depends on a mediating event that is not subject to the same confounders, the problem of causal estimation is greatly simplified. We demonstrate this approach in display advertising using ad viewability as a natural experiment that mediates the impact of served ads. Approximately 45% of display ad impressions never make it into a viewable portion of the user's browser. We show that an analysis based on ad viewability can massively reduce the amount of bias in estimating campaign lift. We integrate the use of negative controls as well as the identification and adjustment for residual confounding to further reduce the bias in estimated lift to less than 10%. A system using these techniques is deployed to monitor the daily causal impact of many large-scale advertising campaigns.

*Corresponding author.

KDD '15 August 11-14, 2015, Sydney, NSW, Australia
ACM 978-1-4503-3664-2/15/08.
DOI: http://dx.doi.org/10.1145/2783258.2788622.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Design, Experimentation, Measurement

Keywords

display advertising, causal analysis, natural experiments, performance measurement, negative controls

1. INTRODUCTION

In today's complex interactive online systems, economically important actions increasingly are taken based on inferences drawn from predictive models. Examples run from classic applications such as automatic credit decisions to contemporary applications such as what content to show a user in a news feed. In this paper we use online display ad targeting as our running example and our domain of application. As has been described comprehensively elsewhere, online display ads are targeted based on various predictive modeling-based strategies, *e.g.*, [13].

When actions are important economically, firms generally would like to assess the impact of the actions. For example, advertisers would like to assess whether their advertisements actually lead to an increase in certain outcomes, such as purchases or other brand actions. Unfortunately, *the use of predictive models substantially increases the difficulty of assessing the true impact of the actions.* In this paper we explain this phenomenon carefully, present a framework for collecting and using data to address this difficulty directly, and provide a demonstration implemented across a variety of real, large-scale ad campaigns.

To understand why predictive modeling makes assessments of impact more difficult, we need to focus on the key difficulty of causal reasoning: confounding. Consider figure 1, which shows a causal directed acyclic graph (DAG) [12] representing a collection of random variables that describe an individual and the causal relationships among them. These

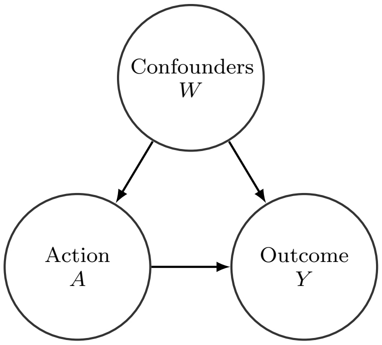


Figure 1: Causal DAG with confounders.

include an action (A), such as whether or not the individual was served an advertisement for a particular product, an outcome of interest (Y), such as whether or not the individual purchased the product, and a set of confounding variables (W). A variable or a set of variables is confounding if it is causally related to both the action and the outcome of interest. In the language of causal reasoning, one must control for confounding if one is to draw valid causal inferences.

We can now state precisely why using predictive modeling creates a problem. Predictive models very often are built to choose actions based on the user's prior inclination towards a particular outcome. For example, an online store may recommend products that the user is modeled as likely to purchase. News feeds may be populated with items that are predicted to be of interest to the user. The very features that predict the outcome for the individual are used to decide actions to further promote that outcome, thus confounding the link between action and outcome. So in our advertising example, predictive models are built based on, say, the web pages that users have visited and those users' purchase behavior. The model identifies the statistical link[1] from web page visits W to purchase behavior Y. Then these predictive models are applied such that users with certain W values are selected for ad targeting. This creates the causal link between W and A, and completes the confounding. The practical upshot is well understood by academics and knowledgeable data scientists, but poorly dealt with in the industry: the models select people who are more likely to purchase even without being shown an ad, a concept known as selection bias. Due to the confounding one cannot judge the effectiveness of the ad campaign easily—estimates that do not control for the confounding will be highly inaccurate [9, 8].

One possible way to address this confounding is by conducting randomized experiments, called A/B tests in the advertising industry. A/B testing can have serious technical challenges, which have been described in detail [6]. Possibly more importantly, A/B testing is expensive. In the case of online display advertising, typically 5-15% of targeted consumers are randomly assigned to a control group and shown campaign-irrelevant public service announcements (PSAs), while the remaining users are assigned to the test group and are shown normal campaign ads. Thus, advertisers or their targeting agents must spend 5-15% of marketing budgets on ads that have no direct effect on outcome, and not surprisingly advertisers are reluctant to spend such sums. Even in cases where they do, usually the A/B tests are conducted over a limited time period that restricts the scope of the assessment. Finally, in certain cases ethics precludes randomized trials. This concern is well known in ethically fraught applications such as medical treatment [1]; it is only beginning to come to attention in online experiments (*e.g.* [7]).

The other major alternative is to use causal analysis to directly adjust for confounders present in observational data. Several factors prevent this from being done efficiently or at all. First, for many modern applications, W can contain tens of millions or more variables, such as all the possible web pages that an ad targeter considers in modeling a user. Correctly controlling for such a huge number of non-independent confounding variables is difficult. Second, for automated actions there is a strong possibility of a violation of the positivity assumption necessary for causal inference [14]. If the action A is taken for every individual with a particular W, then there is no control group for W in which the action was not taken.

Most importantly, the organization interested in the assessment of impact often does not have precise knowledge of the data W on which the actions are taken. For example, it is the brand who cares about the impact of advertisements. A large brand typically will engage a dozen or more targeting firms. For most of those firms, the brand does not know exactly how they are targeting, or even exactly what data they are using. And the brand generally does not have access to those data, which may have been gathered by the targeter, purchased from third parties, or created internally based on inferences from other models. So while targeters themselves can conduct observational causal studies based on their own data [3, 15], the brand generally cannot do so either for individual targeters, or across targeters, *e.g.*, to judge which targeters are generating the best return on ad spend.

Here we offer an alternative, that can complement A/B tests or causal analysis, and that can also be applied in some situations where these techniques break down *or lack statistical significance*. Our framework, as outlined in section 2, is based on the identification and exploitation of an event M that mediates the impact of the action A on the outcome Y (figure 2). When such an event exists and is independent of the predictors W used to decide the action, it can be treated as a natural experiment. This removes or dramatically reduces the need to perform complicated adjustments for W in assessing causal effects. A major contribution of our framework is the proposal to instrument computer systems to observe such mediating events. The framework also integrates the use of negative controls to assess any residual confounding, and the use of methods from causal inference to adjust for residual confounding.

The second main contribution of this paper is the specific application of our framework to assessing the impact of online display advertisements. The key novelty is the instrumentation of the display advertising ecosystem to reveal a mediator M: viewability of ads [11]. More specifically,

[1] Technically, there are some hidden behavioral drivers that cause the individual to have both an increased propensity to visit the web pages and an increased propensity to purchase. We will ignore that distinction for now; it makes the causal graph slightly more complex, but does not affect the confounding by W.

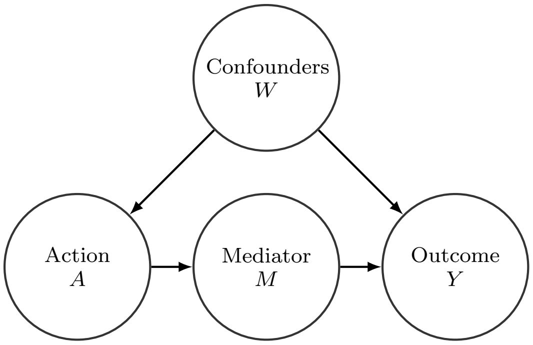


Figure 2: Causal DAG with mediating variable.

it turns out that for various reasons described below, after an ad targeting action A has been taken, online consumers are able to *view* only about 55% of display ads. The reasons for not seeing the ad are largely independent from the predictive modeling and from W, and thus confounding is dramatically reduced. Our results show that indeed the apparent impact of online display ads is dramatically reduced when utilizing viewability as a natural experiment. This concurs with prior results from randomized experiments [2] and from causal modeling with access to W [3, 15]. We then show, by employing negative controls, that the residual confounding is relatively small—though still present. We identify particular confounders that affect the viewability of ads, and further reduce the bias of our estimates through an adjustment. An implementation of our framework that estimates causal performance of display ad campaigns using viewability has been deployed as a product at the corresponding author's company. The system monitors lift for dozens of advertising campaigns on a daily basis. The selection of the right confounders to adjust the performance estimate is a work in progress and will be incorporated into future product releases.

2. MEASUREMENT FRAMEWORK

We now outline our framework for using mediating events to measure the causal impact of an action, as illustrated in figure 3. The workflow is briefly sketched here and then described in more detail in later sections as we work through our application in online advertising.

The process begins with the identification of potential mediators that correspond to the causal model in figure 2. A good candidate must meet several criteria. First is that the event must happen after the action is taken but before the outcome occurs. Second, the mediating event must be necessary for the action to have an effect on the outcome. Third, the event should not be directly influenced by the same predictors W that confound the system in the first place. Finally, the mediating event must fail to occur sometimes for all values of W in order to avoid positivity violations. The mediating event for our application is the viewability of display ads, but another example exists in online advertising that uses real-time bidding (RTB) systems. The effect of targeting a user for ads is mediated by winning an auction. If the auction is lost, the user cannot be influenced by targeting. There are candidates in other domains as well. The effect of adding a recommended song to a playlist is mediated by whether the user remains online long enough for the song to play. The effect of placing an item in a news feed is mediated by whether the user scrolls through all the preceding items.[2] In our framework, once a candidate event is proposed, the next step is to instrument the mediator to collect measurements of M.

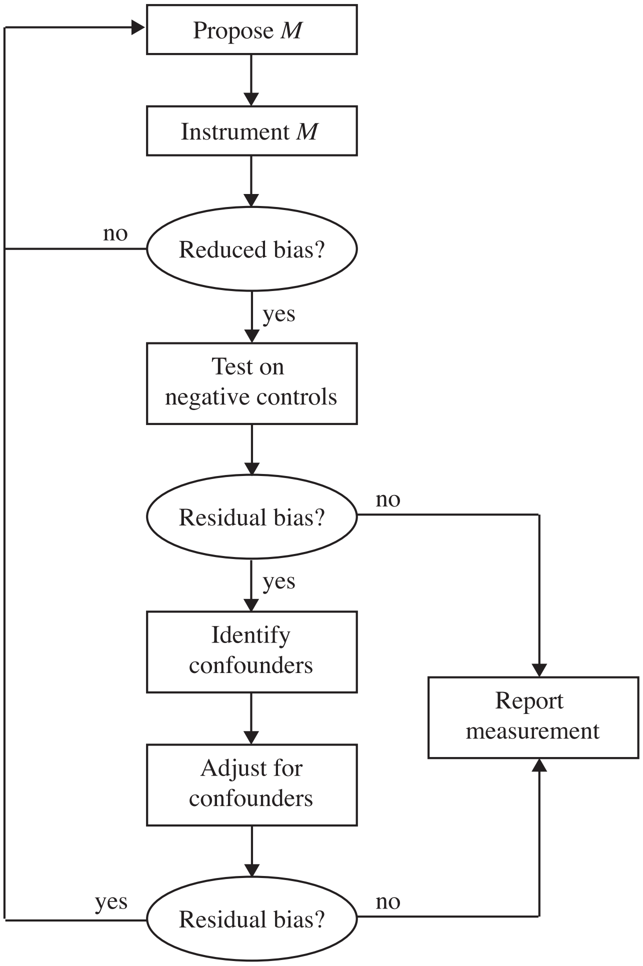


Figure 3: Framework for exploiting mediators.

An ideal or perfect natural experiment results when the treatment status for each user is assigned due to a truly random and naturally occurring event. If the event M results in an ideal natural experiment, then the estimation of a causal effect is straight-forward. For all users who experienced the action A, the probability of outcome Y can be compared directly between users who experienced the mediating event and users who did not, as if a randomized experiment had

[2] Jensen et al. [5] discuss an automated method of discovering natural experiments (what they term "quasi-experimental designs") in observational data. They show that natural experiments can be discovered using a relational database schema, additional information about the temporal durations of specific events, and limited prior knowledge about potential causes.

been conducted.[3] This analysis would be subject to none of the bias associated with a naïve analysis where users who experienced the confounded action A are compared to those who did not experience A. It may be quite helpful for a comparison of both types of analysis to be carried out as a first-pass evaluation of the usefulness of the measurement M as a natural experiment.

It may be the case that M is not a perfect natural experiment, meaning that there exist residual confounders W' between the mediator M and the outcome Y (figure 4). The confounders W' may or may not overlap with W depending on the domain. How can we identify whether there is residual confounding of the mediator? In principle, this could be identified by comparing results from the analysis to that of a randomized experiment. However, it may be that reliable experiments are not available if this framework is being used. If available, a low-cost alternative is to employ *negative controls* as a way to identify confounding of M.

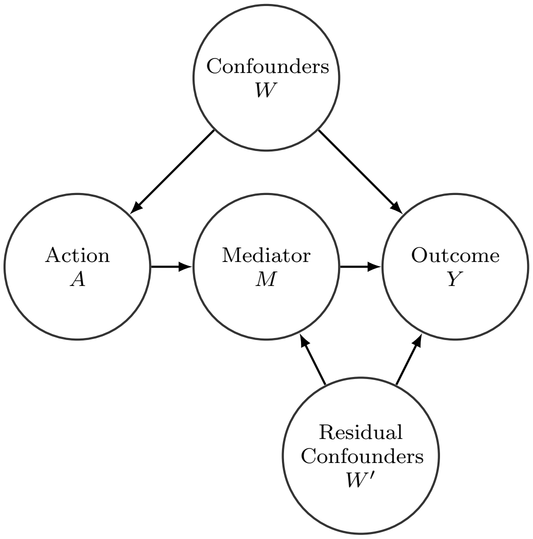


Figure 4: Causal DAG with confounded viewability.

A negative control is an irrelevant but observable outcome Y^- where the action A is known to have no impact (figure 5). Negative controls can help identify bias in causal studies as they provide a test where the analysis should produce an estimate of zero effect [10]. Online entities typically track a large number of actions that a user may take, so negative controls are often readily available. A commonly used negative control in online advertising is a conversion event from an unrelated campaign [3, 15]. In the case of recommender systems, the selection by the user of an unrelated item could serve as a negative control for the recommended item.

A negative control is most useful when it is subject to all of the same confounders W as the original outcome Y. If the negative control conversion has fewer confounders, then it is possible to correctly predict no effect on the negative control while still admitting residual bias on the actual outcome. For this reason, we propose to use a spectrum of negative controls to increase the likelihood that each confounder is covered by at least one control. An additional advantage of using multiple negative controls is the possibility to infer the magnitude of estimation bias. If the negative controls are subject to the same bias as the outcome of interest, they can provide an empirical distribution of this bias.

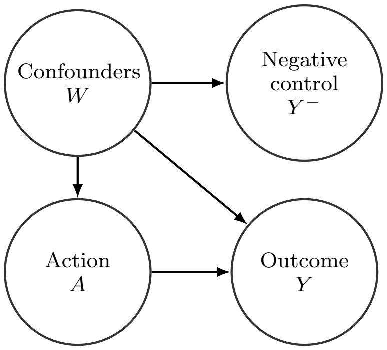


Figure 5: Causal DAG with negative control.

If an analysis based on a mediating event produces significant non-zero effects when analyzed against a spectrum of negative controls, this likely indicates there are residual confounders W'. Below we describe an approach to identifying W' and adjusting for it to revise a causal estimate in the context of online advertising. When successful, adjusting for residual confounders can lead to low-bias estimates of causal effects at the mere cost of instrumenting M. In general, we expect that controlling for confounding of the mediator M to be easier than controlling for the deep confounding of the action A. That being said, if the adjustment for confounders of M estimates large effects on negative controls, then the proposed M should be discarded for a new candidate.

3. AD VIEWABILITY

We apply our framework to online advertising where we identified *ad viewability* as a mediating event. A large fraction of ad impressions that are served to users are never seen because these ads are not loaded to a viewable portion of the user's browser. Unviewable ads cannot have an effect on the user's actions, and so viewability mediates the effect of serving ads. To understand this phenomenon, it is important to know the factors that determine whether an ad becomes viewable.

The most common reason for a display ad to be unviewable is it being served "below the fold." This means the user must scroll down the page for the ad to move into the window of the web browser. If the user leaves the web page before scrolling to the ad's location, the impression will not be seen. Whether an ad is served below the fold is determined by a combination of factors including the web page's layout, the size and location of the ad, the window size of the user's browser, and the monitor's display settings. Each of these factors is subject to a wide degree of variability. For example, ads can be displayed as wide, short banners at the top of a web page, as narrow, tall "skyscrapers" on the border of a page's content, or as rectangular patches distributed at various locations within the page.

Another reason why ads become unviewable is their being "out of focus." Sometimes content is loaded to a browser that is minimized or to a tab that is not currently active. This is commonly a result of auto-refreshing where content is reloaded at regular intervals without the user's intervention.

[3] Please note the similarity between this approach and the front-door criterion described by Pearl [12].

The content is not seen unless the user actively navigates to the hidden or minimized window.

The targeter has very limited information about any of these factors at the time of placing an ad. Typically the targeter bids for placements based on the URL, the ad size, and aspects of the visitor, so detailed information about the likelihood of the ad to be viewable cannot be factored into the decision to serve the ad. This suggests that viewability is a good candidate for our analysis framework since it is not subject to normal targeting bias.

The corresponding author's company has instrumented viewability via JavaScript code that is served along with ads. This code is run in the user's browser, where it queries geometric data and other information from the browser that it uses to determine whether the ad is currently in view. The definition we use for *in view* for this paper is that at least 50% of the ad's pixels are visible on the user's screen. The end result is a measurement of the duration over which the ad was determined to be in view.

Figure 6 shows the distribution of the duration of time in view for a random sample of ads collected from advertising campaigns from a variety of industries. We find that approximately 45% of ads are in view for less than one second. Hereafter, we will refer to such ads as *unviewable*. Impressions that were in view for at least one second will be called *viewable*.

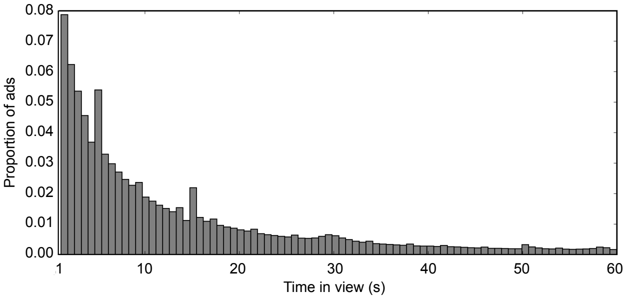


Figure 6: Distribution of time in view for viewable impressions (*i.e.* impressions that have at least 50% of the ad's pixels in a visible portion of the user's screen for at least one second).

Figure 7 summarizes the influence of vertical location of the impression on ad viewability. The median browser window size is depicted as the red box in figure 7a and the frequency of a pixel containing an ad is shown by the green heatmap. We see in figure 7b that less than half of all ads are served above the fold. Furthermore, the probability of the impression being viewable drops quickly around the median location of the bottom of the browser window (figure 7c). In addition, hidden tabs and minimized browser windows account for about 6% of ads being unviewable.

4. REDUCTION IN BIAS

In this section, we use measurements of ad viewability to reduce bias in estimating campaign performance. We compare a viewability analysis to a naïve one where the presence of selection bias is ignored. For both analyses, we use the individual impression as the unit of analysis (see figure 8). The observed data structure is $O = (A, M, Y) \sim P$ where A indicates whether the impression is from the campaign of interest, M indicates whether the impression was viewable, Y indicates whether the user converted in a window after the impression, and P is the underlying probability distribution. We use a conversion window of 7 days from the exact time of the impression. Note that in both analyses it is assumed that the effect of an ad is not modified by preceding or following ads. If a user receives multiple ads, they are considered in isolation of each other. While this assumption is likely to break down in certain scenarios, we use it to simplify our analysis while still addressing the problem of selection bias.

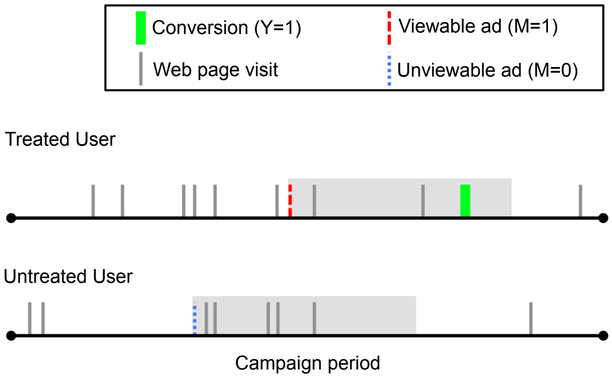


Figure 8: Diagram depicting variable definitions for viewability analysis. Gray boxes indicate the window in which conversions are monitored.

In our naïve analysis, we compare cases where the impression is from the campaign associated with the conversion ($A = 1$) versus cases where the impression is from an unrelated campaign ($A = 0$).[4] We expect this analysis to be biased by strong confounding from ad targeting. The target parameter for this analysis is the campaign lift defined by

$$\begin{aligned}\Phi_{\text{naïve}}(P) &= \frac{E\{Y \mid A=1\}}{E\{Y \mid A=0\}} - 1 \\ &= \frac{p(Y=1 \mid A=1)}{p(Y=1 \mid A=0)} - 1.\end{aligned} \tag{1}$$

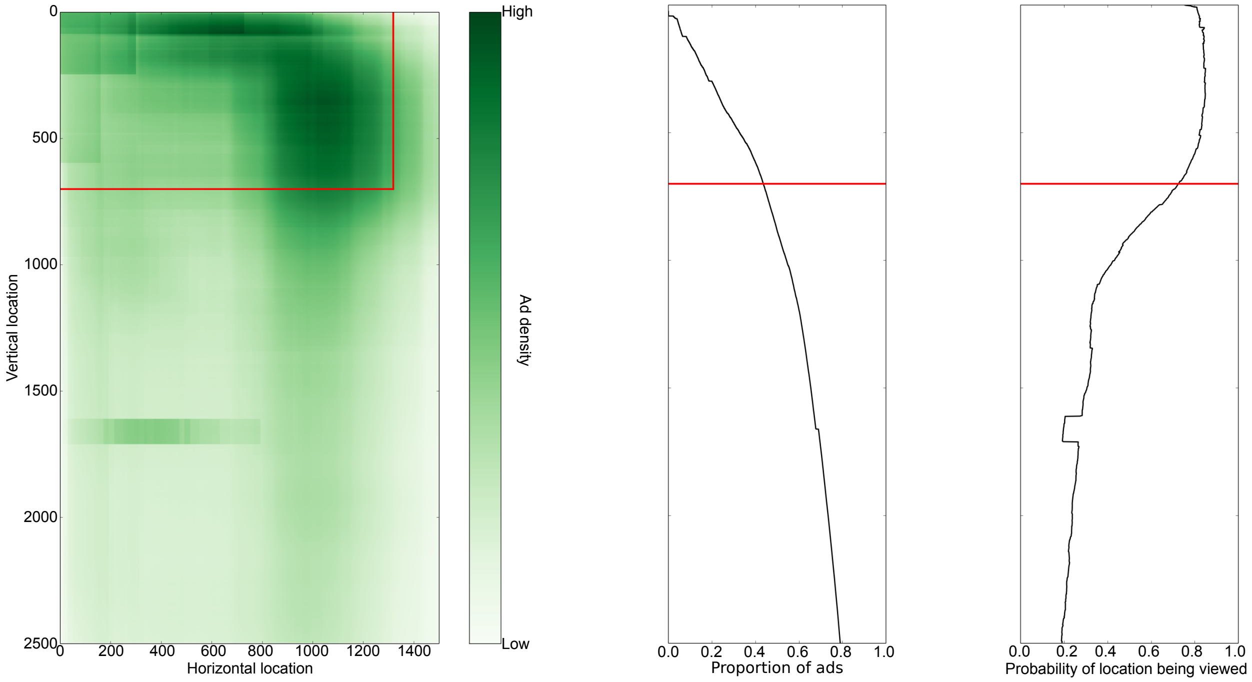
For the viewability analysis we focus only on impressions from the campaign ($A = 1$). We compare conversion rates for users who received viewable ads ($M = 1$) with those who had unviewable ones ($M = 0$). The lift in this case is defined by

$$\begin{aligned}\Phi_{\text{viewability}}(P) &= \frac{E\{Y \mid M=1, A=1\}}{E\{Y \mid M=0, A=1\}} - 1 \\ &= \frac{p(Y=1 \mid M=1, A=1)}{p(Y=1 \mid M=0, A=1)} - 1.\end{aligned} \tag{2}$$

Both equations could be interpreted causally in the absence of confounders. However, as we expect strong selection bias in ad targeting, our hypothesis is that utilizing ad viewability will dramatically reduce such bias.

Our data come from seven display advertising campaigns run during the 4th quarter of 2014. The products advertised represented diverse industries including auto insurance,

[4] All the users in data under consideration received at least one impression from some campaign.



(a) Frequency of ads served at different locations on a page.

(b) Proportion of ads served at or above a particular page height.

(c) Probability of a pixel being brought into a viewable portion of the user's web browser.

Figure 7: Summary of the effect of the page fold on viewability. The red box and line depict the median size of a browser window. Location units are in pixels.

beauty products, finance, and online marketing. Conversions also represented diverse actions such as visiting the advertiser's web page, purchasing a product, or filling out a quote form. The volume of the campaigns ranged from 3 million to 29 million impressions. We define our users via browser cookies. The number of users who converted after receiving an impression ranged between 2,000 and 2 million depending on the campaign. For each campaign, we calculated $\Phi_{\text{naïve}}$ and $\Phi_{\text{viewability}}$ from the empirical probabilities that impressions of each type were followed by conversions. Confidence intervals were derived from the binomial distribution.

We found that across all campaigns values dropped dramatically between $\Phi_{\text{naïve}}$ and $\Phi_{\text{viewability}}$ (figure 9). For example, campaign C saw a drop in lift from 33,000% (95% CI, 28,000%–40,000%) in the naive analysis to 26% (95% CI, 21%–32%) in the viewability analysis. This campaign utilized retargeting practices where users were served ads if they had specifically interacted with the advertising brand in the past. This practice suggests selection bias should be particularly severe. Campaign F saw the smallest difference between the naive and viewability analyses displaying lifts of 65% (95% CI, 62%–69%) and 12% (95% CI, 7%–16%), respectively. The campaign in question was run by a major US consumer banking company which may use rather broad targeting criteria, limiting the amount of selection bias.

While the above data provide compelling evidence that bias can be reduced by using measurements of ad viewability, it is unclear how much residual bias remains. To probe the remaining bias, we turn to negative controls.

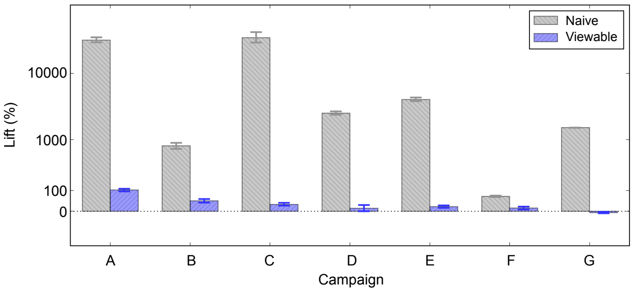


Figure 9: Lifts estimated for the naive analysis and for the analysis using viewable ads to define the treatment group. Note: a log scale is used to depict the difference between lifts.

5. NEGATIVE CONTROLS

As discussed above, negative controls help to identify bias in causal studies as they provide a test where an unbiased estimator should predict zero effect. For each of our seven campaigns, we identified 7–10 conversion events from other campaigns to serve as negative controls. The negative controls were chosen from a different industry from the true conversion in order to minimize any possible interaction. In figure 10 we show the result of campaign B's impressions analyzed against its true conversion event as well as 10 negative controls using the viewability methodology. The lift on the true conversion, 41% (95% CI, 34%–50%), is larger than that of any of the negative controls. However, all but one negative control showed a significant lift despite being irrelevant to the campaign ads ($p < 0.05$).

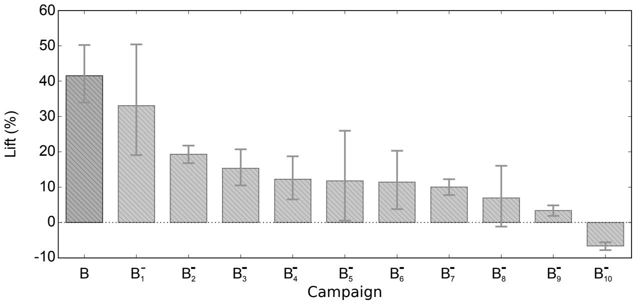


Figure 10: Lift estimated for campaign B's true conversion and several negative controls B_i^- using ad viewability to define treatment.

In total, we had seven campaigns with a combined 60 negative controls. The average absolute effect estimated on the negative controls was 11% ± 8% indicating a modest bias in the viewability analysis. Out of 60 negative controls, 44 showed a significant effect ($p < 0.05$). The most likely explanation for this bias is that ad viewability is not a perfect natural experiment and that there are confounders for the impact of ad viewability on conversion (W' in figure 4). In the next section we proceed to identify and adjust for these confounders.

6. ADJUSTMENT FOR CONFOUNDERS

Confounders are variables that influence both the outcome and the action as described above. A common way to identify such factors is to test for a statistical dependence on a candidate confounder of both the action and outcome. In our case, we are searching for factors X with the properties $M \not\perp X$ and $Y \not\perp X \mid M$. We caution that statistical tests are generally insufficient for identifying a true confounder and that the identification must also rest on true domain knowledge of the system under study [12].

To determine which variables constitute a set of confounders in our new framework, we first identified a set of candidates by drawing on domain knowledge from the online advertising ecosystem. These features are organized into three general categories that encompass user level and campaign level characteristics. A summary of potential confounders is given in table 1. We identified the confounders W' on a campaign-by-campaign basis as those factors that were significantly correlated with both M and Y. This was determined by performing a logistic regression with each factor alone and testing its coefficient for statistical significance ($p < 0.05$).

Category	**Description**	**Examples**
Technical	The user's computer setup.	Operating system, browser, screen resolution, aspect ratio
Behavioral	Determined by the user's actions.	Browsing frequency, average viewability for user, previous brand interaction, time of day
Targeting	Parameters of campaign as set by advertiser.	Prospecting or retargeting strategy, average viewability for publisher, targeted web page

Table 1: Summary of potential confounders.

We adjusted our analysis for these confounders through an extension of our viewability analysis. We continued to use individual impressions as the unit of analysis. The augmented data structure was $O' = (W', A, M, Y) \sim P'$ where the confounders W' have been added while all other variables are as before. The estimation of lift was altered to provide for an adjustment over W'. The equation for the adjusted lift is

$$\begin{aligned}\Phi_{\text{adjusted}}(P') &= \frac{E\{Y \mid M=1, A=1\}}{E\{Y_0 \mid M=1, A=1\}} - 1 \\ &= \frac{p(Y=1 \mid M=1, A=1)}{p(Y_0=1 \mid M=1, A=1)} - 1\end{aligned} \tag{3}$$

where Y_0 indicates whether the outcome would occur if the ad had been unviewable. In the language of causal inference this is a *counterfactual* outcome which can only be identified by performing an adjustment for the confounders W' [12]. This adjustment is given by

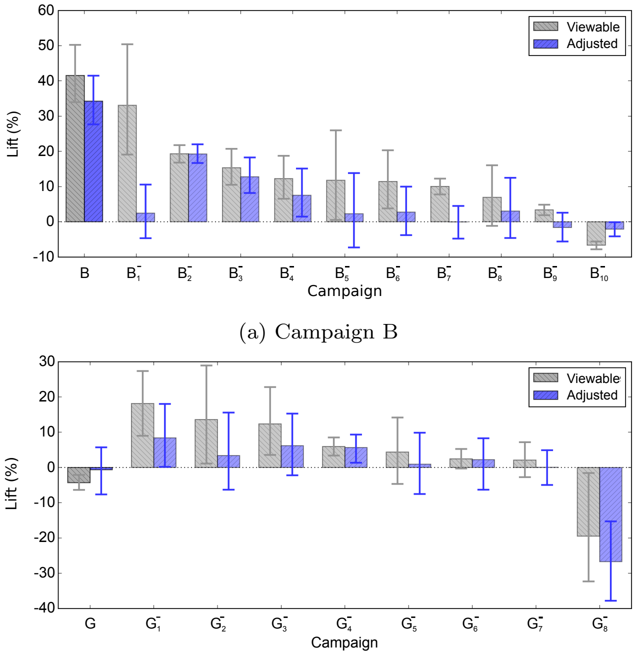
$$\begin{aligned}&p(Y_0 = 1 \mid M=1, A=1) \\ &\quad = \sum_{w'} p(Y=1 \mid W'=w', M=0, A=1) \\ &\qquad\qquad \times p(W'=w' \mid M=1, A=1) \\ &\quad \approx \frac{1}{N}\sum_{i}^{N} p(Y=1 \mid W'=w'_i, M=0, A=1).\end{aligned} \tag{4}$$

The first equality performs an adjustment over all possible values of the confounder $W' = w'$. The second equality makes an empirical approximation for the distribution of W' among viewable impressions. The final summation is over the N viewable impressions in O' where w'_i represents the confounders of viewability for the i^{th} viewable impression.

To estimate Φ_{adjusted}, we need estimates for the probabilities in the numerator and denominator of equation 3. We estimated the numerator as the empirical probability of conversion among viewable impressions. To estimate the denominator we fit the probability $p(Y=1|W', M=0, A=1)$ with a logistic regression trained on data from unviewable impressions. This model was then substituted into equation 4 and applied to the value of W' for all viewable impressions. Confidence intervals were obtained through 1000 bootstrap samples.

Following our framework, we tested our model on each of the seven campaigns and their corresponding set of negative controls. Both the unadjusted and adjusted viewability model results for campaigns B and G are shown in figure 11. In campaign B, we find the lift on the true conversion, 34% (95% CI, 28%–41%), is larger than that of any of its negative controls. Whereas the unadjusted estimate shows a significant effect ($p < 0.05$) in 9 out of 10 negative controls, this number drops to 4 out of 10 when adjusting for W'. In contrast, the adjusted effect of impressions from campaign G on its true conversion is not significant ($p > 0.05$).

Overall, the average lift across negative controls is -2% suggesting an absence of systematic bias. The average absolute value of the lift estimated for negative controls was $9\% \pm 9\%$ which constitutes only a small decrease when compared to the unadjusted estimate. Out of 60 negative controls, 29 still showed a significant non-zero effect ($p < 0.05$). These results indicate that while adjustment for W' were successful in reducing bias, there is still likely some uncontrolled source of confounding.



(a) Campaign B

(b) Campaign G

Figure 11: Unadjusted and adjusted lift for campaigns B and G and their corresponding negative controls.

The effect of each campaign on its true conversion is shown in figure 12. Across campaigns we found that the estimate of campaign lift dropped on average from 30% to 10% when adjusting for W'. The adjustment for confounding therefore created substantial revisions in the estimate of lift. The variation in the effect of adjustment on the estimated lift can likely be attributed to large differences in the degree of confounding present in each campaign.

7. DISCUSSION AND CONCLUSION

The results presented above demonstrate that even with intractable confounding, causal estimation is still possible when the effect of an action is mediated through another event. We applied our framework to display advertising where the effect of serving ads is mediated by whether the ad appears in a viewable portion of the user's screen. Despite huge bias in comparing users who receive ads with those who do not, we found a relatively small bias on the order of 10% when comparing the effect of viewable ads to unviewable ones under proper controls. Approximately 50% of display ads are unviewable, so ad viewability represents a perpetual, balanced, and free natural experiment for measuring the impact of ad impressions. An implementation of our analysis has been released in production to monitor the lift of dozens of advertising campaigns on a daily basis. Extensions of this technique are being used to understand the performance of campaigns on a more granular scale, such as different audience segments or targeting strategies.

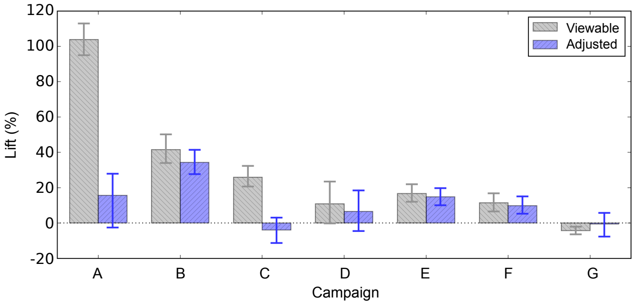


Figure 12: The lift estimates for the campaign's true conversion event when using viewability with and without adjusting for confounding.

We discuss two limitations of our current approach. The first is the assumption that ad impressions can be considered independently from one another. Violations of this assumption may account for some of the residual bias in our analysis. It may be that when a user receives multiple ads from the same campaign that the impressions interact in a non-additive way. Our current research focuses on a *longitudinal* implementation of our viewability analysis [4] in which we consider the evolution of the user's ad experience over the course of the campaign. Second, while we actively collect information related to measuring ad viewability, we do not collect all possible metrics. For example, we do not collect information related to page structure or creative format (*e.g.* static image or animated graphics). With further data mining efforts, other potential confounders can be introduced into our pipeline. The exploration and discovery of these features is the subject of ongoing work.

Our framework has applications beyond advertising as confounding is an almost unavoidable consequence of today's predictive modeling systems. The use of mediating events thus provides an attractive alternative for estimating causal effects when randomized experiments are prohibitive. Companies commonly collect and store an enormous amount of data on their customers. These data could be leveraged to identify causal mediators with minimal additional cost to the company. The instrumentation of mediators has the potential to become a critical step in measuring the performance of automated systems.

8. ACKNOWLEDGMENTS

An earlier version of this work was presented at the Winter Conference on Business Intelligence [11]. We thank Ori Stitelman, Mark van der Laan, Steve Tadelis, Tom Blake, Dominic Coey, Katia Eliseeva, and Gijs Joost Brouwer for useful discussion and suggestions. This work was done while Daniel Hill, Vadim Tsemekhman, and Foster Provost were at Integral Ad Science. Foster Provost thanks Andre Meyer for a Faculty Fellowship.

9. REFERENCES

[1] N. Black. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*, 312(7040):1215–1218, 1996.

[2] T. Blake, C. Nosko, and S. Tadelis. Consumer heterogeneity and paid search effectiveness: A large scale field experiment. Technical report, National Bureau of Economic Research, 2014.

[3] 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.

[4] D. Hill, G. J. Brouwer, A. E. Hubbard, and Tsemekhman. It's about time: A longitudinal model for the causal impact of display ads. *Winter Conference on Business Intelligence*, 2014.

[5] D. D. Jensen, A. S. Fast, B. J. Taylor, and M. E. Maier Automatic identification of quasi-experimental designs for discovering causal knowledge. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 372–380. ACM, 2008.

[6] R. Kohavi, and R. Longbotham. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter*, 12(2):31–35, 2011.

[7] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[8] R. Lewis and J. M. Rao. The unfavorable economics of measuring the returns to advertising. *Available at SSRN*, 2014.

[9] 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*, pages 157–166. ACM, 2011.

[10] M. Lipsitch, E. T. Tchetgen, and T. Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

[11] R. Moakler, V. Tsemekhman, A. Ahuja, D. N. Hill, G. J. Brouwer, F. Provost, and K. Tsemekhman. Causal impact of online advertisements using viewability as a method of treatment. *Winter Conference on Business Intelligence*, 2014.

[12] J. Pearl. *Causality: models, reasoning and inference.* Cambridge University Press, New York, 2000.

[13] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine Learning*, 95(1):103–127, 2014.

[14] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.

[15] O. Stitelman, B. Dalessandro, C. Perlich, and F. Provost. Estimating the effect of online display advertising on browser conversion. *The Fifth International Workshop on Data Mining and Audience Intelligence for Online Advertising at SIGKDD*, 2011.

[16] O. M. Stitelman and M. J. van der Laan. Collaborative targeted maximum likelihood for time to event data. *Int J Biostat*, 6(1):Article 21 2010.