Improving Precision in Clustered Experiments

Guido Imbens* **Lorenzo Masoero*** **James McQueen*** **Thomas Richardson***

Ido Rosen* **Suhas Vijaykumar***

Abstract

Network interference, where observed outcomes are influenced by interaction with nearby units, is a fundamental issue in A/B testing and experimentation in social and economic networks. Clustered randomization is a frequently-used strategy that aims to prevent confounding by limiting interaction between treated and untreated units. We study a model of least-squares estimation under network interference, and give a tight characterization of the mean-squared error as a function of the clustering design, extending prior work that studies the bias alone. Based on this result, we propose a semidefinite relaxation for the error-minimizing clustered design and compare it to standard clustering approaches that target only the bias.

1 Overview

Competition and social interaction are essential features of two-sided markets ranging from online job boards and social networking platforms to traditional retail marketplaces. However, when conducting randomized experiments, such interactions often lead to *spillovers*, where the outcome for an experimental unit may be affected the treatment status of other, interacting units. The classical theory of randomized controlled experiments does not account for these spillovers, and conventional estimates become biased.

This has led to much research focusing both on how to quantify spillovers as well as how to mitigate bias. Several works aim to explicitly account spillovers and control for them, either by modifying the experimental design [3], by using modified estimators that account for exposure to treated units [2, 1], or both [4]. A widely-used strategy is the use of clustered designs, where units are grouped together and assigned a constant level of treatment, so that the frequency of interaction between two nodes in different groups is small [6, 5]. These methods are effective, yet the choice of clustering involves a crucial trade-off between bias and variance that is not fully understood.

In this paper, we consider the performance of a standard, least-squares estimator of the average treatment effect in the presence of spillovers. Building on prior work that studies the bias alone, we tightly characterize the bias and variance of any non-stratified clustered randomized design. We then propose a heuristic algorithm to minimize the mean-squared error in practice.

2 Problem formulation

Similar to Brennan et al. [5], who studies bias-minimizing designs in a related (one-sided bipartite) setting, we suppose the researcher has oracle knowledge of the noise level, spillover size, and treatment size. This allows us to characterize the fundamental limits of the estimation problem. It is

*All authors are affiliated with Amazon.com. Guido Imbens is also affiliated with Stanford University and Thomas Richardson is also affiliated with University of Washington. Correspondence to vijasuha@amazon.com.

Preprint. Under review.

often the case that the researcher has some prior information on the spillover network, and in practice we propose using this information to choose the experimental design.

Model and estimator

We are given n units corresponding to nodes in a graph $G = (V, E)$. We write N_i for the set of neighbors of a given node i in G.

Potential outcomes. For any value of the treatment variables $\mathbf{d} = (d_1, d_2, \ldots, d_n) \in \{0, 1\}^n$ we observe potential outcomes, which may be expressed using the exposure mapping of Aronow and Samii [1]. In our setting, the exposure mapping $\psi_i(\mathbf{d})$ reduces to the number of adjacent treated units if untreated, and 0 if treated.

$$Y_i(\mathbf{d}) = \tau d_i - \rho \psi_i(\mathbf{d}) + \varepsilon_i; \qquad \psi_i(\mathbf{d}) = \sum_{j \in N_i} d_j(1 - d_i),$$

Here, the ε_i are independent, centered random variables with variance σ^2.

In this context, we are interested in learning the *main effect* τ, which corresponds to the unit-level average effect of treating the whole population: $\tau = \frac{1}{n} \sum_i \mathbb{E}[Y_i(\mathbf{1}) - Y_i(\mathbf{0})]$, where $\mathbf{1}$ (resp. $\mathbf{0}$) denotes the all-ones (resp. all-zeros) vector.

We consider a highly restricted form of the spillovers: they have a constant level ρ, are proportional to the number of treated neighbors (linear in sums), and are only realized when a given unit is untreated, while its neighbors are treated. This may be a reasonable approximation in settings such as online markets, where a promotional discount's effect depends upon the relative price of competing products.

Estimator. We consider the following, standard estimator for the main effect, τ:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} 2Y_i(2D_i - 1).$$

This is simply the least-squares regression of Y_i on $(D_i - 1/2)$. Note that $\hat{\tau}$ incorporates no knowledge of spillovers. For designs where $\mathbb{E}[D_i] = 1/2$, it is unbiased when $\rho = 0$ (no spillovers).

Experimental design

We consider the following class of non-stratified, clustered experimental designs. This is similar to Brennan et al. [5], except we allow clusters of non-equal size. Notably, in this setting the number of clusters no longer has a direct correspondence to the estimator's variance, and this fact is captured by our subsequent analysis.

1. Choose $C : V \to [s]$ to be a partition of V, the set of units. For a given unit i, $C(i)$ indexes the cluster containing i.
2. For each cluster $1 \leq k \leq s$ we independently sample $\sigma_k \sim \mathrm{Ber}(1/2)$.
3. We assign $D_i = \sigma_{C(i)}$ for all units $i \in V$.

3 Results

First, we characterize of the bias of $\hat{\tau}$. This coincides with prior results on clustered randomization under interference and will facilitate comparison. To state results, let the clusters be given by $C_1, C_2, \ldots, C_s \subset V$ and write ∂C_k to denote the set of edges with exactly one endpoint in C_k.

Lemma 1 *The bias of $\hat{\tau}$ for τ is given by*

$$\mathbb{E}[\hat{\tau} - \tau] = \frac{\rho}{n} \sum_{k=1}^{s} |\partial C_k|$$

This result agrees with prior results of Brennan et al. [5] and others, and corresponds directly to the min-cut objective proposed in that paper. Next, we characterize the mean-squared error.

Proposition 1 *The mean-squared error of $\hat{\tau}$ admits the upper bound*

$$\mathbb{E}(\hat{\tau}-\tau)^2 \leq \frac{4\sigma^2}{n} + \frac{1}{n^2}\left\{\tau^2 \sum_{k\leq s}|C_k|^2 + \rho^2\left(\sum_{k\leq s}|\partial C_k|\right)^2\right\}.$$

Moreover, this characterization is tight up to a constant factor of at most 5*:*

$$\mathbb{E}(\hat{\tau}-\tau)^2 \geq \frac{4\sigma^2}{n} + \frac{1}{5n^2}\left\{\tau^2 \sum_{k\leq s}|C_k|^2 + \rho^2\left(\sum_{k\leq s}|\partial C_k|\right)^2\right\}.$$

Remark 1 (Interpreting Proposition 1) This shows that the MSE is bounded by two terms: the squared cut size divided by n^2 (second term), and the squared sum of cluster sizes divided by n^2 (first term), which we can call the reciprocal of the "effective sample size." Note that the effective sample size can still be constant even when there are many clusters, for example if the cluster sizes decay as $|C_k| \asymp n/k^2$. When all s clusters have equal size, it is $1/s$.

3.1 Algorithms

We introduce a heuristic algorithm to minimize the MSE objective derived above. The algorithm can be stated for a slightly more general model: $Y_i = \tau_i d_i - \sum_{j\in N_i}\rho_{ij}d_j(1-d_i) + \varepsilon_i$. Let $\boldsymbol{\tau} \in \mathbb{R}^n$ contain the individual treatment effects τ_i, and $\boldsymbol{\rho} \in \mathbb{R}^{n\times n}$ the pairwise spillovers ρ_{ij}. We write

$$L(\boldsymbol{\rho}) = \rho - \operatorname{diag}(\boldsymbol{\rho}^\top \mathbf{1})$$

to denote the graph Laplacian associated to the spillover network. According to Proposition 1, when $\tau_i = \tau$ and $\rho_{ij} = \rho$, the MSE is tightly approximated by

$$\mathbb{E}(\hat{\tau}-\tau)^2 \asymp \left\langle \Sigma,\, \boldsymbol{\tau}\boldsymbol{\tau}^\top \right\rangle + \left\langle \Sigma,\, -L(\boldsymbol{\rho})\right\rangle^2,$$

where Σ is the second moment matrix of the treatment variable, i.e., $\Sigma_{ij} = \mathbb{E}[D_i D_j]$.

For our class of designs, we have $\Sigma = \frac{1}{2}\mathbf{C}\mathbf{C}^\top$, where $\mathbf{C} \in \mathbb{R}^{n\times s}$ has entries given by $C_{ik} = \mathbb{1}\{i \in C_k\}$. Thus, a convex relaxation for the optimal clustering could take the form

$$\begin{aligned} \min_{\Sigma} \quad & \left\langle \Sigma,\, \boldsymbol{\tau}\boldsymbol{\tau}^\top \right\rangle + \left\langle \Sigma,\, -L(\boldsymbol{\rho})\right\rangle^2 \\ \text{s.t.} \quad & \Sigma \succeq 0, \quad \Sigma = \Sigma^\top, \quad \Sigma_{ii} = 1/2, \quad \Sigma_{ij} \geq 0. \end{aligned} \tag{1}$$

Given a solution Σ^*, one can use various heuristics to recover a partitioning of the graph. One potential method, by analogy to spectral clustering, would be to factor $2\Sigma^* = \tilde{\mathbf{C}}\tilde{\mathbf{C}}^\top$ and apply k-means clustering to the rows of $\tilde{\mathbf{C}}$. Comparison and rigorous analysis of such heuristics merits further study.

References

[1] Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912 – 1947, 2017. doi: 10.1214/16-AOAS1005. URL `https://doi.org/10.1214/16-AOAS1005`.

[2] Eric Auerbach and Max Tabord-Meehan. The local approach to causal inference under network interference. *arXiv preprint arXiv:2105.03810*, 2021.

[3] Sarah Baird, J Aislinn Bohren, Craig McIntosh, and Berk Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860, 2018.

[4] Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas S Richardson, and Ido M Rosen. Experimental design in marketplaces. *Statistical Science*, 1(1):1–19, 2023.

[5] Jennifer Brennan, Vahab Mirrokni, and Jean Pouget-Abadie. Cluster randomized designs for one-sided bipartite experiments. *Advances in Neural Information Processing Systems*, 35:37962–37974, 2022.

[6] Michael P Leung. Rate-optimal cluster-randomized designs for spatial interference. *The Annals of Statistics*, 50(5):3064–3087, 2022.