Optimizing Cluster-based Randomized Experiments under Monotonicity

Jean Pouget-Abadie
Harvard University
Cambridge, MA, USA
jeanpougetabadie@g.harvard.edu

Vahab Mirrokni
Google Research, New York
New York City, NY, USA
mirrokni@google.com

David C. Parkes
Harvard University
Cambridge, MA, USA
parkes@eecs.harvard.edu

Edoardo M. Airoldi
Harvard University
Cambridge, MA, USA
airoldi@fas.harvard.edu

ABSTRACT

Cluster-based randomized experiments are popular designs for mitigating the bias of standard estimators when interference is present and classical causal inference and experimental design assumptions (such as SUTVA or ITR) do not hold. Without an exact knowledge of the interference structure, it can be challenging to understand which partitioning of the experimental units is optimal to minimize the estimation bias. In the paper, we introduce a monotonicity condition under which a novel two-stage experimental design allows us to determine which of two cluster-based designs yields the least biased estimator. We then consider the setting of online advertising auctions and show that reserve price experiments satisfy the monotonicity condition and the proposed framework and methodology apply. We validate our findings on an advertising auction dataset.

CCS CONCEPTS

• **Mathematics of computing → Probability and statistics**; • **Computing methodologies → Machine learning**;

KEYWORDS

Causal inference; potential outcomes; violations of SUTVA

ACM Reference Format:
Jean Pouget-Abadie, Vahab Mirrokni, David C. Parkes, and Edoardo M. Airoldi. 2018. Optimizing Cluster-based Randomized Experiments under Monotonicity. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220067

1 INTRODUCTION

Randomized experiments — or A/B tests — are at the core of many product decisions at large technology companies. Under the commonly assumed Stable Unit Treatment Value Assumption (SUTVA), these A/B tests serve to estimate unbiasedly the effect of assigning all units to a particular intervention over an alternative condition [12]. The SUTVA assumption is one of no interference between units: a unit's outcome in the experiment does not depend on the treatment assignment of any other unit.

In many A/B tests however, this assumption is not tenable. Consider an intervention on a user of a messaging platform: the (potential) resulting change in her behavior (e.g. increase in time spent on the platform, in number of messages sent, a decrease in response time) would affect the friends on the platform she chooses to communicate with. The same cascading phenomenon can also occur in more subtle ways in a social feed setting. Changes to a feed ranking algorithm, and the resulting behavioral changes (e.g. a higher click-through rate, feedback, or interaction time with the content on the feed) will invariably affect the content on that unit's friends' social feeds [10, 11].

In particular, the same is true in an advertiser auction setting, where modifications to the ecosystem can impact bidders not originally assigned to the intervention [5]. Suppose that one bidder changes her strategy as a result of being assigned to a higher reserve price, or her usual bid no longer meets the reserve — she is more competitive if she increases her bid to meet the new reserve, or less competitive if she fails to meet the reserve. The bidders she competes with now face a different bid distribution. These bidders might react to this new bid distribution by updating their own bidding strategy, even though they were not originally assigned to the intervention.

When SUTVA does not hold, we say there is *interference* between units, and many fundamental results of the causal inference literature no longer hold. For example, the difference-in-means estimator under a completely randomized assignment is no longer unbiased [12]. When the estimand is the difference of outcomes under two extreme assignments — one assigning all units to the intervention, and the other assigning none — a common approach to mitigating the bias of standard estimators in the face of interference is to run cluster-based randomized designs [9, 22, 25]. These randomized designs assign units to treatment or control in groups to limit the amount of interaction between different treatment buckets.

If it can be shown that there is no interaction across treatment buckets, we recover many of the results stated under SUTVA. In practice, however, such a grouping of units may not exist and A/B test practitioners often settle to find the best possible clustering. In particular, a perfect clustering of units cannot be found in an

KDD '18, August 19–23, 2018, London, United Kingdom
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5552-0/18/08.
https://doi.org/10.1145/3219819.3220067

ad auction context due to large advertisers bidding on very large fraction of keywords. Hence, the problem is often formulated as the balanced clustering of a weighted graph on the experimental units, where an edge is drawn between two units that are liable to interfere with one another. This is a challenging task, both algorithmically and empirically: clustering a graph into balanced clusterings is known to be NP-hard, even if we tolerate some unevenness between clusters [2]; furthermore, the correct graph representation of the interference mechanism is not always clear.

While the literature on finding balanced clustering of weighted graphs and analysing cluster-based randomized designs is extensive [8, 9, 13], there are relatively few prior works that tackle the following question: can we determine which of two balanced clusterings produces less biased estimates of the total treatment effect, without assuming that the exact structure of interference is known? The objective of this paper is to show that we can in fact identify the better of two clusterings through experimentation under an assumption on the interference mechanism, which we call *monotonicity.*

Even when the exact structure of interference is not known, monotonicity can be established under a theoretical model. For example, some interference mechanisms are *self-exciting* — if assigning any unit to the intervention will boost the outcomes of any neighboring units. Examples range from vaccination campaigns to social feed ranking algorithms. In both cases, the units in the vicinity of a unit assigned to the intervention tend to benefit over those surrounded by units in the control bucket. Interference mechanisms that exhibit this self-exciting property are a particular example of monotone mechanisms (cf. Section 2.2). When monotonicity holds, we show that it is feasible to compare two balanced clusterings of the experimental units by running a straightforward modification of an experiment-of-experiments design [15, 18].

We make the following contributions: we present an experiment-of-experiments design for comparing cluster-based randomized designs. We define a monotonicity assumption under which we can determine which clustering induces the least biased estimates of the total treatment effect using this comparative design. We prove that pricing experiments in the context of ad exchanges are monotone, and thus our framework applies to this illustrative example. In particular, we state results for the welfare of a single-item second-price auction and the Vickrey-Clarke-Groves auction in the positional ad setting. Finally, we report an empirical simulation study of our algorithms for a publicly-available dataset for online ads. While pricing experiments are done in the context of ad exchanges [1], we note that our paper is a theoretical study of the subject and does not include any real treatments of ad campaigns.

In Section 2, we establish the theoretical framework by defining the monotonocity assumption, describing the suggested experiment-of-experiments design, and proposing a test for interpreting its results. In Section 3, we explain how this framework can be applied to a real-world setting, by showing that reserve-price experiments on advertising auctions are monotone. Finally, we validate these findings on a Yahoo! ad auction dataset in Section 4.

2 THEORY

In this section, we set the notation for the estimand, estimates, and cluster-based randomized designs that we study. We then define the monotonicity assumption, introduce our experiment-of-experiments design, and suggest an approach to analysing its results.

2.1 Cluster-based randomized designs

Let N be the number of experimental units, let vector $\mathbf{Y}$ denote the outcome metric of interest, and let vector $\mathbf{Z}$ denote the assignment of units to treatment ($Z_i = 1$) or control ($Z_i = 0$). Recall that under the potential outcomes framework, $\mathbf{Y}(\mathbf{Z})$ denotes the potential outcomes of the N units under assignment $\mathbf{Z}$. Under the Stable Unit Treatment Value Assumption (SUTVA), this simplifies to $(Y_i(Z_i))_1^N$. The estimand of interest here is the *Total Treatment Effect* (TTE), defined as the difference of outcomes between one assignment assigning all units to treatment, and another assigning none:

$$TTE = \frac{1}{N}\sum_{i=1}^{N} Y_i(\mathbf{Z} = \vec{1}) - Y_i(\mathbf{Z} = \vec{0}) \tag{1}$$

A completely randomized (CR) design assigns N_T units chosen completely at random to treatment and the remaining $N_C = N - N_T$ units to control. A clustering C is a partition of the N experimental units into M groups or "clusters". A *cluster-based randomized* (CBR) design is a randomized assignment of units to treatment and control at the cluster level: if cluster j is assigned to treatment (resp. control), then all units in cluster j are assigned to treatment (resp. control). We will use the notation $\mathbb{E}_{\mathbf{Z}\sim C}[X]$ to denote the expected value of estimator X under a C-cluster-based randomized design. Recall that $\mathbf{Z} \sim C$ represents the assignment of units to treatment and control, resulting from assigning the *clusters* of C uniformly at random to treatment or control.

Let M_T (resp. M_C) be the number of clusters assigned to treatment (resp. control). Let $z \in \{0,1\}^M$ be the assignment vector over *clusters*, where $M = M_T + M_C$. In practice, we will use the Horvitz-Thompson (HT) estimator, defined below:

$$\hat{\tau} = \frac{M}{N}\left(\frac{1}{M_T}\sum_{j=1}^{M} z_j \sum_{i\in C_j} Y_i(\mathbf{Z}) - \frac{1}{M_C}\sum_{j=1}^{M}(1-z_j)\sum_{i\in C_j} Y_i(\mathbf{Z})\right) \tag{2}$$

Under SUTVA, the HT estimator is an unbiased estimator of the total treatment effect under any C-CBR assignment [13]:

$$\mathbb{E}_{\mathbf{Z}\sim C}[\hat{\tau}] = TTE$$

When SUTVA does not hold, this property is no longer guaranteed, and $\hat{\tau}$ may be biased. Our objective is to minimize the bias, defined below, with respect to the clustering, without assuming any explicit knowledge of the interference mechanism or the value of the estimand TTE:

$$\min_C |\mathbb{E}_{\mathbf{Z}\sim C}[\hat{\tau}] - TTE| \tag{3}$$

2.2 A monotonicity assumption

Choosing the clustering of our experimental units in a way that minimizes the bias of our estimators (cf. Eq. 3) when running a cluster-based experiment is a difficult task: without the ground truth, we cannot observe the bias directly. However, under a specific monotonicity property— common to many randomized experiments —the task of choosing the better of two clusterings becomes straightforward.

Definition 2.1. For a domain $\mathcal{P}$ of clusterings of our N units, we say that the interference model is *$\mathcal{P}$-increasing* if and only if

$$\forall C \in \mathcal{P},\ \mathbb{E}_{\mathbf{Z}\sim C}[\hat{\tau}] \leq \tau,$$

and it is *$\mathcal{P}$-decreasing* if and only if

$$\forall C \in \mathcal{P},\ \mathbb{E}_{\mathbf{Z}\sim C}[\hat{\tau}] \geq \tau$$

A model that is either $\mathcal{P}$-increasing or $\mathcal{P}$-decreasing for all clusterings of $\mathcal{P}$ is *$\mathcal{P}$-monotone.*

A $\mathcal{P}$-monotone model is one for which the expectation of the HT estimator $\hat{\tau}$ is either always a lower bound or always an upper-bound of the estimand under any C-CBR design for $C \in \mathcal{P}$. If a model is $\mathcal{P}$-increasing, $\mathcal{P}$-decreasing, or $\mathcal{P}$-monotone for the trivial set of all possible clusterings $\mathcal{P}$, then we simply say that the model is "increasing", "decreasing", or "monotone" without specifying $\mathcal{P}$. Before delving into examples of monotone interference mechanisms, we introduce the following proposition, which highlights why monotonicity is useful for reasoning about bias.

Proposition 2.2. *If the interference model is $\mathcal{P}$-increasing, then for all $C_1, C_2 \in \mathcal{P}$, it holds that*

$$\mathbb{E}_{\mathbf{Z}\sim C_1}[\hat{\tau}] \leq \mathbb{E}_{\mathbf{Z}\sim C_2}[\hat{\tau}] \implies |\mathbb{E}_{\mathbf{Z}\sim C_1}[\hat{\tau}] - \tau| \geq |\mathbb{E}_{\mathbf{Z}\sim C_2}[\hat{\tau}] - \tau|$$

If the interference model is $\mathcal{P}$-decreasing, then for all $C_1, C_2 \in \mathcal{P}$, it holds that

$$\mathbb{E}_{\mathbf{Z}\sim C_1}[\hat{\tau}] \leq \mathbb{E}_{\mathbf{Z}\sim C_2}[\hat{\tau}] \implies |\mathbb{E}_{\mathbf{Z}\sim C_1}[\hat{\tau}] - \tau| \leq |\mathbb{E}_{\mathbf{Z}\sim C_2}[\hat{\tau}] - \tau|$$

Proof. If the model is $\mathcal{P}$-increasing, for $k \in \{1, 2\}$, and $C_k \in P$,

$$\mathbb{E}_{\mathbf{Z}\sim C_k}[\hat{\tau}] - \tau = -|\mathbb{E}_{\mathbf{Z}\sim C_k}[\hat{\tau}] - \tau|$$

Hence, the inequality sign is flipped when the model is $\mathcal{P}$-increasing. A similar reasoning applies for $\mathcal{P}$-decreasing models. □

Proposition 2.2 is a simple consequence of Definition 2.1: if we know that two cluster-based estimates are both lower bounds of the estimand, then the greater of the two must be less biased. The same reasoning applies if they both upper-bound the estimand. It is sufficient to compare the expectation of our estimators to determine which is less biased.

The crux of our framework therefore relies on reasoning about monotonicity. Many commonly studied parametric models of interference are in fact monotone. Consider the following *linear model of interference* (e.g. studied in [9]):

$$Y_i(\mathbf{Z}) = \alpha_i + \beta_i Z_i + \gamma \rho_i + \epsilon_i, \tag{4}$$

where for all i, $(\alpha_i, \beta_i, \gamma) \in \mathbb{R}^3$, $\epsilon_i \sim \mathcal{N}(0, 1)$ is independent of ρ_i, and $\rho_i = \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} Z_j$ is the proportion of i's neighborhood $\mathcal{N}_i$ that is treated. This expresses each unit's outcome as a linear function of a fixed effect, a heterogeneous treatment effect, and a network effect proportional to the fraction of i's neighborhood that is treated. As shown in the following proposition, this is monotone.

Proposition 2.3. *For a given clustering C, let $\theta_C = \frac{1}{N}\sum_i \frac{|\mathcal{N}_i \cap C(i)|}{|\mathcal{N}_i|}$ be the average proportion of a unit i's neighborhood $\mathcal{N}_i$ included in its assigned cluster $C(i)$. Then,*

$$\tau - \mathbb{E}_{\mathbf{Z}\sim C}[\hat{\tau}] = \frac{\gamma M}{M-1}(1 - \theta_C)$$

It follows that if $\gamma \geq 0$, the interference model is increasing, otherwise it is decreasing.

We can also extend the above for heterogeneous network effect parameters γ_i. A proof can be found in Section 6.

Proposition 2.4. *For a clustering C, let $\theta_{C,i} = \frac{|\mathcal{N}_i \cap C(i)|}{|\mathcal{N}_i|}$. For all possible clusterings C,*

$$\tau - \mathbb{E}_{\mathbf{Z}\sim C}[\hat{\tau}] = \frac{M}{N(M-1)}\sum_i \gamma_i(1 - \theta_{C,i})$$

It follows that if $\sum_i \gamma_i(1 - \theta_i) \geq 0$, then the interference model is increasing, otherwise it is decreasing.

It follows that if $\gamma_i \geq 0, \forall i$, then the interference mechanism is increasing, and if $\gamma_i \leq 0, \forall i$, then it is decreasing. If the sign of γ_i is not consistent, then the monotonicity depends on the clustering: if all units with a given sign are perfectly clustered ($\theta_{C,i} = 1$), e.g. all units with $\gamma_i \geq 0$, then the mechanism is once again monotone.

For complex interference mechanisms, it can be easier to establish the following sufficient (but not necessary) condition:

Proposition 2.5. *We say an interference mechanism verifies the* self-excitation property *for a set of clusterings $\mathcal{P}$, if for all units i and clustering $C \in \mathcal{P}$,*

$$\mathbb{E}_{\mathbf{Z}\sim C}[Y_i(\mathbf{Z}) : Z_i = 0] \geq Y_i(\vec{0})$$
$$\mathbb{E}_{\mathbf{Z}\sim C}[Y_i(\mathbf{Z}) : Z_i = 1] \leq Y_i(\vec{1})$$

A $\mathcal{P}$-self-exciting process is $\mathcal{P}$-increasing. A $\mathcal{P}$-self-deexciting mechanism, *with flipped inequalities, is $\mathcal{P}$-decreasing.*

The proof is included in Section 6. The two inequalities capture the following phenomenon: conditioned on i's treatment status, if i's outcome is greatest when i's neighborhood is entirely in treatment, and lowest when i's neighborhood is entirely in control, then an experiment always under-estimates the true treatment effect. This only needs to be true in *expectation* over the assignments $\mathbf{Z}$. For example, we show that the interference mechanism present in certain reserve price experiments in an advertiser auction setting is self-exciting. See Section 3 for more details.

We say the interference mechanism is self-exciting because these inequalities are verified when units benefit from being surrounded by units in treatment. A successful messaging feature launch is a straightforward example of a self-exciting process, as is any pricing mechanism that penalizes any treated bidders and boosts the utility of their competitors.

2.3 An experiment-of-experiments design

Under monotonicity, Proposition 2.2 states that we can determine the least-biased of two $\mathcal{P}$-increasing or $\mathcal{P}$-decreasing cluster-based designs, without knowledge of the estimand, by comparing the expectation of their estimates. However, only one cluster-based design can ever be applied to the set of experimental units in its

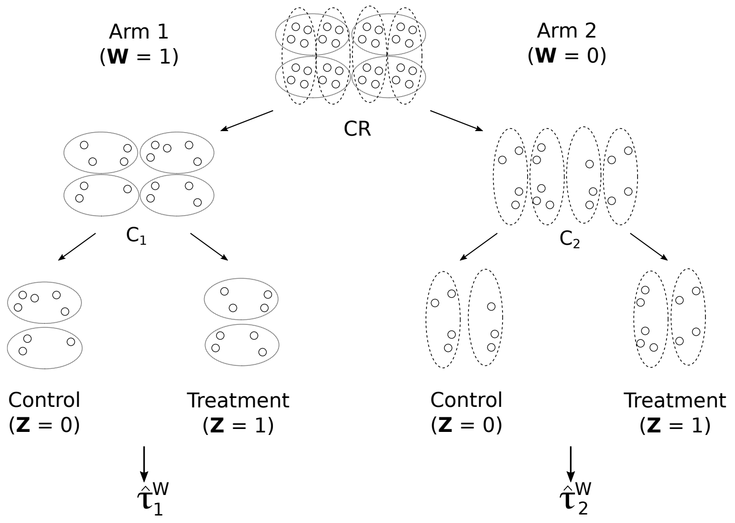

Figure 1: A hierarchical experimental design, which assigns the experimental units to one of two cluster-based randomized designs, C_1 and C_2, completely at random (CR). $\hat{\tau}_1^{\mathbf{W}}$ and $\hat{\tau}_2^{\mathbf{W}}$ represent the treatment effect estimates under each design respectively.

entirety, and the comparison of $\mathbb{E}_{Z\sim C_1}[\hat{\tau}]$ with $\mathbb{E}_{Z\sim C_2}[\hat{\tau}]$ cannot be done directly.

This resembles the fundamental problem of causal inference, which states that units cannot be placed both in treatment and control buckets, and is solved through randomization. Inspired by [15, 18], we suggest to randomly assign different units to either clustering, resulting in a 2-step hierarchical randomized design. The procedure, described in pseudo-code in Algorithm 1, is as follows:

- Assign units completely at random to two design buckets, one for each clustering. Let $\mathbf{W} \in \{1, 2\}^N$ be the vector representing that assignment.
- Within each design bucket, cluster the remaining units together according to the appropriate cluster: if $W_i = W_j = k$ *and* $C_k(i) = C_k(j)$, then i and j belong to the same cluster in design bucket $k \in \{1, 2\}$. The resulting clusterings are $C_1^{\mathbf{W}}$ and $C_2^{\mathbf{W}}$.
- Within each design bucket, assign the resulting clusters to treatment and control. Let $\mathbf{Z}$ be the resulting assignment vector. This is possible because no unit belongs to both $C_1^{\mathbf{W}}$ and $C_2^{\mathbf{W}}$.

Algorithm 1 provides us with two estimates, $\hat{\tau}_1^{\mathbf{W}}$ and $\hat{\tau}_2^{\mathbf{W}}$, of the causal effect, one from each design arm. The resulting clusterings $C_1^{\mathbf{W}}$ and $C_2^{\mathbf{W}}$ may be unbalanced. This is of minor importance as the HT estimator (cf. Eq. 2) is unbiased (under SUTVA) for unbalanced clusterings, and balancedness is required only to control its variance. In practice, C_1 and C_2 are not required to have the same number of clusters, but we expect the clusters sizes to be large enough for each cluster to have at least one unit in each design arm after the first stage with high probability.

From the comparison of $\hat{\tau}_1^{\mathbf{W}}$ and $\hat{\tau}_2^{\mathbf{W}}$, we seek to order $\mathbb{E}_{Z\sim C_1}[\hat{\tau}_1]$ and $\mathbb{E}_{Z\sim C_2}[\hat{\tau}_2]$. Under arbitrary interference structures, these proxy estimates are not guaranteed to have the same ordering, the key condition for Proposition 2.2. Intuitively, $\hat{\tau}_1^{\mathbf{W}}$ and $\hat{\tau}_2^{\mathbf{W}}$ represent the treatment effect estimates for two "weakened" versions of each clustering C_1 and C_2. Because the assignment of units to design arms is done completely at random, it affects each clustering in the same way, and we expect the ordering to stay the same. For the linear model of interference in Prop. 2.4, we have:

```
Algorithm 1: Experiment of experiments design
  Input: Clusterings $C_1$, $C_2$ of the $N$ units into $M_1$, $M_2$ clusters.
  Output: $\mathbf{Z} \in \{0, 1\}^N$ encoding the assignment of each unit to
          a treatment or control bucket.
  Choose $\mathbf{W} \in \{1, 2\}^N$ uniformly at random, encoding the
  assignment of units to design arms 1 and 2;
  for $k \in \{1, 2\}$ do
      Let $C_k^{\mathbf{W}}$ be the clustering on $\{i \in [1, N] : W_i = k\}$ such that
      $C_k^{\mathbf{W}}(i) = C_k^{\mathbf{W}}(j)$ iff $C_k(i) = C_k(j)$;
      Assign units in treatment arm $k$ to treatment and control
      with a $C_k^{\mathbf{W}}$-cluster-based design;
  end
  return the resulting assignment vector $\mathbf{Z}$;
```

Property 1. *An interference mechanism is said to be $\mathcal{P}$-transitive if* $\forall C_1, C_2 \in \mathcal{P}$,

$$\mathbb{E}_{\mathbf{W},Z\sim C_1^{\mathbf{W}}}\left[\hat{\tau}_1^{\mathbf{W}}\right] \le \mathbb{E}_{\mathbf{W},Z\sim C_2^{\mathbf{W}}}\left[\hat{\tau}_2^{\mathbf{W}}\right] \Leftrightarrow \mathbb{E}_{Z\sim C_1}[\hat{\tau}] \le \mathbb{E}_{Z\sim C_2}[\hat{\tau}]$$

If an interference mechanism is transitive for all possible clusterings $\mathcal{P}$, we simply say that it is "transitive" without specifying $\mathcal{P}$. As a sanity check, we can also confirm that the property holds for SUTVA. The property can also be shown for the linear interference mechanisms introduced in Prop. 2.4:

Proposition 2.6. *Under SUTVA, for all C_1, C_2 and $k \in \{1, 2\}$, it holds that*

$$\mathbb{E}_{\mathbf{W},\mathbf{Z}\sim C_k^{\mathbf{W}}}\left[\hat{\tau}_k^{\mathbf{W}}\right] = \mathbb{E}_{\mathbf{Z}\sim C_k}[\hat{\tau}] = \tau.$$

Hence, the no-interference case is trivially transitive. Furthermore, the linear model of interference in Prop. 2.4 is transitive if the same number of units is assigned to each design arm in the first stage of the experiment-of-experiment design: $\sum[W_i = 1] = \frac{N}{2}$.

A full proof can be found in Section 6. For more complex mechanisms of interference, as is the case for reserve price experiments, we use simulations to confirm the intuition that transitivity holds. See Section 4 for more details.

As is common with A/B tests, we do not have access to the expectation of our estimators, and rely on approximations to the variance, such as Neyman's variance estimator. In order to meaningfully compare the estimates we obtain, we must apply our method of choice to determine when their ordering is significant. For example, we can make a normal approximation to the distribution of the estimates— using Neyman's estimator to upper-bound the variance —to estimate the probability that one estimate is greater than the other with a certain significance level:

Proposition 2.7. *Let C_1, C_2 be two clusterings in $\mathcal{P}$. For $k \in \{1, 2\}$, recall the definition of the Neymanian variance estimator for cluster-based randomized designs:*

$$\hat{\sigma}_k^{\mathbf{W}} = \frac{M_k}{N_k}\left(\frac{\hat{S}_{k,t}}{M_{k,t}} + \frac{\hat{S}_{k,c}}{M_{k,c}}\right), \tag{5}$$

where M_k (resp. N_k) is the number of clusters (resp. units) in $C_k^{\mathbf{W}}$, $\hat{S}_{k,t} = var\{Y'_{j,k} : z_j = 1\}$ and $\hat{S}_{k,c} = var\{Y'_{j,k} : z_j = 0\}$, and $Y'_{j,k} = \sum_{C_k^{\mathbf{W}}(i)=j} Y_i$. Assume that the interference mechanism is transitive and $\mathcal{P}$-increasing. If α is the level of significance chosen, we state that C_1 is a significantly better clustering than C_2 if and only if

$$\Phi\left(\frac{\hat{\tau}_1^{\mathbf{W}} - \hat{\tau}_2^{\mathbf{W}}}{\sqrt{\hat{\sigma}_1^{\mathbf{W}} + \hat{\sigma}_2^{\mathbf{W}}}}\right) < \alpha,$$

where Φ is the cdf of the normal distribution.

A similar reasoning applies to $\mathcal{P}$-decreasing mechanisms. If the Gaussian approximation is not appropriate, the distribution of the estimators can equally be approximated by a bootstrap analysis, or a more sophisticated model-based imputation method [12]. More details can be found in Section 6.

3 APPLICATION TO RESERVE PRICE EXPERIMENTS

Online advertising exchanges provide an interface for bidders to participate in a set of auctions for advertising online. These ads can appear within the company's own content, in a social feed, below a search query, or on the webpage of an affiliated publisher. These auctions provide the vast majority of revenue to these platforms, and are thus the subject of experimentation and optimization. Platforms run experiments and monitor different metrics including of revenue and estimates of bidders' welfare.

One possible parameter subject to optimization is the method of determining reserve prices. Online marketplaces can choose to implement a reserve price, which sets the minimum bid required for a bid to be valid and compete with others. It may vary from bidder to bidder, and from auction to auction. A higher reserve may improve revenue, but if it is too high, then too many bids are discarded and ad opportunities can go unsold.

Modifications to a reserve price rule are prime examples of experiments where SUTVA does not hold. A change in reserve price to one bidder affects the bidding problem facing another bidder, even when her reserve is unchanged (e.g., reducing competition when the reserve to the first bidder is higher). We establish conditions under which the resulting interference mechanism within reserve price experiments is monotone, both in the case of a single-item second price auction setting and in the Vickrey-Clarke-Groves auction setting for positional ads. See [24] for a reference.

3.1 Single-item second price auctions

We consider a single-item second-price auction with N bidders $B = \{B_i\}_{i \in N}$ without budget constraints: the highest bidder wins the auction and is charged the maximum of her reserve price and the second-highest bid. The second price auction is truthful (bidding true values is a dominant-strategy equilibrium), and we will assume that the bidders are rational.

Consider two reserve price mechanisms $(r_i)_{i \in B}$ (control) and $(r'_i)_{i \in B}$ (treatment). Suppose that the reserve price mechanism corresponding to treatment always sets a higher reserve price than the reserve price mechanism corresponding to control: $\forall i, r'_i > r_i$. By symmetry, the following argumentation would also work if the treatment and control labels were switched.

We suppose the bidders have values (v_i) for winning the auction. We randomly assign bidders to either the treatment or control reserve price mechanism, with $\mathbf{Z}$ the resulting assignment. The chosen metric of interest is a bidder's utility, denoted by $Y_i(\mathbf{Z})$. For a second-price auction, $Y_i = 0$ if bidder i does not win the auction, and $Y_i = v_i - p$ when she wins the auction and pays price p. The bidder welfare of an auction is the sum of each bidder's utility, $\sum_i Y_i(\mathbf{Z})$, and the estimand is given by: $S = \sum_i Y_i(\vec{1}) - \sum_i Y_i(\vec{0})$

Tthe reserve price experiment for second price auctions verifies the self-excitation property (cf. Prop. 2.5). The idea is that assigning a unit to the intervention can only make them less competitive by discarding their bid from the auction. Thus, the higher the number of treated units, the lower the competition for the remaining bidders, and the higher their utility.

Theorem 3.1. *Consider a set of rational agents with no budget-constraints. Let the outcome of interest be each agent's welfare. The interference mechanism of a reserve price experiment, assigning treated units to a higher personalized reserve price, for a single-item second-price auction is self-exciting, and thus monotone.*

Proof. Consider bidder i's outcome under $\mathbf{Z} = \vec{0}$ and under any assignment $\mathbf{Z}'$ such that $Z_i = 0$. There are three possible cases:

- Bidder i wins the auction in neither assignment. Her utility is therefore constant.
- Bidder i wins the auction in only one assignment. It must be that bidder i wins under $\mathbf{Z}'$ but not $\mathbf{Z}$. Her utility is 0 under $\mathbf{Z}$ and greater than 0 under $\mathbf{Z}'$.

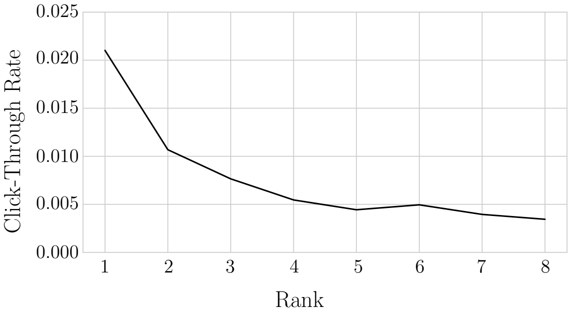

Figure 2: The average click-through rate (CTR) observed in the *Yahoo! Search Auction* dataset, described in Section 4, can be observed to be an approximately decreasing and convex function of the slot rank. The confidence intervals were too small to be meaningfully reported in the figure.

- Bidder i wins the auction under both assignments. If the second highest bid is the same under both assignments, bidder i's utility is constant. Otherwise, the second highest bid under $\mathbf{Z}'$ can only be lower than the second highest bid under $\mathbf{Z}$. Thus bidder i's payment is lower and her utility is higher under assignment $\mathbf{Z}'$ than under assignment $\mathbf{Z}$.

By symmetry, we reach a similar conclusion when comparing assignments $\mathbf{Z} = \vec{1}$ and any assignment $\mathbf{Z}'$ such that $Z'_i = 1$. □

It follows that the reserve price experiment is increasing, and any cluster-based randomized design underestimates the bidder welfare estimand.

3.2 Positional ad auctions

In practice, ad auctions are multi-item, used for selling more than one ad position on a user's view. We now extend the previous results to a multi-item setting, with m items (or "slots"). We assume the common positional ad setting, where each slot has an inherent click-through rate pos_j, which we can suppose is ordered: $pos_1 > pos_2 > \cdots > pos_m$ [23]. Each bidder i is only ever allocated at most one item, with value v_i for getting a click. We assume for simplicity that all bidders have the same ad quality, and thus the same click-through rate for a given ad slot. As a result, bidder i's utility for winning slot j is $v_i \cdot pos_j - p_i$, where p_i is the required payment of bidder i.

The Vickrey-Clarke-Groves (VCG) auction takes place in two parts. First, a value-maximising allocation is chosen (based on bids). Here, the highest bids win the highest slots. Bidders are then charged the externality they impose on all other bidders. In other words, assuming that bidder k obtains the k^{th} slot, bidder k pays:

$$p_k = \sum_{j=k+1}^{m} (pos_{j-1} - pos_j) \cdot v_j \cdot \mathbf{1}_{[v_j \geq r_j]}$$

where r_j is the reserve imposed on bidder j with value v_j. We can prove that the self-excitation property holds under a convexity assumption.

Theorem 3.2. *Consider a set of rational agents with no budget-constraints. Let the outcome of interest be each agent's welfare. The interference mechanism of a reserve price experiment, assigning treated units to a higher personalized reserve price, for a VCG auction in the positional ad setting with no quality effects is self-exciting, and thus monotone if the click-through rate function $pos : i \mapsto pos_i$ is convex:*

$$\forall i > j,\ pos_{i+1} - pos_i \leq pos_{j+1} - pos_j,$$

This convexity assumption is verified empirically in the literature and in the Yahoo! auction dataset[1] introduced in Section 4 (cf. Figure 2). The intuition behind the proof is similar to the single-item setting: for a bidder i, the greater the number of i's competitors are treated, the fewer are able to compete, and thus the higher i's utility. We prove this through a case-by-case analysis. Let $r_i^{\mathbf{Z}}$ be the reserve that bidder k faces under assignment vector Z: $r_i^{\mathbf{Z}} = r_i$ if $Z_i = 0$ and r'_i otherwise.

Proof. Consider the outcomes of bidder i and j under $\mathbf{Z}$ and $\mathbf{Z}'$ such that for all $k \neq j$, $Z_k = Z'_k$, $Z_i = Z'_i = 0$, and $Z_j = 0 < Z'_j = 1$. By transitivity, if we can show $Y_i(Z) \leq Y_i(Z')$, then it follows that $Y_i(\vec{0}) \leq \mathbb{E}_C[Y_i(\mathbf{Z}) : Z_i = 0]$. There are three possible cases:

- The allocation of bidders to slots does not change and thus prices do not change. Bidder i's utility is constant.
- Bidder i is allocated to slot i for both $\mathbf{Z}$ and $\mathbf{Z}'$ assignments, but bidder j's ($j < i$) bid is discarded when j is treated (Z'): $r'_j > v_j > r_j$. The difference of bidder i's outcome under the two treatment assignments is: $Y_i(\mathbf{Z}) - Y_i(\mathbf{Z}') = -\sum_{k \geq j}(pos_{k-1} - pos_k)(v_k \mathbf{1}_{v_k > r_k^{\mathbf{Z}}} - v_{k+1}\mathbf{1}_{v_{k+1} > r_{k+1}^{\mathbf{Z}}})$. This quantity is always negative, hence $Y_i(\mathbf{Z}) \leq Y_i(\mathbf{Z}')$.
- Bidder j's ($j < i$) bid is discarded when j is treated and thus bidder i is allocated to slot $i - 1$. In that case, bidder i's utility under $\mathbf{Z}$ is: $Y_i(\mathbf{Z}) = pos_i v_i - \sum_{k \geq i+1}(pos_{k-1} - pos_k) v_k \mathbf{1}_{v_k > r_k^Z}$. The same bidder $i's$ utility under $\mathbf{Z}'$ is: $Y_i(\mathbf{Z}') = pos_{i-1} v_i - \sum_{k \geq i+1}(pos_{k-2} - pos_k) v_k \mathbf{1}_{v_k > r_k^Z}$.
  It follows that the difference of bidder i's outcomes is equal to: $Y_i(\mathbf{Z}) - Y_i(\mathbf{Z}') = (pos_i - pos_{i-1})v_i - \sum_{k \geq i+1}(pos_{k-2} + pos_k - 2pos_{k-1})v_k$, where the $\mathbf{1}_{v_k > r_k^Z}$ terms are implicit. Note that each individual term of the sum is positive by convexity, such that $Y_i(\mathbf{Z}) \leq Y_i(\mathbf{Z}')$.

□

4 EXPERIMENTAL VALIDATION

In this section, we validate our design strategy for comparing two given graph clusterings for the purpose of experimentation under interference to an advertising auction dataset. For this purpose, we make use of a Yahoo! auction dataset.

4.1 The Yahoo! Search Auction dataset

The *Yahoo! Search Marketing Advertiser Bid-Impression-Click data on competing Keywords* dataset is a publicly-available dataset released by Yahoo![2], containing bid, impression, click, and revenue data between advertiser-keyphrase pairs over a period of 4 months.

[1] Our own dataset could potentially suffer from endogeneity, where weaker bidders are consistently assigned to lower slots. The assumption is, however, supported elsewhere in the literature [7, 16].

[2] Available for download at https://webscope.sandbox.yahoo.com/

Per keyphrase			Per bidder		
nbr of bids	min	1	nbr of bids	min	1
	median	2		median	9
	max	7041		max	$2.1 \cdot 10^4$
bid value	min	.3¢	bid value	min	.5¢
	median	66¢		median	60¢
	max	$320		max	$4700
impressions	min	1	impressions	min	1
	median	3		median	31
	max	$5 \cdot 10^6$		max	$1.4 \cdot 10^6$
clicks	min	0	clicks	min	0
	$cdf(1)$	91.4		$cdf(1)$	93.3
	max	7041		max	$1.1 \cdot 10^4$

Table 1: Summary statistics for the Yahoo! dataset, aggregated by keyphrase or by bidder, per day for the entire 4 month period. Bid values are given in USD unless specified otherwise. $cdf(1)$ is the value of the cumulative distribution function of impressions for a single impression.

The advertiser and keyphrase are anonymized, represented as a randomly-chosen string. A sample line of the dataset is reproduced[3] below:

day	id	rank	keyphrase	bid	impress.	clicks
1	`a3d2`	2	`f3e4,j6r3,...`	100.0	1.0	0.0

The dataset contains 77, 850, 272 bidding activities of 16, 268 different bidders. There are a total of 75, 359 keywords represented, for a total of unique 648, 515 keyphrases (or list of keywords). Table 1 contains a series of summary statistics computed over keyphrase-day pairs and bidder-day pairs, namely the total number of bids, the total bid value, the total number of impressions, and the total number of clicks per keyword (or per bidder) and per day.

We can represent the *Yahoo!* dataset by a set of bipartite graphs between bidders, identified by their `account_id`, and the keyphrases. The *bid* bipartite graph on day t draws a weighted edge of weight w_{ij} between every bidder-keyphrase pair such that bidder i bids w_{ij} on keyphrase j on day t. We can aggregate these graphs over the entire time period (4 months) by summing their edge weights together. We can also consider the impression, rank, and clicks graphs, where the weight of the edge is given by the number of impressions, the rank, or the number of clicks respectively received by bidder i on keyphrase j.

The dataset only provides data aggregated at the granularity of a single day, reporting the average bid and total number of impressions and clicks for each bidder, keyphrase day triplet. Hence, we define a keyphrase-day pair as a single auction, where each bidder's bid is set to the reported average bid for that keyphrase-day pair. For the sake of simplicity, we will only consider a setting with the top four ad positions, which account for the majority of clicks.

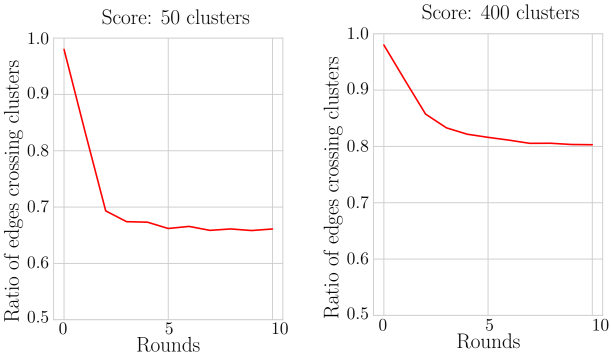

Figure 3: Weighted ratio of edges across clusters for successive runs of the R-LDG algorithm on the weighted bid graph into 50 clusters and 400 clusters respectively.

4.2 Simulating a reserve price experiment

While the *Yahoo! Search Auction* dataset provides us with a set of bidders, keyphrases, and the bids, impressions, and clicks that link them, it does not provide us with an actual intervention on the auction ecosystem. We must therefore simulate the impact of a change in the reserve price given to each bidder.

While many possible units of randomization exist for an auction experiment (keyphrases, bidders, browsers, users, various pairings of these units, etc.), the reserve price experiment we consider randomizes on bidders. On large auction platforms, the reserve price might be set through the application of machine learning methods. In our context, we choose a random non-zero reserve price for each bidder, calibrating the spread of the distribution such that some bidders will not be able to match the reserve price for all auctions. All bidders assigned to the intervention will face their non-zero reserve price, fixed for every auction for simplicity. All bidders assigned to the control bucket will not face a reserve price.

Within the same auction for a given keyphrase, two participating bidders may face distinct reserves and be assigned to different treatment buckets. A bidder-cluster-based randomized experiment is thus used to mitigate the possible interference between bidders, our units of randomization, within a single auction.

To validate our experiment-of-experiments design, we must find candidate balanced graph clusterings to compare, a problem known to be NP-hard — even when we slightly relax the balancedness assumption [2]. In the last several years, there has been good progress in developing scalable distributed balanced clustering algorithms for graphs with billions of edges [3, 20]. These algorithms have enabled practitioners to apply these large-scale graph mining algorithms for large-scale randomized experimental studies [17, 18, 21]. Of the numerous heuristic algorithms for finding such clusterings, the *Restreaming Linear Deterministic Greedy* (R-LDG) algorithm [14] is a popular choice. It consists of repeatedly applying a greedy algorithm, originally proposed in [19], which assigns each node u to

[3] The account ID and keyword ID's have been shortened for the sake of exposition in this sample line. The bid value is given in 1/100¢.

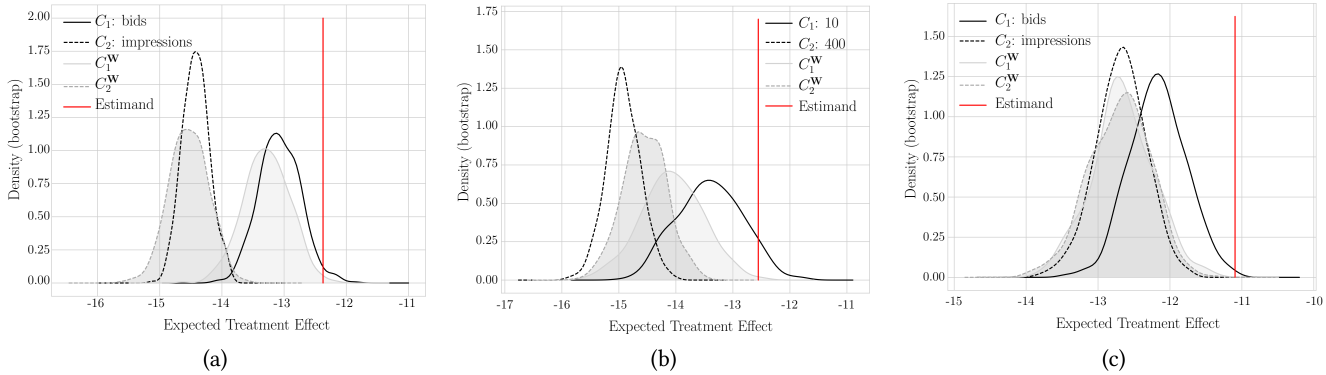

Figure 4: We compare the distribution of the expectation of our Horvitz-Thompson estimator for the total treatment effect (in red) under several cluster-based randomized assignments. In each plot, the solid and dotted lines represent the expectation of the estimator under C_1 and C_2 respectively — the two estimate distributions we wish to compare but cannot simultaneously observe. The shaded distributions correspond to the *observed* distributions of the expectation of the estimator under the induced clusterings $C_1^{\mathbf{W}}$ and $C_2^{\mathbf{W}}$, resulting from our Experiment-of-Experiments design. The red segment represents the total treatment effect estimand. Each plot establishes a comparison of two different clusterings: (a) C_1 is a R-LDG clustering, C_2 is a random clustering ($M_1 = M_2 = 50$); (b) C_1 is a R-LDG clustering with 10 clusters, C_2 is a R-LDG clustering with 400 clusters; (c) C_1 a R-LDG clustering of the bid graph, whereas C_2 is a R-LDG clustering of the impressions graph. ($M_1 = M_2 = 50$). Monotonicity is verified because every distribution is on the same side of the estimand; transitivity is verified because the ordering of the solid and dotted distributions is preserved when going from the unshaded plots to the shaded plots. The loss of power is quantified by the increase in overlap between the solid and dotted distributions, when comparing the unshaded plots with the shaded plots.

.

one of k clusterings according to the following objective:

$$\underset{i\in\{1,\dots k\}}{\arg\max}\ |P_i^t \cap N(u)| \left(1 - \frac{|P_i^t|}{H_i}\right), \tag{6}$$

where P_i^t is the set of nodes assigned to cluster i at step t of the algorithm, H_i is the maximum capacity of cluster $i \in \{1, \dots k\}$, and $N(u)$ is the set of neigbhors of node u in the graph.

We can apply this clustering algorithm to any of the bipartite graphs introduced in Section 4.2, aggregated over the entire time period, resulting in a set of mixed bidder-keyphrase clusters. The bidder-only clusters are obtained from the previous clustering by simpling removing the keyphrase nodes from consideration. The algorithm's objective must be slightly modifed to accomodate weighted graphs, by replacing $|P_i^t \cap N(u)|$ with $\sum_{i,j} w_{ij} \mathbf{1}_{i\in N(u)} \mathbf{1}_{j\in P_i^t}$. Furthermore, we must also modify the balance requirement, since only the bidder side of the bipartite graph clustering is required to be balanced! We therefore replace $\left(1 - |P_i^t|/H_i\right)$ with $\left(1 - |P_{i,c}^t|/H_{i,c}\right)$ where $P_{i,c}^t$ is the set of bidder nodes in cluster P_i^t and $H_{i,c}$ is the maximum number of allowed bidder nodes in cluster P_i^t. The final objective is given by:

$$\underset{i\in\{1,\dots k\}}{\arg\max} \left| \sum_{i\in N(u), j\in P_i^t} w_{ij} \right| \left(1 - \frac{|P_{i,c}^t|}{H_{i,c}}\right) \tag{7}$$

Figure 3 plots the proportion of edges cut, weighted by the bid amount, over consecutive runs of the R-LDG algorithm for 50 and 100 clusters. We adopt three main vectors of comparison between candidate clusterings to determine the efficacy of our proposed experiment-of-experiment design:

- *Quality:* comparing clusterings of the graph that differ in their estimated quality, for example by looking at the number of edges cut, for a fixed number of clusters: we compare a random graph clustering to a clustering obtained by running the R-LDG algorithm to convergence.
- *Number of clusters:* comparing two clusterings of the graph obtained by running the same clustering algorithm for a different number of clusters: we consider a R-LDG clustering with 10 clusters and a R-LDG clustering with 400 clusters.
- *Metric:* comparing clusterings of the graph that are obtained by applying the same algorithm on different bipartite graphs: we compare a R-LDG clustering of the *bid* graph with an R-LDG clustering of the *impressions* graph.

The dataset does not provide the budgets of the bidders or their perceived ad quality, hence we will adopt the same simplifying assumptions as Section 3 of no quality effects between bidders and no budget constraints. Furthermore, we assume bids are unchanged as a result of the experiment (which would be valid for rational, non budget-limited bidders).

4.3 Validating the empirical optimization

We first compare a clustering of the graph obtained by running the modified R-LDG algorithm (cf. Section 4.2) against a completely random balanced clustering of the graph. We fix a subset of auctions with few bidders per auction, in order to showcase the framework and establish the monotonicity and transitivity properties by allowing a setting for which there is a clear difference between the two clusterings. The reduction in cut size — measured by the ratio of the weighted sum of edges inter-clusters over the sum of all edge weights — over the iterations of the algorithm is shown in Figure 3. While the weighted cut of the graph for a random clustering is around 98%, the clustering obtained with the R-LDG algorithm approaches 66% within a few iterations.

We validate the monotonicity assumption, as well as the transitivity assumption, for reserve price experiments. In Figure 4 (a), we plot four distributions as well as the Total Treatment Effect estimand (cf. Eq. 1), obtained by taking the difference between assigning all units to a higher reserve price and assigning none. Namely, we plot the distribution of the HT estimator's expectation (cf. Eq 2) under each cluster-based design: $\mathbb{E}_{\mathbf{Z}\sim\mathcal{C}_k}[\hat{\tau}]$ where $k = 1$ for the R-LDG clustering and $k = 2$ for the random clustering. We also plot the distribution of the expectation of the experiment-of-experiments (EoE) estimators: $\mathbb{E}_{\mathbf{W},\mathbf{Z}\sim\mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}}]$.

We find that they all under-estimate the true treatment effect, as expected from the increasing property. As expected, the HT estimator is more biased under a random clustering than under the R-LDG clustering. Furthermore, we find that the property of transitivity holds (cf. Eq. 1), namely the EoE estimate of the "random estimator" under-estimates the total treatment effect more severely than the EoE estimate of the "R-LDG estimator".

We repeat the experiment to compare a R-LDG clustering with 10 clusters with another R-LDG clustering with 400 clusters (cf. Figure 4 (b)). We find that the clustering with 10 clusters is less biased but exhibits higher variance, and that the transitivity property holds. Finally, in Figure 4 (c), we compare a clustering of the impressions bipartite graph with a clustering of the bid bipartite graph. The transitivity property is again verified. Moreover, we see that clustering the bid bipartite graph may be a better heuristic in this setting, but the difference in the two clusterings is very slight. The code is available for download at https://jean.pouget-abadie.com/kdd2018code.html.

5 FUTURE WORK

We have introduced two properties, monotonicity and transitivity, under which the estimation of causal effects in the presence of interference can be improved by selecting the least-biased of two clusterings. We proved that certain parametric models of interference are monotone and transitive. A more exhaustive examination of other parametric models of interference (e.g. [4, 6]) for these properties was beyond the scope of this work. Furthermore, while we were able to prove monotonicity for certain reserve price experiments, transitivity was established only in simulations. A natural question arising from this work is whether monotonicity and transitivity can be established through empirical means, using an observational method or through a randomized experiment.

Furthermore, while our Experiment-of-Experiment design can improve the bias of subsequent randomized experiments — by selecting which of two clusterings should be used for the cluster-based randomized assignment, the reduction in bias comes at a cost of reduced power in the current experiment: half the units belonging to the more biased clustering are discarded in the final analysis. Hence, an important direction of future work is quantifying and bounding this loss of power, as well as exploring alternate means of choosing a clustering with a smaller power reduction, either through observational data or a less intrusive experimental design.

6 PROOFS

6.1 Proof of Proposition 2.3 and 2.4

Assume that $\forall \mathbf{Z}$, $Y_i(\mathbf{Z}) = \alpha_i + \beta_i \cdot Z_i + \gamma_i \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Recall the definition of the estimand: $\tau = \frac{1}{N} \sum_i Y_i(\vec{1}) - Y_i(\vec{0})$. Plugging in the expression for $Y_i(\vec{Z})$, we obtain: $\tau = \frac{1}{N} \sum_i \beta_i + \frac{1}{N} \sum_i \gamma_i$. The estimator is given by:

$$\hat{\tau} = \frac{M}{N} \sum_i \frac{(-1)^{1-Z_i}}{M_t^{Z_i} M_c^{(1-Z_i)}} Y_i(\mathbf{Z}),$$

where M_t (resp. M_c) is the number of clusters in treatment (resp. control). Plugging in the expression for $Y_i(\vec{Z})$, we obtain:

$$\mathbb{E}_{Z\sim\mathcal{C}}[\hat{\tau}] = \frac{1}{N} \sum_i \beta_i + \frac{1}{N} \sum_i \gamma_i \left(\frac{|\mathcal{N}_i \cap C(i)|}{|\mathcal{N}_i|} - \frac{1}{M-1} \frac{|\mathcal{N}_i \backslash C(i)|}{|\mathcal{N}_i|} \right)$$

We obtain the desired result by taking the difference between these quantities. Prop. 2.3 follows by substituting $\gamma_i = \gamma$.

6.2 Proof of Proposition 2.5

The proposition can be established by rewritting the definition of $\mathcal{P}$-increasing interference mechanisms,

$$\tau - \mathbb{E}_{\mathbf{Z}\sim\mathcal{C}}[\hat{\tau}] = \frac{1}{N} \sum_i \left(Y_i(\vec{1}) - \mathbb{E}_{\mathbf{Z}\sim\mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 1] \right)$$
$$+ \left(\mathbb{E}_{\mathbf{Z}\sim\mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 0] - Y_i(\vec{0}) \right),$$

such that a sufficient condition of the model to be $\mathcal{P}$-increasing is for $Y_i(\vec{1}) > \mathbb{E}_{\mathbf{Z}\sim\mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 1]$ and $Y_i(\vec{0}) < \mathbb{E}_{\mathbf{Z}\sim\mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 0]$. If increasing the number of treated units in that unit's neighborhood increases that unit's outcome — holding that unit's treatment assignment constant — then the two previous inequalities hold.

6.3 Proof of Proposition 2.6

Recall that for $k \in \{1, 2\}$, our estimator can be written as:

$$\hat{\tau}_k^{\mathbf{W}} = \frac{M_k}{N_k} \sum_i W_i Y_i(\mathbf{Z}) \frac{(-1)^{1-Z_i}}{M_{k,t}^{Z_i} M_{k,c}^{1-Z_i}},$$

where $M_{k,t}$ (resp. $M_{k,c}$) is the number of treated (resp. control) clusters in design arm k and N_k is the number of units in design arm k. We begin by first considering the no-interference case. We have that

$$\mathbb{E}_{Z\sim\mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k | \mathbf{W}] = \frac{1}{N_k} \sum_i W_i (Y_i(1) - Y_i(0)).$$

By the law of iterated expectations, we have $\mathbb{E}_{\mathbf{W},Z\sim\mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}}] = \tau$. We now consider the linear model suggested in Eq. 4, where we assume heterogeneous network effects (γ_i). From the proof of Proposition 2.4, we have that

$$\mathbb{E}_{\mathbf{Z}\sim\mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}} | \mathbf{W}] = \bar{\beta} + \frac{M_k}{M_k - 1} \frac{1}{N_k} \sum_i W_i \gamma_i \left(\theta_{\mathcal{C}_k^{\mathbf{W}}, i} - 1 \right)$$

Note that we have

$$\mathbb{E}_{\mathbf{W}}[W_i \theta_{C_k^{\mathbf{W}}, i}] = \frac{N_k(N_k - 1)}{N(N-1)} \theta_{C_k, i}.$$

It follows that, if $M_1 >> 1$, $M_2 >> 1$, and $N_1 = N_2 = \frac{N}{2}$,

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim C_1^{\mathbf{W}}}[\hat{\tau}_1^{\mathbf{W}}] - \mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim C_2^{\mathbf{W}}}[\hat{\tau}_2^{\mathbf{W}}] &\approx \frac{1}{2N} \sum_i \gamma_i \theta_i \\ &\approx \mathbb{E}_{\mathbf{Z} \sim C_1}[\hat{\tau}] - \mathbb{E}_{\mathbf{Z} \sim C_2}[\hat{\tau}] \end{aligned}$$

We conclude that the linear model of interference is transitive.

6.4 Discussion for Proposition 2.7

Under unspecified models of interference, theoretical bounds on the power of even the simplest randomized experiment are hard to come by. While the joint assumption of monotonicity and transitivity allow us to design a sensible test for detecting the better of two partitions, they are not sufficient to bound its power without stronger assumptions. We thus rely on simulations, like the ones run in Section 4, or theoretical approximations, like the ones suggested in Prop. 2.7. It approximates $\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\hat{\tau}_k^{\mathbf{W}}]$, for $k \in \{1, 2\}$ by two independently-distributed Gaussian variables of mean $\hat{\tau}_k^{\mathbf{W}}$ and variance $\hat{\sigma}_k^{\mathbf{W}}$, given in Eq. 5. Their difference therefore has the distribution $\mathcal{N}(\hat{\tau}_1^{\mathbf{W}} - \hat{\tau}_2^{\mathbf{W}}, \hat{\sigma}_1^{\mathbf{W}} + \hat{\sigma}_2^{\mathbf{W}})$. Recall that Neyman's variance estimator is an upper-bound of the true variance, under SUTVA, in expectation over the assignment $\mathbf{Z}$ (cf. [12]). We prove in the lemma below that this still holds true for a hierarchical assignment.

Lemma 6.1. *Under SUTVA, Neyman's variance estimator is an upper-bound in expectation of the true variance of the HT estimator:*

$$\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\hat{\sigma}_k^{\mathbf{W}}] \geq var_{\mathbf{W},\mathbf{Z}}[\hat{\tau}_k^{\mathbf{W}}]$$

Proof. By Eve's law,

$$var_{\mathbf{W},\mathbf{Z}}[\hat{\tau}_k^{\mathbf{W}}] = \mathbb{E}_{\mathbf{W}}[var_{\mathbf{Z} \sim C_k^{\mathbf{W}}}[\tau_k^{\hat{\mathbf{W}}} | \mathbf{W}]] + var_{\mathbf{W}}[\mathbb{E}_{\mathbf{Z} \sim C_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}}]].$$

From [12], the first term can is equal to:

$$\frac{M_k}{N_k} \left(\frac{var(Y'(1))}{M_{k,t}} + \frac{var(Y'(0))}{M_{k,c}} - \frac{var(Y'(1) - Y'(0))}{M_k} \right),$$

where $Y'_j(Z) = \sum_{i \in C_k^{\mathbf{W}}(j)} Y_i(Z)$, the cluster-level outcomes. The second term can be shown to be equal to $\frac{var(Y(1) - Y(0))}{N}$. Since we have that:

$$\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\hat{\sigma}_k^2] = \frac{M_k}{N_k} \left(\frac{var(Y'(1))}{M_{k,t}} + \frac{var(Y'(0))}{M_{k,c}} \right),$$

we must prove:

$$\frac{var(Y'(1) - Y'(0))}{N_k} \geq \frac{var(Y(1) - Y(0))}{N}.$$

This follows from an application of the Cauchy-Schwarz inequality for balanced clusters: $\sum_j (\sum_i Y_i)^2 \leq \sum_j |C_j| \sum_i Y_i^2$, where C_j are the cluster sizes, equal to $\frac{N}{N_k}$ in the balanced case. □

In order to determine the greater of two clusterings, we can perform two one-sided t-tests. The Bayesian approach is to compute the posterior distribution of the difference of the two estimates, using a conjugate Gaussian prior. In order to assess the impact of assuming the two estimates are independent Gaussians, we suggest running a sensitivity analysis, by considering the result of the test for different values of the correlation coefficient.

ACKNOWLEDGEMENTS

This work was partially supported by National Science Foundation grants IIS-1149662, IIS-1409177, by an Office of Naval Research grant N00014-17-1-2131, by a Google Research Award, and by a Siebel Scholarship.

REFERENCES

[1] Ad exchange auction model. https://support.google.com/adxseller/answer/152039?hl=en&ref_topic=2904831, February 2018.
[2] Konstantin Andreev and Harald Racke. Balanced graph partitioning. *Theory of Computing Systems*, 39(6):929–939, 2006.
[3] Kevin Aydin, Mohammad Hossein Bateni, and Vahab S. Mirrokni. Distributed balanced partitioning via linear embedding. In *WSDM*, 2016.
[4] Guillaume W. Basse and Edoardo M. Airoldi. Model-assisted design of experiments in the presence of network correlated outcomes. *arXiv preprint arXiv:1507.00803*, 2015.
[5] Guillaume W. Basse, Hossein Azari Soufiani, and Diane Lambert. Randomization and the pernicious effects of limited budgets on auction experiments. In *Artificial Intelligence and Statistics*, pages 1412–1420, 2016.
[6] Niloy Biswas and Edoardo M. Airoldi. Estimating peer-influence effects under homophily: Randomized treatments and insights. In *Complex Networks IX*, pages 323–347. Springer International Publishing, 2018.
[7] Nico Brooks. The atlas rank report: How search engine rank impacts traffic. *Insights, Atlas Institute Digital Marketing*, 2004.
[8] Allan Donner and Neil Klar. Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, 94(3):416–422, 2004.
[9] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.
[10] Dean Eckles, René F. Kizilcec, and Eytan Bakshy. Estimating peer effects in networks with peer encouragement designs. *PNAS*, 113(27):7316–7322, 2016.
[11] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. Network a/b testing: From sampling to estimation. In *WWW*, 2015.
[12] Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.
[13] Joel A Middleton and Peter M Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *SSRN:1803849*, 2011.
[14] Joel Nishimura and Johan Ugander. Restreaming graph partitioning: simple versatile algorithms for advanced balancing. In *KDD*, 2013.
[15] Jean Pouget-Abadie, Martin Saveski, Guillaume Saint-Jacques, Weitao Duan, Ya Xu, Souvik Ghosh, and Edoardo M. Airoldi. Testing for arbitrary interference on experimentation platforms. *arXiv:1704.01190*, 2017.
[16] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW*, 2007.
[17] David Rolnick, Kevin Aydin, Shahab Kamali, Vahab S. Mirrokni, and Amir Najmi. Geocuts: Geographic clustering using travel statistics. *arxiv:1611.03780*, 2017.
[18] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M. Airoldi. Detecting network effects: Randomizing over randomized experiments. In *KDD*, 2017.
[19] Isabelle Stanton and Gabriel Kliot. Streaming graph partitioning for large distributed graphs. In *KDD*, 2012.
[20] Charalampos E. Tsourakakis, Christos Gkantsidis, Bozidar Radunovic, and Milan Vojnovic. FENNEL: streaming graph partitioning for massive scale graphs. In *WSDM*, 2014.
[21] Johan Ugander and Lars Backstrom. Balanced label propagation for partitioning massive graphs. In *WSDM*, 2013.
[22] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *KDD*, 2013.
[23] Hal R. Varian. Position auctions. *International Journal of industrial Organization*, 25(6):1163–1178, 2007.
[24] Hal R. Varian and Christopher Harris. The vcg auction in theory and practice. *American Economic Review*, 104(5):442–45, 2014.
[25] David Walker and Lev Muchnik. Design of randomized experiments in networks. *Proceedings of the IEEE*, 102(12):1940–1951, 2014.