A Simulated-Counterfactual Based Experiment Design to Estimate Treatment Effect in Online Ad Auctions

SUBMISSION 133

We introduce a new methodology to estimate the treatment effect in two-sided advertising markets, where a group of advertisers compete in online ad auctions. We consider the case of an advertiser facing treatment, such as changing or introducing a recommendation to optimize the performance of an ad campaign. In this case, advertiser randomized experiments are unable to provide accurate estimates of the treatment effect due to spillovers caused by competition between advertisers. In this paper, we tackle this challenge of estimating the treatment effect by introducing a simulated-counterfactual (*SC*) market where treatment and control advertisers have similar performance, thus yielding negligible spillover effect. The *SC* market is then used to transform the problem into an auction randomized experiment that does not suffer from spillovers and allows us to accurately estimate the treatment effect.

Manuscript submitted for review to the 24nd ACM Conference on Economics & Computation (EC'23).

1 INTRODUCTION

In online advertising, advertisers set up ad campaigns that specify where their ads can appear, and how much they are willing to pay. When there is an ad opportunity, all eligible ads participate in an ad auction to determine a winning ad that gets shown. In the most basic version, advertisers specify bids for different types of ad opportunities, and a first-price auction determines the winner for each opportunity. To help set up performant ad campaigns, an aggregator provides campaign recommendations, e.g., the aggregator recommends a bid for each pair of ad opportunity-type. In this paper, we address the problem of estimating the global treatment effect (GTE), where the treatment is a new set of campaign recommendations.

Ad opportunities are everywhere on the internet, alongside search results, social media, videos, apps, games, and almost every type of content. Advertisers range from large companies with multi-billion dollar annual ad budgets[1] to individuals who spend a few dollars a day. Aggregators could be ad agencies that help advertisers run their ad campaigns, demand side platforms (DSP), which are automated services providing access to a wide range of ad opportunities, or owned and operated (OnO) ad platforms that own a huge supply of ad opportunities such as those by Google and Meta. In 2022, the worldwide spend on online ads was estimated at $566 billion [2].

Suppose that the metric we care about is ad revenue. The GTE is the difference in ad revenue between two worlds: one where all advertisers are given the new recommendations (the *Treatment* T), and one where all advertisers are given the old recommendations (the *Control* C). The gold standard for experiments to estimate GTEs is the randomized control trial (RCT), more commonly called A/B testing. The main difficulty in using RCTs in this setting is the *spillover* effect, resulting from advertisers in Control and Treatment competing with each other in the same ad auctions; we call such auctions *mixed auctions*. This violates the Stable Unit Treatment Values Assumption (SUTVA) [?] that underpins RCTs. In general, there is no simple relationship between the revenue of a mixed auction, and the revenues in the two worlds of all Control, and all Treatment. The example below illustrates this.

Example of mixed auctions. There are only two ads, and one type of ad opportunity. The auction ranks the ads by their bid, and the highest bidder wins and pays bid. The Treatment results in advertisers changing their bids: Ad 1 changes it from $1 to $1.10, while Ad 2 changes it from $0.75 to $1.20. There are two possible outcomes for a mixed auction, depending on which ad is in C and which is in T. Table 1 shows all possible outcomes.

	Ad 1	**Ad 2**	**Bid 1**	**Bid 2**	**Winner**	**Ad revenue**
All Control	C	C	$1.00	$0.75	Ad 1	$1.00
All Treatment	T	T	$1.10	$1.20	Ad 2	$1.20
Mixed auction	C	T	$1.00	$1.20	Ad 2	$1.20
Mixed auction	T	C	$1.10	$0.75	Ad 1	$1.10

Table 1. Example illustrating how the revenue of a mixed auction is in general unrelated to the revenue in Control and in Treatment.

Suppose that we have many copies of this example, and we conduct an RCT where each advertiser is in C and T independently with equal probability. The average revenue per ad in T $= (1.2 + 1.2 + 1.1)/4 =$ \$0.875, since advertisers in T win in 3 of the 4 possibilities listed above. The average revenue per ad in $C = 1/4 =$ \$0.25. Estimating the GTE using 2× the difference in average

[1] https://www.statista.com/statistics/191998/ad-spending-of-procter-and-gamble-in-the-us/.

[2] https://www.statista.com/topics/1176/online-advertising/

revenue per ad in T vs. C gives an estimate of \$1.25, while the actual GTE is just the difference between row 2 and row 1 = $(1.2 - 1)$ = \$0.2, resulting in a 6.25x overestimation.

While this is a hypothetical example, this phenomenon is general, and the problem is prevalent. ? give empirical evidence from an experiment sending marketing emails to eBay sellers that the treatment effect on the auction win rate was almost 0, while an RCT showed statistically significant positive effect. Companies need to continuously innovate to improve their advertisers' and users' experiences, as well as grow their revenue. They release tens of new advertiser features every year, while many more improvements happen behind the scenes. The many types of campaign recommendations include setting performance targets such as a target cost per click (CPC), cost per action (CPA), or cost per conversion value, setting budgets, and setting targeting criteria such as search keywords or audience segments. In addition to recommendations, there are also user interface (UI) changes such as surfacing forecasts or trending terms. They also introduce new features such as an option to pay per action instead of paying per click. The challenge of estimating GTEs is common to all such advertiser facing changes.

In addition to ad revenue, we care about metrics that capture the value to advertisers such as their return on investment. OnO platforms also care about the impact on their users, such as the impact on daily average users, time spent, or the number of clicks and "like"s. It is well understood that the complex nature of these marketplaces require data driven decisions [?], therefore designing reliable advertiser facing experiments is an important problem for the online advertising industry.

2 CONTRIBUTIONS

We introduce a new experiment design framework based on the concept of a *simulated counterfactual*, which simulates the auction outcome under a different input to an ad campaign, for a random fraction of the auctions. This essentially results in an RCT on the auction side, where for a control group of auctions, we use the original inputs to the ad campaign, and for a treatment group of auctions, we simulate counterfactual inputs. Advertisers only see the total outcome of all the auctions, and not the counterfactual inputs used. How we simulate a counterfactual input depends on the type of input, as shown in the examples below.

Bids: Ad platforms offer auto-bidding strategies that let the platforms automatically raise or lower the bids based on real-time signals such as the probability of an action[3]. Suppose that an ad campaign has auto-bidding enabled, and has set a bid of \$1. For a random fraction of the auctions, we can simulate a counterfactual bid of \$0.8 by letting auto-bidding place bids in these auction as if the advertiser set bid was \$0.8.

Performance targets: Auto-bidding campaigns may require as input a performance target such as a target CPA. Suppose that the actual target CPA is \$2 and we want to simulate a target CPA of \$2.5. We run two independent instances of the auto-bidding algorithm, one with target CPA of \$2 on a control group of auctions, and another with a target CPA of \$2.5 on a treatment group of auctions.

Budgets: Suppose that an ad has a budget of \$100, and we need to simulate a budget of \$150, on a random 10% of the auctions. We enforce a budget of \$15 on one treatment group ($T1$) of 10% of auctions, a budget of \$10 on another treatment group ($T2$) of 10% of auctions, and a budget of \$75 on the remaining 80% of auctions. While $T1$ simulates the counterfactual budget of \$150, $T2$ simulates the original budget of \$100.

Targeting: We can simulate counterfactual targeting criteria in cases where the counterfactual targets a subset of the auctions, by stopping the ad from appearing in those auctions that

[3] E.g., Google offers *enhanced CPC*: https://support.google.com/google-ads/answer/6239141

are not targeted by the counterfactual. Since most changes introduced influence targeting inputs to expand rather than restrict the criteria (e.g., add more keywords), we can use such counterfactuals to simulate C when advertisers are in T.

2.1 Proposed Approach

We propose an experiment design framework, the Simulated Counterfactual based Treatment Effect (SCOTE), whose basic unit of measurement is an *auction side RCT* that measures the impact of replacing the campaign inputs with a counterfactual. If we could do this when campaign inputs are a response to Treatment, and the counterfactual is a response to Control, then this would give us the GTE. E.g., in the example in Table 1, if we knew the bids under both C and T, we would run half the auction with bids under C (row 1), and other half of them with bids under T (row 2), and the difference would give us the correct GTE. Since we could only ever observe one of these responses at any time, the main idea is to use a simulated counterfactual (SC) to bridge the gap. We run two auction side RCTs, one measuring the impact of T vs. SC (the measurement phase), and another measuring the impact of SC vs. C (the calibration phase), which together give us the T vs. C GTE.

The key to the success of this approach is to get a SC that mimics C sufficiently well. We propose two generic approaches to defining a SC. While the measurement phase is identical for both, the calibration phase is different.

Transformation based SC. This approach is useful for inputs such as targeting criteria where we need to make sure that the SC inputs err on one side of the real inputs. In this approach, the counterfactual input is obtained by applying a transformation to the campaign input in response to T (See Figure 1). As a result, we change the campaign inputs only for advertisers in T. E.g., in SC, for advertisers in T, we drop those keywords that are recommended in T but not in C. The changes are made in the back-end at the auction level and are not noticeable by the advertisers. The exact transformation is learnt during the calibration phase, which is run concurrently to an advertiser side RCT. In this case, we cannot exactly measure SC vs. C, but we try to make it as close to 0 as possible.

More formally, suppose that we have split the advertisers into two groups, A and B, chosen at random with 50% probability each. We consider three situations, 1) all advertisers are in C, 2) all advertisers are in T, and 3) advertisers in A are in C and advertisers in B are in T. On the auction side, we have two possibilities, where either we change the campaign inputs for the simulated counterfactual (SC), or we do not change the campaign inputs, which we denote by Z. Denote by $R_i(x, y, z)$ the revenue generated by advertiser i when advertisers in A are in x, advertisers in B are in y, where $x, y \in \{C, T\}$, and the campaign inputs are in $z \in \{SC, Z\}$. We wish to measure the global treatment effect GTE $= \sum_{i \in A \cup B} R_i(T, T, Z) - R_i(C, C, Z)$, which can be decomposed as follows:

$$\underbrace{\sum_{i \in A \cup B} R_i(T,T,Z) - R_i(T,T,SC)}_{\chi_1} + \underbrace{\sum_{i \in A \cup B} R_i(T,T,SC) - R_i(C,C,SC)}_{\chi_2} + \underbrace{\sum_{i \in A \cup B} R_i(C,C,SC) - R_i(C,C,Z)}_{\chi_3}. \tag{1}$$

In the measurement phase, we measure χ_1 using an auction side RCT with two treatments Z and SC. During this phase, all advertisers are in T. By definition of SC, $\chi_3 = 0$ because the changes in the campaign inputs in the SC auctions are applied only to advertisers in T. E.g., we only change the keywords of treated advertisers. We use the calibration phase to define SC so that $\chi_2 \approx 0$,

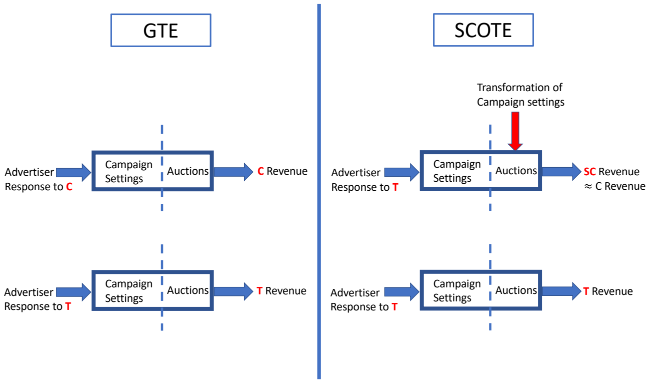

Fig. 1. GTE is the difference in revenue of two hypothetical worlds: (1) all advertisers in T vs. (2) all advertisers in C. SCOTE is the difference in revenue of two hypothetical worlds: (1) all advertisers in T vs. (2) all advertisers in T with their campaign inputs transformed to make SC revenue close to C revenue.

which gives us GTE $\approx \chi_1$. In other words, we can measure GTE using an auction side RCT, with all the advertisers in T.

The question is, "how to define SC?" Intuitively, we want to reverse the effect that T has on the inputs, e.g., if T caused advertisers to choose 10% more keywords, we want SC to drop those 10% keywords. We run an advertiser side RCT to measure the impact that T has on campaign inputs. Using this, we determine qualitatively which inputs need to be changed. For the exact definition of SC, we use tunable parameters, such as the % of keywords to be dropped or the factor by which bids are decreased. We calibrate these parameters so that advertisers in C and T have roughly the same ad revenue in the SC auctions, i.e., $\sum_{i\in B} R_i(C,T,SC) - \sum_{i\in A} R_i(C,T,SC) \approx 0$. As we will show, making T advertisers similar to C advertisers in terms of revenue weakens drastically the spillover effects and ensures that

$$\chi_2 \approx 2\left(\sum_{i\in B} R_i(C,T,SC) - \sum_{i\in A} R_i(C,T,SC)\right) \approx 0. \tag{2}$$

We have implicitly assumed that the marketplace does not change between the calibration and measurement phases. Marketplace variation over time is one of the biggest challenges in estimating GTE. If not we could simple do a pre-post analysis to get the GTE, i.e, measure $\sum_{i\in A\cup B} R_i(T,T,Z)$ and $\sum_{i\in A\cup B} R_i(C,C,Z)$ during different time periods and take the difference.

We use simulations as well as simple stochastic models to show the following for the transformation based approach to SC:

- We show that as long as there exists a multiplier that transforms the bid distribution of T advertisers to that of C advertisers (Matching Assumption), our SC based estimator of GTE is strongly consistent. Moreover, the rate of convergence of the estimator to GTE is $O(N^{-\frac{1}{2}+\delta})$, where N is the number of auctions.
- We show via a numerical example that our SC based estimator outperforms by a large margin advertiser side RCT, cluster RCT, and pre-post experiment in terms of bias and precision. The key feature of this high performance is the ability of the simulated counterfactual market to

suppress the spillover effects. Our simulations show that this feature makes our estimator robust w.r.t. the violation of the Matching Assumption. Moreover, the results show that the *SC* based estimator is robust to marketplace variations over time, which pose a real challenge to pre-post experiments.

Model based SC. This approach is suitable for campaign inputs for which we can simulate both an increase as well as a decrease in the input value, such as budgets. In this approach, the counterfactual input is obtained from a machine learned (ML) model that uses historic data on what campaign inputs advertisers have set under C, to predict what campaign inputs advertisers would set in the future if they were still in C. Unlike the transformation based approach, we can replace the campaign inputs with model inputs even for advertisers in C. Indeed, we first run the calibration phase when all advertisers are still in C: half the auctions use the actual advertiser inputs in response to C, while the other half use inputs predicted by the ML model. Ideally we would like to observe a 0 GTE in the calibration phase, which would happen if the ML model perfectly predicted the advertiser inputs. The errors in the model prediction result in a non zero *SC* vs. C impact, but since we can measure it, we can correct for this. We may repeat this phase multiple times to tune the ML model to get *SC* vs. C as close to 0 as possible. We then expose all advertisers to T and run the measurement phase, which measures T vs. *SC*. E.g., for the example in Table 1, suppose that the model predicts that the bids of the two advertisers are $1 + \epsilon_1$ and $0.75 + \epsilon_2$ resp. As long as the ϵs are small enough, under *SC*, Ad 1 still wins and the ad revenue is $1 + \epsilon_1$. An auction side RCT to measure *SC* vs. C shows a difference of $1 + \epsilon_1 - 1 = \epsilon_1$ ad revenue per auction, and the T vs. *SC* auction side RCT shows an increase in ad revenue per auction of $1.2 - (1 + \epsilon_1) = 0.2 - \epsilon_1$, adding which we get the true GTE of \$0.20 ad revenue per auction.

Going back to the GTE decomposition given in (1), as before in the measurement phase, we estimate χ_1 using an auction side RCT. Unlike previously, in the calibration phase we measure χ_3 using an auction side RCT because we can use model inputs as *SC* even when the advertiser is in C. In this case, $\chi_2 = 0$ because once you replace the advertiser inputs with the model inputs, it does not matter whether the advertiser is in C or T. Thus $\chi_1 + \chi_3$ is an unbiased estimator of GTE. This is a surprising conclusion, in that it does not depend on the model accuracy! Nonetheless, model accuracy still matters because it affects the variance of this estimator. In addition, model accuracy is also important to make the estimator robust against marketplace changes. We use simulations to show the following for the model based approach to *SC*, for changes in campaign daily budgets:

- As the number of days increases, the estimated ATE converges to the true ATE, while being much more accurate than other methods even with few days of data.
- However, model accuracy is still a desirable property, as the rate of convergence of the estimated ATE to true ATE is determined by the mean squared error (MSE) of the ML model.
- The error in the estimate is monotone with the magnitude of marketplace changes between the calibration phase and the measurement phase, but the estimates are still much more accurate than other methods.

2.2 Pros and cons of the proposed approach

Pros. All other methods to estimate GTE suffer from high bias or high variance or both. Our approach gets the best of both worlds.

- We avoid the spillover effects by design, and give estimates that are unbiased or have low bias. Our simulations show that our estimates are robust even when there are large spillover effects.
- Furthermore, the number of auctions is typically much larger than the number of advertisers, yielding high statistical power compared to advertiser randomized experiments.

Cons.

- Additional effort is required to implement our proposed design. We need to develop ML models to predict advertiser inputs, or run calibrations to identify good input transformations. Implementing a simulated counterfactual requires effort, e.g., to simulate different budgets we need the capability of tracking spend and enforcing budget by auction treatment groups. In addition to an advertiser side RCT, we need to run two auction side RCTs.

Potential Cons that are mitigated.

- Simulating counterfactual inputs may affect advertiser performance, such as resulting in fewer impressions, clicks, etc. We mitigate this by having SC run on a small percentage of traffic. Since the number of auctions is typically very large (in many millions and often in billions), even a small percentage of auctions is sufficient to get statistically significant results. (We use 50% in the paper just for the ease of presentation.)
- We make certain assumptions, such as the Matching assumption that a bid multiplier transforms the T bid distribution to the C distribution, to show our theoretical guarantees. Even if there exists such a multiplier, we may not be able to find it exactly. However, our simulations show that the estimation accuracy degrades gracefully with the calibration error, either in finding the right transformation or in the ML model.
- Since we run the calibration and measurement phases at different times, the marketplace might change in between. Indeed, if there were no such changes, we could get an accurate estimate of GTE using a simple pre-post analysis: expose all advertisers to T starting some time period t, and use the difference between the ad revenue in time t and that in time $t-1$. However, our simulations indicate that our experiment design is robust to marketplace changes. We do not currently have any theoretical analysis of this robustness, and this is an interesting avenue for future research.

3 RELATED WORK

Network spillover effects. The spillover effect in the online advertising setting that we consider is a special case of network spillover effects, where there is a graph that captures interactions among treatment units. This includes other two sided marketplaces such as ride sharing, as well as social networks. Significant academic literature exists in measuring spillover effects in the context of social networks. Typical suggestions include assigning weights based on an ML model that estimates the impact of the influence of the admixture from historical data, similar to synthetic control methods [?]. However, this cannot be done for two-sided markets as the effects of mixed auction are not estimable easily, in particular, we cannot even conclude that mixed auction effects are monotone in the proportion of Treatment vs Control in the experiment, which is typically the minimal assumption required [??].

Cluster randomized experiments. The most common approach to address network spillover effect is cluster randomization, where the graph is clustered so that most of the interaction occurs within a cluster, and each cluster is entirely assigned to either C or T [??]. The main drawback of this is that it reduces statistical power: ? demonstrate a method to increase statistical power in cluster-randomized experiments via a method called *independent block randomization* (IBR). Nonetheless, this cannot address the main reason for the drop in statistical power, which is that we reduce typically millions of advertisers to maybe tens of thousands of clusters. Another drawback of this design is that cross-cluster leakage leads to biased estimators. Recently, ? show that in the context of two-sided marketplaces, standard clustering objectives are not aligned with minimizing bias, and derive a bias minimizing objective. This cannot address the underlying issue where in the online advertising setting, the graph is so well connected that there simply are no good clusters, especially when you need thousands of them. Another extension of cluster randomized experiments is geo-randomization [??], which uses geographic regions as natural clusters. However, this does not mitigate the issue of statistical power.

Double randomization. Double randomization is another experiment design technique used to mitigate spillover effects in two sided markets [??]. The main difference is that these are meant for treatments that could be applied to a pair of advertiser and auction. E.g., the treatment could be to show extra information when an ad is viewed, which could be shown for all the ads for a given auction, for some ads for all the auctions, or for any combination of ad, auction pairs. In this setting, **?** provide guidance on bias-variance trade offs in choosing advertiser side vs. auction side randomization, as well as choosing the proportion of treated units. Our treatments are inherently only applicable to ads therefore these are not applicable.

Switch-back based design. Switch-back is another popular method used to avoid spillover, where we switch back and forth between control and treatment (for all advertisers) at fixed or random times [??]. However, switch-back experiments have the issue that past interventions are likely to impact future outcomes (referred to as a carryover effect). To mitigate this the experiment may need to be run for a really long time. That is to say, when we estimate the impact in Treatment triggered the week after Control, the temporal effect of Control recommendations will continue to impact the decisions that customers make when Treatment is launched. As a consequence, the experiment may need to be run for a really long time. Moreover, switch-back is not possible for UI changes that are apparent to the advertiser, since it would lead to a bad advertiser experience. In addition, inference from switch-back experiments often require strong assumptions, which are often not realistic [??]. Mitigating these issues require extensive and continuous offline testing and modeling.

Non randomized methods. One can also use methods that rely on observational data alone to make inference of causal impact. E.g., we can just switch over from C to T for all advertisers at a particular point of time, and use the difference in the metrics pre and post switching to estimate the impact. Advanced techniques such as difference in differences, synthetic controls, and debiased ML **?** could be potentially applied here, although these do not directly address the spillover effect. Observational methods are typically inferior to RCTs, and **?** show using experiments on Facebook that observational methods often diverge from the results of RCTs. Our method has some similarity to synthetic controls, but unlike those we make extensive use of RCTs.

4 CHANGING CAMPAIGN BIDS

4.1 Model and Problem Statement

We consider a two-sided market where a group of n ads are competing in N ad auctions. Advertisers decide their bids before the start of any auction. We denote by b_i^C the bid of ad i, and assume that b_i^C are drawn independently from a probability distribution with cumulative distribution function (CDF) F_C and support in $[0, \infty)$. At each time step $t \in \{1, \ldots, N\}$, we conduct an ad auction, which has a participation rate ρ_t drawn from a uniform distribution $U(a, b)$, where $0 < a \leq b \leq 1$. The participation rate ρ_t represents the percentage of ads whose targeting criteria includes this auction. Each ad participates in the auction w.p. ρ_t, independently. To simplify the analysis, we consider *first-price auctions*: the ad with the highest bid wins the auction and pays her bid. All the results can be easily extended to second price auctions.

We intervene on a random fraction ρ_I of the ads and increase their bids, for example, by recommending competitive bids. The intervened ads will have their bids b_i^T drawn independently from a distribution with CDF F_T and support in $[0, \infty)$. We assume that the mean of F_T is greater than that of F_C. Our goal is to estimate the GTE of bid increase, i.e.

$$\begin{aligned}\text{GTE} = &\text{(spend of intervened + spend of non-intervened, when all intervened are in } T\text{)}\\ &\text{- (spend of intervened + spend of non-intervened, when all intervened are in } C\text{).}\end{aligned} \tag{3}$$

4.2 SCOTE Estimation of GTE

Following Section 2.1, our simulated counterfactual based experiment has two phases. In the first phase we calibrate the auction treatment *SC*, which is defined as follows. In *SC* auctions, we multiply by $m > 0$ all the bids of treated advertisers that participate in the auction. The treated advertisers will not notice the changes in their bids as the changes are made in the back-end at the auction level. The parameter m is calibrated by running an advertiser side RCT, where 50% of intervened advertisers are in T. On the auction level, we apply and calibrate m so that

$$\begin{aligned}\widehat{C(m)} := &\text{ average spend of an intervened advertiser in } T\\ &- \text{average spend of an intervened advertiser in } C = 0.\end{aligned} \tag{4}$$

We use binary search to the find the multiplier $\widehat{m}$ that satisfies equation (4). Once we find $\widehat{m}$, we launch the advertiser Treatment (i.e. bids $\sim F_T$) to all intervened advertisers and run an auction side RCT where in 50% of auctions (*SC* group) we apply the multiplier found in the first phase, while in the rest (*Z* group), we do not change the bids set by advertisers (i.e. multiplier = 1). The experiment is summarized in Algorithm 1.

ALGORITHM 1: Simulated-Counterfactual Based Experiment Design

```
Input: Binary search error $\epsilon_D$
Output: Estimate $\widehat{\mathcal{E}}$ of GTE
$m_l = 0$; $m_h = 1$;
First Phase: Calibration: 50% of intervened advertisers are in $T$;
repeat
    $m_c = (m_l + m_h)/2$;
    Bids of $T$ advertisers are multiplied by $m_c$;
    $\widehat{C(m_c)}$ = average spend of intervened advertisers in $T$ - average spend of intervened advertisers in
      $C$;
    if $\widehat{C(m_c)} > 0$ then
        $m_h = m_c$;
    else
        $m_l = m_c$;
    end
    $error = m_h - m_l$
until $error \leq \epsilon_D$;
$\hat{m} = m_c$;
Second Phase: Estimation: all intervened advertisers are in $T$;
Auction side RCT: 50% of auctions have multipliers $\hat{m}$ and the rest $m = 1$;
$\widehat{\mathcal{E}}$ = 2(revenue generated by auctions with $m = 1$) - 2( revenue generated by auctions with multiplier $\hat{m}$)
```

In the following, we derive a theoretical bound on the accuracy $|\widehat{\mathcal{E}} - \text{GTE}|$ of the estimator $\widehat{\mathcal{E}}$ described in Algorithm 1. Moreover, we show that the estimator is strongly consistent. The proofs of the results are given in the Appendix. We distinguish between the bid b_i set by an advertiser, and her actual bid y_i that enters the auction. The actual bid y_i depends on whether the advertiser participates in the auction or not (it is equal to 0 if i does not participate), and on the bid multiplier m of the auction. Formally, y_i can be written as follows:

$$y_i = p_i((1 - I_i)b_i^C + I_i(T_i m b_i^T + (1 - T_i)b_i^C)), \tag{5}$$

where $p_i = 1$ if i participates in the auction and 0 otherwise, $I_i = 1$ if i is intervened and 0 otherwise, $T_i = 1$ if i is treated and 0 otherwise. We denote by ρ_T the fraction of intervened advertisers in T (exposed to the new bids F_T). We introduce the following quantities, which will be used later. Given that the fraction of intervened advertisers in T is ρ_T, and that the auctions are treated with multiplier m, the mean revenue generated by an auction is $r(m, \rho_T) = \mathbb{E} \max_i y_i$, the mean spend of a T advertiser is $s_T(m, \rho_T) = N\mathbb{E}(y_i 1(y_i \geq \max_j y_j)|T_i = 1, I_i = 1)$, and the mean spend of a control advertiser is $s_C(m, \rho_T) = N\mathbb{E}(y_i 1(y_i \geq \max_j y_j)|T_i = 0, I_i = 1)$.

Assumption 4.1 (Matching Assumption). There exists a multiplier m^* such that the distributions of $m^* b_i^T$ and b_i^C are equal, i.e. $F_{m^*} = F_C$.

Assumption 4.1 is satisfied, for example, by exponential and uniform distributions. We can relax Assumption 4.1 by requiring that $\|F_{m^*} - F_C\|_{TV} \leq \epsilon$ for some small $\epsilon > 0$, without changing drastically the results of this section. We will show later in the simulations that when Assumption 4.1 is violated, we can still get good estimates of GTE. In our future work, we will analyze the more general case where Assumption 4.1 is satisfied by general bid transformations $B : b_i^T \to B(b_i^T)$.

We define the calibration function as the mean of $\widehat{C}$ defined in (4), i.e. difference between the mean spend of a treated advertiser and mean spend of a control advertiser in an advertiser side RCT with auctions treated with multiplier m:

$$C(m) = s_T(m, 0.5) - s_C(m, 0.5). \tag{6}$$

The following Lemma states if we find a multiplier m^* that satisfies the calibration condition (4) in expectation, then the estimator $\widehat{\mathcal{E}}$ of GTE is unbiased, meaning that $\mathcal{E} = \mathbb{E}\widehat{\mathcal{E}} = \text{GTE}$. The mean estimate $\mathcal{E}$ is the difference between the mean of the total spend of advertisers when all intervened advertisers are in T and auctions have multipliers 1 and the total spend of advertisers when all intervened advertisers are in T and auctions have multipliers the solution of the first phase:

$$\mathcal{E} = N \times (r(1, 1) - r(m^*, 1)). \tag{7}$$

Lemma 4.2. *The following statements hold:*

(1) The calibration function (6) *is increasing and has a unique zero at m^*.*
(2) The mean of the estimate $\mathcal{E}$ is equal to GTE.

GTE decomposition described in Section 2.1 is:

$$\text{GTE} = N \times (r(1, 1) - r(1, 0)) = \mathcal{E} + \underbrace{N \times (r(m^*, 1) - r(m^*, 0))}_{\chi_2} + \underbrace{N \times (r(m^*, 0) - r(1, 0))}_{\chi_3}. \tag{8}$$

Following Section 2.1, the last term χ_3 is zero for any multiplier m. The term χ_2 is the difference between the spend of all advertisers when all intervened advertisers are in T and the spend of all advertisers when all intervened advertisers are in C, given that all the auctions have multipliers m^*. This is a bit surprising! We are saying that in a market with m^* auctions only, the treatment effect χ_2, can be perfectly measured by an advertiser side RCT (both the impact χ_2 and the mean of $\widehat{C(m^*)}$ are equal to 0), despite the competition between the advertisers. To see what is happening, let us decompose the term χ_2 for a general multiplier m. We have

$$\begin{aligned} N \times (r(m,1) - r(m,0)) = n\rho_I \Big[& \underbrace{(s_T(m,1) - s_T(m,0.5))}_{\gamma_T(m)} + \underbrace{(s_T(m,0.5) - s_C(m,0.5))}_{C(m)} \\ & - \underbrace{(s_C(m,1) - s_C(m,0.5))}_{\gamma_C(m)} \Big] + n(1-\rho_I) \underbrace{(s_C(m,1) - s_C(m,0))}_{\gamma_o(m)} . \end{aligned} \tag{9}$$

γ_T is the difference between the spend of a T advertiser when all intervened advertisers are in T and her spend when 50% of intervened advertisers are in T. Thus, this term measures the spillover effects from a treated advertiser perspective. Similarly, γ_C measures the spillover effects from a C advertiser perspective, and γ_o measures the spillover effect from a non-intervened advertiser perspective. When Assumption 4.1 is satisfied, the spillover terms vanish at m^*. In other words, by making the treated advertisers similar to control advertisers, we create an *SC* market without spillover effects. This allows us to accurately estimate GTE by running an auction side RCT. In the simulations, we will provide an example where Assumption 4.1 is violated but the spillover terms remain small at m^*, yielding accurate estimate of GTE.

Lemma 4.2 states that if we find m^* in the first phase, then the estimation is unbiased. However, we don't have access to the calibration function (6), but samples $\widehat{C}$ with mean C that we obtain from the advertiser side RCT in the first phase. Moreover, binary search runs for a finite number of steps. The following theorem accounts for these facts by giving a bound on the accuracy of the estimation.

Theorem 4.3. *There exist constants C_1 and C_2 that do not depend on the number of auctions N, such that, for all $\epsilon_D > 0$ and $\epsilon > 0$ with $\epsilon > C_2\epsilon_D$,*

$$\begin{aligned} \mathbb{P}(|GTE - \widehat{\mathcal{E}}| \leq N\epsilon) \geq & \left(1 - 2\exp\left(-\frac{N(\epsilon - C_2\epsilon_D)^2}{2B_m^2}\right)\right)^2 \left(1 - \exp\left(-\frac{NC_1^2\epsilon_D^2}{16B_m^2}\right)\right)^2 \\ & \times \prod_{s=1}^{\lceil \log_2(1/\epsilon_D) \rceil - 2} \left(1 - \exp\left(-\frac{NC_1^2(1+2^s)^2\epsilon_D^2}{16B_m^2}\right)\right), \end{aligned} \tag{10}$$

where $\widehat{\mathcal{E}}$ is the estimate of GTE computed by our method described in Algorithm 1. ϵ_D is the maximum error at the output of the binary search in Algorithm 1.

Corollary 4.4. *$\frac{1}{N}\widehat{\mathcal{E}}$ is a strongly consistent estimator of $\frac{1}{N}$GTE. In particular, for any $0 < \delta < 1/2$, $\frac{1}{N}|GTE - \widehat{\mathcal{E}}| = O(N^{-\frac{1}{2}+\delta})$ w.p. 1.*

We discuss in Section A.2 the implications of the lower bound in (10).

4.3 Simulations

We consider $n = 150000$ advertisers with bids drawn from an exponential distribution F_C with mean 1. We wish to increase the bids of intervened advertisers by 10%, i.e. the bids of intervened advertisers would be drawn from an exponential distribution F_T with mean 1.1. The fraction of intervened advertisers is $\rho_I = 0.4$. We assume that the market is changing on a weekly basis. In particular, during week t_1 the participation rate ρ_t is drawn from $U(0, 0.4)$, while during the following week t_2 it follows $U(0.1, 0.5)$ (this assumption does not affect our theoretical results). We run the calibration phase during week t_1 and use binary search to compute the optimal multiplier, i.e. the multiplier of *SC* auctions $m = 0.90909$. We run the measurement phase during week t_2.

				SCOTE	SCOTE	SCOTE
	GTE	**Adv RCT**	**Pre-post**	**MEE = 0%**	**MEE = 5%**	**MEE = -5%**
Mean estimate	5763.922	57703.824	9818.573	5763.921	4409.04	6694.42
Relative bias	0	901.12%	70.34%	-1.06×10^{-5}%	-23.50%	16.14 %

Table 2. Estimation of GTE for changing campaign bids, where F_C and F_T are exponential. MEE = Multiplier Estimation Error

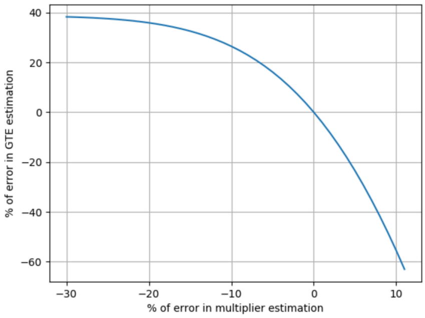

Fig. 2. GTE estimation error as a function of the error of estimating the optimal multiplier

We compare our method SCOTE to advertiser side RCT, and pre-post method. In the pre-post experiment, we launch the new bid recommendations to all intervened advertiser during week t_2. In this case, the impact is estimated as the difference between the revenue during t_2 and t_1. It should be noted that the advertiser interaction network (graph modelling who compete with whom in the auctions) is sampled from the set of complete subgraphs of the complete graph including all advertisers, where the probability of a subgraph with k advertisers is equal to $[0.5(a+b)]^k > 0$. Thus, it is impossible to define clusters of competing advertisers and run cluster randomized experiments. Moreover, since the changes introduced are observable by the advertisers (new bid suggestions, for example), running switch-back based experiment would lead to a bad advertiser experience.

Table 2 summarizes the mean of SCOTE, advertiser side RCT, and pre-post estimators (mean estimate). For each method, we report the relative bias, which is equal to (mean estimate)/GTE-1. We see that the SCOTE estimator outperforms the two other estimators, even when there is an error in the estimation of the multiplier in the first phase. Figure 2 shows how the estimation error of the optimal multiplier m^* during the first phase translates into estimation error of GTE during the second phase. In particular, overestimation of the multiplier leads to underestimation of GTE and vice versa.

Figure 3 gives the profile of the spillover effects terms γ_T, γ_C, and γ_o as functions of the multiplier m. We see that all the terms vanish at the optimal multiplier m^*, which is the key property of SC that makes SCOTE estimate of GTE accurate. Interestingly, γ_T has a non-monotonic behavior, which can be explained as follows. When m is very small, the winning probability of a T advertiser is almost zero. As m increases, this probability increases at higher rate when $\rho_T = 1$ than when $\rho_T = 0.5$, because T advertisers don't have very competitive bids when m is small. Thus, γ_T increases. When m becomes large enough, T advertisers' bids become very competitive causing the rate of

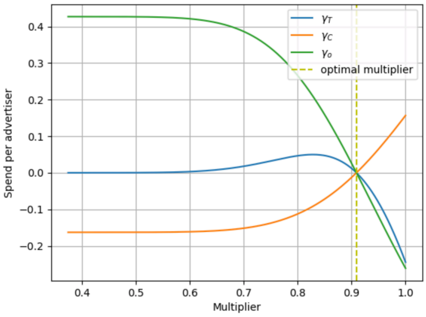

Fig. 3. Spillover effects vanish at the optimal multiplier

	GTE	Adv RCT	Pre-post	SCOTE MEE = 0%	SCOTE MEE = 5%	SCOTE MEE = -5%
Mean estimate	16035.85	172054.17	24155.62	16466.12	13797.68	18244.62
Relative bias	0	972.93%	50.63%	2.68%	-13.95%	13.77 %

Table 3. SCOTE is robustness w.r.t. violation of Assumption 4.1. Estimation of GTE for changing campaign bids, where F_C and F_T are χ^2

increase of the winning probability of a T advertiser to increase at higher rate when $\rho_T = 0.5$ than when $\rho_T = 1$. Thus, γ_T starts decreasing.

Finally, we illustrate through an example the effects of violating Assumption 4.1. In particular, we assume that $b_i^C \sim \chi^2(2)$ and $b_i^T \sim \chi^2(3)$. Table 3 shows that SCOTE estimation is robust to violation of Assumption 4.1 (Relative bias equal to 2.68%) and it outperforms by a large margin the two other methods. The robustness of the method comes from the fact that the calibration phase is weakening the spillover effect terms despite the violation of the Matching Assumption. In fact, the spillover effect terms are close to zero at the optimal m^* computed during the first phase: $\gamma_T = -0.0028$, $\gamma_C = 0.0019$, and $\gamma_o = -0.0032$.

4.3.1 Comparison with Cluster Randomized Experiments. To compare SCOTE with cluster randomized experiments, we consider the same example at the beginning of this section, but we assume 15000 advertisers partitioned into 100 cliques (each clique is a complete graph), each containing approximately 150 advertisers. We assume that each clique has x% of its advertisers shared with some other cliques. x is called the percentage of spillovers (we give in the appendix (Figure 11) an example of a graph with 200 advertisers, 5 social cliques and 10% spillovers). We consider 50000 ad auctions. Each ad auction involves only one clique, i.e. before the start of the auction, a clique is chosen randomly, and each advertiser belonging to that clique will participate w.p. 0.4 in the auction. For the cluster randomized experiment, the clusters are defined to be equal to the social cliques, while making sure that the advertisers that belong to more than one clique are assigned to only one of the corresponding clusters. We repeat 50 times the cluster randomized experiment, advertiser RCT, and SCOTE. Tables 4 and 5 report the average estimates, as well as the average confidence intervals (CI) for different percentages of spillovers x. Although cluster RCT performs better than advertiser RCT, SCOTE outperforms cluster RCT in terms of bias and variance. For more details about the distribution of the results for the 50 performed experiments, refer to Section A.3.

	GTE	Adv RCT	Cluster RCT	SCOTE MEE = 0%	SCOTE MEE = 5%	SCOTE MEE = -5%
Avg Estimate (5% Spillover)	9707.06	265798.25	14125.92	10117.37	5323.86	14697.04
Relative Bias	0%	2638.2%	45.5%	4.23%	-45.15%	51.41%
Avg Estimate (10% Spillover)	10073.52	263938.77	20140.67	10453.88	5442.88	14985.85
Relative Bias	0%	2520.12%	99.94%	3.78%	-45.97%	48.76%
Avg Estimate (20% Spillover)	10494.08	329906.34	22988.86	10811.83	5599.46	15654.41
Relative Bias	0%	3043.73%	119.06%	3.03%	-46.64%	49.17%

Table 4. Comparison of SCOTE and cluster randomized experiments - Changing bids - Average estimates

Spillover	Adv RCT CI	Cluster RCT CI	SCOTE CI (MEE = 0%)
5%	[83387.22 , 448209.3]	[-13547.86 , 41799.71]	[8948.3 , 11286.45]
10%	[59779.86 , 468097.69]	[-15907.45 , 56188.8]	[9254.46 , 11653.31]
20%	[83771.47 , 576041.22]	[-25445.25 , 71422.98]	[9616.29 , 12007.37]

Table 5. Comparison of SCOTE and cluster randomized experiments - Changing bids - Average CIs

	GTE	Adv RCT	Pre-post	SCOTE TREE=0%	SCOTE TREE=5%	SCOTE TREE=-5%
Mean estimate	3701.194	40584.47	14587.36	3701.194	3483.36	3923.42
Relative bias	0	996.52%	294.12%	-1.13×10^{-5}%	-5.88%	6.004 %

Table 6. Estimation of GTE for changing campaign targeting. TREE = Throttling Rate Estimation Error

5 CHANGING CAMPAIGN TARGETING

We consider the model described in Section 4. Instead of changing the bids of intervened advertisers, we wish to increase their participation rate. Each auction has two participation rates ρ_t^C and ρ_t^T drawn from two different distributions. Accordingly, a non-intervened advertiser will participate in the auction w.p. ρ_t^C, while an intervened one w.p. ρ_t^T. One can think of this change as suggesting to the intervened advertisers new keywords or broader audience segments. In order to simulate a counterfactual market, we throttle in auctions the bids of treated advertisers with some throttling rate m, i.e. if a treated advertiser is participating in an auction, we set her bid to zero w.p. $1 - m$, and w.p. m we don't change her bid. During the first phase, we calibrate m to satisfy (4). All the results of Section 4 can be extended to the targeting case. Table 6 gives the results of the simulations, where we consider the same example of Section 4.3, but instead of changing the bids, we change the participation rate distribution of intervened advertiser to $U(0.2, 0.6)$. Moreover, we assume that the market is evolving over time, where the bids increase by 10% from week t_1 to week t_2. According to the table, SCOTE is much more accurate than the two other methods.

5.0.1 Comparison with Cluster Randomized Experiments. To compare the performance of SCOTE with cluster RCT in case of changing targeting, we add to the model the same graph structure,

	GTE	Adv RCT	Cluster RCT	SCOTE TREE=0%	SCOTE TREE=5%	SCOTE TREE=-5%
Avg Estimate (5% Spillover)	9074.53	473681.91	13797.50	9100.12	8035.61	10109.19
Relative Bias	0%	5119.9%	52.05%	0.28%	-11.45%	11.4%
Avg Estimate (10% Spillover)	9041.86	474731.37	18950.04	9021.99	8034.26	10055.30
Relative Bias	0%	5150.38%	109.58%	-0.22%	-11.14%	11.21%
Avg Estimate (20% Spillover)	9466.72	587526.29	19746.69	9523.15	8430.15	10545.38
Relative Bias	0%	6106.23%	108.59%	0.6%	-10.95%	11.39%

Table 7. Comparison of SCOTE and cluster randomized experiments - Changing targeting - Average estimates

Spillover	Adv RCT CI	Cluster RCT CI	SCOTE CI (TREE = 0%)
5%	[259685.85 , 687677.98]	[-13927.53 , 41522.53]	[7963.49 , 10236.74]
10%	[236605.02 , 712857.73]	[-18423.33 , 56323.41]	[7886.53 , 10157.44]
20%	[294021.56 , 881031.02]	[-33390.99 , 72884.37]	[8334.93 , 10711.37]

Table 8. Comparison of SCOTE and cluster randomized experiments - Changing targeting - Average CIs

number of advertisers and ad auctions described in Section 4.3.1, and we assume that the participation rate changes from 0.4 for C to 0.6 for T. Tables 7 and 8 show that outperforms advertiser RCT and cluster RCT in terms of bias and precision (CI).

6 CHANGING CAMPAIGN BUDGETS

6.1 Model and Problem Statement

Advertisers have a number of constraints, budgets being one of them. Budgets refer to the maximum amount an advertiser can spend in a day. As advertisers continue to spend throughout the day (via auction wins), they may fully expend their budget before the day is over, which makes them ineligible to bid in relevant auctions once they run out of budget. In this section, we discuss a method to use SCOTE to measure the impact of budget recommendations in two-sided markets.

As before, we consider a two-sided market where a group of n ads compete in N ad auctions in a day. Advertisers have fixed bids and budgets prior to an auction. Assume the budgets of the advertisers are currently generated from a distribution $F_C = F(\mu_C, \sigma_C^2)$, where (μ_C, σ_C^2) is determined by the platform's current recommendations. Now, via a model change, assume the recommendations are changed to (μ_T, σ_T^2), which leads to budgets now being generated from $F_T = F(\mu_T, \sigma_T^2)$. Assume advertisers generate a bid from the distribution $G_b(.)$, and the bid generation distribution remains invariant over time, and is the same for all advertisers. Assume that both the budgets and the bid distribution have their supports contained in $\mathbf{R}^+$. Our goal, as before is to estimate the GTE of the change in the budgets distribution.

As in section 4, assume at each time point, we run an ad auction with a participation rate ρ_t, where each eligible advertiser (i.e., an advertiser whose spend upto time t has not exceeded their budget) participates with probability ρ_t. The advertiser with the highest bid wins, and pays their bid if their current spend plus the winning bid does not exceed their budget, else they pay the difference of their budget and current spend. If no advertiser is eligible, no auction is conducted and therefore no ad is served to the audience. The participant rate distribution can vary from week to week, reflecting changes in the marketplace level. After N requests have arrived, the day ends. The same process is repeated for T days.

6.2 SCOTE Estimation of GTE for Budgets

For an advertiser randomized experiment, if treated advertisers do increase their budgets, they are able to remain in budget longer compared to their non-treated counterparts, and hence face less competition in latter hours of the day. Advertiser side RCT estimate are therefore inaccurate, and overestimate the GTE.

As we can simulate both an increase and a decrease in budgets via ML models, the model-based SC design is ideal. Assume that we have built an ML model using historical data based on campaign inputs in C, such that the Budget in SC for the corresponding advertiser is given by $B_i^{ML} = B_i^C + \epsilon_i$, where ϵ_i is generated from a distribution $G_{ML}(\mu_{ML}, \sigma^2_{ML})$. Here μ_{ML} denotes the bias of the ML model, and σ^2_{ML} the error. We aim to train the model such that the bias and the variance are low, so that the true budget in Control is very close to the ML predicted budget.

Algorithm 2 is the algorithm for obtaining SCOTE Based estimate for Budgets. The SCOTE estimate for the GTE is given by

ALGORITHM 2: Simulated-Counterfactual Based Experiment Design

```
Input: Model Bias and Error ($\mu_{ML}, \sigma^2_{ML}$)
Output: SCOTE estimate $\hat{\mathcal{E}}$
 First Phase: Calibration: 100% of intervened advertisers are in $C$;
Auction side RCT: Two treatments $T_{ML}$: 50 % of auctions are assigned to budgets $B_i^{ML}$
$T_C$: the rest 50% are assigned to budgets $B_i^C$.
Advertisers are allowed to bid in $T_{ML}$ as long as spend of campaigns in $T_{ML}$, $S_{i,ML} \leq 0.5 \times B_i^{ML}$.
  Similarly, advertisers are allowed to bid in $T_C$ as long as $S_{i,T} \leq 0.5 \times B_i^C$.
Repeat for $T$ times.;
Calibration Estimate $\widehat{\mathcal{E}}_C$ = 2(Revenue generated by auctions assigned to $T_{ML}$) - 2(Revenue generated by
  auctions assigned to $T_C$)
Second Phase: Estimation: all intervened advertisers are in $T$;
$T_{ML}$: 50% of auctions are assigned to budgets $B_i^{ML}$, $T_T$: the rest 50% are assigned to budgets $B_i^T$.
  Advertisers are allowed to bid in $T_{ML}$ as long as spend of campaigns in $T_{ML}$, $S_{i,ML} \leq 0.5 \times B_i^{ML}$.
  Similarly, advertisers are allowed to bid in $T_C$ as long as $S_{i,T} \leq 0.5 \times B_i^T$.
Repeat for $T$ times;
ML SCOTE Estimate $\widehat{\mathcal{E}}_{ML}$ = 2(Revenue generated by auctions assigned to $T_T$) - 2(Revenue generated by
  auctions assigned to $T_{ML}$);
```

$$\widehat{\mathcal{E}} = \widehat{\mathcal{E}}_{ML} + \widehat{\mathcal{E}}_C.$$

Essentially, while all advertisers are in C, the experiment is run to obtain the calibration estimate for χ_3 defined in (1). After the calibration estimate is obtained, then all advertisers are promoted to T, after which the Measurement is conducted to obtain an estimate for χ_1. χ_2 is 0 by the experiment structure. The whole procedure is repeated T times. In the following section we prove some desirable properties of this estimate.

Notice here, at time t, the bid $b_{it} = b_i$ for advertiser i at time t is replaced by y_{it}, which is the true bid entered. There are two factors which can impact the bid at time t, the first being whether the advertiser partakes in the auction or not (bid is 0 if it they do not participate), and b_i if they do. For the r^{th} auction, the maximum bid is both a function of the bid distribution as well as the (random) count of advertisers (n_r) who still have enough funds remaining to partake in the auction. Then, the following statement holds:

Lemma 6.1. *When the underlying marketplace remains static:*

(1) the ML based SCOTE estimate is an unbiased estimator of the GTE,
(2) If the underlying budgets distribution is light-tailed and stable, the SCOTE estimate is strongly consistent in T.

Notice that light-tailed and stable are sufficient conditions for strong consistency, not necessary. In practice we notice that with heavy tailed budgets distributions such as log-normal, convergence is quite rapid. While SCOTE estimate being an unbiased estimator of the GTE is a strong benefit, there needs to be some guarantees on the convergence. This is followed up in Theorem 6.2, under the conditions of the second point of Lemma 6.1.

Theorem 6.2. *Assume the bids are generated from i.i.d. $Exp(\lambda)$, with the error distribution $N(\mu_{ML}, \sigma^2_{ML})$. N auctions are conducted in a day, repeated over T days. The marketplace participation rate is assumed to be generated from $U(0, 2\rho_I)$. Then,*

$$\mathbb{P}(|GTE - \widehat{\mathcal{E}}| > T\delta) \leq \frac{1 - (T\lambda^{-Tn\rho_I}\exp\{\sigma^2_{ML}\lambda^2 - \mu_{ML}\lambda\})}{T\delta\exp T\delta}$$

.

Note that the above bound is quite loose. We can use concentration inequalities to obtain tighter bounds, but that would require more assumptions in the distribution (specifically the existence of moments > 1 of budgets distribution, which may not be realistic).

While simply by virtue of Strong Law of Large Numbers (SLLN), the estimate is $o(T)$ (where T is the number of days the experiment is run), faster convergence is ensured by having a well-calibrated ML model. This is all the more crucial for estimation, as while the marketplace may remain static over a short duration of time, in the long run the marketplace evolves, and hence any long-running estimate would be affected by the non-stationarity of the marketplace. Indeed, the above bound relies on significant assumption that the marketplace remains invariant over time. This assumption is often violated. The next section is devoted to simulations that compare and contrast the performance of SCOTE against that of GTE, along with other standard marketplace estimates, in case such violations occur.

6.3 Simulations

We consider $n = 1500$ advertisers with bids simulated from an Exponential distribution $Exp(1)$. The control budgets are generated from an lognormal distribution with parameter $(1, 1)$. The budgets of the intervened campaigns are raised by roughly 10%, to be generated from lognormal distribution with parameter $(2.1, 1)$. Initially, we assume that the marketplace remains invariant over time, and the participation rate is $\rho_I = 0.4$. The calibration and the measurement phase is run for T days each, where T is varied. We consider 10000 requests, and the bid distribution is $Exp(1)$. Table 9 demonstrates, while all SCOTE estimates converge over time, the lower model error leads to a faster convergence. In particular, Table 9 is a best case scenario where the marketplace remains invariant over time. However, in reality, the marketplace evolves over time, e.g., via an increase in participation rates. To model this, consider the following scenario: the entire process takes place over two weeks. In week 1, the participation rate is 0.4, and in the second week it is 0.5. In week 1, the pre-measurement is taken, and the Calibration phase of the SCOTE is conducted. In the second

Days	GTE	SCOTE (ϵ = 0.5)	SCOTE (ϵ = 0.25)	SCOTE (ϵ = 0.125)	Adv RCT	Pre-post
1	1051.003	4040.212	1391.767	1227.912	28087.882	1077.921
2	1051.003	1415.432	1341.863	1201.129	29989.123	1087.983
3	1051.003	1359.098	1295.423	1198.082	21167.099	1024.448
4	1051.003	1105.775	1098.972	1091.093	21098.453	1030.321
5	1051.003	1077.822	1067.099	1058.991	22347.076	1041.092
6	1051.003	1062.096	1060.091	1049.234	25327.653	1047.912
7	1051.003	1059.219	1053.547	1050.982	24035.242	1058.992

Table 9. SCOTE Estimate as a function of the number of days with different model errors, when the marketplace is static

Days	GTE	SCOTE (ϵ = 0.5)	SCOTE (ϵ = 0.25)	SCOTE (ϵ = 0.125)	Adv RCT	Pre-post
1	1051.003	6215.433	4436.31	3192.122	35365.245	4331.871
2	1051.003	3952.04	2309.879	2010.228	30135.764	4309.548
3	1051.003	3112.338	1521.201	1239.291	32008.548	4301.829
4	1051.003	2908.921	1160.401	1104.131	32789.341	4302.021
5	1051.003	2894.321	1154.098	1090.125	30865.977	4301.998
6	1051.003	2771.092	1159.213	1082.871	31675.887	4302.128
7	1051.003	2770.213	1151.765	1081.982	29299.566	4301.997

Table 10. SCOTE Estimate as a function of the number of days with different model errors, when the marketplace is non-stationary

week, the post-measurement is taken, the advertiser randomized experiment is conducted and the second SCOTE measurement is taken.

In Table 10, we notice that the Pre-post estimate converges quickly, but it significantly overestimates the real impact (as does the advertiser side RCT). On the other hand, while the SCOTE estimates perform poorly as well in the beginning, it converges much faster and the extent of overestimation is much lower.

Participation Rate	GTE	SCOTE (ϵ = 0.5)	SCOTE (ϵ = 0.25)	SCOTE (ϵ = 0.125)	Adv RCT	Pre-post
0.8 $\longrightarrow$ 1.0	1731.122	2212.232	2191.474	2092.144	45912.231	8321.232
0.5 $\longrightarrow$ 0.6	1097.199	2594.123	2108.176	1198.918	29299.566	3922.8
0.1 $\longrightarrow$ 0.3	209.019	401.03	311.982	290.124	9082.546	500.434

Table 11. SCOTE Estimate on 7 days as a function of Marketplace change for different model errors

Table 11 demonstrates the relative performance of the SCOTE estimator with variations in the participation rate after $T = 7$ days of run. While all estimates inaccurately estimate the GTE, SCOTE has a lower extent of overestimation.

6.3.1 Comparison with Cluster Randomized Experiments. To compare the performance of SCOTE with cluster RCT in case of changing budgets, we add to the model the same graph structure

Avg Estimate	GTE	Adv RCT	Cluster RCT	SCOTE $\epsilon = \mathbf{0.5}$	SCOTE $\epsilon = \mathbf{0.25}$	SCOTE $\epsilon = \mathbf{0.125}$
(20% Spillover: LN)	2079.91	12091.91	3011.128	2224.98	2199.78	2131.92
(20% Spillover: EXP)	1993.82	10081.18	2518.19	2191.93	2050.67	2001.29
(10% Spillover: LN)	2018.192	10241.082	2391.988	2291.28	2109.764	2078.291
(10% Spillover: EXP)	1901.121	9726.102	2010.92	2010.92	1988.12	1951.198
(5% Spillover: LN)	2015.29	11029.192	2348.01	2189.63	2120.45	2081.24
(5% Spillover: EXP)	1923.63	9681.28	1999.92	2001.18	1993.83	1951.22

Table 12. Comparison of SCOTE and cluster randomized experiments - Average estimates

described in Section 4.3.1. We assume that the advertisers are partitioned into 10 cliques, each containing approximately 150 advertisers. We assume that $x\%$ is the percentage of spillovers. We repeat 150 times the cluster randomized experiment, advertiser RCT, and SCOTE. Tables 12 report the average estimates for different percentages of spillovers x, when the budget distributions are generated from Log-Normal and Exponential Distributions. Although cluster RCT performs better than advertiser RCT, SCOTE outperforms cluster RCT in terms of bias and variance for low model errors.

7 CONCLUSIONS AND FUTURE DIRECTIONS

Accurately estimating the effect of advertiser facing treatments in the presence of spillover effects is an important problem in advertising markets. We propose a new methodology to do this, with the key feature being the introduction of a simulated-counterfactual market that allows us to transform the problem into an auction side RCT, thus bypassing the spillover effects challenge. In addition, due to higher number of auctions, we get low variance, thus achieving the best of both worlds with low bias and low variance. We further give evidence that the approach is robust to various sources of errors, using simplifying distributional assumptions as well as a variety of simulations. The most important contribution is the fundamental design of SCOTE, while we acknowledge that there is plenty of room for additional analysis, theoretical as well as empirical.

There are two main sources of errors in the SCOTE approach, due to calibration and due to marketplace changes. Our theoretical and simulation results give preliminary evidence that the approach is robust to both error sources. However, there is a lot to explore in both directions.

- For calibration in the transformation based approach, we considered simple multipliers in this paper. Potentially, we could use arbitrary transformers that are learned using the advertiser side RCT. We could match the entire input distribution in C, or only certain statistics like we do here. Our calibration strategy equalizes ad spend for advertisers under C and SC, but there could be other calibration strategies that are more robust to errors. We can draw an analogy with the de-biased ML approach of **?**, where errors in ML models that are nuisance estimators are orthogonal to errors in the treatment effect estimators. Is there an equivalent notion here, where the first stage estimators (either ML models or transformations) are in a sense orthogonal?
- We do not have any theoretical guarantees about robustness to marketplace changes. Formalizing this and proving such robustness is an important next step. We currently do not explicitly account for such changes in the estimator. Potentially, one could employ a variety of causal ML techniques such as diff-in-diff in combination with SCOTE to make the estimates even more robust to marketplace changes.

An important moral of our paper is that while the problem of spillover effects may be hard to solve in general, the special case of online advertising has additional properties that allows us to design innovative solutions. This special setting of online advertising has immense practical applications, and is worthy of more attention from the economics and computation community. We expect that our paper inspires additional future work in this direction.

A BIDS

A.1 Proofs and Technical Assumptions

We give explicit formulas for the terms introduced in Section 4, to be used in the proofs of the main results:

$$r(m, \rho_T) = \mathbb{E}\max_i y_i = \int_0^\infty xd\left(\frac{a+b}{2}(\rho_I\rho_T F_m(x) + ((1-\rho_T)\rho_I + 1 - \rho_I)F_C(x)) + 1 - \frac{a+b}{2}\right)^n, \tag{11}$$

$$\begin{aligned} s_T(m, \rho_T) &= N\mathbb{E}(y_i 1(y_i \geq \max_j y_j)|T_i = 1, I_i = 1) \\ &= \int_0^\infty N\frac{a+b}{2}x\left(\frac{a+b}{2}(\rho_I\rho_T F_m(x) + ((1-\rho_T)\rho_I + 1 - \rho_I)F_C(x)) + 1 - \frac{a+b}{2}\right)^{n-1} dF_m(x), \end{aligned} \tag{12}$$

$$\begin{aligned} s_C(m, \rho_T) &= N\mathbb{E}(y_i 1(y_i \geq \max_j y_j)|T_i = 0, I_i = 1) \\ &= \int_0^\infty N\frac{a+b}{2}x\left(\frac{a+b}{2}(\rho_I\rho_T F_m(x) + ((1-\rho_T)\rho_I + 1 - \rho_I)F_C(x)) + 1 - \frac{a+b}{2}\right)^{n-1} dF_C(x), \end{aligned} \tag{13}$$

where F_m is the CDF of mb_i^T.

To prove the following Lemma and Theorem, we assume the following assumptions:

Assumption A.1. The distributions F_T and F_C have continuously differentiable density functions f_T and f_C, respectively.

Assumption A.2. The bids b_i^T and b_i^C belong to the interval $[0, B_m]$.

Assumption A.2 can be relaxed by requiring the distributions F_T and F_c to be Sub-Gaussian.

A.1.1 Proof of Lemma 4.2. We start by proving the first point. It is easy to see that under Assumption 4.1, m^* is a solution of $C(m) = 0$. Next, we show that C is increasing. Let $m_1 < m_2$. For advertiser i and $J \subset \{1, \ldots, n\}$, we define the event

$$A_i(J) = \{\text{advertisers in } J \text{ are in Treatment}\} \cup \{T_i = 1, I_i = 1\}.$$

The index m_a in the expectation indicates that all the requests have multiplier m_a. We have

$$\begin{aligned} \mathbb{E}_{m_2}(y_i 1(y_i \geq \max_j y_j)|A_i(J)) &= \mathbb{E}\left(p_i m_2 b_i^T 1\left(p_i m_2 b_i^T \geq \max(\{p_j m_2 b_j^T\}_{j\in J}, \{p_k b_k^C\}_{k\notin J})\right)|A_i(J)\right) \\ &= \frac{m_2}{m_1}\mathbb{E}\left(p_i m_1 b_i^T 1\left(p_i m_2 b_i^T \geq \max(\{p_j m_2 b_j^T\}_{j\in J}, \{p_k b_k^C\}_{k\notin J})\right)|A_i(J)\right) \\ &\geq \frac{m_2}{m_1}\mathbb{E}\left(p_i m_1 b_i^T 1\left(p_i m_1 b_i^T \geq \max(\{p_j m_1 b_j^T\}_{j\in J}, \{p_k b_k^C\}_{k\notin J})\right)|A_i(J)\right) \\ &= \frac{m_2}{m_1}\mathbb{E}_{m_1}(y_i 1(y_i \geq \max_j y_j)|A_i(J)). \end{aligned}$$

Noticing that $\cup_{J\subset\{1,\ldots,n\}}A_i(J) = \{T_i = 1, I_i = 1\}$, we obtain

$$s_T(m_2, 0.5) = N\mathbb{E}_{m_2}(y_i 1(y_i \geq \max_j y_j)|T_i = 1, I_i = 1) \geq$$
$$N\frac{m_2}{m_1}\mathbb{E}_{m_1}(y_i 1(y_i \geq \max_j y_j)T_i = 1, I_i = 1) = \frac{m_2}{m_1}s_T(m_1, 0.5).$$

Similarly, we can show that $s_C(m_2, 0.5) \leq s_C(m_1, 0.5)$. This show that C is increasing. It remains to show that m^* is a unique zero of C. Suppose that $m_1 < m_2$ are zeros of C. Thus,

$$0 = s_T(m_2, 0.5) - s_C(m_2, 0.5) \geq \frac{m_2}{m_1} s_T(m_1, 0.5) - s_C(m_1, 0.5) = \left(\frac{m_2}{m_1} - 1\right) s_C(m_1, 0.5).$$

Thus, $s_C(m_1, 0.5) = s_T(m_1, 0.5) = 0$. Using $s_T(m_1, 0.5) = 0$, we obtain,

$$\begin{aligned} 0 &= \mathbb{E}_{m_1}\left(\mathbb{E}_{m_1}\left(m_1 b_i^T 1\left(m_1 b_i^T \geq \max_{j \neq i}(y_i)\right) | T_i = 1, I_i = 1, p_i = 1\right)\right) \\ &= \mathbb{E}_{m_1}\left(m_1 b_i^T 1\left(m_1 b_i^T \geq \max_{j \neq i}(y_i)\right)\right) 0.5 \rho_I \frac{a+b}{2}. \end{aligned}$$

Thus, $\mathbb{E}_{m_1}\left(m_1 b_i^T 1\left(m_1 b_i^T \geq \max_{j \neq i}(y_i)\right)\right) = 0$ (There is nothing special about the index i, it can be any advertiser i). By conditioning on events like $\{T_j = 1, I_j = 1, p_j = 1\}$ and $\{I_j = 0, p_j = 0\}$, one can show that either (i) the event $\{b_j^T > b_i^T\}$ has probability 1, or (ii) the event $\{m_1 b_i^T < b_j^C\}$, for all i and j. The first one implies that w.p. 1, $b_j^T > b_i^T$ and $b_j^T < b_i^T$, which is impossible. Thus, w.p. 1, $m_1 b_i^T < b_j^C$ for all i and j. Similarly, using $s_C(m_1, 0.5) = 0$, we obtain that w.p. 1, $m_1 b_i^T > b_j^C$ for all i and j. This leads to a contradiction, and proves the first point.

To prove the second point, observe that the impact can be written as in (8). Following (11), the last two terms of the rhs of the second equality in (8) are zero. This proves the second point.

A.1.2 Proof of Theorem 4.3. We prove this result in two steps. First, we derive a bound on the estimation error of m^* during the calibration phase. This error is caused by (i) the fact we are running Binary Search method a finite number of steps only, and (ii) we are using a finite number of samples to evaluate the calibration function. In the second step, we derive (10).

We start by the first step. We assume wlog that $m \in [0, 1]$. We divide the unit interval $[0, 1]$ into 2^k intervals, each of length 2^{-k}. k is chosen such that $2^{-k} < \epsilon_D$. We denote by $U_i = (i/2^k, (i+1)/2^k)$. We assume wlog that m^* is the midpoint of some interval U_{i^*}. Let us describe the Binary Search method when we know C perfectly. One can look at Binary Search as the process of removing iteratively the half interval that does not contain U_{i^*} until we are left with U_{i^*}. For example, if $k = 3$ and $i^* = 5$, we start by removing the left half of $[0, 1]$, i.e. $U_0 \cup U_1 \cup U_2 \cup U_3$, then the right half of $(1/2, 1)$, and finally the left half of $(1/4, 3/4)$. Of course, we don't know m^* and what we are actually doing is, at each step s, we evaluate C at the mid-point c_s of the remaining interval, if $C(c_s) > 0$ we remove the right half, and if $C(c_s) < 0$ we remove the left half. Using this description of Binary Search, one can show that it always takes k steps to find the interval containing m^*.

The problem is that we don't know $C(c_s)$, but rather an estimate $\widehat{C(c_s)}$ from the advertiser randomized experiment. In this case, for Binary Search to succeed in finding the U_{i^*}, at each step, $\widehat{C(c_s)}$ must have the same sign as $C(c_s)$. Thus, if $C(c_s) < 0$, then the probability of success at step is

$$P_s = \mathbb{P}\left(\widehat{C(c_s)} - C(c_s) < -C(c_s)\right) = \mathbb{P}\left(\widehat{C(c_s)} - C(c_s) < |C(c_s)|\right).$$

Using similar arguments than those used in Lemma 4.2, one can show that for $m_2 > m_1$, if $s_T(m_1, 0.5) = s_T(m_2, 0.5)$, then $s_T(m_1, 0.5) = s_T(m_2, 0.5) = 0$. But, from the expression of $s_T(m_1, 0.5)$ (12), one can see that $s_T(m, 0.5) > 0$ for $m > 0$. As a result, C is strictly increasing. Under Assumption A.1, C has a continuously smooth derivative which is lower bounded by NC_1 on $[0, 1]$, where $C_1 > 0$ is a constant that does not depend on N. It depends on the problem's parameters $\{F_T, F_C, n, a, b, \rho_I, \rho_T\}$. Thus, $|C(m)| \geq NC_1|m - m^*|$, which implies,

$$P_s \geq \mathbb{P}\left(\widehat{C(c_s)} - C(c_s) \leq NC_1|c_s - m^*|\right).$$

We get similar result when $C(c_s) > 0$. We have

$$\begin{aligned}\widehat{C(c_s)} &= \sum_{r=1}^{N} (\text{average spend in r of advertisers in } T - \text{average spend in r of advertisers in } C) \\ &:= \sum_{r=1}^{N} \left(\widehat{s_T^r} - \widehat{s_C^r}\right).\end{aligned}$$

where r is he index of the auction. Under Assumption A.2, $\{\widehat{s_T^r} - \widehat{s_C^r}\}_r$ are between 0 inequality, we get

$$P_s \geq 1 - \exp\left(-\frac{2NC_1^2|c_s - m^*|^2}{B_m^2}\right).$$

Recalling that $\hat{m}$ is the estimate of m^* computed using Binary Search. We obtain

$$\mathbb{P}(|\hat{m} - m^*| \leq \epsilon_D) \geq \prod_{s=1}^{\lceil \log_2(1/\epsilon_D) \rceil} \left(1 - \exp\left(-\frac{2NC_1^2|c_s - m^*|^2}{B_m^2}\right)\right).$$

It remains to come up with a worst case scenario for the mid-points c_s. The probabilities of success P_s are worse when c_s are close to m^*. So we wish to find an example where at each step of the Binary Search the distance $|c_s - m^*|$ is minimal. We will reason backwards to come up with the worse case scenario. After the final step k, we are left with U_{i^*}. At the beginning of step k, there 2 options, either we are left with U_{i^*-1} and U_{i^*} or U_{i^*+1} and U_{i^*}. Both give the same distance $|c_k - m^*| = 2^{-(k+1)}$. Thus lets assume we are left with $U_{i^*-1} \cup U_{i^*} = V_k$. At step k-1, we are left with one of the following options: (i) V_k plus two U intervals to the left of V_k, or (ii) V_k plus two U intervals to the right of V_k. The optimal choice is (ii), which gives $|c_{k-1} - m^*| = 2^{-(k+1)}$. One can show iteratively, that switching between left and right at each step is the worse case scenario. In this case, we get $\{|c_s - m^*|, s = 1, \ldots, k\} = \{\frac{1}{2^{k+1}}, \frac{1}{2^{k+1}}, \frac{1+2}{2^{k+1}}, \frac{1+2^2}{2^{k+1}}, \ldots, \frac{1+2^{k-2}}{2^{k+1}}\}$. Thus,

$$\mathbb{P}(|\hat{m} - m^*| \leq \epsilon_D) \geq \left(1 - \exp\left(-\frac{NC_1^2\epsilon_D^2}{16B_m^2}\right)\right)^2 \prod_{s=1}^{\lceil \log_2(1/\epsilon_D) \rceil - 2} \left(1 - \exp\left(-\frac{NC_1^2(1+2^s)^2\epsilon_D^2}{16B_m^2}\right)\right).$$

Now, we can derive (10). Following Assumption A.1, $r(m, 1)$ is continuously differentiable on $[0, 1]$. Thus, there exists a constant C_2 independent of N (it depends only on $\{F_T, F_C, n, a, b, \rho_I\}$) such that $|r(m, 1) - r(m', 1)| \leq C_2|m - m'|$, for all m, m' in the unit interval.

We have that

$$\begin{aligned}&\mathbb{P}(|\widehat{Nr(1,1)} - \widehat{Nr(\hat{m},1)} - \text{GTE}| \leq N\epsilon) = \mathbb{P}(|\widehat{Nr(1,1)} - \widehat{Nr(\hat{m},1)} - (Nr(1,1) - Nr(m^*,1))| \leq \epsilon) \geq \\ &\mathbb{P}\left(|\widehat{Nr(1,1)} - Nr(1,1)| \leq N((\epsilon - C_2\epsilon_D)/2\right) \mathbb{P}\left(|\widehat{Nr(\hat{m},1)} - Nr(\hat{m},1)| \leq N((\epsilon - C_2\epsilon_D)/2\right) \\ &\times \mathbb{P}(|Nr(m^*,1) - Nr(\hat{m},1)| \leq NC_2\epsilon_D)\end{aligned}$$

where the first equality follows from the second point of Lemma 4.2. Using Hoeffding's inequality [?] and Assumption A.2, one can show that the product of the first two terms is greater than

$$\left(1 - 2\exp\left(-\frac{N(\epsilon - C_2\epsilon_D)^2}{2B_m^2}\right)\right)^2.$$

Morover,

$$\mathbb{P}(|Nr(m^*,1) - Nr(\hat{m},1)| \leq NC_2\epsilon_D) \geq \mathbb{P}(|m^* - \hat{m}| \leq \epsilon_D).$$

This proves the result.

A.1.3 Proof of Corollary 4.4. Using (10), one can show that

$$\mathbb{P}(|\mathrm{GTE} - \widehat{\mathcal{E}}| \leq N\epsilon) \geq \left(1 - 2\exp\left(-\frac{N(\epsilon - C_2\epsilon_D)^2}{2B_m^2}\right)\right)^2 \left(1 - \exp\left(-\frac{NC_1^2\epsilon_D^2}{16B_m^2}\right)\right)^{\lceil \log_2(1/\epsilon_D) \rceil}.$$

One can show by iteratively that, for $0 < a_i < 1$, $1 - \prod_{i=1}^{k}(1 - a_i) \leq \sum_{i=1}^{k} a_i$. Thus,

$$\mathbb{P}(|\mathrm{GTE} - \widehat{\mathcal{E}}| \geq N\epsilon) \leq 4\exp\left(-\frac{N(\epsilon - C_2\epsilon_D)^2}{2B_m^2}\right) + \lceil \log_2(1/\epsilon_D) \rceil \exp\left(-\frac{NC_1^2\epsilon_D^2}{16B_m^2}\right).$$

Setting $\epsilon = N^{-\frac{1}{2}+\delta}$ and $\epsilon_D = (1/2C_2)N^{-\frac{1}{2}+\delta}$, we get

$$\mathbb{P}\left(\frac{1}{N}|\mathrm{GTE} - \widehat{\mathcal{E}}| \geq N^{-\frac{1}{2}+\delta}\right) \leq 4\exp\left(-\frac{N^\delta}{8B_m^2}\right) + \left(\log_2(2C_2)\frac{1-2\delta}{2\log 2}\log_2 N + 1\right)\exp\left(-\frac{N^\delta C_1^2}{64C_2^2B_m^2}\right).$$

Finally, using Borel-Cantelli's Lemma [?, Lemma 2.3.1], we get the result.

A.2 Minimum Number of Auctions to Achieve High Estimate Accuracy

To understand the implications of (10), we analyze the following functions: (i) $N_{min}(\epsilon_D)$ = the minimum number of auctions such that the estimator $\widehat{\mathcal{E}}$ is 0.5−accurate, i.e $|\widehat{\mathcal{E}} - \mathrm{GTE}| \leq N\epsilon = 0.5N$, (ii) $N^2_{min}(\epsilon_D)$ = the minimum number of auctions such that $\widehat{\mathcal{E}}$ is 0.5−accurate, given that the estimate of m in the first phase is ϵ_D−accurate, and (iii) $N^1_{min}(\epsilon_D)$ = the minimum number of auctions such that the estimator of m in the first phase is ϵ_D−accurate. N_{min}, N^2_{min}, and N^1_{min} are computed so that the corresponding probabilities are greater than 0.95. Their graphs are given in Figure 4. N_{min} is decreasing when $\epsilon_D < 0.35$, but surprisingly increasing when $\epsilon_D > 0.35$. The decreasing part follows from the following fact. Given a small ϵ_D and ϵ_D−accurate m, very little noise carries over from the estimation of m in the first phase to the estimation of GTE in the second phase. Thus, most of the noise in the estimation of GTE in the second phase comes from the estimation itself, which is small because of the independence of the auctions. As a result, given an ϵ_D−accurate m, the minimum number of auctions to insure 0.5−accuracy in the second phase is low ($N^2_{min} \leq 14k$). This means that when ϵ_D is small, N_{min} is dictated by the first phase, i.e. it is computed to satisfy a small ϵ_D. One can argue that imposing a very small ϵ_D and asking for a large number of auctions ($\sim 100k$ when $\epsilon_D = 0.05$) is unnecessary to reach an ϵ−accuracy in the second phase. All what we need is $\epsilon_D = 0.35$ and 14k auctions? This is true, but we cannot compute the optimal ϵ_D value (0.35 in our example), since it depends on the problem's parameters which we don't know. When ϵ_D is large ($C_2\epsilon_D$ is close ϵ), a low number of auctions ($N^1_{min} \sim 2k$) is sufficient to guaranty an ϵ_D-accuracy in the first phase. However, the large amount of noise in estimating m carries over to the second phase. It increases with ϵ_D, thus requiring an increasing number of auctions to wash out the noise and satisfy an ϵ−accuracy of the GTE estimate. In this case, the minimum number of auctions is dictated by the second phase ($N_{min} \approx N^2_{min}$).

A.3 Comparison of cluster RCT and SCOTE

In this section, we give the distribution of the results for the 50 performed experiments to compare the performance of cluster randomized experiments and SCOTE. In particular, we report the distribution of real GTE and its estimates using advertiser RCT, cluster RCT, and SCOTE (Figures 5, 7, and 9). Moreover, we report the distribution of the width of CIs for the different methods (Figures 6, 8, and 10).

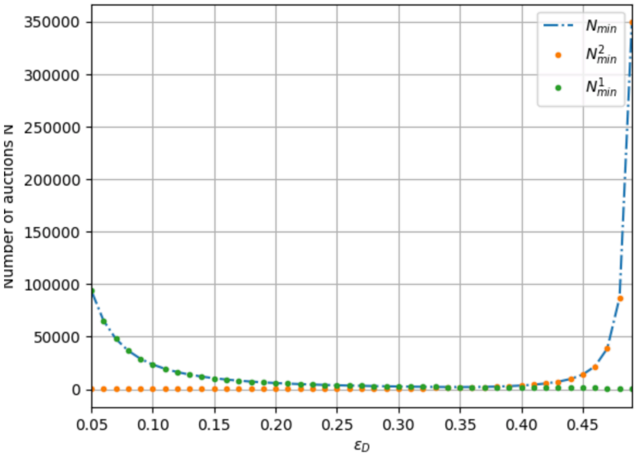

Fig. 4. Minimum number of auctions to achieve high estimate accuracy

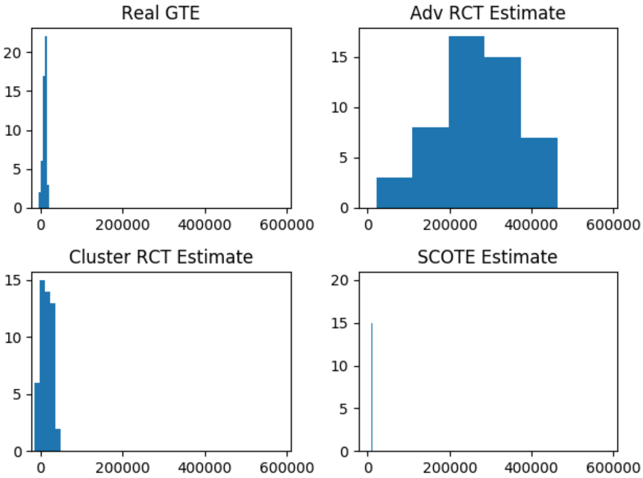

Fig. 5. Distribution of GTE and GTE estimates with spillovers equal to 5%

B PROOFS AND TECHNICAL ASSUMPTIONS FOR BUDGETS

In this section we provide proofs of the results stated in Section 6.

Assumption B.1. The budgets distribution on T and C have finite mean and variance.

B.0.1 Proof of lemma 6.1.

Proof. Notice that for all auction treatments, the spend is monotone non-decreasing over time (spend in a treatment wins if the advertiser wins the auction, by the winning bid. Otherwise it

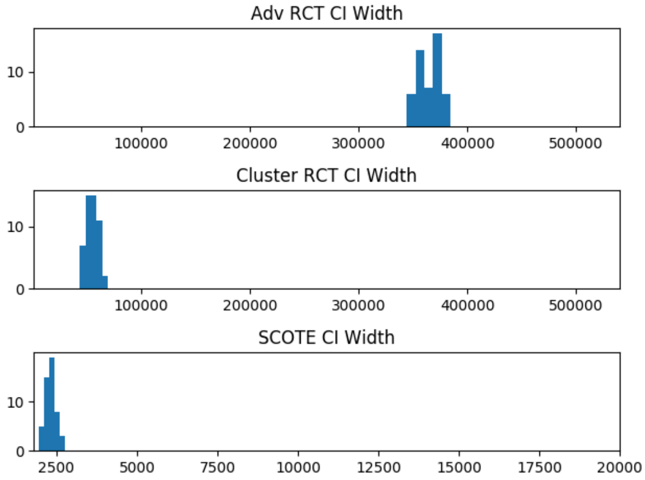

Fig. 6. Distribution of CI width with spillovers equal to 5%

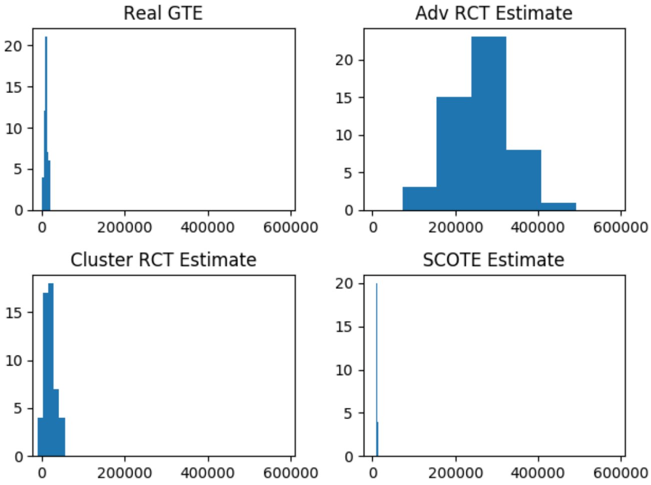

Fig. 7. Distribution of GTE and GTE estimates with spillovers equal to 10%

remains the same). For treatment k, let $S_{i,r,k}$ denote the spend of advertiser i until auction r. Let $B_{i,k}$ be the corresponding budget.

$$0 \leq S_{i,0,k} \leq S_{i,1,k} \leq S_{i,2,k} \leq \cdots \forall k$$

and

$$\mathbb{E}(S_{i,r,k}) \leq \mathbb{E}(B_{i,r,k}) \forall \{i, r, k\} pointwise.$$

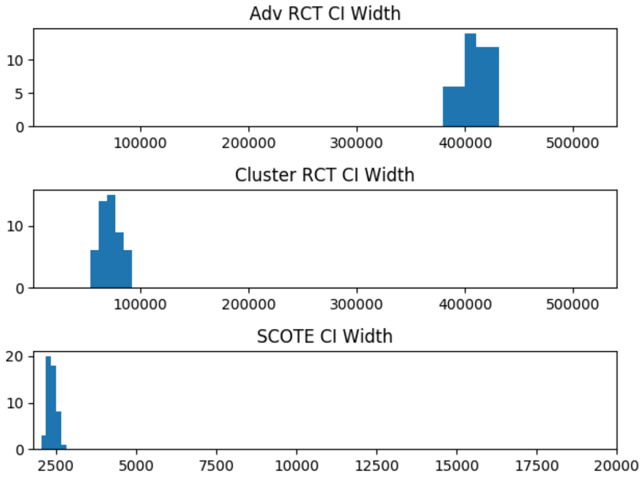

Fig. 8. Distribution of CI width with spillovers equal to 10%

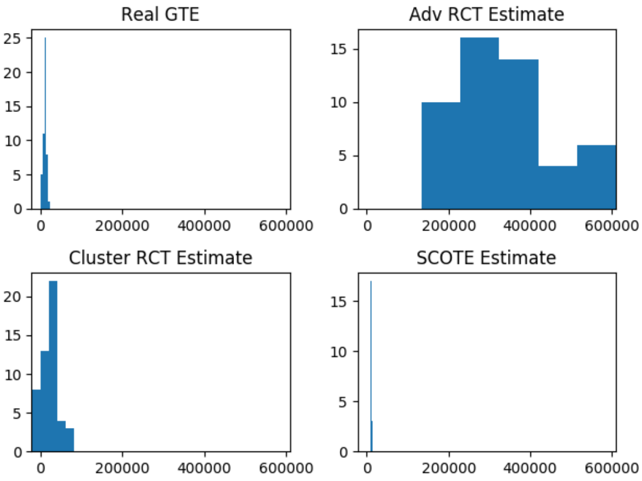

Fig. 9. Distribution of GTE and GTE estimates with spillovers equal to 20%

As per assumption B, $S_{i,r,k}$ is a nonnegative random variable which is point wise bounded. Note that the total spend after n ad auctions is given by

$$S_k = \sum_{i=1}^{N} \sum_{r=1}^{n} S_{i,r,k}$$

, which is still point wise bounded (by $\sum_{i=1}^{N} B_{i,r,k}$). Now, if the experiment is repeated over T days independently, then the mean spend over T days for treatment k is given by $\bar{S_k} = \frac{\sum_{t=1}^{T} S_k^t}{T}$, which is a mean of i.i.d. random variables which are non-negative. Writing out SCOTE estimate, we notice

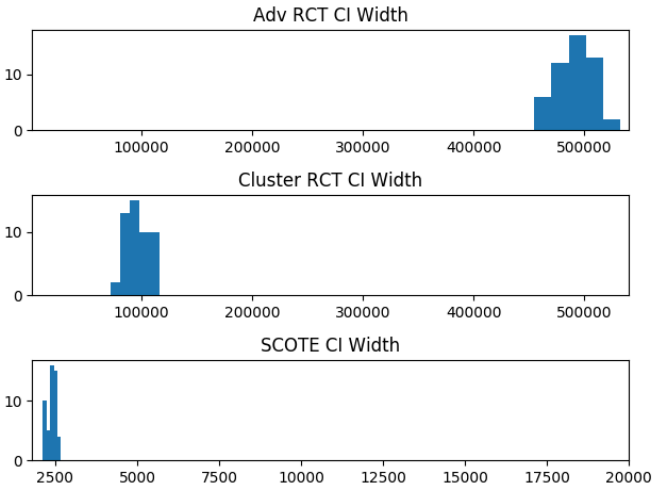

Fig. 10. Distribution of CI width with spillovers equal to 20%

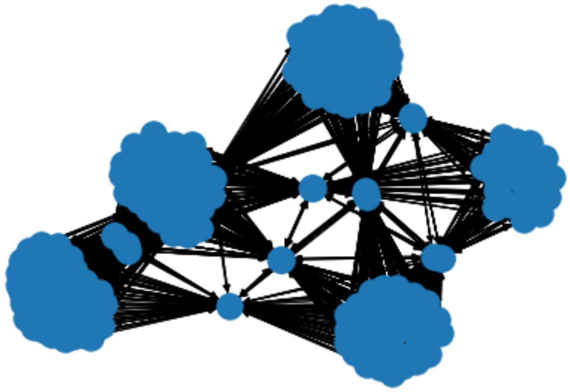

Fig. 11. An example of graph with 5 social cliques, 200 advertisers, and 10% spillovers

that the estimate is given by $\widehat{\mathcal{E}} = 2 \times [(\bar{S_T} - \bar{S}_{ML}) + (\bar{S_{ML}} - \bar{S_C})]$. Taking expectation, we notice that the estimate is unbiased.

For proof of the second part, notice if the budgets distribution is light tailed and stable, so is the spend distribution. Hence, the Strong Law of Large numbers applies, which provides strong consistency. □

B.1 Proof of 6.2

Proof. Notice that the maximum bid of an auction is decreasing over time for any bid distribution, as the number of eligible advertisers decrease. To see this, assume $N(t)$ is the number of advertisers that have remained eligible to bid at period t. Consider first the simple case that participation rate is 1 (i.e., all advertisers are eligible). The winning bid at t, assigned $WB(t)$, has the distribution $G_{WB_t}(x) = [G_B(x))]^{N_t}$. As $0 \leq G_B(x) \leq 1$ and $N_t \geq N_{t+1}$, we have

$$\mathbb{E}(WB(t)) = \int_0^\infty (1 - G_B(x)^{N_t}) \geq \mathbb{E}(WB(t)).$$

Indeed, if the participation rates are drawn i.i.d from the same distribution, the result also follows.

That is to say, $\mathbb{E}(B_{max}(n+1))) \geq \mathbb{E}(B_{max}(n)))$. Next, notice, in the spend difference, we can separate out two sets of advertisers: first, advertisers who are out-of-budget in both, and advertisers who are out-of-budget in only 1.

As the spend difference in the estimates is only the spend difference for advertisers who have gone out of budget in one treatment, but not the other.

$$\mathbb{P}(|\text{GTE} - \widehat{\mathcal{E}}| > t\delta) \leq \frac{\mathbb{E}|\text{GTE} - \widehat{\mathcal{E}}|}{t\delta} \leq \frac{\mathbb{E}|\widehat{\mathcal{E}_{ML}}| + \mathbb{E}|\widehat{\mathcal{E}_C} - \text{GTE}|}{t\delta}.$$

The decomposition is by virtue of Markov inequality and triangle inequality. For the first term, note that

$$\begin{aligned}
\widehat{E_M L} = &\sum_{i \in \{1,2,\cdots,N\}} (S_{i,ML} - S_{i,C}) = \\
&\sum_{i \in \{1,2,\cdots,N\}} (S_{i,ML} - S_{i,C}) 1\{\text{i goes out of budget (OOB) in one treatment and not the other}\} \\
&+ \sum_{i \in \{1,2,\cdots,N\}} (S_{i,ML} - S_{i,C}) 1\{\text{Does not go OOB in either}\}.
\end{aligned}$$

That is to say (with a slight notational abuse), for a particular day $\exists t_C \leq n$ such that $S_{i,C,t_C} = B_{i,C}$, but $S_{i,ML,n} \leq B_{i,ML}$ (or vice versa). Further assume that $B_{i,ML} = B_{i,C} + \epsilon_i$. Consider the following event:

$$\begin{aligned}
&\{\forall t > t_C \text{ s.t. } S_{i,t,ML} < B_{i,ML}, S_{i,t_C,C} = B_{i,C}\} \\
&= \{\forall t > t_C, S_{i,t,ML} - S_{i,t_C,C} < B_{i,ML} - B_{i,C}\} \\
&\geq \{\forall t > t_C, S_{i,t,ML} - S_{i,t_C,ML} + S_{t,t_C,ML} - S_{i,t_C,C} < \epsilon\} \\
&\geq \{\forall t > t_C, (t - t_C) \sum_{k=t}^{t_C} WB_k + S_{t,t_C,ML} - S_{i,t_C,C} < \epsilon\} \\
&\geq \{\forall t > t_C, n \sum_{k=t}^{t_C} WB_k + S_{t,t_C,ML} < \epsilon\}.
\end{aligned}$$

Noting that the maximum bid distribution is stochastically decreasing,

$$\begin{aligned}
\mathbb{P}(|\text{GTE} - \widehat{\mathcal{E}}| > T\delta) &\leq \mathbb{P}(\sum_t WB(n) \leq \sum_t \epsilon + T\delta - \sum_t B_{t,C}) \\
&\leq \frac{\mathbb{E}(\lambda^{-n\rho_T T} \exp -(T\epsilon + T\delta - TB_C))}{T\delta} \leq \frac{(1 - \lambda^{Tn\rho_T} \exp^{-\lambda\mu + \lambda^2\sigma^2/2})}{T \exp^{-\delta} T\delta}
\end{aligned}$$

The second inequality follows from Markov inequality and noting that total is bounded by the maximum bid in that round, which is in turn bounded by the maximum bid in round 1 (when all advertisers are present, discounted by the participation rate). The final inequality follows by iteratively calculating the expectation, noting that the bid, budgets and error distributions are independent, and exploiting the (known) normal and exponential structures of the bid. □

B.1.1 More Simulations for Budgets. In the main section, we have only computed simulation results when the calibration model has 0 bias. In this section we will follow up first with cases where the mean is nonzero. The scenario is as before. The marketplace participation rate is 0.4. The following table demonstrates the performance of SCOTE when there is a large mean difference between Model and Calibration.

Mean	**GTE**	**SCOTE** ($\epsilon = 0.5$)	**SCOTE** ($\epsilon = 0.25$)	**SCOTE** ($\epsilon = 0.125$)	**Adv RCT**	**Pre-post**
$\mu = 1$	1051.003	4040.212	1391.767	1227.912	28087.882	1077.921
$\mu = 5$	1051.003	1415.432	1341.863	1201.129	29989.123	1087.983
$\mu = 10$	1051.003	1359.098	1295.423	594.312	27167.099	1024.448

Table 13. SCOTE Estimate as a function of the number of days with different model bias, when the Marketplace is Static

Note that the consistency results in the paper for budgets is demonstrated for light tailed distributions, but the simulations in the main section was only for log-normal distribution, which are heavy tailed. Table 14 replicates the same results, but when the budgets are generated from exponential distribution with means 10 and 20. The bids distribution are assumed to be the same.

Days	**GTE**	**SCOTE** ($\epsilon = 0.5$)	**SCOTE** ($\epsilon = 0.25$)	**SCOTE** ($\epsilon = 0.125$)	**Adv RCT**	**Pre-post**
1	6746.553	9282.275	8012.823	3192.122	30239.239	8012.232
2	6746.553	7822.181	7633.128	2010.228	34092.783	8220.176
3	6746.553	7019.223	6901.235	1239.291	29928.164	8194.877
4	6746.553	6903.092	6910.238	1104.131	31092.128	8203.484
5	6746.553	6847.647	6772.921	1090.125	33991.129	8109.283
6	6746.553	6790.934	6769.498	1082.871	32532.698	8129.349
7	6746.553	6762.145	6759.391	1081.982	30923.085	4301.997

Table 14. SCOTE Estimate as a function of the number of days with different model errors, when the marketplace is stationary

Table 15 exhibits the simulation results when the marketplace is non-stationary, as in the previous setting.

Days	**GTE**	**SCOTE** ($\epsilon = 0.5$)	**SCOTE** ($\epsilon = 0.25$)	**SCOTE** ($\epsilon = 0.125$)	**Adv RCT**	**Pre-post**
1	6746.553	12034.181	10023.991	9810.283	38254.312	10029.124
2	6746.553	11023.124	9901.234	8018.471	39235.443	9919.127
3	6746.553	8712.182	7819.998	7714.099	40034.126	9878.342
4	6746.553	8810.029	7192.281	7084.283	39982.288	9901.187
5	6746.553	7890.231	7099.712	6981.185	38301.127	9857.096
6	6746.553	7991.231	7081.892	6990.341	39918.778	9885.482
7	6746.553	7823.238	6992.124	6982.192	38034.348	9876.389

Table 15. SCOTE Estimate as a function of the number of days with different model errors, when the Marketplace is non-stationary