Optimizing Randomized and Deterministic Saturation Designs under Interference

Chencheng Cai [1] Jean Pouget-Abadie [2] Edoardo Airoldi [1]

[1]Fox School of Business, Temple University [2]Google Research, NYC

GraphEx 2023

Introduction

- Interference in potential outcomes is ubiquitous including examples from social networks, transportation network, crop field study, etc...
- Randomized saturation design is widely adopted to estimate effects involving interference on networks.
- We consider a potential outcome model, which is based on the local proportion of treated neighbors. The model is more realistic than the partial interference model and relaxes certain assumptions on networks.
- We consider the optimal randomized saturation design, compared to the two widely-used special cases: the cluster-based design and the stratified one.
- Furthermore, we show a deterministic saturation design would further improve the performance.

Background

Interference: one unit's treatment status would affect another unit's outcome.
Graph Representation: $\mathcal{G} = (V, E)$ with V the units and $(i, j) \in E$ if units i and j interfere with each other.
Assumption on potential outcomes: For any unit i, let $\mathcal{N}_i$ be the neighbors of i in $\mathcal{G}$. Then we assume

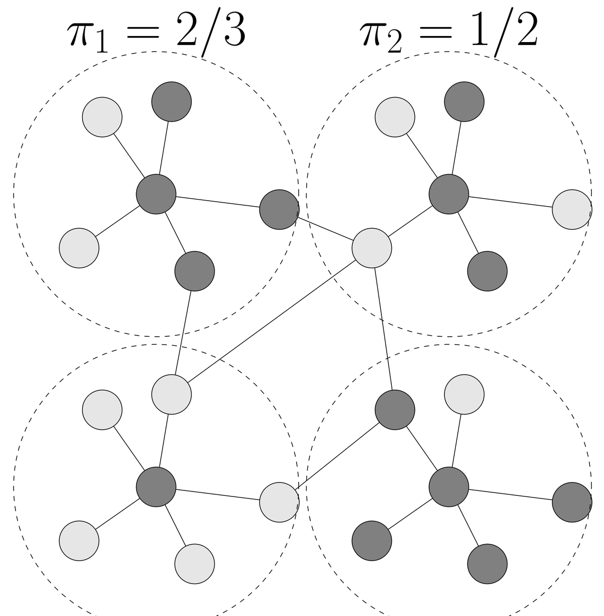
$$Y_i(Z_1, \ldots, Z_N) = Y_i(Z_1', \ldots, Z_N') \text{ whenever } Z_i = Z_i' \text{ and } Z_k = Z_k' \; \forall k \in \mathcal{N}_i.$$

Further assumption on proportion of treated neighbors:

$$Y_i(Z_1, \ldots, Z_N) = Y_i(Z_1, \ldots, Z_N) \text{ whenever } Z_i = Z_i' \text{ and } \rho_i = \rho_i',$$

where $\rho_i = |\mathcal{N}_i|^{-1} \sum_{k \in \mathcal{N}_i} Z_k$ is the proportion of treated neighbors.

Randomized Saturation Design



Population: A collection of J clusters of units.

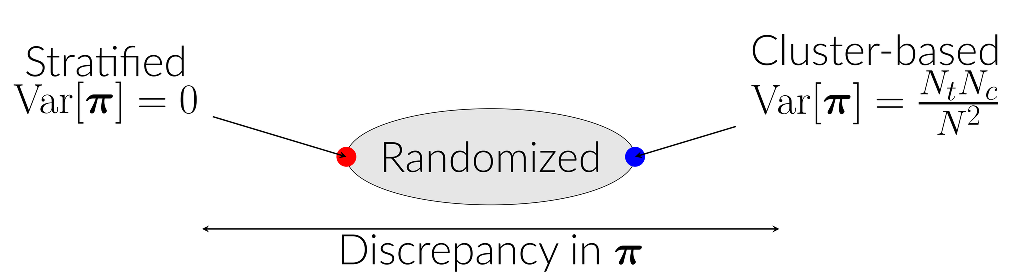
Two-step Randomization:

1. Randomly generate a *proportion vector* $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_J]$ from $\boldsymbol{\Pi}$.
2. Randomly assign $n_j = \lfloor \pi_j N_j \rfloor$ units in cluster j to treatment.

Example: A realization of treatment assignment generated by a randomized saturation design where the realized proportion vector is $\boldsymbol{\pi} = \left[\frac{2}{3}, \frac{1}{2}, \frac{1}{6}, \frac{5}{6}\right]$.
●: treated units ○: control units

Two special cases:

- Cluster-based: $\pi_j \in \{0, 1\} \; \forall j$.
- Stratified: $\pi_j = N_t/N \; \forall j$.



Analysis Setup

Assumptions: arbitrary network with equally sized clusters
Potential Outcomes: Linear interference model (SUTVA when $\gamma_i \equiv 0$).

$$Y_i(\boldsymbol{Z}) = \alpha_i + \beta_i Z_i + \gamma_i \rho_i, \quad \rho_i = \sum_{k \in \mathcal{N}_i} Z_i / |\mathcal{N}_i|,$$

ρ_i: proportion of treatment in the neighbor of unit i.
Estimator: Difference-in-means estimator. $\hat{\tau} = N_t^{-1} \sum_i Y_i^{obs} Z_i - N_c^{-1} \sum_i Y_i^{obs}(1 - Z_i)$
Estimand: Total treatment effect (TTE): $\mathrm{TTE} = \overline{Y_i(\boldsymbol{1})} - \overline{Y_i(\boldsymbol{0})} = \overline{\beta} + \overline{\gamma}$

Theoretical Results

Bias under interference

$$\mathbb{E}_{\boldsymbol{Z}}[\hat{\tau}] = \overline{\beta} + \frac{N^2}{N_t N_c} \left(\gamma' - \frac{\overline{\gamma} - \gamma'}{J - 1} \right) \mathrm{Var}[\boldsymbol{\pi}] \leqslant \mathrm{TTE},$$

where $\gamma' = N^{-1} \sum_j \sum_{i \in \mathcal{C}_j} \gamma_i \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|}$ is the down-weighted averaged γ.

In terms of minimizing the bias $(\mathrm{TTE} - \mathbb{E}_{\boldsymbol{Z}}[\hat{\tau}])$,
(i) if $\gamma' > \overline{\gamma}/J$, the cluster-based design ($\mathrm{Var}[\boldsymbol{\pi}] = N_t N_c / N$) is optimal;
(ii) if $\gamma' < \overline{\gamma}/J$, the stratified design ($\mathrm{Var}[\boldsymbol{\pi}] = 0$) is optimal;
(iii) if $\gamma' = \overline{\gamma}/J$, any randomized based design is optimal.

Variance under interference

$$\mathrm{Var}_{\boldsymbol{Z}}[\hat{\tau}] = V_0 + V_1 \mathrm{Var}[\boldsymbol{\pi}] + V_2 \mathrm{Var}^2[\boldsymbol{\pi}] + V_3 \mu_{3c}[\boldsymbol{\pi}] + V_4 (\mu_{4c}[\boldsymbol{\pi}] - \mathrm{Var}^2[\boldsymbol{\pi}]) + o(N^{-1}),$$

where $\mu_{3c}[\boldsymbol{\pi}]$, $\mu_{4c}[\boldsymbol{\pi}]$ are the third and forth central moments of $\boldsymbol{\pi}$, and $V_0, \ldots, V_4$ are some extremely complicated coefficients related to the potential outcomes (α, β, γ) and the network. The variance can be optimized within the following symmetric proportion vector family:

$$\mathcal{F} := \left\{ (\pi_1, \ldots, \pi_J) \in [0, 1]^J : \pi_1 \leq \cdots \leq \pi_J \text{ and } \pi_j + \pi_{J+1-j} = \frac{2N_t}{N} \; \forall j \in \{1, \ldots, J\} \right\}.$$

If $N_t \leq N/2$, the **optimal proportion vector** $\boldsymbol{\pi}^*$ is

(i) if $V_4 \geqslant 0$,

$$\pi_j^* = \begin{cases} \frac{N_t}{N} - d & 1 \leq j \leq \frac{J}{2} \\ \frac{N_t}{N} + d & \frac{J}{2} < j \leq J \end{cases},$$

where

$$d = \arg \min_{x \in [0, N_t/N]} V_1 x^2 + V_2 x^4.$$

(ii) if $V_4 < 0$,

$$\pi_j^* = \begin{cases} 0 & 1 \leq j \leq \frac{N^2 d^2 J}{2N_t^2} \\ \frac{N_t}{N} & \frac{N^2 d^2 J}{2N_t^2} < j \leq J - \frac{N^2 d^2 J}{2N_t^2} \\ \frac{2N_t}{N} & J - \frac{N^2 d^2 J}{2N_t^2} < j \leq J \end{cases},$$

where

$$d = \arg \min_{x \in [0, N_t/N]} \left(V_1 + \frac{N_t^2}{N^2} V_4 \right) x^2 + (V_2 - V_4) x^4.$$

Deterministic Design

- An *optimal deterministic* design selects $\boldsymbol{\pi}^* \in \mathcal{S}$ that minimizes f $\boldsymbol{\pi}^* \in \arg\min_{\boldsymbol{\pi} \in \mathcal{S}} f(\boldsymbol{\pi}, \mathcal{C}, \Theta)$
- **Example**: objective function

$$f : (\boldsymbol{\pi}, \mathcal{C}, \{\boldsymbol{Y}(1), \boldsymbol{Y}(0)\}) \mapsto \mathrm{MSE}_{\boldsymbol{Z}}[\hat{\tau} | \boldsymbol{\pi}] = (\mathrm{TTE} - \mathbb{E}_{\boldsymbol{Z}}[\hat{\tau} | \boldsymbol{\pi}])^2 + \mathrm{Var}_{\boldsymbol{Z}}[\hat{\tau} | \boldsymbol{\pi}]$$

- The optimization is $\text{minimize} \quad \boldsymbol{\pi}^T \left(\tilde{\boldsymbol{W}}^+ [\tilde{\boldsymbol{W}}^+]^T - \mathcal{S}^+ \right) \boldsymbol{\pi} + \boldsymbol{\pi}^T \boldsymbol{S}^+$

$$s.t. \sum \pi_j = J \overline{\pi} = \frac{N_t J}{N}, \; \forall j = 1, \ldots, J$$

, where $W_i := \frac{N_t}{N} Y_i(0) + \frac{N_c}{N} Y_i(1)$, $\tilde{\boldsymbol{W}}^+ = \{W^{(j)} - N_j \overline{W}\}_{j=1}^J$, $\boldsymbol{S}^+ = \{N_j \mathbb{S}[\boldsymbol{W}^{(j)}]\}_{j=1}^J$, and $\mathcal{S}^+$ is a diagonal matrix with diagonal $\boldsymbol{S}^+$.

Example: Optimal Randomized Saturation Design

Population 40 cluster. 50 units in each cluster.
Network Stochastic Block Model with probability

$$P_{jl} = \exp\{-0.5|j - l|\}.$$

Outcome

$$\alpha_i \sim N(0, \sigma_\alpha) \text{ i.i.d. and } \sum_{i \in \mathcal{C}_j} \alpha_i = 0 \; \forall j$$

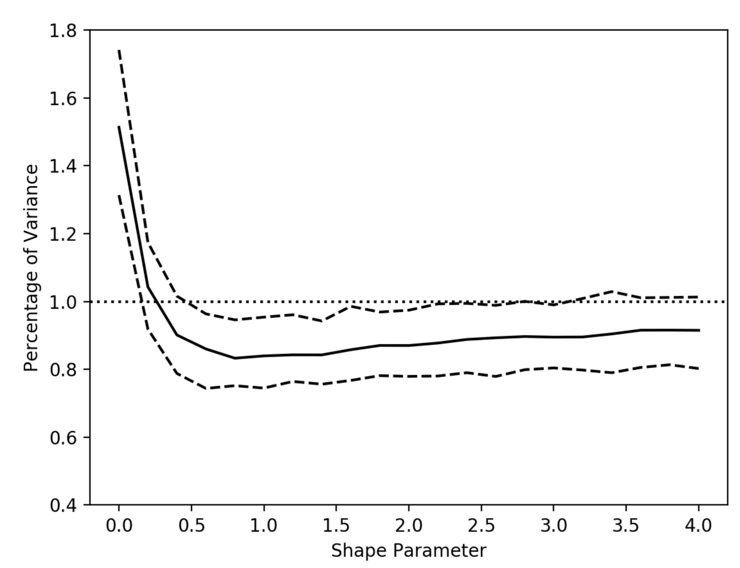
$$\beta_i = \gamma_i = 1 \; \forall i$$

Proportion vector

$$\pi_j = F_\lambda^{-1}(j/(J+1)), j = 1, \ldots, J,$$

where F_λ is the c.d.f. function of $\mathrm{Beta}(\lambda, \lambda)$. λ ranges from 0 (cluster-based) to 4, including ∞ (stratified)
Control σ_α is chosen so that $V_1 \approx -V_2/6$.

Plot Variance vs shape parameter λ



- variance normalized by the variance in stratified design ($\lambda = \infty$)
- 100 repetitions of experiment
- Solid: mean variance. Dashed: lower and upper 2.5% quantile.

Example: Deterministic Saturation Design

Population 40 cluster. 50 units in each cluster.
Network Stochastic Block Model with probability

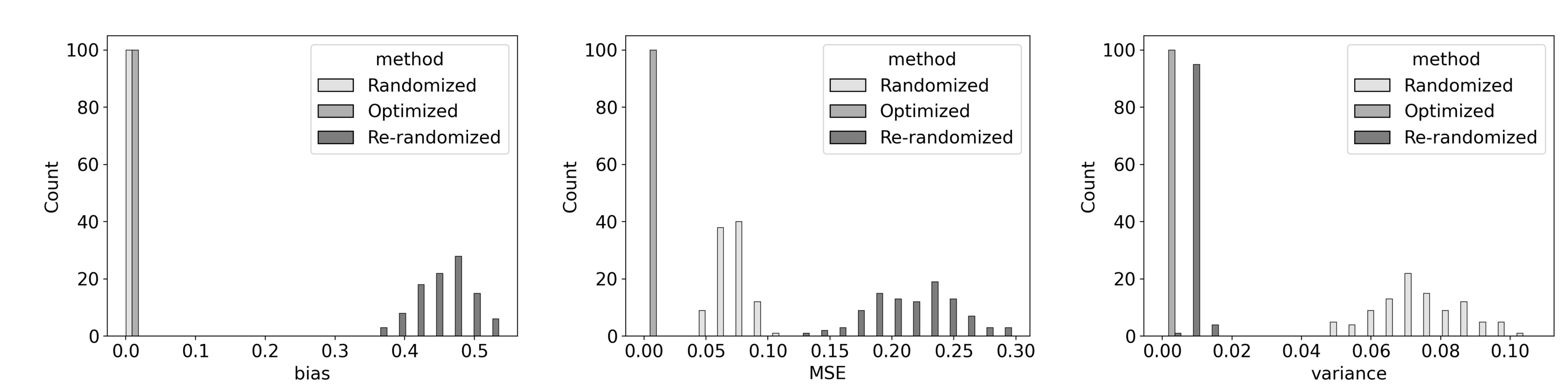
$$P_{jl} = 0.5 \cdot \mathbf{1}_{\{j=l\}}.$$

Outcome

$$\alpha_{[j]} \sim \mathrm{Unif}(0, 3)$$
$$\alpha_i = \alpha_{[j]} + \mathcal{N}(0, 0.01) \; \forall i \in \mathcal{C}_j$$
$$\beta_i = 1$$
$$\gamma_i = \gamma_{[j]} \sim \mathrm{Unif}(0, 1) \; \forall i \in \mathcal{C}_j$$

Treatment $N_t = N_c = N/2$

Three designs are simulated and compared:
Randomized Saturation Design. The optimal randomized saturation design using $\boldsymbol{\pi} \in \{0, 1\}^J$ with $\sum_j \pi_j = J/2$.
Optimized Deterministic Design. The optimal deterministic saturation design with a fixed proportion vector $\boldsymbol{\pi}^*$, obtained by optimizing minimizing $\mathrm{MSE}_{\boldsymbol{Z}}[\hat{\tau} \mid \boldsymbol{\pi}]$.
Re-randomized Saturation Design. The randomized saturation design using the proportion vector $\boldsymbol{\pi}^*$ from the optimal deterministic saturation design.



- Bias: $\text{Optimized} \approx \text{Randomized} \ll \text{ReRandomized}$
- Variance: $\text{Optimized} < \text{ReRandomized} \ll \text{Randomized}$
- MSE: $\text{Optimized} \ll \text{Randomized} \ll \text{ReRandomized}$

Manuscript

Optimizing Randomized and Deterministic Saturation Designs under Interference, arXiv:2203.09682

Contact: chencheng.cai@temple.edu