Randomization and The Pernicious Effects of Limited Budgets on Auction Experiments

Guillaume W. Basse
gbasse@fas.harvard.edu
Statistics Department
Harvard University

Hossein Azari Soufiani
azari@google.com
Google Research
New York, NY 10011, USA

Diane Lambert
exdlambert@gmail.com
Summit, NJ

Abstract

Buyers (e.g., advertisers) often have limited financial and processing resources, and so their participation in auctions is throttled. Changes to auctions may affect bids or throttling and any change may affect what winners pay. This paper shows that if an A/B experiment affects only bids, then the observed treatment effect is unbiased when all the bidders in an auction are randomly assigned to A or B but it can be severely biased otherwise, even in the absence of throttling. Experiments that affect throttling algorithms can also be badly biased, but the bias can be substantially reduced if the budget for each advertiser in the experiment is allocated to separate pots for the A and B arms of the experiment.

1 INTRODUCTION

A search query generates a request for ads to show with search results. A user's visit to a webpage generates a request to an ad exchange for page ads. In either case, advertisers are chosen to participate in an auction that chooses the ad to be shown. Advertisers often cannot pay for or process all auction requests they are eligible for, so the requests passed on to them are throttled to meet their quota constraints.

Auction parameters like reserve prices and throttling schemes can affect an advertiser's payments and an exchange's revenue, so A/B experiments are run to test ideas for improving outcomes for advertisers and the ad exchange. Lucking-Reiley (1999) and Einav et al. (2011) experimented with auction formats, Reiley (2006) and Ostrovsky and Schwarz (2011) experimented with reserve prices, and Ausubel et al. (2013) experimented with budget constraints.

Usually, A/B experiments are analyzed by assuming that the two experiment arms, traditionally called treatment and control, are independent. But, as Blake and Coey (2014) explains, independence fails when the demands of the treatment and control arms affect each other. Such *interference* is unavoidable when some advertisers in an auction are in treatment and some in control. Kohavi et al. (2009) recognizes that some randomization schemes can give misleading treatment estimates for auction experiments. For more insight into interference in other applications see Halloran and Struchiner (1995), Hudgens and Halloran (2012), Rosenbaum (2007), Tchetgen and VanderWeele (2010), Aronow and Samii (2013), Eckles et al. (2014), Athey et al. (2015) and Airoldi et al. (2012). Optimal experiment design in the presence of interference has been explored by David and Kempton (1996), Eckles et al. (2014) and Walker and Muchnik (2014).

This paper uses the framework of causal models and interference to shed light on auction experiments. Section 2 introduces the main elements of our model:

1. potential outcomes that describe what each advertiser would bid if assigned to treatment or if assigned to control,
2. throttling algorithms that determine which of the advertisers *eligible* to respond to a request for an ad are called to its auction,
3. *bid treatments* that affect what advertisers bid and *throttling treatments* that affect when they are called to bid, and
4. effects on the total daily payment of an advertiser or the total daily revenue of the ad exchange.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

Section 3 defines two randomization schemes for auction experiments. In *query randomization* each request for an ad is randomized to treatment or control, so all participants in an auction are in treatment or all are in control. In *(query, advertiser) randomization*, each advertiser that is eligible for a query is independently assigned to treatment or control, so treatment bidders and control bidders can compete in the same auction. With either kind of randomization, an advertiser can be assigned to treatment for some queries and to control for others. (Query, advertiser) randomization is undesirable because it introduces interference, but it may be unavoidable if only some advertisers are included in an experiment, perhaps to avoid revealing a possible change in auction algorithms before it has proven useful. The remainder of the paper explores bias and variance of estimated effects for bid and throttling treatments under query and (query, advertiser) randomization.

To establish the ideas, Section 4 shows what can go wrong when treatment and control bidders compete in the same auction. Section 5 introduces budget (processing or financial) throttling. In *Split quota* experiments each advertiser has separate budgets for treatment and control queries. In *joint quota* experiments the advertiser draws against the same budget for all its queries. Simulations in Section 6 suggest that estimated treatment effects can be severely biased under (query, advertiser) randomization regardless of whether the budget is split or joint for both bid and throttling experiments. Estimated treatment effects for throttling experiments are biased for both query and (query, advertiser) randomization, but the bias is much smaller for split quota than for joint quota experiments.

2 CAUSAL MODELS FOR AUCTIONS

This section casts auction experiments as causal models in which each advertiser has two potential bids for each query: the bid that would be made if the advertiser were assigned to treatment for that query and the bid if assigned to control. Of course, only one potential bid can be observed. This section points out further subtleties that result from advertiser competition and quota throttling.

2.1 Potential Bids

To start, suppose there is no throttling, so advertisers bid in all auctions for which they are eligible. The raw data for a sample of auctions is then a set of vectors (q, a, B, Y), where q denotes a query, a an advertiser, B the advertiser's bid, and Y the advertiser's payment. We consider only auctions like first and second price auctions in which the payment is positive if the advertiser wins the auction and zero otherwise. Define N_q to be the number of unique queries and N the total number of (query, advertiser) pairs, typically considered over the course of one day.

In an experiment, a (query, advertiser) pair is assigned to either treatment or control. Let $Z = (Z_1, \ldots, Z_N)$ be the treatment assignment vector, where $Z_i = 1$ if (query, advertiser) pair i is assigned to treatment, and $Z_i = 0$ if assigned to control. Following the potential outcomes framework of Rubin (1990), each eligible advertiser for a query has a bid and payment that will be observed if the pair is assigned to control and a possibly different bid and payment that will be observed if assigned to treatment. That is, the *potential bids* for the N (query, advertiser) pairs are $B(Z) = (B_1(Z), \ldots, B_N(Z))$ and the potential payments are $Y(Z) = (Y_1(Z), \ldots, Y_N(Z))$.

An advertiser does not know which other advertisers are eligible for a query so its bid is independent of all other bids for the query, which implies that the following assumption given in Rubin (1980) holds.

Assumption 1 (Stable Unit Treatment Value (SUTVA))**.** *There is only one version of the treatment and there is no interference in the potential bids for eligible advertisers in treatment and control.*

Because SUTVA holds for bids, the potential bids for all query, eligible advertiser pairs under treatment $B(\vec{1})$ and under control $B(\vec{0})$ can be considered separately.

2.2 Potential payments

If bidder i's payment is positive, then any other bidder in the auction pays zero, so a treatment that affects what advertisers bid can affect the payment for both treatment and control advertisers. For example, suppose there are three advertisers in an auction with the potential bids given in the following table and the winner pays what it bid.

advertiser	B(1)	B(0)
b_1	5	2
b_2	4	4
b_3	3	1

If all three advertisers participate in the auction and advertiser 2 is the only one assigned to treatment, then it wins and pays $Y_2((0, 1, 0)) = 4$. However, if advertiser 1 is also assigned to treatment and participates in the auction then advertiser 2 loses and pays nothing. This is an example of interference: the payment of an advertiser is affected by the assignment of other advertisers in the auction to treatment and control. Hence, SUTVA does not hold for payments.

We make the following assumption about payments.

Assumption 2. *The potential payment of an advertiser eligible for a query depends only on its assignment to treatment or control and the assignments of the other eligible advertisers for the query. That is*

$$Y_i(Z) = Y_i(Z_{q[i]}) \tag{1}$$

where $Z_{q[i]}$ is the subset of the assignment vector Z on (query, advertiser) pairs for which the query is $q[i]$.

We extend this assumption to auctions with quota throttling in Section 2.3.

The interference structure allowed by Assumption 2 is similar to that given in Rosenbaum (2007) and Hudgens and Halloran (2012) and can be seen as a special type of *effective treatment* (Manski (2013)) or *exposure mapping* (Aronow and Samii (2013); Eckles et al. (2014)). However, those papers do not consider quota throttling.

2.3 Quota throttling

Advertisers are generally *quota constrained*, meaning that they have insufficient budget or infrastructure to process all the queries they could be sent (Chakraborty et al. (2010)). In that case, some queries are dropped or *throttled* to meet the quota constraints. That is, there is a vector $W(Z) = (W_1(Z), \ldots, W_N(Z))$ with $W_i(Z) = 0$ if (query, advertiser) pair i is dropped and $W_i(Z) = 1$ otherwise. Note that throttling can depend on the assignment Z to treatment and control. We assume random dropping according to a *throttling distribution* $p(W(Z)|Z)$. In the presence of throttling, Assumption 2 means that the potential payment of an advertiser depends only on its assignment to treatment or control, and on the assignment of the advertisers for the same query who were not throttled. Henceforth, advertisers who participate (bid) in auctions are called bidders.

Because how much a bidder in an auction pays depends on the other bidders in the auction, the potential payment for any advertiser a is random when at least one other advertiser for the query is quota constrained, even if advertiser a is unconstrained.

2.4 Bid treatments and throttling treatments

Loosely speaking, an experiment about bids is designed to test whether a change to bidding rules, such as a change to the reserve or "floor" price for auctions, matters when throttling rules are unchanged. (See Reiley (2006) and Ostrovsky and Schwarz (2011) for examples.)

Definition 1 (Bid Treatment)**.** *Under a* bid treatment*, for all Z, the throttling distribution W satisfies*

$$W(Z)|Z \sim (W(Z)|Z = \vec{1}) \sim (W(Z)|Z = \vec{0}). \tag{2}$$

In words, a bid treatment affects only potential bids, so a given (query, advertiser) for an eligible advertiser has the same probability of being dropped regardless of how other eligible bidders are assigned to treatment and control. That condition holds if queries are throttled before the advertiser's bid is known.

An experiment about throttling is designed to understand whether changing the rules for meeting quota constraints matters if bidding parameters like reserve prices are unchanged. (See the selective callout algorithm in Chakraborty et al. (2010).)

Definition 2 (Throttling Treatment)**.** *Under a* throttling treatment*, for all Z, the potential bids satisfy*

$$B(Z) = B(\vec{0}) = B(\vec{1}). \tag{3}$$

In a throttling experiment, any eligible advertiser would bid the same amount, whether it is throttled or not. This is true if advertisers do not reveal their bids before throttling.

2.5 Effects on revenue

If no advertiser is quota constrained, then no advertiser is dropped from a query and the effect τ of the treatment on the total revenue of the ad exchange compares the revenue when every (query, advertiser) pair is treated to the revenue when every (query, advertiser) pair is in control:

$$\tau = \sum_{i}^{N} (Y_i(\vec{1}) - Y_i(\vec{0})) \tag{4}$$

where Y_i is the payment of (query, advertiser) pair i. If there are quota constraints, then τ is affected by random dropping so the effect on total revenue is $\tau^* = E(\tau)$, taking the expectation under the dropping scheme $p(W)$. Note that the sum in this case is taken over the N (query, eligible advertiser) pairs.

The effect τ_a of treatment on the revenue generated by a given advertiser a when there are no quota constraints is given by

$$\tau_a = \sum_{i:a[i]=a} (Y_i(\vec{1}) - Y_i(\vec{0})). \tag{5}$$

Again, with random throttling the effect of interest is $\tau_a^* = E(\tau_a)$, taking the expectation under the dropping scheme $p(W)$. Note that τ_a considers the effect of treating all eligible advertisers on the revenue generated *only* by advertiser a, rather than the effect of treating only the queries for advertiser a.

3 RANDOMIZATION

With *query randomization*, each query is randomized to treatment or control and then all the eligible advertisers for the query are assigned to control if the query is assigned to control or to treatment if the query is assigned to treatment. Formally, an auction experiment is query-randomized if

1. $P(Z_i = 1) = p_i$ and $P(Z_i = 0) = 1 - p_i$, and
2. $Z_i = 1$ implies that $\vec{Z}_{q[i]} = \vec{1}_{q[i]}$.

With *(query, advertiser)* randomization, each (query, advertiser) pair is randomized independently to treatment or control. Auctions now may have a mix of treated and control bidders, which is not representative of the behavior of future auctions. However, (query, advertiser) randomization may be necessary if only some advertisers are allowed to be in the experiment, and hence many of the auctions with treated advertisers will also have untreated advertisers. Both query randomization and (query, advertiser) randomization happen before any form of throttling takes place.

Our parameters of interest are total differences over a day, not a mean per-query difference. Here the total under treatment is estimated by inversely weighting each observation in treatment by its probability of occurrence, and the total under the control is estimated by inversely weighting each observation in control by its probability of occurrence.

$$\hat{\tau}(Z) = \left(\sum_{i:Z_i=1} \frac{Y_i(Z)}{p_i} - \sum_{i:Z_i=0} \frac{Y_i(Z)}{1-p_i} \right) \tag{6}$$

$$\hat{\tau}_a(Z) = \left(\sum_{i:a[i]=a, Z_i=1} \frac{Y_i(Z)}{p_i} - \sum_{i:a[i]=a, Z_i=0} \frac{Y_i(Z)}{1-p_i} \right). \tag{7}$$

The estimators $\hat{\tau}$ and $\hat{\tau}_a$ would be unbiased for τ and τ_a respectively, if the SUTVA assumption (2.1) held for payments $Y(Z)$, but SUTVA does not hold for payments. Nonetheless, Section 5 shows that $\hat{\tau}$ and $\hat{\tau}_a$ are unbiased under query randomization.

4 TWO EXAMPLES OF BIAS

To illustrate the issues, two toy examples show that the estimated treatment effect for an experiment with just one auction can be severely biased.

4.1 Identical but independent bidders

Suppose K bidders participate in a first price auction, and they all have the same bid under treatment and the same bid under control. That is,

$$\begin{aligned} B_i(1) &= R_1, \;\; i = 1 \dots K \\ B_i(0) &= R_0, \;\; i = 1 \dots K \\ R_1 &> R_0. \end{aligned}$$

Also suppose each advertiser was independently assigned to treatment with probability $\frac{1}{2}$; this is an example of (query, advertiser) randomization. The goal is to estimate the effect of treatment on this auction alone: $\tau = \sum_{i=1}^{K} (Y_i(\vec{1}) - Y_i(\vec{0}))$. How ties are decided does not matter. The expected value of the estimator defined in (6) is

$$\begin{aligned} E(\hat{\tau}) &= \frac{1}{2^K}(\hat{\tau}(\vec{1}) + \hat{\tau}(\vec{0})) + \frac{1}{2^K} \sum_{Z \neq \{\vec{1},\vec{0}\}} \hat{\tau}(Z) \\ &= \frac{1}{2^K}(R_1 - R_0) + \frac{1}{2^K} \sum_{Z \neq \{\vec{1},\vec{0}\}} R_1 \\ &= \frac{1}{2^K}\tau + \frac{1}{2^K}(2^K - 2)R_1 \\ &= \frac{1}{2^K}\tau + (1 - \frac{1}{2^{K-1}})R_1 \end{aligned}$$

where all 2^K possible assignments are equiprobable. The bias of $\hat{\tau}$ is

$$\text{Bias}(\hat{\tau}, \tau) = (\frac{1}{2^K} - 1)\tau + (1 - \frac{1}{2^{K-1}})R_1. \tag{8}$$

Intuitively, what happens is that the estimator $\hat{\tau}$ takes the value R_1 for every assignment Z, except for $Z = \vec{0}$ where it takes the value $-R_0$. Thus, as the number K of bidders grows, the bias approaches R_0, so the bias becomes as large as the control bid as the size of the auction grows. For example, with $K = 4$ bidders, treatment bids of $R_1 = 6$ and control bids of $R_0 = 5$, the true effect is $\tau = 1$ while equation (8) shows that the bias is about 4.3. The same result would hold for second price auctions in this scenario. The bias would not vanish if we only allowed randomizations with given numbers N_0 and N_1 of control and treated bidders respectively. Indeed, for any such randomization scheme with $0 < N_0, N_1 < K$, the bias would be exactly R_0 for any K. However, as will be shown in Section 5, the estimator is unbiased under query randomization.

4.2 Treatment dominates control

Suppose K bidders participate in an auction and each bidder under treatment bids more than every bidder under control. If we label the bidders according to their bids under treatment, then

$$B_i(1) > B_j(0) \quad \text{for all } i, j$$
$$B_i(1) > B_j(1) \text{ if } i < j$$

Then (see the supplementary material) the bias of the estimated treatment effect is bounded below by $\tau(\frac{1}{2^K} - 1) + A_K$, where A_K approaches $B_K(1)$ as the number K of eligible bidders grows. Thus, the bias grows at least as large as $\max_i(B_i(0)) - (B_1(1) - B_K(1))$.

The limiting bias is especially large if the smallest and largest bids under treatment are close, even if K is small. When $K = 4$, $(B_1(0), B_2(0), B_3(0), B_4(0)) = (4, 4.25, 4.50, 4.75)$ and $(B_1(1), B_2(1), B_3(1), B_4(1)) = (6, 5.50, 5.25, 5)$, the true effect is $\tau = 1.25$, and the exact bias is 3.8, which is about three times as large as the effect itself.

5 BIAS WITH QUOTA THROTTLING

5.1 Joint and split throttling

Let $N_q[a]$ be the total number of queries that advertiser a is eligible for and let $Q[a]$ be the *quota* or number of queries that advertiser a can process. If $N_q[a] > Q[a]$, then the queries for advertiser a must be throttled so some are randomly dropped. An experiment can use advertiser a's quota for both treatment and control, or it can split the quota into two pieces, using one piece to service the a's queries assigned to treatment and the other piece to service a's queries assigned to control. We do not consider mixed experiments in which some advertisers are assigned to joint throttling and others to split throttling.

Definition 3. *Under* joint throttling*,*

$$\sum_{i:a[i]=a} W_i(Z) \leq Q[a] \tag{9}$$

for all assignments Z to treatment and all advertisers a.

Definition 4. *Under* split throttling*,*

$\sum_{i:a[i]=a} I(Z_i = 1) W_i(Z) \leq Q^{(1)}[a]$,

$\sum_{i:a[i]=a} I(Z_i = 0) W_i(Z) \leq Q^{(0)}[a]$, *and*

$Q^{(1)}[a] + Q^{(0)}[a] = Q[a]$

for all assignment Z to treatment and every advertiser a, where $Q^{(1)}[a]$ and $Q^{(0)}[a]$ denote the quotas *in treatment and control respectively.*

5.2 Bid treatments and joint throttling

Suppose that the ad exchange always fulfills each advertiser's quota entirely and that the number of queries that could be sent to advertiser a is more than it can process, so $N_q[a] > Q_q[a]$. Then joint throttling implies that

$$\sum_{i:a[i]=a} W_i(Z) = Q[a] \tag{10}$$

for all advertisers a and all assignments Z to treatment. Further suppose that throttling is random, so an advertiser is randomly dropped from queries to meet its quota constraints. Then for all vectors w satisfying (10),

$$P(W(Z) = w|Z) = \binom{N[a]}{Q[a]}^{-1} \quad \text{for all } Z. \tag{11}$$

Under these assumptions, the following theorem is proved in the supplementary material.

Theorem 5.1. *With joint throttling and query randomization, the estimator $\hat{\tau}$ is unbiased for τ^* for any bid treatment when every advertiser is quota constrained and there are sufficient queries for every advertiser's budget quota to be fulfilled.*

Unsurprisingly, the same result holds for query randomization when all auctions are unconstrained, that is, when no advertiser is quota throttled. The corresponding theorem is stated and proved in the supplementary material. In summary, query randomization leads to unbiased estimates for bid treatments both with and without throttling.

5.3 Quota treatment and split throttling

A quota experiment is designed to test whether a change to the algorithm for dropping advertisers from queries to satisfy quota constraints affects revenue. Suppose that the change is in addition to the standard throttling algorithm (the control case) and that it is applied before the standard throttling algorithm is applied. For example, some feature of the query, such as the user's locale, could be used to drop queries, reducing the need to randomly drop queries without regard to their value to the advertiser. To be specific, the treatment drops the (query, eligible advertiser) pair i before the control throttling algorithm is applied when a binary random variable $x[i]$ is zero and the control throttling algorithm alone is applied if $x[i]$ is one. Then the treatment throttling enforces the constraint:

$$(Z_i = 1) \cap (x_i = 0) \Rightarrow W_i(Z) = 0. \qquad (12)$$

The simulations in Section 6 show that the estimated effect of the quota treatment on revenue is biased under joint throttling for both query randomized and (query, advertiser) randomized experiments. The question is what happens if separate budgets are maintained for treatment and control for each advertiser (i.e., under split throttling)? Theorem 5.2, which is proved in the supplementary material, shows that the estimated effect can be unbiased under split-quota throttling under a set of conditions. Unfortunately, the conditions are unlikely to hold in practice.

To state Theorem 5.2, let $N_a^{(1)}(x = 1)(Z)$ be the number of eligible queries under treatment for advertiser a when $x = 1$, and $Z = 1$. Define $N_a^{(0)}(x = 1)(Z)$ analogously. Also, define $N(x = 1)$ to be the total number of (query, eligible advertiser) pairs for which $x = 1$.

Theorem 5.2. *Let $\mathcal{Z}$ be the assignments of (query, eligible advertiser) pairs allowed by query randomization. If*

$x_i = 0$ *implies that the bid B_i for advertiser i is 0,*

$\frac{Q_a^{(0)}}{N_a^{(0)}(Z)} = \frac{Q_a^{(0)}+Q_a^{(1)}}{N_a}$ *for all advertisers $a \in \mathcal{A}$ and assignments of (query, bidder) pairs $Z \in \mathcal{Z}$ to treatment and control, and*

$\frac{Q_a^{(1)}}{N_a^{(1)}(x=1)(Z)} = \frac{Q_a^{(1)}+Q_a^{(0)}}{N_a(x=1)}$ *for all $a \in \mathcal{B}$ and $Z \in \mathcal{Z}$*

then the estimator $\hat{\tau}$ is unbiased for τ^* under query randomization with split throttling.

6 SIMULATIONS

This section reports the results of simulating the bias and variance of revenue estimators under quota throttling for query randomized experiments and (query, advertiser) randomized experiments. The simulated query randomized experiments are *balanced* in the sense that they have the same number of queries in treatment and control. Similarly, the simulated (query, advertiser) randomized experiments balance the number of (query, advertiser) pairs assigned to treatment and control. Without such balance, effect estimates for total revenue would be much more variable and that additional variability would obscure differences in the randomization and quota sharing schemes for reasonably sized simulations.

There are two main conclusions. First, when estimating *bid treatment* effects in constrained auctions, the estimated effects under balanced query randomization are not only unbiased (as proved in Theorem 5.1), but also have smaller variance than those obtained with balanced (query, advertiser) randomization. Second, even if the conditions of Theorem 5.2 are violated, query randomization combined with split throttling can dramatically reduce the variance of the estimated *quota treatment* effect, compared to the variance under joint throttling.

Each simulated experiment considers auctions with three eligible advertisers. These advertisers all compete in $N_q = 90$, 120 or 150 auctions unless they are throttled. Their quota limits are either $N_q/3$ or $2N_q/3$. For bid treatments, potential bids are Lognormal($\mu_0 = 1, v = 0.1$) under control and Lognormal($\mu_1, v = 0.1$) under treatment where μ_1 is either 1.05, 1.1, or 2. All the treatment bids in a simulation are drawn from the same distribution, and all the control bids are drawn from the same distribution. The control and treatment bids for a (query, advertiser) pair are correlated, as described in the supplementary material. The potential bids define the true treatment effect on total revenue.

For each distribution of potential bids and quota limit, we generated 20,000 random query experiments with exactly half the N_q queries in each experiment assigned to treatment and half to control. We also generated 20,000 random (query, advertiser) pair experiments, with half the pairs in treatment and half in control. The bid treatment effect or quota treatment effect, depending on the nature of the simulation, was estimated in each experiment. Because the bias depends on the treatment effect for a lognormal, we report simulated bias relative to the true effect.

6.1 Bid treatments

Figure 1 describes the simulated relative bias (ratio of the bias to the true effect), where the rows correspond to the different quota settings and the columns to the different mean log treatment bids. The standard errors reported are computed over the 200 draws from the bid distribution and are divided by $\sqrt{200}$ to reflect uncertainty about the simulated mean bias.

Figure 1 shows that the relative bias is nearly zero for randomized query experiments, while the relative bias has a mean as high as 1.5 for randomized (query, advertiser) pair experiments. That is, the penalty for allowing control and treated advertisers to compete in the same auction is a 50% increase in relative bias. Moreover, the simulated variance of the bid treatment estimate under the random query experiment is no more than its variance under the random (query, advertiser) experiment (see Figure 2.) The ratio of simulated variances for (query, advertiser) randomization versus query randomization is above one for all bid and

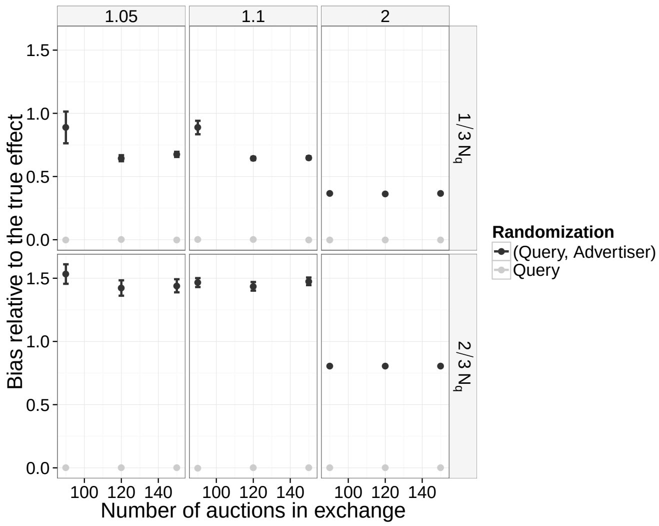

Figure 1: Simulated bias of the effect of a bid treatment relative to the true effect under balanced query randomization and balanced (query, advertiser) randomization. The dots show the mean relative bias, and the endpoints show the mean $\pm$ twice its simulated standard error. Columns show the mean log bid under treatment and the rows show the throttling rate.

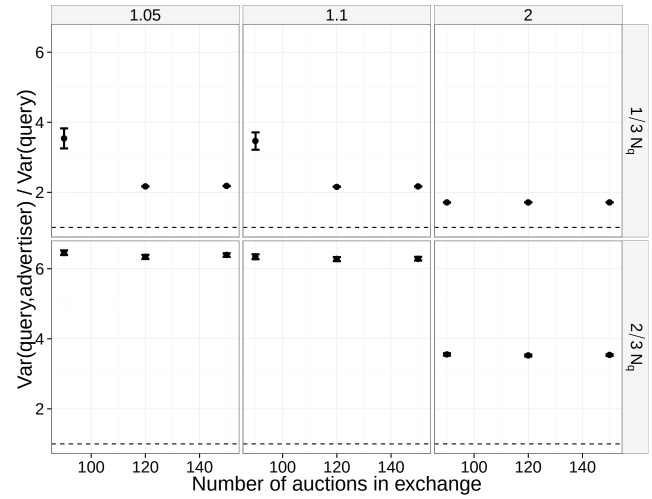

Figure 2: Ratio of the variance of estimated treatment effects under (query, advertiser) randomization to the variance of the estimated treatment effects under balanced query randomization for different lognormal bid distributions and throttling rates. The dots show the mean ratios and the endpoints show the mean $\pm$ twice its simulated standard error. The horizontal dotted line lies at one.

throttling combinations considered here, and as high as 6 when the treatment bids are 5% higher than the control bids, and the quota is around 66%. Note that the number of auctions in the experiment has little effect on the relative bias or variance.

6.2 Quota treatments

Because a quota treatment does not affect bids, the potential bids with a quota treatment are the same under treatment and control. Here $B_i(0) = B_i(1) \sim$ Lognormal$(\mu_0 = 1, v = 0.1)$. For each random query or random (query, advertiser) pair, we draw a covariate $x_i \sim Bernoulli(p_x)$, where $p_x = 0.1$ or $p_x = 0.5$ or $p_x = 0.9$ in different simulations. Figure 3 shows the relative bias of the estimate of total revenue under balanced query randomization with joint throttling and with split throttling. The results for the variance are shown in the supplementary material. Clearly, relative bias is close to zero for split throttling (even if the conditions of theorem 5.2 are violated), whereas it is around -1 for joint throttling. This means that the estimator under joint throttling estimates 0 regardless of the true effect on total revenue! So the effect on total revenue of changing the throttling mechanism can never be estimated from an experiment that uses joint throttling to meet quota constraints. The supplementary material shows that this conclusion about joint quota also holds for (query, advertiser) randomization.

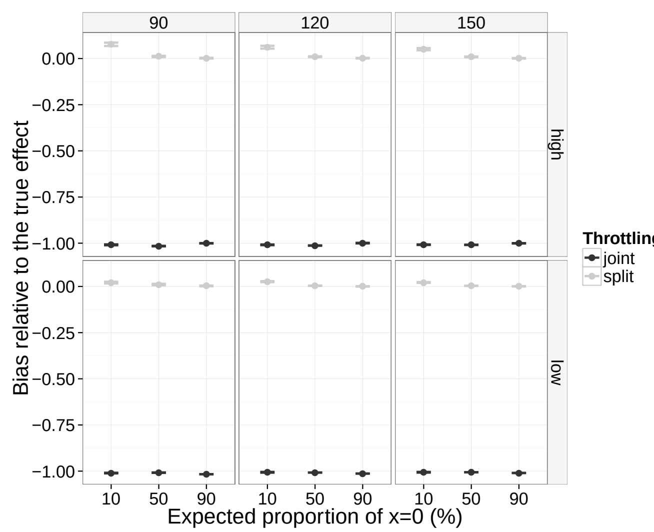

Figure 3: Bias of joint quota throttling and split quota throttling estimators relative to the true effect under query randomization. The dots are the mean relative bias and the segments show two simulated standard errors around that mean.

6.3 Extension to larger samples

Our simulations generate only a few auctions relative to the millions of auctions that might participate in a real experiment every day. Figures 1 and 3 hint that the relative biases are close to constant for larger numbers of auctions, while Figure 2 hints that the ratio of the variances is constant, or even slightly increasing as the number of auctions in an experiment increases. Together, these figures suggest that the variance of the estimator of the effect of a treatment on total revenue increases faster under balanced (query, advertiser) randomization than under balanced query randomization. We ran additional simulations (see supplementary material) that give further evidence for these trends.

7 CONCLUSION

Every day, companies like Google, Facebook and Yahoo run billions of auctions in ad exchanges, and every day they experiment to improve the exchange. This paper shows how casting experiments in the framework of potential outcomes clarifies many practical issues, such as the consequences of the choice of randomization. Bias has been emphasized throughout because when it is large it makes experiments misleading.

As a general policy, query randomization should be preferred over (query, advertiser) randomization when experimenting with bid treatments because it allows an unbiased estimator of the true treatment effect, without exceeding the variance of the same estimator under (query, advertiser) randomization. Split-quotas are preferable to joint quotas when experimenting with throttling treatments because split quotas have reduced bias and RMSE in simulations.

Admittedly, we do not have a complete solution to the problems that arise in practice, such as the fact that advertisers are often provided information about the user, such as location, that the advertiser may use to decide whether to bid or the bid amount. Experiments that take account of such covariates might be better analyzed with statistical models rather than with the simple estimators proposed here. Nor do we have analytical results for the variance of the estimators or formal procedures for hypothesis testing. These are all possible directions for future work.

Acknowledgments:

Guillaume Basse is the 2014 Google Fellow in Statistics for North America. This work was completed at Google Research. The authors thank Google's Max Lin and George Levitte for their help, as well as Google NYC's statistics group for early feedback. We also thank three anonymous reviewers for their constructive feedback.

References

E. M. Airoldi, Panos Toulis, Edward Kao, and D. B. Rubin. Causal estimation of peer influence effects. In *NIPS Workshop on Social Network and Social Media Analysis*, 2012.

Peter M Aronow and Cyrus Samii. Estimating average causal effects under interference between units. *arXiv preprint arXiv:1305.6156*, 2013.

Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. Technical report, National Bureau of Economic Research, 2015.

Lawrence M Ausubel, Justin E Burkett, and Emel Filiz-Ozbay. An experiment on auctions with endogenous budget constraints. *Available at SSRN 2331419*, 2013.

Thomas Blake and Dominic Coey. Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 567–582. ACM, 2014.

Tanmoy Chakraborty, Eyal Even-Dar, Sudipto Guha, Yishay Mansour, and S Muthukrishnan. Selective call out and real time bidding. In *Internet and Network Economics*, pages 145–157. Springer, 2010.

Olivier David and Rob A Kempton. Designs for interference. *Biometrics*, pages 597–606, 1996.

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*, 2014.

Liran Einav, Theresa Kuchler, Jonathan D Levin, and Neel Sundaresan. Learning from seller experiments in online markets. Technical report, National Bureau of Economic Research, 2011.

M Elizabeth Halloran and Claudio J Struchiner. Causal inference in infectious diseases. *Epidemiology*, pages 142–151, 1995.

Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 2012.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

David Lucking-Reiley. Using field experiments to test equivalence between auction formats: Magic on the internet. *American Economic Review*, pages 1063–1080, 1999.

Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

Michael Ostrovsky and Michael Schwarz. Reserve prices in internet advertising auctions: A field experiment. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 59–60. ACM, 2011.

David H Reiley. Field experiments on the effects of reserve prices in auctions: More magic on the internet. *RAND Journal of Economics*, pages 195–211, 2006.

Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477), 2007.

D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

Donald B Rubin. Comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, page 0962280210386779, 2010.

David Walker and Lev Muchnik. Design of randomized experiments in networks. *Proceedings of the IEEE*, 102(12):1940–1951, 2014.