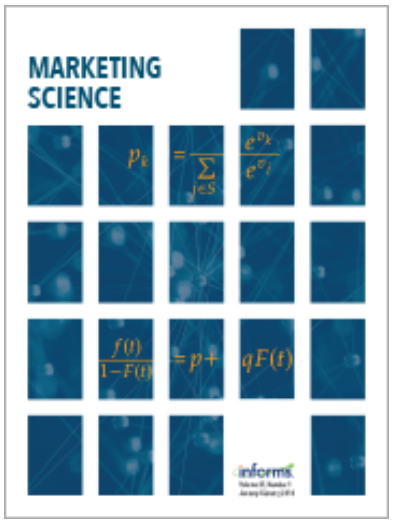


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces

Joel Barajas, Ram Akella, Marius Holtan, Aaron Flores

To cite this article:
Joel Barajas, Ram Akella, Marius Holtan, Aaron Flores (2016) Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces. Marketing Science 35(3):465-483. https://doi.org/10.1287/mksc.2016.0982

MARKETING SCIENCE
Vol. 35, No. 3, May–June 2016, pp. 465–483
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2016.0982

Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces

Joel Barajas
University of California, Santa Cruz, California 95064, jbarajas@soe.ucsc.edu

Ram Akella
School of Information, University of California, Berkeley, California 94720; and
University of California, Santa Cruz, California 95064, akella@ischool.berkeley.edu

Marius Holtan, Aaron Flores
AOL Research, Palo Alto, California 94306
{marius.holtan@teamaol.com, aaron.flores@teamaol.com}

Online Display Advertising's importance as a marketing channel is partially due to its ability to attribute conversions to campaigns. Current industry practice to measure ad effectiveness is to run randomized experiments using placebo ads, assuming external validity for future exposures. We identify two different effects, i.e., a strategic effect of the campaign presence in marketplaces, and a selection effect due to user targeting; these are confounded in current practices. We propose two novel randomized designs to: (1) estimate the overall campaign attribution without placebo ads, (2) disaggregate the campaign presence and ad effects. Using the Potential Outcomes Causal Model, we address the selection effect by estimating the probability of selecting influenceable users. We show the ex-ante value of continuing evaluation to enhance the user selection for ad exposure mid-flight. We analyze two performance-based (CPA) and one Cost-Per-Impression (CPM) campaigns with 20 million users each. We estimate a negative CPM campaign presence effect due to cross product spillovers. Experimental evidence suggests that CPA campaigns incentivize selection of converting users regardless of the ad, up to 96% more than CPM campaigns, thus challenging the standard practice of targeting most likely converting users.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mksc.2016.0982.

Keywords: online advertising; experimental design; user targeting; field experiments; Bayesian estimation; econometrics

History: Received: December 31, 2013; accepted: October 28, 2015; Pradeep Chintagunta, Dominique Hanssens, and John Hauser served as the special issue editors and Carl Mela served as associate editor for this article. Published online in *Articles in Advance* May 3, 2016.

1. Introduction and Problem Context

According to the Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers (PWC), Internet display related advertising revenues in the United States totaled $6.5 billion during the first six months of 2014. This revenue represents 28% of the total online advertising ($23.1 billion) and constitutes an increase of 6% over the $6.1 billion reported over the same period in 2013. Because of the proliferation of the online user activity tracking, performance-based, or Cost-Per-Action (CPA), campaigns accounted for 65% of the campaigns run for the same period. On the other hand, 34% of campaigns were run under the more traditional Cost-Per-Impression (CPM) business model.[1] In this context, determining the effectiveness of an online campaign in achieving increased user commercial actions is usually used to give credit to CPA campaigns. This process is termed *campaign attribution*.

[1] IAB internet advertising revenue report. 2014 first six months' results. http://www.iab.com/wp-content/uploads/2015/05/IAB_Internet_Advertising_Revenue_Report_HY_2014_PDF.pdf.

The advertising industry has developed methods for online conversion attribution such as Last-Touch Attribution. In this framework, the full conversion credit is given to the last campaign that a converting user is exposed to (i.e., the touch point). Another method is the Multi-Touch Attribution (MTA), where the conversion credit is heuristically split across the touch points in the path to conversion (Atlas Institute 2008). Data-driven MTA approaches have been proposed to model interacting channel effects (Shao and Li 2011, Li and Kannan 2014). However, these methods assign attribution credit to every exposed and converting user while ignoring the counterfactual response without ad exposures. Also, these approaches incentivize selection for ad exposure of baseline users (Berman 2015), those who convert regardless of the touch point (*always-buy* users).

Running randomized experiments (or field experiments) is becoming the standard approach to measuring the marginal effectiveness of online campaigns (Chittilappilly 2012, Lewis et al. 2011, Yildiz and Narayanan 2013, Johnson et al. 2016). In this practice,

the ad is assumed to be the *treatment* to evaluate, and users are randomly separated into study and control groups. Hence, when a targeting engine selects a visiting user for exposure, the campaign ad is displayed to users in the study group, or a placebo ad is displayed to users in the control group (Yildiz and Narayanan 2013, Lewis et al. 2011). Full deployment of this framework is limited by the cost of displaying placebo ads, and the potential revenue loss resulting from yielding the opportunity to advertise to control users. As a consequence, current industry practice is to run a low-budget CPM campaign and measure its effectiveness, which is also assumed to hold (external validity) for a larger budget CPA campaign (Yildiz and Narayanan 2013, Chittilappilly 2012).

Today, ad exchange platforms facilitate marketplaces where advertising spaces on websites are bought and sold. A survey of 49 media buyers indicates that 87.8% intended to purchase digital advertising via real-time bidding (RTB) by 2011 (Digiday and Google 2011). Similarly, outside RTB exchanges, ad networks run internal auctions (Broder and Josifovski 2011). Because media buying is done endogenously in a competitive market, the user selection for ad exposure complicates the evaluation using placebo ads. Moreover, to display a placebo ad, the opportunity to advertise must be *consumed* and the campaign must exist in the marketplace (i.e., campaign presence effect). Johnson et al. (2016) acknowledge the bias induced by endogenous user selection when running a placebo campaign. Similar to propensity-score based corrections, their proposed solution predicts ad exposures of users in the control group based on their features, which are often noisy and fragmented. Also, their approach relies on auction simulations assumed to be stationary. Most important, the effect of the campaign presence in the marketplace[2] (i.e., strategic effect) is ignored by current practices and literature.

User targeting is one of the most important decisions in running a campaign. A survey of 100 marketers, agencies, and media planners indicates that user targeting and campaign optimization capabilities are perceived as the main differentiators among ad networks (Morrison and Coolbirth 2008). In campaign attribution, ad exposures are often considered a consequence of user activity (Lewis et al. 2011), or even a potential "coincidence" (Yildiz and Narayanan 2013). In reality, deployment of CPA campaigns has produced increasingly sophisticated targeting engines that aim to display ads to converting users (i.e., selection effect) (Pandey et al. 2011, Aly et al. 2012). As a result, the external validity of nonoptimized CPM campaign effects to CPA campaigns, assumed by current industry evaluation practice, is prone to inaccuracies.

In a recent economic literature review, Goldfarb (2014) surveys the online advertising literature based on the decreasing cost of user targeting. However, most of the literature on ad effectiveness based on field experiments evaluates focused and specific targeting practices (Lambrecht and Tucker 2013, Goldfarb and Tucker 2011, Lewis and Reiley 2014). To our knowledge, comparing the selection policy performance of CPA and CPM campaigns, and the implications for current practices have not been adequately addressed by previous literature.

1.1. Our Contribution

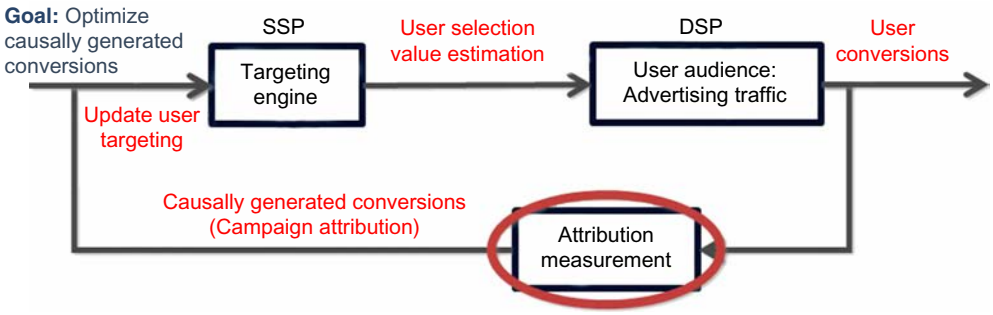
We focus on the marginal causal attribution of single-product online conversions to online display campaigns (i.e., single channel) run on hundreds of publisher websites, given all other advertising channel exposures or prior branding effect. We find that current industry practice often confounds three campaign effects, i.e., the ad effect on exposed users, the strategic impact of the campaign presence in a competitive market, and the selection effect of the media buyer. We summarize the elements of our contribution below in this context.

Expand the scope of attribution in marketplaces to the overall campaign. We propose to perform continuing evaluation and estimate the campaign attribution for the current running conditions instead of isolating the ad effect. In this new perspective, the entire campaign, including the campaign presence in the marketplace and the ad, is now the treatment to evaluate. Consequently, we propose a new randomized design that considers all of the visiting users and does not display placebo ads to users of the control group. We argue that this control group represents the right campaign counterfactual in a marketplace. This design cost, which is minimal as to revenue loss, enables us to perform continuing evaluation and attempts to close the feedback loop for causal campaign optimization displayed by Figure 1. The proposed design is simple to implement and does not suffer from endogenous user selection.

Capture the effect of the campaign presence in the marketplace. We propose a second randomized design that separates the ad effect from the impact of the campaign presence in the marketplace. Without relying on noisy user features, we develop a method to estimate the user conversion probability of the statistically equivalent users in the control group to those exposed to the ad in the study group. We show the risks of inducing a selection effect in the standard evaluation practice of using placebo ads in a marketplace. Contrary to paid search marketplaces (Blake et al. 2015), we report evidence of a campaign presence effect. This effect is

[2] Blake and Coey (2014) identify test-control interference in marketplaces where users bid on scarce products. In the marketplace we address, advertisers bid on ad slots assuming there is enough inventory to satisfy the demand.

Figure 1 **(Color online) Online Advertising Optimization Loop**



Note. The focus of this paper is the measurement (i.e., attribution) box.

largely ignored in the literature and can significantly change the campaign attribution.

Characterize the user selection based on user influenceable classes. We present a method to characterize the user selection of influenceable user classes using a Potential Outcomes causal model and Principal Stratification (Frangakis and Rubin 2002). By comparing the probability of selecting *always-buy* users, we report evidence supporting the hypothesis that CPA campaigns incentivize the selection of these users when compared with CPM campaigns (Berman 2015). Based on user demographics, we test different user selection policies for mid-flight campaign optimization in the context of the control loop in Figure 1. Our results suggest that optimizing user selection for ad exposure has a significant impact on campaign effects. These findings raise questions about the external validity of ad effects estimated by a CPM experiment to the CPA campaigns (current practice).

We approach the problem in two phases: (1) the randomized design, and (2) the causal estimation given this design. Hence, in §2, we analyze the targeted display advertising process in a marketplace and describe the proposed design. We present the methodology to characterize the users based on the potential causal effects on them, and the user selection effect in §3. In §4, we cover the estimation model validation, attribution results for two CPA campaigns, the market presence effect for one CPM campaign, and the user selection characterization for these campaigns. We show the value of continuing evaluation in §5 by optimizing user selection for ad exposure assuming short-term ex-ante external validity of the effects. Finally, in §6, we discuss the main findings and their managerial implications. Tables 1 and 2 define the notation used in this paper.

2. Experimental Design for Attribution in Marketplaces

2.1. Targeted Display Advertising in Marketplaces: Overview

In Targeted Display Advertising, marketing campaigns are often run by advertisers working closely with a given ad network. The mechanism for displaying an ad is depicted by the decision tree shown in Figure 2(a). This process is based on conducting an auction for

Table 1 **Description of the Variables Used in This Paper**

Term	Description	Term	Description
$Z \in \{\text{C}, \text{P}, \text{S}\}$	Random treatment assignment	$Y \in \{0, 1\}$	Converting user indicator
$B \in \{\text{No}, \text{Yes}\}$	Decision to bid indicator	$A \in \{\text{Lose}, \text{Win}\}$	Auction output indicator
$D \in \{0, 1\}$	Selected for ad exposure indicator	$W \in \{0, 1\}$	Ad exposure indicator
$i \in \mathbb{N}$	Variable index for the ith user	$X \in \mathbb{Z}^p$	User feature vector
$\theta_{dz} \in [0, 1]$	Probability of $Y = 1$ given $D = d$, $Z = z$	$p_{sel} \in [0, 1]$	Probability of $D = 1$
$\theta_{0z}^{(s)}, p_{sel,z}^{(s)} \in [0, 1]$	Parameters obtained by repeated randomization for validation	$\Delta_{psel}^{(s)}, \Delta_{\theta 0}^{(0)} \in [-1, 1]$	Difference statistics between repeated randomized groups
$\Delta^{select} \in [-1, 1]$	Statistic to test for equivalent user selection for placebo ads	$\Delta^{convert} \in [-1, 1]$	Statistic to test for equivalent user populations for placebo ads
$N_{dz}^{y} \in \mathbb{N}$	User count given $Y = y$, $D = d$, $Z = z$	N_{obs}, N_{samp}	Observed/sampled count sets
$N_{burnin}, N_s \in \mathbb{N}$	Burn-in/Gibbs number of samples	$a_0, b_0 \in (0, \infty]$	Beta prior parameters
$U \in \{\text{Per}^{+}, \text{Per}^{-}, \text{AB}, \text{NB}\}$	Influenceable user category indicator	$\beta_{sel}, \gamma_{dz} \in \mathbb{R}^{p+1}$	Regression parameters to model $P(Y \mid X, D, Z, \gamma_{dz}), P(D \mid \beta_{sel}, X)$
Θ	Parameters Equation (5): $\{\theta_0, \theta_{1C}, \theta_{1S}, p_{sel}\}$	Θ_X	Parameters Equation (14): $\{\gamma_0, \gamma_{1C}, \gamma_{1S}, \beta_{sel}\}$
$F_{targ}(X_i), F_{sig}(X_i), F_{sign}^{ATE}(X_i), \mathbf{w}^{sig}$		Exposure selection optimizing functions (Algorithm 3)	

Table 2 Performance Metrics Used in This Paper

Metric	Lift	Description
ATE_{Camp}	lift_{Camp}	Overall campaign average effect on all visiting users
ATE_{Ad}	ACL_{Ad}	Average effect of the ad on selected users for ad exposure
ATE_{Market}	ACL_{Market}	Average campaign presence in the marketplace effect on exposure-selected users
$\text{ATE}^{D=1}_{Camp}$	$\text{lift}^{D=1}_{Camp}$	Average treatment effect of the campaign on selected users
SelEff	lift_{sel}	User selection effect introduced by the targeting engine
$P(D=1 \mid U)$	Probability of selecting user influenceable category U for ad exposure	
$\text{ATRB}^{D=1}_{Camp}$	Campaign attributed converting users, with respect to $N^1_{0S}+N^1_{1S}$, estimated based on $\text{ATE}^{D=1}_{Camp}$	
ATRB_{Camp}	Campaign attributed converting users, with respect to $N^1_{0S}+N^1_{1S}$, estimated based on ATE_{Camp}	

Notes. Lifts $\in [-1,\infty)$, $P(D=1 \mid U) \in [0,1]$. Other metrics $\in [-1,1]$.

every visiting user who is provided by a supply-side platform (SSP) or publisher websites. To target users, advertisers develop user profiles of the target market segment based on demographics and other features. However, in practice, the ad network uses a highly sophisticated algorithm, illustrated by the decision node B of Figure 2(a), to determine if a user should be targeted. In CPA campaigns, this decision is based on user behavior and history, and how likely the user is to convert, among other features (Pandey et al. 2011, Aly et al. 2012). If the campaign decides to bid through a demand-side platform (DSP) in the ad exchange ($B=$ Yes), it submits the bid through RTB (Spencer et al. 2011). The chance (endogenous) node A of Figure 2(a) represents this auction output. If the campaign wins the advertising slot ($A=$ Win), the campaign ad is displayed to the user. Otherwise, another advertiser shows an ad. For CPM campaigns, the decision to bid is set to $B=$ Yes. Moreover, the bidding strategy is determined by guaranteed delivery contracts or by the spot market (Ghosh et al. 2009). Outside of ad exchanges, these targeting and auction processes are routinely run by large ad networks (Broder and Josifovski 2011). For the effects of the current paper, we consider the aggregate targeting engine output (chance node D of Figure 2(b)) to refer to users selected for ad exposure, where $D=1$ if the user is selected, i.e., if $B=$ Yes and $A=$ Win, and $D=0$ otherwise. Referring to selected ad-exposed users as targeted users is typical in the targeted advertising literature. However, we note the case where the user is targeted $B=$ Yes but not exposed to the ad if $A=$ Lose.

2.2. Campaign Evaluation Using Placebos: The Standard Practice

The standard approach to evaluating online marketing campaigns is to use randomized experiments assuming the ad design is the *treatment* to evaluate. Lewis et al. (2011) propose randomly assigning visiting users at serving time to see the focal ad (study) or the placebo ad assumed to be unrelated to the brand (control). Figure 3(a) illustrates this process. In this model, none of the components of standard Targeted Display Advertising in a marketplace are considered. Moreover, randomizing user visits limits this design power; a given user might be assigned to both treatment arms during different visits.

Current industry practice is to randomize the visiting users once and keep them in the same arm throughout the experiment, as depicted in Figure 3(b) (Yildiz and Narayanan 2013). Because media buying is endogenously performed in a competitive market, user selection for ad exposure indicator D becomes a post-treatment variable. Conditioning the analysis on its realization might introduce a post-treatment bias.[3] Moreover, the targeting engine routinely incorporates user activity feedback, such as user clicks and visits, to improve user selection for ad exposure (Aly et al. 2012), which would not be the case for the placebo ad.

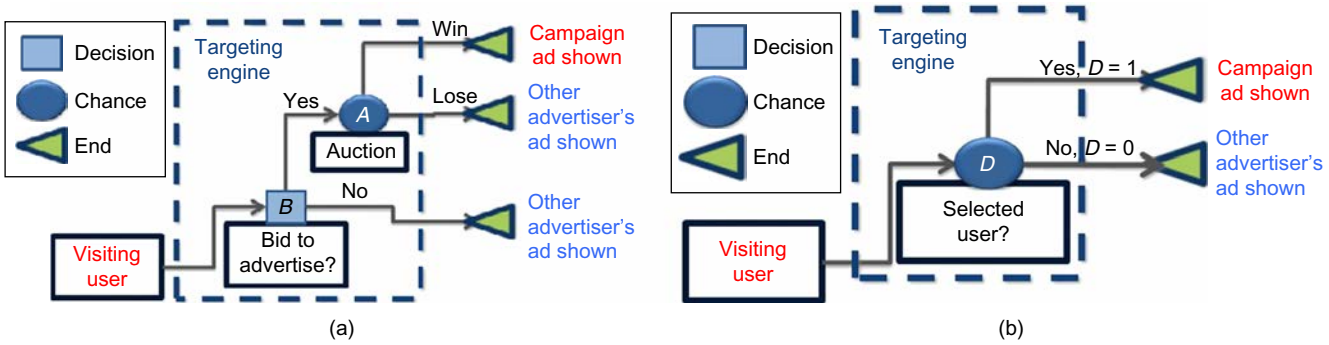
These practices focus on the ad evaluation, without considering the effect of the campaign presence in the marketplace. Also, the ad is often evaluated with a low-budget CPM campaign; the effects are assumed to hold for larger-budget CPA campaigns (Chittilappilly 2012 describes a general industry practice). However, the external validity of CPM campaign effects to CPA campaigns is prone to inaccuracies due to different user selection incentives (Berman 2015), and market interactions (Morrison and Coolbirth 2008).

2.3. Proposed Randomized Design

We propose evaluating the overall campaign, including the ad and the campaign presence in the marketplace. This new perspective implies that the campaign is now the *treatment* to evaluate. We randomize the visiting users before any decision has been made in the decision tree shown in Figure 3(c), and keep them in the same group for the campaign duration. As a result, users in the control group are not exposed to placebo ads. This design aggregates the ad and campaign presence in the marketplace effects analyzed in detail below. Our goal for this randomized design is *not* to predict or generalize campaign performance for future long-term exposures, which is the objective of randomized experiments. Our goal is to evaluate campaign performance under the current conditions to attribute credit to its overall performance, which is the key attribution problem of interest to online

[3] A post-treatment variable is a random variable whose realization is available after performing the randomized assignment. As a result, the treatment could affect this realization (Frangakis and Rubin 2002).

Figure 2 **(Color online) Online Targeted Display Advertising Flow for a Given User Visit**



advertisers. In the context of the campaign loop shown in Figure 1, our focus is short-term (mid-flight) ad prediction where both effects are stable.[4]

To disaggregate the proposed design in Figure 3(c), we consider the design of Figure 3(d), where $Z \in \{\text{Control, Placebo, Study}\} = \{C, P, S\}$. To avoid a selection effect, two assumptions of the observed selection in the study and placebo arms need to be tested:

Assumption 1. *Statistically equivalent user selection; the marginal probability of user selection for ad exposure is the same for both treatment arms.*

Assumption 2. *Statistically equivalent selected populations; the marginal conversion probability of the nonselected users for ad exposure is the same for both treatment arms.*

Testing Assumption 1 indicates whether the selection policy (aggregated over user segments) is the same for placebo and study arms. Testing Assumption 2 indicates whether the user selection process (aggregated over user segments) provides statistically equivalent populations based on conversion probabilities. If the nonselected populations are equivalent, in terms of conversion probability, then the complementary populations are statistically equivalent as a consequence of user randomization. Although rejecting Assumption 1 suggests nonequivalent user selection, testing Assumption 2 determines the presence of a selection effect (bias) on the observed conversion data.[5]

Let $Y_i(Z_i)$ be the ith user conversion indicator under the treatment Z_i, and assume Assumption 2 holds. Similarly, assume $P(Y_i(C) \mid D_i = 1, Z_i = C)$ is known for the control group, in which the user selection indicator D_i is not observed; we address this estimation in §3. Thus, the ad average treatment effect $\text{ATE}_{Ad,i}$, and the average treatment effect of the campaign presence in the marketplace $\text{ATE}_{Market,i}$ are defined as follows:

$$\begin{aligned} \text{ATE}_{Ad,i} &= \text{E}[Y_i(S) \mid D_i = 1, Z_i = S] \\ &\quad - \text{E}[Y_i(P) \mid D_i = 1, Z_i = P], \\ \text{ATE}_{Market,i} &= \text{E}[Y_i(P) \mid D_i = 1, Z_i = P] \\ &\quad - \text{E}[Y_i(C) \mid D_i = 1, Z_i = C]. \end{aligned} \tag{1}$$

The proposed randomized design in Figure 3(c) takes the entire campaign as *treatment* and estimates the campaign average treatment effect ($\text{ATE}_{Camp,i}$) as follows:

$$\begin{aligned} \text{ATE}_{Camp,i} &= \text{E}[Y_i(S) \mid Z_i = S] - \text{E}[Y_i(C) \mid Z_i = C] \\ &= \sum_{d \in \{0,1\}} P(D_i = d) \times \{\text{E}[Y_i(S) \mid D_i = d, Z_i = S] \\ &\quad - \text{E}[Y_i(C) \mid D_i = d, Z_i = C]\}. \end{aligned} \tag{2}$$

Given that Y_i is affected only for the users to whom the ad is displayed, i.e., $(D_i = 1)$, all other terms of Equation (2) cancel out. Thus, by substituting for $\text{ATE}_{Ad,i}$ and $\text{ATE}_{Market,i}$ from Equation (1), and defining $\text{ATE}^{D=1}_{Camp,i}$ to be the campaign local effect given $D_i = 1$ we have

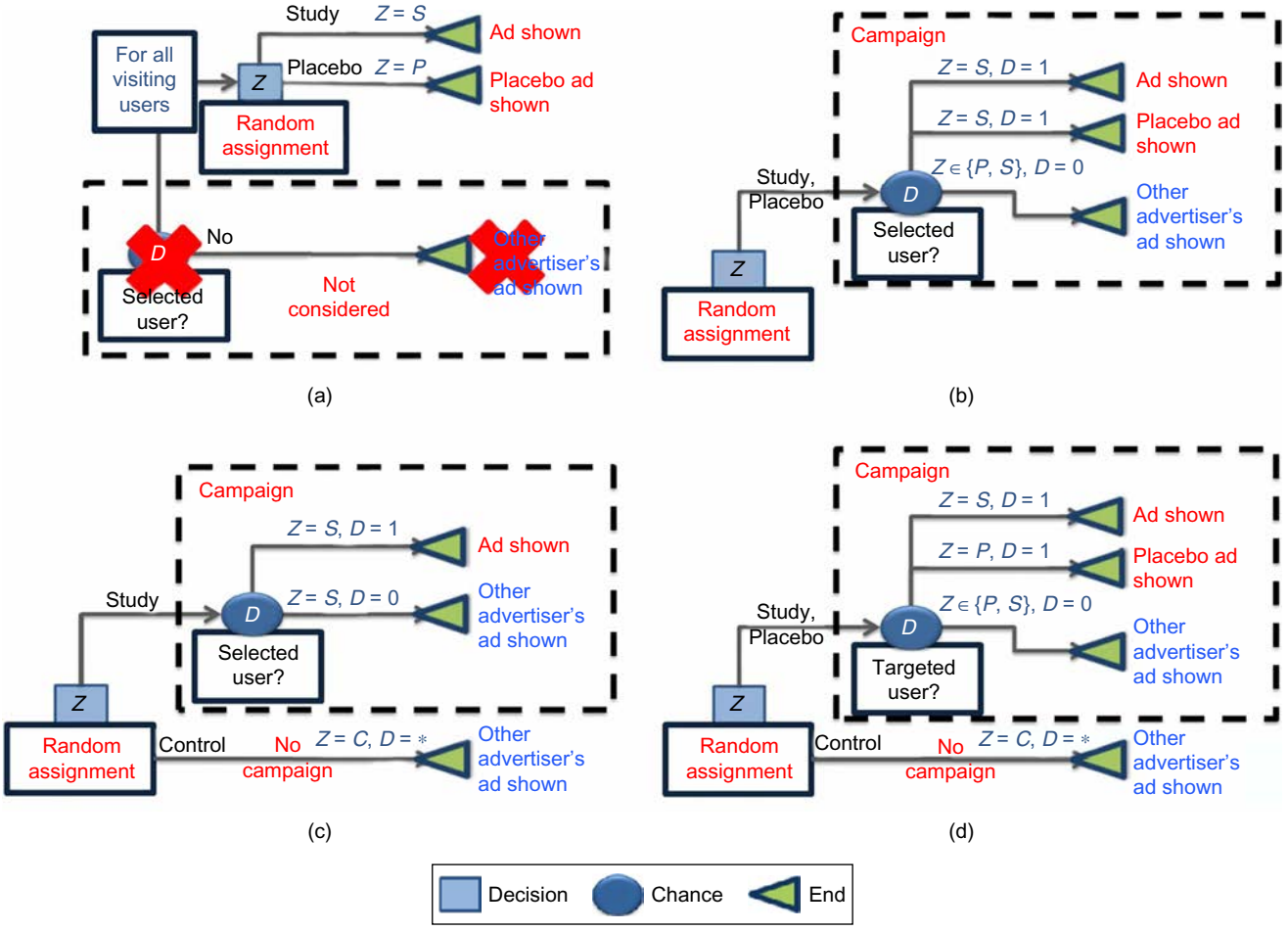
$$\begin{aligned} \text{ATE}_{Camp,i} &= P(D_i = 1) \times \{\text{E}[Y_i(S) \mid D_i = 1, Z_i = S] \\ &\quad - \text{E}[Y_i(C) \mid D_i = 1, Z_i = C]\} \\ &= P(D_i = 1) \times \{\text{ATE}_{Ad,i} + \text{ATE}_{Market,i}\} \\ &= P(D_i = 1) \times \text{ATE}^{D=1}_{Camp,i}. \end{aligned} \tag{3}$$

Therefore, the campaign effect of the proposed design in Figure 3(c) provides the aggregated ad and campaign presence effect, $\text{ATE}^{D=1}_{Camp,i}$. The weighting term, $P(D_i = 1)$, is a consequence of a larger user population considered by the campaign (i.e., all visiting users), rather than the subpopulation of exposed users to whom the ad is displayed.

[4] External validity of effects is implicitly assumed by most evaluation practices in literature (Lewis et al. 2011, Yildiz and Narayanan 2013, Johnson et al. 2016). Given evolving user tastes, marketplace dynamics, and the sequential learning of targeting algorithms, even medium-term effect generalizations could be highly inaccurate.

[5] A typical belief is that user pre-treatment feature based balancing is the only way to show that control and study populations are statistically equivalent (Johnson et al. 2016). However, Assumptions 1 and 2 do not require these features as long as the user assignment is independent of the effect, i.e., random.

Figure 3 **(Color online) (a) User Randomization Framework Proposed by Lewis et al. (2011) Without User Targeting-Engine Selection; (b) Standard Industry Randomization Practice with Placebo Ads; (c) Proposed Randomization Design for Campaign Attribution; (d) Randomization Framework with Disaggregated Campaign Effects**



The standard evaluation using placebo ads identifies $\text{ATE}_{Ad,i}$ as the "campaign" effect. However, the estimation of the economic value (campaign attribution) based on $\text{ATE}_{Ad,i}$ alone does not incorporate $\text{ATE}_{Market,i}$, which is a consequence of displaying the ad. Therefore, the summation of these *two effects*, $\text{ATE}^{D=1}_{Camp,i}$, must be considered. We analyze values of $\text{ATE}_{Market,i}$ for different scenarios in Appendix A. In Appendix B we show that the proposed design has the lowest potential revenue loss when compared with the standard practice, and is the most suitable for continuing evaluation.

Remark 1. The design in Figure 3(c) identifies the right counterfactual to calculate ATE_{Camp} (Equation (2)) when the objective is to estimate the campaign attribution. The design in Figure 3(d) disaggregates ATE_{Camp} into two effects: ATE_{Ad} and ATE_{Market} (Equation (3)).

Remark 2. To estimate ATE_{Market} (Equation (1)), the expected conversion probability of control users who would be exposed to the ad if they were in the study group, $\text{E}[Y_i(C) \mid D_i = 1, Z_i = C]$, has to be inferred (missing D_i if $Z_i = C$). In §3, we address this estimation.

Remark 3. One might believe that the three-arm design described in Figure 3(d) is easily analyzed. We can perceive this model as an extension of the standard randomized experiment of Figure 3(b), which includes a placebo arm. We reiterate that the error in that logic, and the reason for a different counterfactual and estimation method, is that the publisher slot must be captured and assigned to the campaign or placebo ad.

3. Estimation Methodology

Given the randomized design shown in Figure 3(c), the estimation of the campaign attribution is straightforward (Equation (6)). However, by Remark 2, the conversion probability of the users of the control group who would be selected for ad exposure must be inferred. We calculate the local campaign effect on this subpopulation and characterize them based on their response. We develop this methodology in the

Potential Outcomes causal model via the Principal Stratification framework.[6]

3.1. Causal Modeling: Campaign Effect on the Users Exposed to the Ad

The Potential Outcomes Causal Model analyzes the individual potential outcomes for each of the treatments (Rubin 2005). For two treatment arms, this framework implies that half of the data is missing because we never observe a unit response in both arms. If the treatment assignment is independent of the treatment effect (i.e., random assignment), then the causal estimates are unbiased. The Stable Unit Treatment Value Assumption (SUTVA) is necessary for this causal model; it implies that the treatment status of any unit does not affect the potential outcomes of the other units (i.e., no user interference). Also, the user indicator events are modeled to be random and conditionally independent among users given a predetermined probability.

Principal Stratification modeling provides a framework to estimate treatment effects conditional on post-treatment (nonignorable) variables, which might be affected by the treatment (Frangakis and Rubin 2002). The key element in this context is identification of user classes, or strata, with equal treatment effects and probability of treatment assignment. Given the proposed randomized design in Figure 3(c), where $Z_i \in \{\text{Control}, \text{Study}\} = \{C, S\}$, user exposure to the ad is a post-treatment variable. Here, the exposure selection process is performed in the study group and not performed in the control group.[7] Let $W_i(Z_i)$ indicate if the ad is shown to the user ($W_i = 1$) or not ($W_i = 0$) under treatment Z_i. To define the principal strata, we model the potential outcomes for $W_i(Z_i)$, $Z_i \in \{C, S\}$. Because the ad is never shown to the users of the control group ($W_i(C) = 0$), we define the user principal strata, W_i^P, as follows:

$$W_i^P = \left\{ \begin{pmatrix} W_i(C) \in \{0\} \\ W_i(S) \in \{0,1\} \end{pmatrix} \right\} = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\},$$
$$D_i = \begin{cases} 0 & \text{if } W_i^P = (0,0)' \\ 1 & \text{if } W_i^P = (0,1)' \end{cases}. \quad (4)$$

Table 3 shows the observed and missed data in the potential outcomes notation. This definition guarantees that the selection effect in the control group is the same as that of the study group (Assumption 1). In the definition of Equation (4), D_i indicates whether the user is exposed to the ad had the user been assigned to the study group (*exposed-if-assigned*, $D_i = 1$), or not (*never-exposed*, $D_i = 0$). Consequently, we do not observe D_i in the control group (Figure 4(a)).

Table 3 Observed User Counts Based on the User Potential Outcomes

User counts	Potential outcomes: Control		Potential outcomes: Study		Treatment	Principal stratum	
N^y_{dz}	$W_i(C)$	$Y_i(C)$	$W_i(S)$	$Y_i(S)$	assignment Z_i	$(W_i(C), W_i(S))$	D_i
$N^0_{\{0,1\}C}$	0	0	$*$	$*$	C	$(0,*)$	$*$
$N^1_{\{0,1\}C}$	0	1	$*$	$*$	C	$(0,*)$	$*$
N^0_{0S}	0	$*$	0	0	S	$(0,0)$	0
N^1_{0S}	0	$*$	0	1	S	$(0,0)$	0
N^0_{1S}	0	$*$	1	0	S	$(0,1)$	1
N^1_{1S}	0	$*$	1	1	S	$(0,1)$	1

Notes. N^y_{dz}, where $D_i = d$, $Z_i = z$, $Y_i = y$, are user counts for the given values of Y, Z, D. Missing values are presented as $*$.

We define the probability of D_i to be Bernoulli distributed with parameter p_{sel}, and the probability of user conversion Y_i to be Bernoulli distributed with parameters θ_{dz} for the four combinations $D_i = d$, $Z_i = z$, and $Y = \{Y_i\}$, $Z = \{Z_i\}$, $D = \{D_i\}$. Let the user selection for ad exposure indicator for those assigned to the control arm be D_i^C and for those assigned to the study arm be D_i^S. Therefore, assuming $\Theta = \{\theta_{dz}, p_{sel};\ d \in \{0,1\},\ z \in \{C, S\}\}$ are random variables, we have

$$\begin{aligned} &P(Y, Z, D, \Theta) \\ &\quad = P(\Theta) \prod_{i=1}^{|Z|} P(D_i = d \mid p_{sel}) \\ &\qquad \cdot P(Y_i(Z_i) \mid D_i = d, Z_i = z, \theta_{dz}) P(Z_i = z). \end{aligned} \quad (5)$$

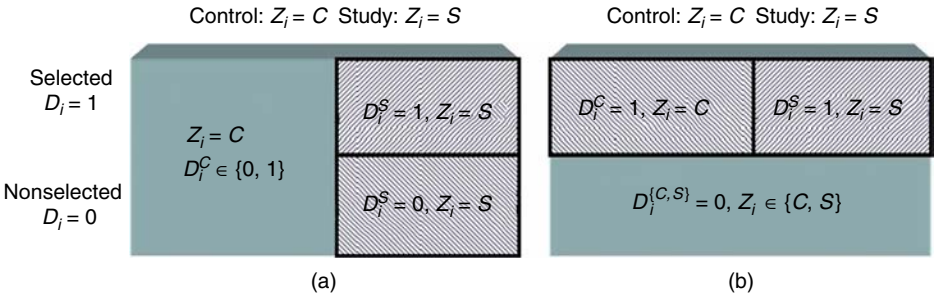
One concern with the model described by Equation (5) is that distribution parameters $\{\theta_{0C}, \theta_{1C}\}$ are not identifiable. That is, for given values of θ_{0C} and θ_{1C}, the same likelihood value ($P(Y, Z, D \mid \Theta)$) is produced if we switch these parameter values. Thus, we require a constraint based on identifiable parameters $\{\theta_{0S}, \theta_{1S}\}$ to guarantee a unique solution. By Assumption 2, the treatment assignment is independent of the potential outcomes of *never-selected* users ($Y_i \perp Z_i \mid D_i = 0$). Therefore, we do not consider any campaign effect on this subpopulation as depicted in Figure 4(b) leading to: $\theta_{0S} = \theta_{0C} = \theta_0 \Rightarrow \Theta = \{\theta_0, \theta_{1C}, \theta_{1S}, p_{sel}\}$.

Note also that, in a sequential setting, the targeting engine uses user conversion probability estimates to determine the selection probability of the next visiting user. However, based on SUTVA and conditionally independent user conversions of each other given a predetermined probability, the user selection for ad exposure indicators D_i; for all i are conditionally

[6] The campaign effect estimation of Equation (6), which is similar to the Intention-to-Treat (ITT) estimation, might be perceived as a noisy estimation. Note that the campaign budget, a crucial decision, is captured in the campaign attribution by this estimator. Also, we emphasize that only the visiting users to a given set of publisher websites are potentially selected for ad exposure, and not all of the online users as stated by Johnson et al. (2016). Although other causal frameworks have been developed, mainly the Structural Equation framework (Heckman 2008), we use Potential Outcomes to model post-treatment variables with experimental data.

[7] The current analysis holds for the randomized design in Figure 3(d) if only the control and study arms are analyzed.

Figure 4 **(Color online) User Segments Based on Control/Study (Z_i) and Nonselected/Selected (D_i) Groups**



Notes. (a) Observed segments. (b) Idealized segments to estimate the campaign effects on the selected users for ad exposure.

independent from each other, given p_{sel}. As a result, p_{sel} represents the aggregate selection probability during the time of the analysis.

Algorithm 1 (Gibbs sampling algorithm based on the joint distribution of Equation (5))

1: **Input:** $N_{obs} = \{N_{dS}^y, N_{\{0,1\}C}^y;\ d \in \{0,1\}, y \in \{0,1\}\}$ from Table 3
2: Define $N_{samp} = \{N_{dC}^y;\ d \in \{0,1\}, y \in \{0,1\}\}$
3: Set $a_0 = 0.5$, $b_0 = 0.5$
4: Initial guess $\Theta^0 = \{\theta_{1z}, \theta_0, p_{sel}\}^0, \quad z \in \{C, S\}$
5: **for** $i \leftarrow 1$ to $N_{burnin} + N_s$ **do**
6: Set $P(D_i^{Cy} = 1 \mid \Theta, N_{obs}) = \dfrac{p_{sel}(\theta_{1C})^y(1-\theta_{1C})^{1-y}}{p_{sel}(\theta_{1C})^y(1-\theta_{1C})^{1-y} + (1-p_{sel})(\theta_0)^y(1-\theta_0)^{1-y}}, \quad y \in \{0,1\}$
7: Draw $N_{1C}^y \mid \Theta, N_{obs} \sim \text{Binomial}(N_{\{0,1\}C}^y, P(D_i^{Cy} = 1 \mid \Theta, N_{obs})), \quad y \in \{0,1\}$
8: Set $N_{0C}^y = N_{\{0,1\}C}^y - N_{1C}^y, \quad y \in \{0,1\}$
9: Draw $\theta_{1z}^{(i)} \mid \Theta_{-\theta_{1z}}, N_{samp}, N_{obs} \sim \text{Beta}(a_0 + N_{1z}^1, b_0 + N_{1z}^0), \quad z \in \{C, S\}$
10: Draw $\theta_0^{(i)} \mid \Theta_{-\theta_0}, N_{samp}, N_{obs} \sim \text{Beta}(a_0 + N_{0C}^1 + N_{0S}^1, b_0 + N_{0C}^0 + N_{0S}^0)$,
11: Draw $p_{sel}^{(i)} \mid \Theta_{-p_{sel}}, N_{samp}, N_{obs} \sim \text{Beta}(a_0 + \sum_{z\in\{C,S\}, y\in\{0,1\}} N_{1z}^y, b_0 + \sum_{z\in\{C,S\}, y\in\{0,1\}} N_{0z}^y)$
12: **end for**
13: **return** $\Theta^{N_{burnin}+1:N_{burnin}+N_s}$

The inference objective of the joint distribution of Equation (5) is to estimate the posterior distribution of the parameters Θ given the observed data from Table 3. Estimating this posterior distribution in closed form is intractable because D^C must be observed. Thus, we implement a Markov Chain Monte Carlo (MCMC)-based approach using Gibbs sampling depicted by Algorithm 1. We denote the set of observed counts as N_{obs} (step 1). Given an initial guess for Θ^0 (step 4), we sample D^C and estimate the counts N_{dC}^y; $d \in \{0,1\}$, $y \in \{0,1\}$ based on the probability of D_i^{Cy} (steps 6–8). We denote these sampled counts as $N_{samp} = \{N_{dC}^y;\ d \in \{0,1\}, y \in \{0,1\}\}$ (step 2). Given the augmented user counts, $\{N_{obs}, N_{samp}\}$, we sample each parameter of Θ conditional on $\Theta_{-\theta}$, which is the set Θ without θ (steps 9–11). The sampling distributions of the parameters $\{\theta_0, \theta_{1C}, \theta_{1S}, p_{sel}\}$, are Beta$(a_0, b_0)$ distributions with Jeffreys conjugate prior parameters, $\{a_0 = 0.5, b_0 = 0.5\}$ (step 3). We test other prior parameters in Appendix C. This sampling process is repeated for $N_{burnin} + N_s$ times (steps 5–12). After discarding a set of burn-in samples, N_{burnin}, a set of samples of the posterior distribution is obtained, $\Theta^{1:N\,samples}$. These samples are used to estimate the variability of the effects and the analysis of §§3.2 and 3.3.

Remark 4. We use the randomization power to estimate the conversion probability of the statistically equivalent users in the control group to those exposed to the ad in the study group. We leverage the fact that there is no campaign effect on the nonselected users. Also, the proportion of users statistically equivalent to those selected in the study group must be the same in both treatment groups. Therefore, the proposed model guarantees that the conversion probabilities are balanced for control and study arms.[8]

3.2. Campaign Effect Estimation

We estimate the average treatment effect by the campaign (ATE_{Camp}) on the overall visiting users and the lift (lift_{Camp}) as follows:

$$\text{ATE}_{Camp} = \text{E}[Y_i(S) \mid Z_i = S] - \text{E}[Y_i(C) \mid Z_i = C],$$
$$\text{lift}_{Camp} = \frac{\text{ATE}_{Camp}}{\text{E}[Y_i(C) \mid Z_i = C]}. \tag{6}$$

Assuming a Jeffreys conjugate prior distribution, $\{a_0 = 0.5, b_0 = 0.5\}$, the posterior distribution becomes Beta$(a_0 + N_z^1, b_0 + N_z^0)$ where N_z^1, N_z^0 are the number of converting and nonconverting users of the z group. We sample from these posterior distributions to provide credible intervals for both ATE_{Camp} and lift_{Camp}.

[8] By focusing on average effect estimates, we obviate the need to predict individual selection for ad exposure indicators, as performed by Johnson et al. (2016), with associated prediction errors.

The campaign average treatment effect on the users selected for ad exposure ($\text{ATE}^{D=1}_{Camp}$), and the lift ($\text{lift}^{D=1}_{Camp}$) are estimated from the posterior distribution of Θ as follows:

$$\begin{aligned}\text{ATE}^{D=1}_{Camp} &= \text{E}(Y_i(S) \mid D_i = 1, Z_i = S) \\ &\quad - \text{E}(Y_i(C) \mid D_i = 1, Z_i = C), \\ \text{ATE}^{D=1}_{Camp} &= \theta_{1S} - \theta_{1C}, \quad \text{lift}^{D=1}_{Camp} = (\theta_{1S} - \theta_{1C})/\theta_{1C}.\end{aligned} \tag{7}$$

Based on the samples $\Theta^{(1:N_s)}$ obtained by the Gibbs sampling procedure in §3.1, credible intervals are estimated from the set $\{\text{ATE}^{D=1}_{Camp}, \text{lift}^{D=1}_{Camp}\}^{(1:N_s)}$.

We estimate the proportion of converting users attributed to the campaign with respect to those in the study group based on ATE_{Camp} and $\text{ATE}^{D=1}_{Camp}$ (ATRB_{Camp}, $\text{ATRB}^{D=1}_{Camp}$)

$$\begin{aligned}\text{ATRB}_{Camp} &= \text{ATE}_{Camp} \times \frac{\sum_{d\in\{0,1\},\, y\in\{0,1\}} N^y_{dS}}{N^1_{0S} + N^1_{1S}}, \\ \text{ATRB}^{D=1}_{Camp} &= \text{ATE}^{D=1}_{Camp} \times \frac{N^0_{1S} + N^1_{1S}}{N^1_{0S} + N^1_{1S}}.\end{aligned} \tag{8}$$

Given that only the users exposed to the ad are impacted by the campaign, these metrics must match. They represent the campaign value (causally generated conversions) and the output of the measurement block shown in Figure 1.

3.3. User Selection Characterization

To characterize user selection for ad exposure of converting users performed by the targeting engine, we estimate the user selection effect (SelEff) and the lift (lift_{sel}) as follows:

$$\begin{aligned}\text{SelEff} &= \text{E}(Y_i(C) \mid D_i = 1, Z_i = C) \\ &\quad - \text{E}(Y_i(C) \mid D_i = 0, Z_i = C), \\ \text{SelEff} &= \theta_{1C} - \theta_0, \quad \text{lift}_{sel} = (\theta_{1C} - \theta_0)/\theta_0.\end{aligned} \tag{9}$$

Note that selecting converting users, whose performance is measured by SelEff, is a common objective of the targeting engine (Pandey et al. 2011) because of the industry business model for CPA campaigns, i.e., last-touch and multitouch attribution (Atlas Institute 2008). Thus, being part of a converting user path is enough to attribute credit to the campaign.

To characterize the causal user selection process, we partition the users into four influenceable categories (Chickering and Heckerman 2000), U_i as follows: Per^+, positively influenced user, *persuadable*; Per^-, negatively influenced user, *anti-persuadable*; AB, converting user with no effect, *always-buy*; NB, nonconverting user with no effect, *never-buy*. Given the selection for ad exposure indicator D_i, the probability of a category U_i is defined as

$$\begin{aligned}P(U_i = \text{Per}^+ \mid D_i, \Theta) &\propto P(Y_i(S) = 1 \mid D_i, Z_i = S, \Theta) \\ &\quad \cdot P(Y_i(C) = 0 \mid D_i, Z_i = C, \Theta), \\ P(U_i = \text{Per}^- \mid D_i, \Theta) &\propto P(Y_i(S) = 0 \mid D_i, Z_i = S, \Theta) \\ &\quad \cdot P(Y_i(C) = 1 \mid D_i, Z_i = C, \Theta), \\ P(U_i = \text{AB} \mid D_i, \Theta) &\propto P(Y_i(S) = 1 \mid D_i, Z_i = S, \Theta) \\ &\quad \cdot P(Y_i(C) = 1 \mid D_i, Z_i = C, \Theta), \\ P(U_i = \text{NB} \mid D_i, \Theta) &\propto P(Y_i(S) = 0 \mid D_i, Z_i = S, \Theta) \\ &\quad \cdot P(Y_i(C) = 0 \mid D_i, Z_i = C, \Theta).\end{aligned} \tag{10}$$

We estimate the probability of selecting a user given U_i by Bayes theorem as follows:[9]

$$\begin{aligned}&P(D_i = 1 \mid U_i, \Theta) \\ &\quad = \frac{P(D_i = 1 \mid \Theta) P(U_i \mid D_i = 1, \Theta)}{\sum_{d\in\{0,1\}} P(D_i = d \mid \Theta) P(U_i \mid D_i = d, \Theta)}, \\ &P(Y_i(Z_i) = y \mid Z_i, D_i, \Theta) = \theta^y_{dz}(1 - \theta_{dz})^{1-y}, \\ &\quad P(D_i = d \mid, \Theta) = p^d_{sel}(1 - p_{sel})^{1-d}.\end{aligned} \tag{11}$$

Remark 5. We estimate the probabilities of *persuadable*, *anti-persuadable*, *always-buy*, and *never-buy* user categories, despite not using user features because we observe the counterfactual user response in both control and study treatment groups.

4. Results

In this section, we discuss data collection and processing. We also validate the model assumptions based on user randomization. Then, we present the analysis of two CPA campaigns (Figure 3(c) design),[10] and one CPM campaign where placebo ads were displayed (Figure 3(d) design). Finally, we analyze the selection for ad exposure policy for these campaigns.

4.1. Data Collection and Description

We ran two large scale randomized (or field) experiments (Figure 3(c) design) collaboratively with two European advertisers in the mobile communications and the public transportation service sectors. The user selection for ad exposure was optimized in real time by a sophisticated targeting engine that valued the user and managed the bidding process for both CPA campaigns. User conversions were economically equivalent for both campaigns. For privacy reasons, we are not allowed to disclose the ad content or the identity of the advertiser.

[9] The campaign effect is not considered for the nonselected users, thus $P(U_i = \text{Per}^+_i \mid D_i = 0) = P(U_i = \text{Per}^-_i \mid D_i = 0)$.

[10] We present a power analysis of the campaign effect estimation in Appendix D, which illustrates the difficulty in measuring this effect in Targeted Advertising even when tens of millions of users are part of the experiment.

Table 4 Campaign Data Based on Notation of Table 3

Count	$N^0_{[0,1)C}$	$N^1_{[0,1)C}$	N^0_{0S}	N^1_{0S}	N^0_{1S}	N^1_{1S}
Campaign 1	1,560,146	400	12,010,058	2,387	5,708,558	2,599
Campaign 2	2,803,640	734	18,681,097	3,170	2,584,728	2,685

Note. Duration for Campaign 1, 30 days, Campaign 2, 28 days.

We randomly assigned the visiting users using the last two digits of the time their cookies were created. This rule separated the users and kept them in their assigned group while the campaign was active. To avoid user contamination and guarantee that we do not miss user tracking due to cookie deletion, we only consider users whose cookies were born before the campaign started and remained active in the ad network.[11]

Given a user timeline of events, we focus on those events recorded after the first visit to any publisher website where the ad was potentially displayed. We mark the user as selected and exposed in the study group ($Z_i = S, W_i = 1, D_i = 1$) if at least one ad exposure was recorded (otherwise $W_i = 0, D_i = 0$). If one conversion was recorded after at least one ad exposure and before the campaign ended, the user is considered to be selected and a converter ($W_i = 1, D_i = 1, Y_i = 1$). No ad exposure was performed for the users in the control group ($Z_i = C, W_i = 0$). Thus, the user selection indicator is missing in this group ($D_i = *$). User counts based on the notation in Table 3 are displayed in Table 4; Table 5 shows user activity statistics.

4.2. Estimation Model Validation

The estimation approach in §3.1 relies on equal probabilities of selection for both treatment groups, and the condition of no campaign effects on the nonselected users for ad exposure (Figure 4(b)). To test these conditions, we randomly partition the users of the study group of the CPA campaigns (Table 4) into simulated control ($Z_i = C$) and study ($Z_i = S$) groups, where D_i^C and D_i^S are observed. We define $p_{sel,z}$ to be the selection for ad exposure probability, p_{sel}, for the z random group. We perform this partition 3,000 times, obtain the method-of-moments (MM) estimate for $\{p^{(s)}_{sel,z}, \theta^{(s)}_{0z}\}$ independently of our proposed model, and calculate the empirical distribution of: $\Delta^{(s)}_{psel} = p^{(s)}_{sel,S} - p^{(s)}_{sel,C}$, $\Delta^{(s)}_{\theta 0} = \theta^{(s)}_{0S} - \theta^{(s)}_{0C}$.

Zero values for Δ_{psel} and $\Delta_{\theta 0}$ verify the conditions of the model (Assumptions 1 and 2). Table 6 reports the credible intervals for these statistics and shows that they are centered at 0 for both campaigns. Therefore, we conclude that p_{sel} is the same for both treatment arms, and that no campaign effect is present in the nonselected users.[12]

4.3. Campaign Effect Results

Figure 5 depicts the estimation results for the CPA campaigns shown in Table 4. Here, we use $N_{burnin} = 2{,}000$ burn-in iterations and $N_s = 10{,}000$ samples for the Gibbs sampling framework of Algorithm 1. As illustrated, the posterior distribution for $\text{lift}^{D=1}_{Camp}$ is skewed because $\text{lift}^{D=1}_{Camp}$ is a ratio of random variables. The posterior distributions for $\{\theta_0, \theta_{1C}, \theta_{1S}\}$ are illustrated by the box plots in Figures 5(a) and 5(b). A significant difference is evident between the conversion rates for the selected for ad exposure (θ_{1C}, θ_{1S}) and the nonselected (θ_0) groups, which is measured by SelEff and lift_{sel} of Equation (9). As indicated by Table 7, we obtain a median $\text{lift}_{sel} = \{89\%, 444\%\}$ for Campaign 1 and 2, respectively.

For comparison, we estimate the campaign effect on the selected users by assuming that we do not observe the control group response, ATE^{I2C}_{Camp}. This naïve estimation is used by last-touch (I2C: impression-to-conversion) or multitouch attribution when only the focal campaign is run (i.e., single channel). Similarly, we estimate the campaign effect without correcting for post-treatment bias, ATE^{post}_{Camp}. These effects are defined as follows:

$$\begin{aligned} \text{ATE}^{I2C}_{Camp} &= \text{E}[Y_i \mid W_i(S) = 1, Z_i = S] \\ &\quad - \text{E}[Y_i \mid W_i(S) = 0, Z_i = S], \\ \text{ATE}^{post}_{Camp} &= \text{E}[Y_i \mid W_i(S) = 1, Z_i = S] \\ &\quad - \text{E}[Y_i \mid W_i(C) = 0, Z_i = C]. \end{aligned} \tag{12}$$

Table 7 shows the campaign effects on the overall user population, ATE_{Camp}, and on the selected for ad exposure population, $\text{ATE}^{D=1}_{Camp}$. Here, the zero effect is not included in the 90% credible intervals for Campaign 1. Campaign 2 leans toward positive values but with a small negative range in the credible interval. In addition, we observe variations of less than 0.2% between median $\text{ATRB}^{D=1}_{Camp}$ and ATRB_{Camp}: {9.05%, 8.90%} for Campaign 1 and {5.05%, 4.91%} for Campaign 2, respectively. This result shows consistency between $\text{ATE}^{D=1}_{Camp}$ and ATE_{Camp}, and confirms the campaign effect analysis of §2.3. Note the severe overestimation by last-touch attribution, and by the effect without correcting for post-treatment bias compared with the causal lift (lift^{I2C}_{Camp} and $\text{lift}^{post}_{Camp}$ versus $\text{lift}^{D=1}_{Camp}$);

[11] We assume that the cookie deletion event is independent of the campaign effect (ignorable or exogenous). Thus, no bias is introduced by focusing on users with stable cookies.

[12] User principal strata are defined based on the observed selection indicator of the study group. Thus, our testing procedure focuses on this group. Generating actual control groups with one less bidder might produce spillover effects among bidders. We assume these (unlikely) effects to be negligible on average. Also, the larger the user population, the more likely this assumption holds.

Table 5 User Activity Statistics for the Campaigns of Table 4

	Campaign 1						Campaign 2					
	$Z_i = C$		$Z_i = S$		$Z_i = S, D_i = 1$		$Z_i = C$		$Z_i = S$		$Z_i = S, D_i = 1$	
Variable	Mean	St dev	Mean	St dev	Mean	St dev	Mean	St dev	Mean	St dev	Mean	St dev
Visits/user	36.25	162.21	36.49	175.98	83.50	332.74	37.32	218.13	37.16	223.23	160.93	637.71
Convs $\mid Y_i = 1$	1.14	0.40	1.19	0.59	1.19	0.59	1.32	0.73	1.33	0.84	1.35	0.87
Imps/user	—	—	—	—	3.47	8.41	—	—	—	—	2.63	5.71

Notes. Mean and standard deviation (St dev) are displayed. *Visits/user* is the number of visits per user. *Convs* $\mid Y_i = 1$ is the number of conversions per converting user. *Imps/user* is the number of ad exposures per selected user ($D_i = 1$).

Table 6 Validation of Model Conditions Expressed by Figure 4(b)

	Campaign 1			Campaign 2				Campaign 1			Campaign 2		
	Low	Med	High	Low	Med	High		Low	Med	High	Low	Med	High
Δ_{psel}(1E–3)	−2.65	0.02	2.71	−0.89	0.01	0.91	$\Delta_{\theta 0}$(1E–5)	−1.71	0.03	1.70	−1.22	0.02	1.23

Notes. The testing procedure is detailed in §4.2. 90% credible intervals are reported. {*Low*, *Med*, *High*} are the {0.05, 0.5, 0.95} quantiles.

that is, for Campaign 1: 129% and 77.54% versus 21.04%; for Campaign 2: 511% and 296% versus 12.36%.[13]

4.4. Comparison with Campaign Evaluation Using Placebo Ads

To illustrate the effect of campaign presence in the marketplace and the risk of conditioning the ad effect on post-treatment (endogenous) variables, we ran a large scale experiment considering three treatment groups, $Z_i \in \{\text{Control, Placebo, Study}\} = \{C, P, S\}$ (Figure 3(d) design), collaboratively with an advertiser in the financial information services sector. We implemented the standard practice to evaluate online campaigns and ran a low-budget CPM campaign, where user conversions are economically equivalent, without optimizing the ad delivery process. Consequently, the decision to bid was always affirmative (Figure 2(a): $B_i = \text{Yes}$), and the auction was run for all visiting users to satisfy the budget contractual schedule. This auction took place inside the ad network where simultaneous campaigns of the same brand were run to market other products, among other competing campaigns. Table 8 shows the aggregated data (Campaign 3) based on the notation in Table 3, and Table 9 shows user activity statistics. To verify that there was no selection effect, we now test Assumptions 1 and 2.

Define the selection indicator D_i under the treatments, $Z_i = \{P, S\}$, to be $\{D_i^P, D_i^S\}$. To estimate the ad effect conditional on the observed D_i^z, we define Δ_i^{select} and $\Delta_i^{convert}$ as

$$\Delta_i^{select} = P(D_i^S = 1 \mid Z_i = S) - P(D_i^P = 1 \mid Z_i = P),$$

$$\Delta_i^{convert} = P(Y_i(S) = 1 \mid D_i^S = 0, Z_i = S) - P(Y_i(P) = 1 \mid D_i^P = 0, Z_i = P). \quad (13)$$

Then, we define the hypotheses: H_0^{select}: $\Delta_i^{select} = 0$, $H_0^{convert}$: $\Delta_i^{convert} = 0$. We test these hypotheses, and estimate their lifts (Δ_i^{select} Lift, $\Delta_i^{convert}$ Lift) by sampling the Beta distribution as in the case of the lift_{Camp} estimation in §3.2.[14] The testing results in Table 10 suggest rejecting H_0^{select} (Δ_i^{select} Lift = [−2.84%, −2.75%, −2.65%]), and not rejecting $H_0^{convert}$ ($\Delta_i^{convert}$ Lift = [−2.80%, 4.12%, 11.41%]). As a result, the change of user selection probability is not enough to reject the assumption that the sampled placebo and campaign populations are equivalent in conversion rates.[15]

We estimate the lift effect of the ad ACL_{Ad}, based on ATE_{Ad} from Equation (1), which is the standard "campaign" attributed effect. We obtain a positively leaning effect ($\text{ACL}_{Ad} = [-2.78\%, 6.74\%, 17.97\%]$). We analyze §3.1 to calculate $E[Y_i(C) \mid D_i = 1, Z_i = C] = \theta_{1C}$, and estimate ATE_{Market} lift, ACL_{Market}, based on Equation (1). We estimate a negative effect of the campaign presence in the marketplace and discard the zero effect of the 90% credible interval ($\text{ACL}_{Market} = [-24.02\%, -15.06\%, -3.70\%]$).

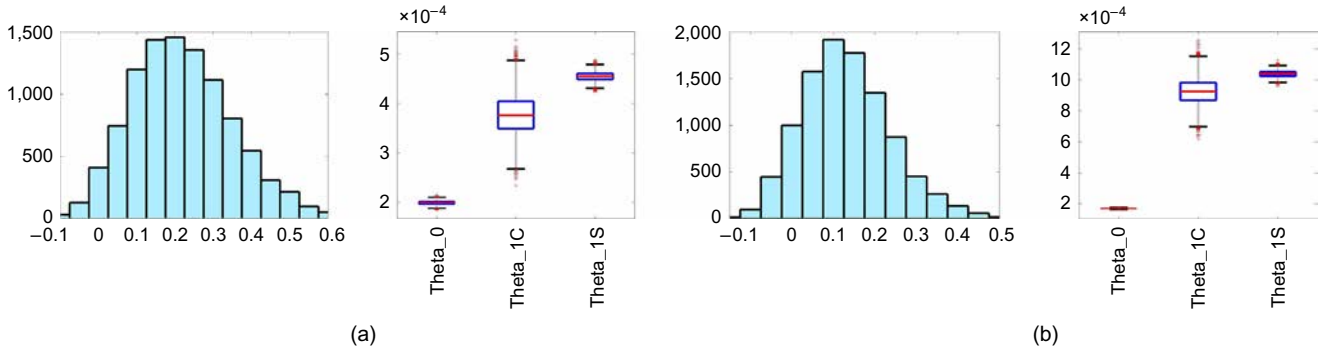
We know that the focal campaign competed in the marketplace against campaigns run to advertise other products of the same brand. We also know that the product being promoted is a free trial of one of the other products. As a result, we expect significant spillover effects from other brand campaigns. In this

[13] Intervals of $\{\text{ATE}_{Camp}^{I2C}, \text{ATE}_{Camp}^{post}\}$ are estimated using their t-statistics. Lifts are the average point estimates. Thus, ATE_{Camp}^{I2C}(1E–4), Campaign 1: [2.40, 2.56, 2.73], Campaign 2: [8.34, 8.68, 9.01]; ATE_{Camp}^{post}(1E–4), Campaign 1: [1.73, 1.99, 2.24], Campaign 2: [7.39, 7.76, 8.13].

[14] We estimate a t-statistic for these conversion probability differences and the results are equivalent. However, estimation of the lifts requires further approximations.

[15] We expect larger effects for CPA campaigns where delivery of placebo ads must be equally optimized.

Figure 5 (Color online) Model Fitting Results for (a) Campaign 1 and (b) Campaign 2



Notes. From left to right, posterior distribution for $\text{lift}^{D=1}_{Camp}$, and the box plot for θ_0, θ_{1C}, θ_{1S} where y-axis is the conversion probability. Gibbs sample size $N_s = 10{,}000$.

Table 7 Attribution Results Using 90% Credible Intervals

	Campaign 1			Campaign 2				Campaign 1			Campaign 2		
	Low	Med	High	Low	Med	High		Low	Med	High	Low	Med	High
lift^{I2C}_{Camp} (%)	—	129	—	—	**511**	—	$\text{lift}^{post}_{Camp}$ (%)	—	77.54	—	—	**296**	—
lift_{Camp} (%)	0.84	9.71	19.61	−1.35	5.15	12.19	ATRB_{Camp} (%)	0.85	**8.90**	16.51	−1.36	**4.91**	10.96
$\text{lift}^{D=1}_{Camp}$ (%)	1.89	**21.04**	46.33	−3.00	12.36	32.43	$\text{ATRB}^{D=1}_{Camp}$ (%)	0.96	**9.05**	16.59	−1.42	**5.05**	11.26
lift_{sel} (%)	55	**89**	126	359	**444**	534							

Note. {*Low*, *Med*, *High*} are the {0.05, 0.5, 0.95} quantiles.

Table 8 Campaign 3 Data (Design of Figure 3(d), $Z_i \in \{C, P, S\}$), Based on Notation of Table 3

Count	$N^0_{\{0,1\}C}$	$N^1_{\{0,1\}C}$	N^0_{0P}	N^1_{0P}	N^0_{1P}	N^1_{1P}	N^0_{0S}	N^1_{0S}	N^0_{1S}	N^1_{1S}
Campaign 3	57,492,247	8,131	9,817,552	1,182	3,713,430	583	9,938,896	1,246	3,618,467	607

Note. Duration, 16 days.

Table 9 User Activity Statistics for Campaign 3 of Table 8

	$Z_i = C$		$Z_i = P$		$Z_i = S$		$Z_i = P, D_i = 1$		$Z_i = S, D_i = 1$	
Variable	Mean	St dev	Mean	St dev	Mean	St dev	Mean	St dev	Mean	St dev
Visits/user	18.18	93.67	18.22	93.40	18.22	94.04	54.42	132.91	54.40	133.12
Convs $\mid Y_i = 1$	1.03	0.36	1.03	0.18	1.04	0.33	1.02	0.16	1.05	0.46
Imps/user	—	—	—	—	—	—	1.68	1.35	1.70	1.39

Notes. Mean and standard deviation (St dev) are displayed. *Visits/user* is the number of visits per user. *Convs* $\mid Y_i = 1$ is the number of conversions per converting user. *Imps/user* is the number of ad exposures per selected user ($D_i = 1$).

Table 10 Campaign Disaggregated Results, and Validation of the Placebo Campaign Based on 90% Credible Intervals

Δ^{select} Lift (%)			$\Delta^{convert}$ Lift (%)			ACL_{Ad} (%)			ACL_{Market} (%)			$\text{lift}^{D=1}_{Camp}$ (%)		
Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
−2.84	−**2.75**	−2.65	−2.80	4.12	11.41	−2.78	**6.74**	17.97	−24.02	−**15.06**	−3.70	−18.88	−**9.15**	2.62

Note. {*Low*, *Med*, *High*} are the {0.05, 0.5, 0.95} quantiles.

scenario, other campaigns generate the user visit (lead) to the advertiser website, where the users are more likely to sign up for a free trial product than for the promoted paid service. Therefore, the mere presence of the focal campaign prevented the other ads of the same brand from being displayed. This strategic effect significantly moves the net campaign effect ($\text{lift}^{D=1}_{Camp} = [-18.88\%, -9.15\%, 2.62\%]$). Similar spillovers across product campaigns have been detected before by Sahni et al. (2015) in the context of email coupon promotions. Note that the user selection effect of this CPM campaign significantly contributes to the negative presence effect. However, this user selection can be improved to identify the positively influenceable population.

4.5. User Selection Characterization Results

Table 11 shows the user selection characterization results, based on the analysis of §3.3, for CPA Campaigns 1 and 2 in Table 4, and for CPM Campaign 3 in Table 8. The probability of *never-buy* users is large in the selected population ($P(U_i = \text{NB} \mid D_i = 1) > 0.99$ for all campaigns); this is a consequence of low conversion rates. Using Bayes theorem as in Equation (11), we observe that the probability of selecting a *never-buy* user is the lowest as there is no incentive to display the ad to this user category ($P(D_i = 1 \mid U_i = \text{NB}) = \{0.32, 0.12, 0.27\}$ for Campaign $\{1, 2, 3\}$, respectively). Similarly, the probability of selecting a *persuadable* user is significantly lower for CPM Campaign 3 than for CPA Campaigns 1 and 2 by as much as 37% ($0.52 - 0.33 = 0.19$ with respect to 0.52, where $P(D_i = 1 \mid U_i = \text{Per}^+) = \{0.52, 0.46, 0.33\}$ for campaigns $\{1, 2, 3\}$, respectively), showing the positive effect of optimized user selection of ad exposure.

As discussed in §3.3, lift_{sel} provides the conversion probability change in the selected population (i.e., selection effect). The CPA last-touch business model suggests that increasing this difference is beneficial for the overall campaign effect. We find that Campaign 2 performance ($\text{lift}_{sel} = 444\%$) is superior to Campaign 1 ($\text{lift}_{sel} = 89\%$) under the CPA policy of selecting converting users. However, we estimate a significantly larger probability of selecting an *always-buy* user for Campaign 2 than for Campaign 1 ($P(D_i = 1 \mid U_i = \text{AB}) = \{0.82, 0.67\}$ for campaigns $\{1, 2\}$, respectively). Therefore, although Campaign 2 is more useful in optimizing user conversions than Campaign 1 by a factor of five (444% vs. 89%), Campaign 2 is 22% ($0.82 - 0.67 = 0.15$ with respect to 0.67) and more likely to select *always-buy* users. This analysis shows that the well accepted policy of selecting users with the highest conversion probability does not necessarily improve the campaign value to the advertiser. Moreover, we find that this probability of selecting *always-buy* users is as much as **96%** larger for CPA Campaign 2 when compared to CPM Campaign 3 ($0.82 - 0.418 = 0.402$ with respect to 0.418). This evidence shows that CPA campaigns incentivize the selection for ad exposure of *always-buy* users when compared with CPM campaigns (Berman 2015). Also, the generalization of the ad effect estimated for a CPM campaign to a CPA campaign, assumed under the standard evaluation practice, is highly prone to inaccuracies.

By analyzing the marginal probabilities $P(U_i)$, we note that the population size of *always-buy* users is three to four orders of magnitude smaller than the size of the *persuadable* and *anti-persuadable* user segments. As a result, the impact of a large $P(D_i = 1 \mid U_i = \text{AB})$ is attenuated by the visiting population size of *always-buy* users.

5. Campaign Mid-Flight Optimization: Leveraging User Features

5.1. Methodology

We illustrate the value of continuing evaluation in the context of Figure 1 by leveraging user features (X_i) in the effect estimation. We develop user selection rules to achieve mid-flight campaign optimization. Here, we replace the Bernoulli distributions in Equation (5) with probit regressions conditional on X_i. Thus, we estimate the campaign effects conditional on X_i to guide the targeting engine. Let $\Phi(x)$ be the standard Normal cumulative density function, $X = \{X_i\}$, and $\Theta_X = \{\gamma_0, \gamma_{1C}, \gamma_{1S}, \beta_{sel}\}$, then

$$
\begin{aligned}
&P(Y, Z, D, \Theta_X \mid X) \\
&\quad = P(\Theta_X) \prod_{i=1}^{|Z|} P(D_i = d \mid \beta_{sel}, X_i) \\
&\qquad \cdot P(Y_i(Z_i) \mid D_i = d, Z_i = z, \gamma_{dz}, X_i) P(Z_i = z), \qquad (14) \\
&P(D_i \mid \beta_{sel}, X_i) = \Phi(\eta_i^{\beta}), \quad \eta_i^{\beta} = X_i' \beta_{sel}, \\
&P(Y_i \mid D_i, Z_i, \gamma_{dz}, X_i) = \Phi(\eta_i^{dz}), \quad \eta_i^{dz} = X_i' \gamma_{dz}.
\end{aligned}
$$

This model exploits the power of randomization and balances the treatment groups in the inference of the indicator $D_i^C \mid X_i$ based on the propensity of being selected.[16]

We find the user counts of Control and Study treatment arms (Table 3) for all user feature combination segments, which are assumed to be finite and countable

$$
\begin{aligned}
N^{obs}_{Camp} = \{&N^{y}_{dS} \mid X_i, N^{y}_{\{0,1\}S} \mid X_i; \\
&d \in \{0, 1\}, y \in \{0, 1\}, X_i \in \{X\}\}, \qquad (15)
\end{aligned}
$$

whose cardinality becomes $\#\{N^{obs}_{Camp}\} = 6 \times \#\{X\}$. To estimate the model described by Equation (14), we

[16] Johnson et al. (2016) balance the user features in the prediction of the selection indicator first, and then compare this prediction with the selected users in the study group. In our approach, we model user randomization and user feature balancing jointly, which is more powerful than stepwise model fittings.

Table 11 User Selection Median Probabilities Based on Equations (10)–(11)

	Campaign 1	Campaign 2	Campaign 3		Campaign 1	Campaign 2	Campaign 3
$P(U_i = \text{Per}^+ \mid D_i = 1)$	**4.55e–4**	**1.04e–3**	1.68e–4	$P(U_i = \text{Per}^+)$	2.81e–4	2.75e–4	1.37e–4
$P(U_i = \text{Per}^- \mid D_i = 1)$	**3.76e–4**	**9.24e–4**	1.85e–4	$P(U_i = \text{Per}^-)$	2.56e–4	2.62e–4	1.41e–4
$P(U_i = \text{AB} \mid D_i = 1)$	1.71e–7	9.59e–7	3.11e–8	$P(U_i = \text{AB})$	8.19e–8	1.42e–7	1.98e–8
$P(U_i = \text{NB} \mid D_i = 1)$	0.9992	0.9980	0.9996	$P(U_i = \text{NB})$	0.9995	0.9995	0.9997
$P(D_i = 1 \mid U_i = \text{Per}^+)$	**0.5211**	0.4583	**0.3276**	$P(D_i = 1 \mid U_i = \text{AB})$	**0.6728**	**0.8217**	**0.4180**
$P(D_i = 1 \mid U_i = \text{Per}^-)$	0.4732	0.4296	**0.3497**	$P(D_i = 1 \mid U_i = \text{NB})$	**0.3221**	0.1215	0.2669

Note. Campaigns 1 and 2 are CPA (optimized selection for ad exposure), and Campaign 3 is CPM (nonoptimized selection for ad exposure).

propose a variant of the Gibbs sampling of Algorithm 1, depicted by Algorithm 2 and detailed in Appendix E. We calculate posterior credible intervals of the effect estimates based on the set of Gibbs samples returned by Algorithm 2, $\Theta_X^{(1:N_s)}$.

5.2. User Selection Optimization Results

We leverage demographic user features to optimize user selection for ad exposure mid-flight, i.e., in the middle of the campaign. For visiting users of CPM Campaign 3, we know the gender, age, and income. These features are segmented by ranges to make them finite and countable (e.g., *Male*, 35–44 *years old*, 50,000–75,000 *income*). We partition the campaign data in duration by half and train the model in Equation (14) using Algorithm 2 for the first half. During the second half of the campaign, we test different user selection policies based on user response categories from Equation (10) for each user segment X_i.

To simulate a given selection policy, we execute Algorithm 3, which requires: (1) a selection function $F_{sel}(X_i)$; (2) a non-zero effect indicator function $F_{sig}(X_i)$, and (3) $\text{ATE}_{Camp}^{D=1}$ sign function $F_{sign}^{ATE}(X_i)$. We discuss this simulation in detail in Appendix F. Given the posterior samples ($\Theta_X^{N_{burnin}+1:\ N_{burnin}+N_s}$), we estimate the median probability of influenceable user categories ($P(U_i \mid D_i = 1, X_i)$) in the ad-exposured population. We avoid classifying the users into these categories because this approach requires a set of fine-tuned thresholds. We choose $F_{sel}(X_i)$ as the ratio of probabilities of desirable over nondesirable classes as indicated in Table 12. We define $F_{sig}(X_i)$ as the inclusion ($F_{sig}(X_i) = \textbf{true}$) or noninclusion ($F_{sig}(X_i) = \textbf{false}$) of the zero $\text{ATE}_{Camp}^{D=1} \mid X_i$ effect in the 90% credible intervals. Similarly, we set $F_{sign}^{ATE}(X_i)$ as the sign of $\text{ATE}_{Camp}^{D=1} \mid X_i$. The intuition behind these functions is to incorporate the degree of uncertainty of the estimated average campaign effects for each user segment. We fix $\mathbf{w}^{sig} = \{w^-, w^{\pm}, w^+\}$ to be a set of certainty weights. These weights are chosen based on whether median $\text{ATE}_{Camp}^{D=1} \mid X_i$ is positive or negative and whether a zero effect lies outside or inside the credible interval. This process generates the count set in Table 3, which we use to run Algorithm 1 to estimate $\text{lift}_{Ad}^{D=1}$ (Equation (7)).

Table 12 shows the results of testing four selection policies. Our benchmark is the optimization of conversion probability ((d) $Y = 1$ vs. $Y = 0$, $\mathbf{w}^{sig} = \{1, 1, 1\}$). This selection policy is the standard industry practice given observational data. Results show that this practice is reasonably effective compared with other policies ((d) 11.72% versus (b) 11.93% or (a) 9.86% average $\text{lift}_{Camp}^{D=1}$).[17] However, the highest performance is achieved when we optimize (c) Per^+ versus $\neg\text{Per}^+$ given $\mathbf{w}^{sig} = \{1, 1, 1\}$ (14.28% $\text{lift}_{Camp}^{D=1}$). Note that user selection of the current CPM campaign is exploratory; consequently the selection effect is significantly smaller than the one for CPA campaigns. Hence, the performance of the standard practice (d) is likely to be inferior to the one we report when CPA campaign data is used to fit the prediction model. We test three weighting frameworks based on the 90% credible intervals of $\text{ATE}_{Camp}^{D=1} \mid X_i$. Intuition suggests that eliminating segments with negative-only intervals and boosting segments with positive-only intervals would dramatically increase the performance. However, we find that a modest decrease of negative-only and an increase of positive-only segment intervals are more effective. We find that $\mathbf{w}^{sig} = \{0.8, 1, 1.1\}$ shows the highest performance of the weighting frameworks we test ((c) 14.89% average $\text{lift}_{Camp}^{D=1}$). Optimizing $\mathbf{w}^{sig}$ represents a line for further research.

The current analysis demonstrates the value of the experimental design and the effect estimation to optimize the user selection in Figure 1. Limitations of this study include: the quality of the cookie-based user features, the percentage of users with missing features estimated to be at 75%, and the assumptions of the selection policy simulation.

6. Conclusion and Managerial Implications

We have shown that evaluating an online advertising campaign involves more than evaluating just the ad.

[17] Credible intervals are in the range of ±20% for all selection functions evaluated. The short evaluation time, seven days, and the observed budget, which is kept constant in the simulation, are among the reasons.

Table 12 Averaged Campaign Effect Results, $\text{lift}^{D=1}_{Camp}$ (%), for Different Selection Functions Based on Algorithm 3 Using the First Half of Campaign 3 as Training and Testing in the Second Half

Selection function, $F_{sel}(X_i)$	$\mathbf{w}^{sig} = \{1, 1, 1\}$	$\mathbf{w}^{sig} = \{0.6, 1, 1.1\}$	$\mathbf{w}^{sig} = \{0.8, 1, 1.2\}$	$\mathbf{w}^{sig} = \{0.8, 1, 1.1\}$
(a) $\frac{P(U_i = \text{Per}^+ \mid D_i = 1, X_i)}{P(U_i = \text{Per}^- \mid D_i = 1, X_i)}$	9.86	8.53	11.49	14.42
(b) $\frac{P(U_i = \text{Per}^+ \mid D_i = 1, X_i)}{P(U_i = \text{Per}^- \cup \text{AB} \mid D_i = 1, X_i)}$	11.93	10.80	12.52	12.63
(c) $\frac{P(U_i = \text{Per}^+ \mid D_i = 1, X_i)}{1 - P(U_i = \text{Per}^+ \mid D_i = 1, X_i)}$	**14.28**	13.40	14.74	**14.89**
(d) $\frac{P(Y_i = 1 \mid D_i = 1, Z_i = S, X_i)}{P(Y_i = 0 \mid D_i = 1, Z_i = S, X_i)}$	**11.72**	—	—	—

Notes. Selection policies: (a) Per^+ vs. Per^-, (b) Per^+ vs. $\{\text{Per}^- \cup \text{AB}\}$, (c) Per^+ vs. $\neg\text{Per}^+$, (d) $Y = 1$ vs. $Y = 0$. Second half campaign duration, 7 days.

Marketplace interactions imply that the final decision to display the campaign/placebo ad is not entirely controllable (i.e., endogenous) in the randomized experiment. We have discussed (§2.3) and demonstrated this endogeneity with the evaluation of a campaign using placebo ads in §4.4. We do not expect that an ad tested in a controlled environment, as assumed by the exploratory evaluation of CPM campaigns, will yield the same performance in a real marketplace. As demonstrated in §2.3 and supported with the results in §4.4, the effects of being in the marketplace are ineluctable if the ad is to be displayed. Consequently, the right placebo is the complete absence of the campaign. Given the difficulty of predicting the ad effect in marketplaces, the most suitable approach is to assign credit to the overall campaign for the time it is run and rely on short-term effect predictions. Therefore, the randomized experiment and the effect estimation together become a measuring tool. Further research involves determining the optimal time span of these effect predictions.

We have illustrated how the strategic campaign presence effect reveals other competing campaigns effects on the focal brand sales. We have analyzed a particular instance where the standing brand campaigns are more beneficial than the new focal campaign. These potentially significant spillover effects provide evidence to determine the right strategic settings to run the campaign. These settings include moderating campaign interactions, adjusting the user reach of the focal campaign, and defining the user selection policy. Explicitly accounting for competitors ad exposures is a further line of research.

By characterizing the user population selected for ad exposure in the CPM and CPA campaigns in §4.5, we have demonstrated that the purported external validity of ad effects tested under CPM selection policy to CPA selection may be invalid. In CPA campaigns, the decision to select users for ad exposure is often driven by the user propensity to convert. As a result, we have found evidence that CPA campaigns incentivize the selection of users who would buy in any case. Selecting these noninfluenceable users does not add any value to the advertiser. We note that ad networks obtain revenue based on user conversions, as in the case of Last-Touch or Multitouch Attribution. On the other hand, purely nonoptimized CPM campaigns are less effective than CPA campaigns in selecting users with positive effect. The current results provide a potential opportunity for advertisers to act on and improve the user selection policy to improve causal estimates.

We have demonstrated the value of characterizing the user selection for ad exposure, and the leverage of user features to improve this selection. In a measurement-optimization cycle, the proposed randomized design may enable the transfer of learning from attribution to user targeting and ex-ante optimization. However, the dynamic campaign effects must be analyzed and understood to achieve an effective fast in-flight campaign optimization. Overall, this assessment takes us a step closer to a commercially valuable use of the experimental data in the user targeting and bidding processes.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2016.0982.

Acknowledgments

The authors thank the anonymous reviewers and the associate editor for their constructive comments and the patience to improve this paper. The authors also thank Jaimie Kwon, Victor Andrei, Professor Philip B. Stark, and James G. Shanahan for their contribution to this paper. This work is partially funded by CONACYT UC-MEXUS [Grant 194880], CITRIS, NSF IIP-0934364 [Subaward 0000015277], and AOL Faculty Award.

Appendix A. Effect of the Campaign Presence in the Marketplace Analysis

Based on the three-arm design in Figure 3(d), $Z_i \in \{\text{Control}, \text{Placebo}, \text{Study}\} = \{C, P, S\}$, we define $\pi_i(Z_i)$ as the competitors' selection for ad exposure policy. Let $\pi_{0,i}$ denote the competitors policy if the focal campaign does not exist ($\pi_i(C) = \pi_{0,i}$). Let $\pi_{1,i}$ be the alternative policy competitors execute with probability α as a consequence of the campaign

presence in the marketplace. If competitors are not interested in user i with probability $1-\alpha$, they will not compete to select this user and $\pi_i(Z_i)=\pi_{0,i}$; $Z_i\in\{P,S\}$. Let β represent the probability that competitors would win the opportunity to advertise in the control group but lose against the focal or placebo campaigns, and their ads have an effect on Y_i. These definitions lead to the distributions

$$P(\pi_i(Z_i)=\pi_{0,i}\mid Z_i)=\begin{cases}1 & \text{if } Z_i=C\\ 1-\alpha & \text{if } Z_i\in\{P,S\},\end{cases}$$
$$P(\pi_i(Z_i)=\pi_{1,i}\mid Z_i)=1-P(\pi_i(Z_i)=\pi_{0,i}\mid Z_i), \quad \text{(A1)}$$
$$P\{E[Y_i(C)\mid \pi_i(C)=\pi_{0,i}]-E[Y_i(P)\mid \pi_i(P)=\pi_{1,i}]\neq 0\}=\beta.$$

The parameter $\beta\in[0,1]$ is related to $\alpha\in[0,1]$ through a competitors policy change function, $\beta=f_\pi(\alpha)\in[0,1]$. Similarly, the effect $\text{ATE}_{Market,i}$ is related to β based on a competitors effect function, $\text{ATE}_{Market,i}=f_{ATE}(\beta)\in[-1,1]$. Some individual cases include (proof of these cases is trivial based on Equation (A1)):

- $\alpha=0\Rightarrow\beta=f_\pi(0)=0\Rightarrow\text{ATE}_{Market,i}=0$: average competitors policy is not affected by the campaign.
- $\alpha>0\wedge\beta=f_\pi(\alpha)=0\Rightarrow\text{ATE}_{Market,i}=0$: competitors advertising will not have any effect on Y_i.
- $\beta=f_\pi(\alpha)>0\Rightarrow\alpha>0$: A competitors effect greater than zero is likely only if the focal campaign is likely to affect their average ad delivery policy.
- $\beta=f_\pi(\alpha)>0\Leftrightarrow\text{ATE}_{Market,i}\neq 0$: An average campaign presence effect implies a non-zero probability of competitors effect on Y_i and vice versa.

Appendix B. The Cost of the Randomized Design

We analyze the cost of the proposed design of Figure 3(c) where no placebo ad is displayed, and $Z_i\in\{\text{Control, Study}\}=\{C,S\}$. Let $N_{D=1}$ be the number of users for whom the opportunity to advertise is won. For the control group, there is a potential revenue loss, proportional to the campaign effect value on this subpopulation ($Val(\text{ATE}^{D=1}_{Camp,i})$), if these users were exposed to the ad. Because no ad impression is displayed to these users, a campaign budget surplus remains from not displaying these ads ($Cost$(AdDisplay)). Thus, the design cost ($Cost$(Design)) becomes

$$Cost(\text{Design})=P(Z_i=C)\times N_{D=1}\times[Val(\text{ATE}^{D=1}_{Camp,i})-Cost(\text{AdDisplay})]. \quad \text{(B1)}$$

Note that $P(Z_i=C)\times N_{D=1}\times Cost(\text{AdDisplay})$ represents a budget surplus for not showing the campaign ad to the users of the control group. If this budget surplus is used to display campaign ads to a larger population in the study group, we have $\text{ATE}^{D=1,\Delta}_{Camp,i}$ as the average campaign effects on these additional exposed users. As a result, the design cost ($Cost(\text{Design}^\Delta)$) results in

$$Cost(\text{Design}^\Delta)=P(Z_i=C)\times N_{Exp}\times Val(\text{ATE}^{D=1}_{Camp,i}-\text{ATE}^{D=1,\Delta}_{Camp,i}). \quad \text{(B2)}$$

Let $\text{ATE}^{D=1}_{Camp,i}-\text{ATE}^{D=1,\Delta}_{Camp,i}=\epsilon$. Given an optimal user selection policy, where the users with highest potential causal impact are most likely to be selected, then $\epsilon>0$ and $\epsilon\ll\text{ATE}^{D=1}_{Camp,i}$. Therefore, the cost of experimentation is reduced to a *function of a small number*: ϵ. Note that the larger $P(Z_i=C)$, the larger the effect difference ϵ.

Appendix C. The Prior Probability and a Method of Moments: Robustness Checks

Given the Bayesian method from §3, we analyze the effect of different Beta prior parameters and compare them with a method of moments that is derived now. Since D_i is observed for the study group, the estimation of p_{sel} and θ_{1S} in the study group is straightforward based on the method of moments. Similarly, θ_0 is approximated using the observed conversions of the nonselected users for ad exposure in the study group. As the observed conversion probability of the control group is a mixture of θ_0 and θ_{1C} weighted by $1-p_{sel}$ and p_{sel}, respectively, and $\{\theta_0,p_{sel}\}$ are shared by both arms (approximation), the estimation of θ_{1C} becomes

$$\hat{p}_{sel}=\frac{N^1_{1S}+N^0_{1S}}{N^1_{1S}+N^0_{1S}+N^1_{0S}+N^0_{0S}},\quad \hat{\theta}_{1S}=\frac{N^1_{1S}}{N^1_{1S}+N^0_{1S}},$$
$$\hat{\theta}_0=\frac{N^1_{0S}}{N^1_{0S}+N^0_{0S}},\quad \hat{\theta}_{1C}=\frac{1}{\hat{p}_{sel}}\left[\frac{N^1_{\{0,1\}C}}{N^1_{\{0,1\}C}+N^0_{\{0,1\}C}}-\hat{\theta}_0(1-\hat{p}_{sel})\right]. \quad \text{(C1)}$$

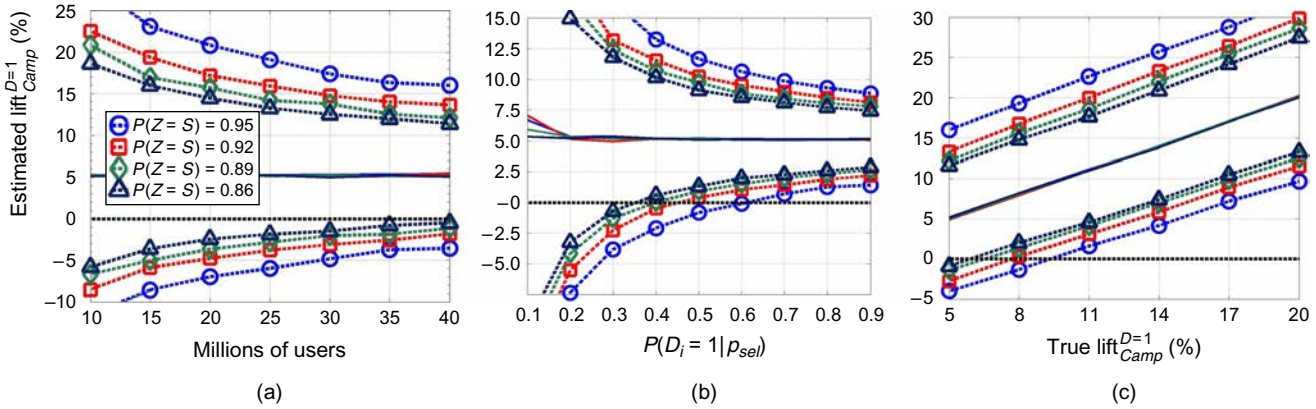
This approach does not account for the data sample size and requires several approximations. Despite these limitations, we provide a robustness check based on this estimator. Table C.1 compares this point estimator with the Bayesian method from §3 for different prior rates: $a_0/(a_0+b_0)$; assuming a prior sample size: $a_0+b_0=1$. Results show that more intuitive prior rate choices for low conversion rates $\{0.01,0.001\}$ do not affect results more than 0.9% in median $\text{lift}^{D=1}_{Camp}$ and its credible interval. We use the Jeffreys prior $\{a_0=0.5,b_0=0.5\}$ because increasingly skewed prior distributions are more likely to be numerically unstable in the Gibbs sampling. The method of moments of Equation (C1) shows discrepancies of less than 1% $\text{lift}^{D=1}_{Camp}$ when compared with this prior choice.

Table C.1 **Prior Rate Effect on the $\text{lift}^{D=1}_{Camp}$ (%) Estimation Given a Prior Sample Size: $a_0+b_0=1$, Based on Algorithm 1, Compared with the Method of Moments of Equation (C1) (Moments)**

Prior rate	Campaign 1			Campaign 2			Campaign 3		
$a_0/(a_0+b_0)$	Low	Med	High	Low	Med	High	Low	Med	High
0.5	2.15	21.15	46.51	−2.64	12.10	31.10	−19.09	−9.32	2.79
0.01	2.63	21.89	47.20	−2.26	12.99	31.86	−18.98	−9.01	3.10
0.001	2.51	21.53	47.55	−2.34	12.57	31.37	−19.58	−9.48	2.51
Moments	—	20.55	—	—	11.99	—	—	−9.59	—

Notes. $N_{burnin}=2{,}000$. $N_s=10{,}000$. {*Low*, *Med*, *High*} are the {0.05, 0.5, 0.95} quantiles.

Figure D.1 **(Color online) Estimation Power as a Function of (a) Total User Population in Millions, (b) User Selection Probability, (c) Campaign Lift on the Users Selected for Ad Exposure (%)**



Notes. 90% credible intervals are displayed. y-axis represents estimated $\text{lift}^{D=1}_{Camp}$. Parameters: Population 37,158,296, $\theta_{1C} = 1.48\text{E}{-3}$, $\theta_{1S} = 1.56\text{E}{-3}$, $\text{lift}^{D=1}_{Camp} = 5.40\%$, $P(Z_i = S) = 0.95$, $p_{sel} = 0.3$, $\theta_0 = 1\text{E}{-3}$.
Source. Lewis et al. (2011).

Appendix D. Estimation Power Analysis

We have observed in major firms that the proportion of users used as a control group is intuitively determined based on the belief that large user populations are readily available. However, in the targeted advertising framework we study in this paper, poorly designed experiments lead to wide credible intervals containing the zero effect. Given the parameter values in Figure D.1, we estimate $\text{lift}^{D=1}_{Camp}$ as a function of the total user population, the user selection probability p_{sel}, and a set of true $\text{lift}^{D=1}_{Camp}$ values. We generate the counts in Table 3 assuming that the point estimate from Equation (C1) is perfect. Given these count sets, we fit the model using the Bayesian approach from §3. Figure D.1(a) shows that even when the user population is 40 million, the credible interval includes zero for all of the randomized designs analyzed, $P(Z_i = S) = \{0.95, 0.92, 0.89, 0.86\}$. If we naïvely set 5% of the users as the control group ($P(Z_i = S) = 0.95$), a typical industry practice, the experiment will be useless. When the user selection probability is $p_{sel} = 0.4$, we observe that the zero effect is discarded of the 90% credible interval when 11% ($P(Z_i = S) = 0.89$) or higher user population is used as the control group, which is depicted by Figure D.1(b). Figure D.1(c) shows that true $\text{lift}^{D=1}_{Camp}$ values as low as 6% are detected when 14% ($P(Z_i = S) = 0.86$) of users are assigned to the control group. This analysis indicates the need to perform a similar analysis at the time of designing the experiment.

Appendix E. Model Fitting Based on User Features

To estimate the model of Equation (14), we provide a variant of the Gibbs sampling of Algorithm 1 depicted by Algorithm 2. We obtain the user counts in Table 3 for all user feature combination segments, assumed to be finite and countable

Algorithm 2 (Gibbs sampling algorithm based on the joint distribution of Equation (14))

Input: $N_{obs} \mid X_i = \{N^y_{dS} \mid X_i, N^y_{\{0,1\}S} \mid X_i;\ d \in \{0,1\},\ y \in \{0,1\}\}$ from Table 3, $X_i \in \{X\}$
Define $N_{samp} \mid X_i = \{N^y_{dC} \mid X_i;\ d \in \{0,1\},\ y \in \{0,1\}\},\ X_i \in \{X\}$
Initial guess $\Theta^0_X = \{\gamma_0, \gamma_{1z}, \beta_{sel}\}^0, \quad z \in \{C, S\}$
for $i \leftarrow 1$ to $N_{burnin} + N_s$ **do**
  Set $P(D_i = d \mid \beta_{sel}, X_i) = (\Phi(\eta^\beta_i))^d (1 - \Phi(\eta^\beta_i))^{1-d}, \quad \eta^\beta_i = X_i' \beta_{sel}, \quad X_i \in \{X\}$
  Set $P(Y_i(Z_i) = y \mid D^z_i = d, Z_i = z, \gamma_{dz}, X_i) = (\Phi(\eta^{\gamma_{dz}}_i))^y (1 - \Phi(\eta^{\gamma_{dz}}_i))^{1-y}, \quad \eta^{dz}_i = X_i' \gamma_{dz}, \quad X_i \in \{X\}$
  Set $P(D^{Cy}_i = 1 \mid \Theta_X, D^s, Y, Z, X_i) = \dfrac{P(D_i = 1 \mid \beta_{sel}, X_i) P(Y_i(C) = y \mid D_i = 1, Z_i = C, \gamma_{dz}, X_i)}{\sum_{d \in \{0,1\}} P(D_i = d \mid \beta_{sel}, X_i) P(Y_i(C) = y \mid D_i = d, Z_i = C, \gamma_{dz}, X_i)}, \ X_i \in \{X\}$
  Draw $N^y_{1C} \mid \Theta_X, N_{obs}, X_i \sim \text{Binomial}(N^y_{\{0,1\}C} \mid X_i, P(D^{Cy}_i = 1 \mid \Theta, N_{obs}, X_i)), \quad y \in \{0,1\}, X_i \in \{X\}$
  Set $N^y_{0C} \mid X_i = N^y_{\{0,1\}C} \mid X_i - N^y_{1C} \mid X_i, \quad y \in \{0,1\}, X_i \in \{X\}$
  Set $\{\hat{\gamma}_{1z}, \hat{\Sigma}_{1z}\} = \text{glmfit}([N^1_{1z} \mid X_i, N^0_{1z} \mid X_i],\ X_i \in \{X\}), \quad z \in \{C, S\}$
  Set $\{\hat{\gamma}_0, \hat{\Sigma}_0\} = \text{glmfit}([N^1_{0C} + N^1_{0S} \mid X_i, N^0_{1C} + N^0_{0S} \mid X_i],\ X_i \in \{X\})$
  Set $\{\hat{\beta}_{sel}, \hat{\Sigma}_{sel}\} = \text{glmfit}([\sum_{z \in \{C,S\}, y \in \{0,1\}} N^y_{1z} \mid X_i, \sum_{z \in \{C,S\}, y \in \{0,1\}} N^y_{0z} \mid Xi],\ X_i \in \{X\})$
  Draw $\gamma^{(i)}_{1z} \mid \Theta_{X,-\gamma_{1z}}, N_{samp}, N_{obs}, X \sim \text{MVN}(\hat{\gamma}_{1z}, \hat{\Sigma}_{1z}), \quad z \in \{C, S\}$
  Draw $\gamma^{(i)}_0 \mid \Theta_{X,-\gamma_0}, N_{samp}, N_{obs}, X \sim \text{MVN}(\hat{\gamma}_0, \hat{\Sigma}_0)$
  Draw $\beta^{(i)}_{sel} \mid \Theta_{X,-\beta_{sel}}, N_{samp}, N_{obs}, X \sim \text{MVN}(\hat{\beta}_{sel}, \hat{\Sigma}_{sel})$
end for
return $\Theta^{N_{burnin}+1:\ N_s}_X$

(step 1: $N_{obs} \mid X_i;\ X_i \in \{X\}$; whose cardinality $\#\{N_{obs} \mid X_i;\ X_i \in \{X\}\} = 6 \times \#\{X\}$). We sample the missing selection indicator, $D_i^C \mid X_i;\ X_i \in \{X\}$, following a similar logic to that of Algorithm 1 (steps: 5–9). We fit binomial probit regression functions based on these counts using a standard fitting function. We calculate the maximum-likelihood estimate (MLE) of the regression coefficients and its covariance matrix (steps 10–12: $\{\hat{\gamma}, \hat{\Sigma}\} = \text{glmfit}([N^1 \mid X_i, N^0 \mid X_i];\ X_i \in \{X\})$). This fitting strategy avoids the fitting of probit regressions with millions of data points. Based on these estimates, the regression parameters are sampled from multivariate normal distributions (steps 13–15: $\text{MVN}(\hat{\gamma}, \hat{\Sigma})$) by Laplace approximation (Geisser et al. 1990). We use $\Theta_X^{(1:N_s)}$ samples to generate credible intervals for the effect estimates conditional on user features X_i.

Appendix F. User Selection Response Simulation

Algorithm 3 (User selection response simulator for campaign effectiveness optimization)

1: **Input:** Selection function $F_{sel}(X_i)$, Non-zero Effect Indicator function $F_{sig}(X_i)$, $\text{ATE}_{Camp}^{D=1}$ Sign function $F_{sign}^{ATE}(X_i)$, Sign Certainty weights $\mathbf{w}^{sig} = \{w^-, w^{\pm}, w^+\}$, User Counts N_{Camp}^{obs} as defined by Equation (15).

2: *// Set segment weighting function $D_w^{sel}(X_i)$, based on inputs: $F_{sel}(X_i)$, $F_{sig}(X_i)$, $F_{sign}^{ATE}(X_i)$, $\mathbf{w}^{sig}$*

3: Define $D_w^{sel}(X_i)$

$$= \begin{cases} w^{\pm} \times F_{sel}(X_i) & \text{if } F_{sig}(X_i) = \textbf{false} \\ w^{+} \times F_{sel}(X_i) & \text{if } F_{sig}(X_i) = \textbf{true and } F_{sign}^{ATE}(X_i) = + \\ w^{-} \times F_{sel}(X_i) & \text{if } F_{sig}(X_i) = \textbf{true and } F_{sign}^{ATE}(X_i) = - \end{cases}$$

4: Set $N_{S,agg}^{new}$ to the output of Algorithm 4 with inputs: $F_{sel}(X_i) = D_w^{sel}(X_i)$, $N_z^{obs} = N_S^{obs} \mid X_i,\ X_i \in \{X\}$
// Simulate Campaign Selection, $Z_i = S$

5: Set $N_{Camp,agg}^{new} = \{\sum_{X_i \in \{X\}} N_C^{obs} \mid X_i, N_{S,agg}^{new}\}$
// Aggregate User Counts

6: **return** $N_{Camp,agg}^{new}$

Algorithm 4 (User selection response simulator)

1: **Input:** Selection function $F_{sel}(X_i)$, User Counts $N_z^{obs} = \{N_{dz}^{y} \mid X_i;\ d \in \{0,1\},\ y \in \{0,1\},\ X_i \in \{X\}\}$.

2: **Output:** Aggregated User Counts After Selection $N_{z,agg}^{new} = \{N_{dz}^{y,new};\ d \in \{0,1\},\ y \in \{0,1\}\}$

3: Set $[\hat{\gamma}_{0z}, \hat{\gamma}_{1z}] = [\text{glmfit}([N_{0z}^1 \mid X_i, N_{0z}^0 \mid X_i],\ X_i \in \{X\}), \text{glmfit}([N_{1z}^1 \mid X_i, N_{1z}^0 \mid X_i],\ X_i \in \{X\})]$
// Probit Approximation

4: Set $[\hat{\theta}_{0z}, \hat{\theta}_{1z}] \mid X_i = [\Phi(X_i'\hat{\gamma}_{0z}), \Phi(X_i'\hat{\gamma}_{1z})],\quad X_i \in \{X\}$
// Observed Conversion Propensity

5: Set $N_z^{Visit} \mid X_i = N_{1z}^1 + N_{1z}^0 + N_{0z}^1 + N_{0z}^0 \mid X_i,\quad X_i \in \{X\}$
// Audience per Segment X_i

6: Set $N_{1z}^{budget} = \sum_{X_i \in \{X\}} (N_{1z}^1 + N_{1z}^0) \mid X_i$ *// Observed Budget*

7: Set $N_{1z}^{1,new} \mid X_i = N_{1z}^{0,new} \mid X_i = 0,\quad X_i \in \{X\}$ *// Set Counts*

8: Set $N_{remain}^{budget} = N_{1z}^{budget}$ *// Initialize Remaining Budget*

9: **while** $N_{remain}^{budget} > 0$ **do**

10: Set $P(X_i) = N_{remain}^{Visit} \mid X_i / \sum_{X_i \in \{X\}} N_{remain}^{Visit} \mid X_i,\quad X_i \in \{X\}$

11: Set $\lambda = N_{remain}^{budget} / (\sum_{X_i \in \{X\}} N_{remain}^{budget} \times F_{sel}(X_i) \times P(X_i) \mid X_i)$
// Budget Multiplier

12: Set $[N_{1z}^{1,new}, N_{1z}^{0,new}] \mid X_i = [N_{1z}^{1,new}, N_{1z}^{0,new}] \mid X_i + \min(\lambda \times F_{sel}(X_i) \times N_{remain}^{budget} \times P(X_i), N_{remain}^{Visit} \mid X_i) \times [\hat{\theta}_{1z}, 1 - \hat{\theta}_{1z}] \mid X_i,\quad X_i \in \{X\}$
// User Selection for ad exposure

13: Set $N_{remain}^{Visit} \mid X_i = N_z^{Visit} - (N_{1z}^{1,new} + N_{1z}^{0,new}) \mid X_i,\ X_i \in \{X\}$ *// Remaining Audience*

14: Set $N_{remain}^{budget} = N_{1z}^{budget} - (\sum_{X_i \in \{X\}} [N_{1z}^{1,new} + N_{1z}^{0,new} \mid X_i])$
// Remaining Budget

15: **end while**

16: Set $[N_{0z}^{1,new}, N_{0z}^{0,new}] \mid X_i = N_{remain}^{Visit} \times [\hat{\theta}_{0z}, 1 - \hat{\theta}_{0z}] \mid X_i,\ X_i \in \{X\}$ *// Nonselected User Counts*

17: Set $N_{z,agg}^{new} = \{\sum_{X_i \in \{X\}} N_{dz}^{y,new} \mid X_i;\ d \in \{0,1\},\ y \in \{0,1\}\}$
// Aggregate User Counts

To simulate a given selection function, we execute Algorithm 3, which aggregates the user counts of the study group ($Z_i = S$) given: (1) a selection function $F_{sel}(X_i)$; (2) a non-zero effect indicator function $F_{sig}(X_i)$; (3) $\text{ATE}_{Camp}^{D=1}$ sign function $F_{sign}^{ATE}(X_i)$; and (4) a sign certainty weighting set $\mathbf{w}^{sig} = \{w^-, w^{\pm}, w^+\}$. These functions are combined into a segment weighting $D_w^{sel}(X_i)$ (steps: 1–3). We use $D_w^{sel}(X_i)$ as compound selection function (step: 4). We simulate this user selection for the users of the study group, $N_S^{obs} \mid X_i,\ X_i \in \{X\}$, by executing Algorithm 4. We aggregate the user counts of the control group over X_i, $N_C^{obs} \mid X_i$, and concatenate them to the aggregated study user counts after selection, $N_{S,agg}^{new}$ (step: 5).

We model the user response of the selected and nonselected populations for a given treatment arm, ($\theta_{0z} \mid X_i, \theta_{1z} \mid X_i;\ X_i \in \{X\}$), using a probit transformation as illustrated by steps 3–4 of Algorithm 4. We consider the audience-by-segment constraint $N_z^{Visit} \mid X_i$, and the observed ad-exposed users as a fixed campaign budget N_{1z}^{budget} (steps: 5–6). We define a budget multiplier λ to guarantee that all this budget is consumed by the user selection, which includes the probability of user segments $P(X_i)$ (steps: 10–11). The min function enforces the visiting population segment constraints ($N_{remain}^{Visit} \mid X_i$). The while loop of steps 9–15 redistributes the remaining budget in case $N_{remain}^{Visit} \mid X_i$ is exhausted for any segment. We aggregate the user counts over X_i to generate the four counts given $Z_i = z$: $N_{z,agg}^{new} = \{N_{dz}^{y,new};\ d \in \{0,1\},\ y \in \{0,1\}\}$ (steps: 16–17).

References

Aly M, Hatch A, Josifovski V, Narayanan VK (2012) Web-scale user modeling for targeting. *Proc. 21st Internat. Conf. World Wide Web* (ACM, New York), 3–12.

Atlas Institute (2008) Engagement mapping: A new measurement standard is emerging for advertisers. White paper.

Berman R (2015) Beyond the last touch: Attribution in online advertising. Working paper, University of Pennsylvania, Philadelphia.

Blake T, Coey D (2014) Why marketplace experimentation is harder than it seems: The role of test-control interference. *Proc. 15th ACM Conf. Econom. Computation* (ACM, New York), 567–582.

Blake T, Nosko C, Tadelis S (2015) Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *Econometrica* 83(1):155–174.

Broder A, Josifovski V (2011) *Introduction to Computational Advertising* (Stanford University Press, Redwood City, CA).

Chickering DM, Heckerman D (2000) A decision theoretic approach to targeted advertising. Boutilier C, Goldszmidt M, eds. *Proc. 16th Uncertainty Artificial Intelligence* (Morgan Kaufmann, Stanford, CA), 82–88.

Chittilappilly A (2012) Using experiment design to build confidence in your attribution model. *Online Metrics Insider* (July 11).

Digiday, Google (2011) Real-time display advertising state of the industry. http://doubleclickadvertisers.blogspot.com/2011/02/real-time-display-advertising-state-of.html.

Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. *Biometrics* 58(1):21–29.

Geisser S, Hodges JS, Press SJ, ZeUner A (1990) The validity of posterior expansions based on Laplace's method. *Bayesian Likelihood Methods Statist. Econometrics* 7:473–488.

Ghosh A, McAfee P, Papineni K, Vassilvitskii S (2009) Bidding for representative allocations for display advertising. Leonardi S, ed. *Proc. Internet Network Econom.: 5th Internat. Workshop, WINE 2009*, Lecture Notes Comput. Sci., Vol. 5929 (Springer-Verlag, Berlin Heidelberg), 208–219.

Goldfarb A (2014) What is different about online advertising? *Rev. Indust. Organ.* 44(2):115–129.

Goldfarb A, Tucker C (2011) Online display advertising: Targeting and obtrusiveness. *Marketing Sci.* 30(3):389–404.

Heckman JJ (2008) Econometric causality. *Internat. Statist. Rev.* 76:1–27.

Johnson GA, Lewis RA, Nubbemeyer EI (2016) Ghost ads: A revolution in measuring ad effectiveness. Working paper, University of Rochester, Rochester, NY.

Lambrecht A, Tucker C (2013) When does retargeting work? Information specificity in online advertising. *J. Marketing Res.* 50(5):561–576.

Lewis R, Reiley D (2014) Online ads and offline sales: Measuring the effects of retail advertising via a controlled experiment on Yahoo!. *Quant. Marketing Econom.* 12(3):235–266.

Lewis RA, Rao JM, Reiley DH (2011) Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. *Proc. 20th Internat. Conf. World Wide Web* (ACM, New York), 157–166.

Li H, Kannan PK (2014) Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *J. Marketing Res.* 51(1):40–56.

Morrison W, Coolbirth R (2008) IAB marketplace: Networks and xchanges. Event Recap. http://www.slideshare.net/tinhanhvy/iab-marketplace-networks-and-xchanges-2008.

Pandey S, Aly M, Bagherjeiran A, Hatch A, Ciccolo P, Ratnaparkhi A, Zinkevich M (2011) Learning to target: What works for behavioral targeting. *Proc. 20th Internat. Conf. Inform. Knowledge Management* (ACM, New York), 1805–1814.

Rubin DB (2005) Causal inference using potential outcomes. *J. Amer. Statist. Assoc.* 100(469):322–331.

Sahni N, Zou D, Chintagunta PK (2015) Do targeted discount offers serve as advertising? Evidence from 70 field experiments. Research Paper, Stanford University Graduate School of Business, Palo Alto, CA.

Shao X, Li L (2011) Data-driven multi-touch attribution models. *Proc. 17th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 258–264.

Spencer S, O'Connell J, Greene M (2011) The arrival of real-time bidding. IAB, Google, Forrester. http://www.slideshare.net/IABmembership/the-arrival-of-realtime-bidding-hosted-by-iab-google-forrester.

Yildiz T, Narayanan S (2013) Star digital: Assessing the effectiveness of display advertising. Case Study, Harvard Business Review, Cambridge, MA. https://hbr.org/product/star-digital-assessing-the-effectivness-of-display-advertising/M347-HCB-ENG.