Just Ramp-up: Unleash the Potential of Regression-based Estimator for A/B Test under Network Interference

Qianyi Chen

Tsinghua University, School of Economics and Management

QR code of paper

Background: A/B Testing on the Network

- A/B test is the gold standard for modern platforms to support data-driven decision making on launching new product features, e.g., new algorithm/UI.

- Many platforms involve a network structure connecting its users, e.g., social network (LinkedIn, WeChat), two-sided market (Taobao, eBay).

- We call the network interference exists, when the outcome of certain units can also be influenced by the treatments allocated to its neighbors.

- Network interference is common in practice of large platforms and introduce substantial bias that blurs the conclusion of A/B testing.

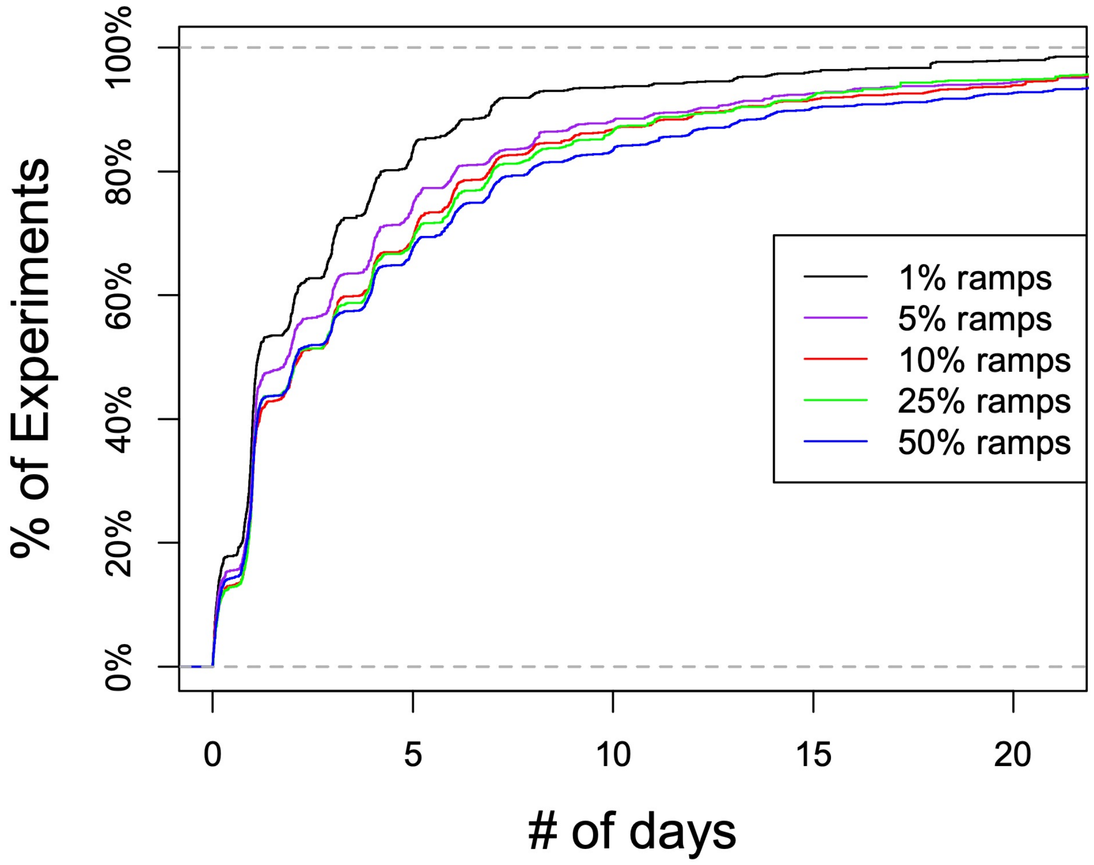
Background: Tackling Network Interference

- Experimental design
  - Cluster-level randomization (Hudgens 2008, Ugander 2013)
  - Refined covariance design of treatment vector (Candogan 2023, Chen 2023)

- Estimation
  - Network-adaptive estimator (Liu 2022&2024, Ugander 2023)
  - Counterfactual prediction with regression model (Leung 2024, Wu 2025)

- Our position: data-centric engineering for better regression-based estimation.

Ramp-up Process



Cumulative distribution of ramp duration, by ramp%

Ramp-up: gradually increasing the traffic to experiments.

- Resource constraint: more than 2,000 experiments are newly launched every week.

- Risk control: many new product features are useless or even harmful for user experience.

Ref: SQR: Balancing Speed, Quality and Risk in Online Experiments, KDD 2018

Basic Setting

- We consider binary treatment vector

$$\mathbf{z} = (z_1, z_2, \ldots, z_n) \in \{0, 1\}^n.$$

- The estimand in the A/B test is global average treatment effect (GATE)

$$\tau := \frac{1}{n} \sum_{i \in [n]} (Y_i(\mathbf{1}) - Y_i(\mathbf{0}))$$

- We consider unit-level complete randomization for analytical tractability

$$\sum_{i=1}^{n} z_i = d \quad \mathbb{E}[z_i] = \frac{d}{n}$$

- Ramp-up: multiple experiments with increasing treatment proportions

$$c_1 \leq c_2 \leq \cdots \leq c_T \quad c_t = \frac{d_t}{n}$$

Potential Outcome Model

- To enable exact bias/variance analysis, we need a parametric form of potential outcomes, which we call general linear interference model

$$Y(\mathbf{z}) = \beta_0 + \beta_1 \mathbf{z} + B\mathbf{z} + \epsilon$$

- It allows for general long-distance interference, in contrast to traditional 1-hop interference.

- Example: Linear-in-means model ($B = D^{-1}A$, normalized adjacency)

$$Y_i(\mathbf{z}) = \beta_0 + \beta_1 z_i + r\frac{\sum_{j\in\mathcal{N}(i)} z_j}{\deg_i} + \epsilon$$

Understand the GATE Estimation

- Estimation of GATE: an extrapolation task
  - The available data is only experimental data with small treatment proportions, e.g., 5%, 10%, etc.
  - The target is the mean outcomes under global treatment and global control.

- Estimation strategy
  - Macro-level: views the mean outcomes as 1-d function of treatment proportion p, $M(p)$. It's almost impossible to predict $M(1)$ with $M(0.05)$, $M(0.1)$.
  - Micro-level: the treated neighbors of some units can approach the case of $p = 1$ locally. Our regression are run on the outcomes of units.

Regression-based Estimator

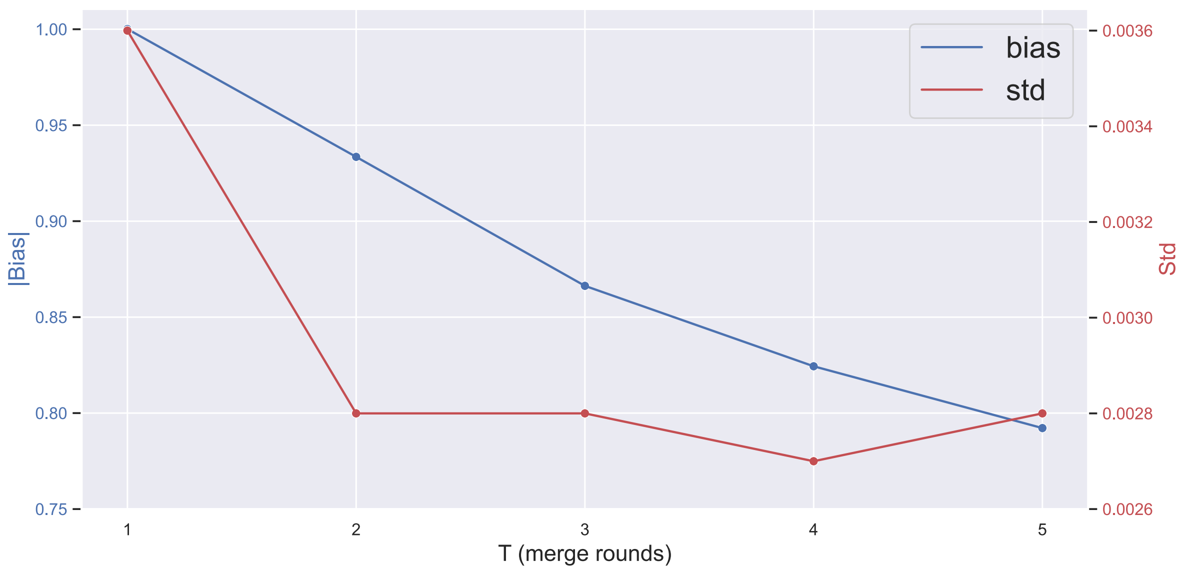
- We then formally define the regression-based estimator:
  A prediction function f that maps the treatment vector z and adjacency matrix A into the outcomes of each unit.

$$f : \{0,1\}^n \times \mathscr{A} \to \mathbb{R}^n$$

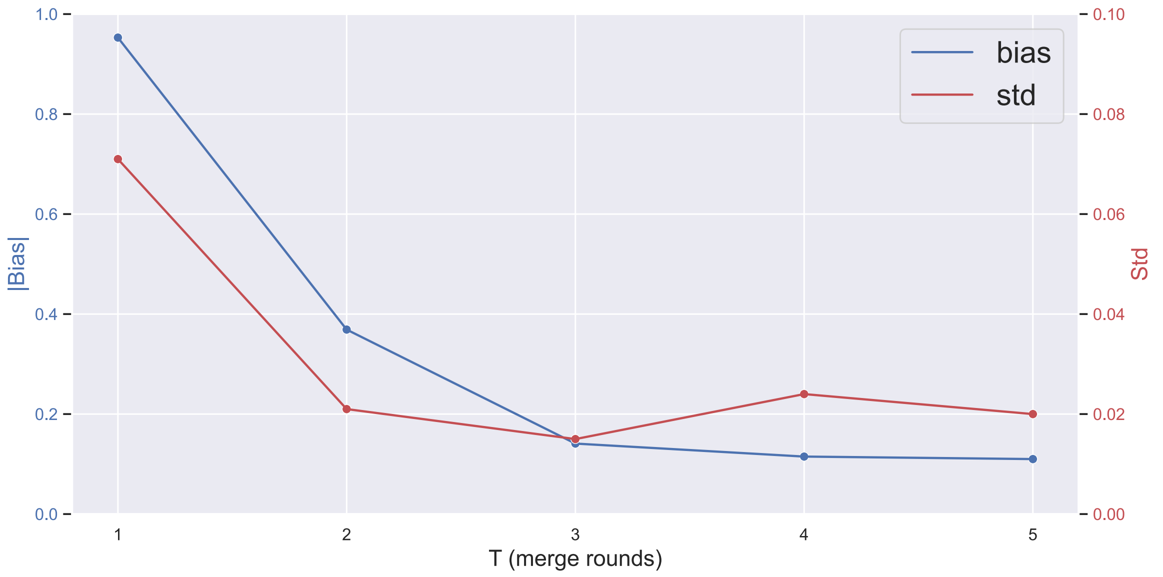
- Given this regression function, we give the GATE estimator as the difference between two predicted mean outcomes:

$$\hat{\tau}(f) = \frac{1}{n}\mathbf{1}^\top(f(\mathbf{1}, A) - f(\mathbf{0}, A))$$

Preview: Power of Merging



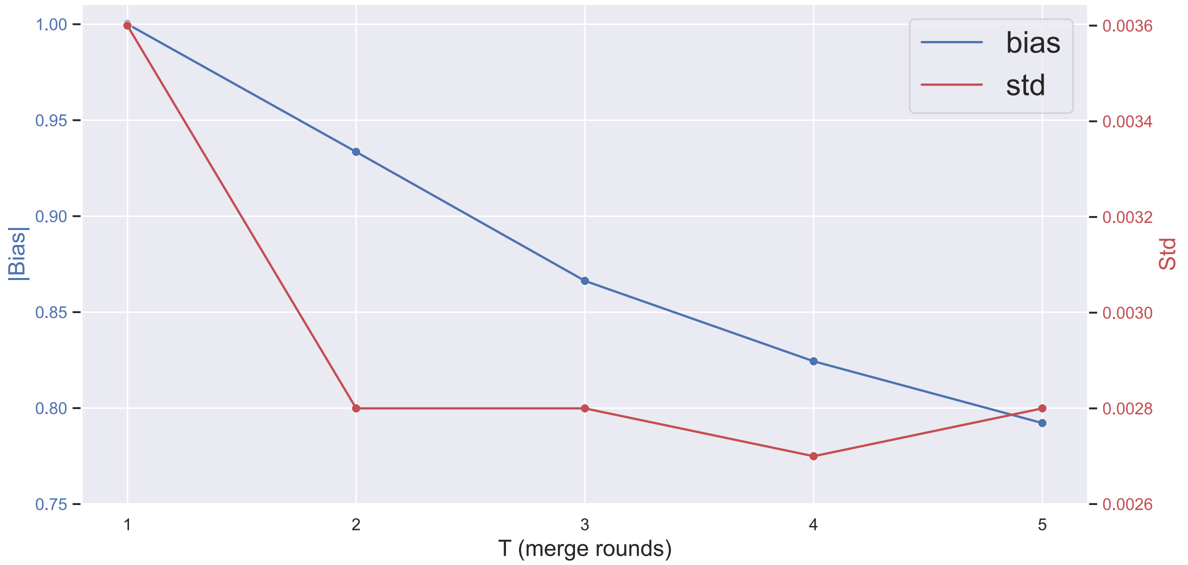
Linear Regression



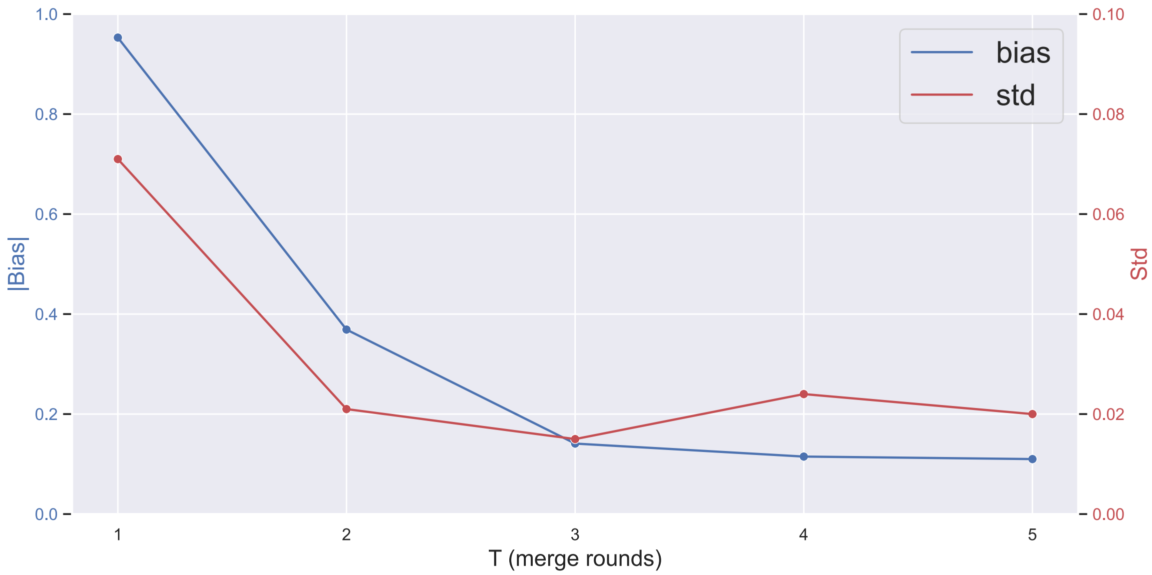
Graph Neural Network

Our methodology: merging experimental data at previous ramp-up steps to train the regression model, instead of only the current step.

Preview: Power of Merging



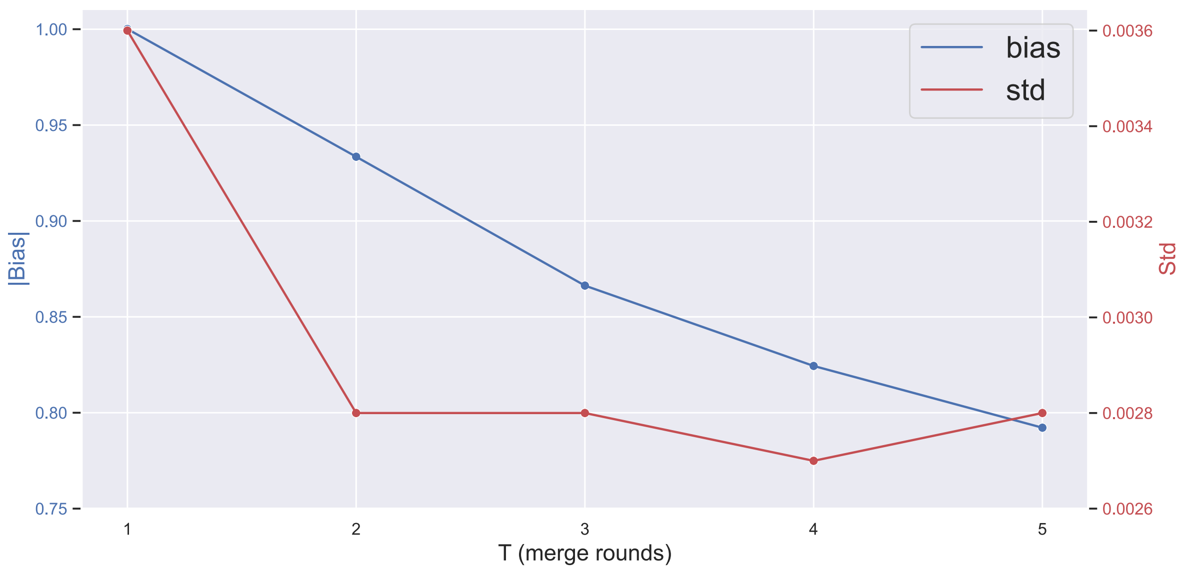
Linear Regression



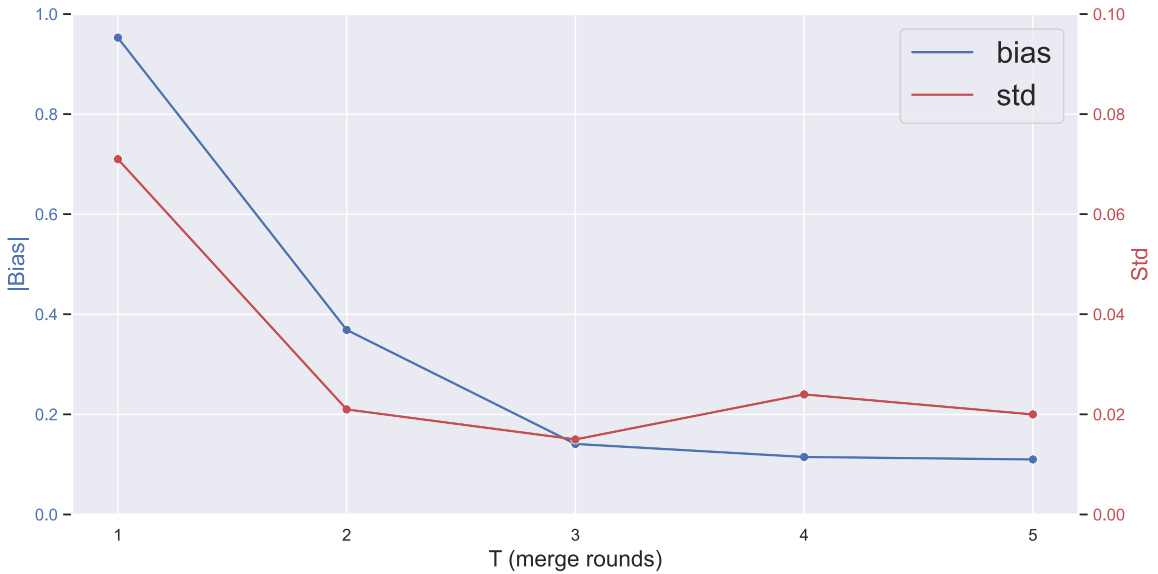
Graph Neural Network

Merging setting: $(c_1, c_2, \ldots, c_5) = (2\%, 5\%, 10\%, 25\%, 50\%)$. The t-th point corresponds to the result with merging the steps $(c_{6-t}, \ldots, c_5)$.

Preview: Power of Merging



Linear Regression



Graph Neural Network

Main messages:

- Bias dominates in this trade-off, even for the complex regression function like GNN.
- Substantial bias reduction is achieved through training regression model on merged data.

Linear Regression Estimator

- Why we choose linear regression as starting point
  - The empirical risk minimizer admits closed-form.
  - The conclusion derived from it can be empirically generalized to other advanced regression functions, e.g., GNN.

- Linear regression function:

$$f(\mathbf{z}, A) = X(\mathbf{z})\hat{\theta}$$

- Here, we do not incorporate network-dependent feature and use OLS

$$X(\mathbf{z}) = (\mathbf{1}, \mathbf{z}) \qquad \hat{\theta} = \left(X(\mathbf{z})^\top X(\mathbf{z})\right)^{-1} X(\mathbf{z})^\top Y$$

Further Tractability Issues

- The key for exact analysis of bias/variance lies in resolving the randomness of design matrix $X(\boldsymbol{z})$

- The matrix $\left(X(\boldsymbol{z})^\top X(\boldsymbol{z})\right)^{-1}$ involves a determinant in the denominator, making the analysis intractable if the determinant $\det(X^T X)$ is random.

$$X^\top X = \begin{pmatrix} n & \mathbf{z}^\top \mathbf{1} \\ \mathbf{z}^\top \mathbf{1} & \|\mathbf{z}\|_2^2 \end{pmatrix} = \begin{pmatrix} n & d \\ d & d \end{pmatrix}$$

Complete randomization: allocate exact d treatments to n units

One-step Experiment

Theorem 1 The bias and variance of the linear regression estimator under general linear interference are given by:

$$\text{Bias}\left(\hat{\tau}_{(1)}\right) = -\frac{\left(\sum_{i,j} B_{ij}\right)}{n}\left(\frac{1}{n(n-1)}+1\right)$$

Recall: B_{ij} denotes impact imposed by j on i

and

$$\begin{aligned}\text{Var}\left(\hat{\tau}_{(1)}\right) = \left(\frac{1}{nc(1-c)}\right)^2 \Bigg(& \left(\sum_{i,j} B_{ij}\right)^2 \left(\frac{c^3(c-1)}{n} + O\left(\frac{1}{n^2}\right)\right) + \left(\sum_{i,j,k} B_{ij}B_{kj}\right)\left(\frac{c(c-1)}{n} + O\left(\frac{1}{n^2}\right)\right) \\ & + \left(\sum_{i,j,l} B_{ij}B_{il}\right)\left(c^3(1-c) + O\left(\frac{1}{n}\right)\right) + \left(\sum_{i,j} B_{ij}^2\right)\left(c^2(1-c) + O\left(\frac{1}{n}\right)\right) \\ & + \left(\sum_{i,j} B_{ij}B_{ji}\right)\left((1-c)^2c^2 + O\left(\frac{1}{n}\right)\right)\Bigg) - \left(\sum_{i,j} B_{ij}\right)^2 \frac{1}{(n(n-1))^2} + \frac{\sigma_e^2}{nc(1-c)}\end{aligned}$$

Intensity of Regular Interference

Assumption 1 (Intensity of regular interference) *Based on the proposed general linear interference model, we further assume that for all* $i \in [n]$

$$\sum_{j=1}^{n} |B_{ij}| = O(1)$$

Moreover,

$$\sum_{i,j} B_{ij} = \Theta(n)$$

and

$$\sum_{j} \left(\sum_{i} B_{ij} \right)^2 = O(n)$$

Intensity of Regular Interference

Assumption 1 (Intensity of regular interference) *Based on the proposed general linear interference model, we further assume that for all* $i \in [n]$

$$\sum_{j=1}^{n} |B_{ij}| = O(1)$$

Recall: B_{ij} denotes impact imposed by j on i

Interference does not overshadow direct effect

Moreover,

$$\sum_{i,j} B_{ij} = \Theta(n)$$

Bias is considerable (otherwise bias would diminish; trivial case)

and

$$\sum_{j} \left(\sum_{i} B_{ij} \right)^2 = O(n)$$

Limited cumulative influence of opinion leaders

Intensity of Regular Interference

Assumption 1 (Intensity of regular interference) *Based on the proposed general linear interference model, we further assume that for all $i \in [n]$*

$$\sum_{j=1}^{n} |B_{ij}| = O(1)$$

Moreover,

$$\sum_{i,j} B_{ij} = \Theta(n)$$

and

$$\sum_{j} \left(\sum_{i} B_{ij} \right)^2 = O(n)$$

Example: Linear-in-means model

$$Y_i(\mathbf{z}) = \beta_0 + \beta_1 z_i + r \frac{\sum_{j \in \mathcal{N}(i)} z_j}{\deg_i} + \epsilon$$

For the instance $B = D^{-1}A$:

- The first two assumptions always hold.
- The third assumption imposes substantial restriction on adjacency matrix A.

A sufficient but not necessary condition: restricted growth rate (common in literature)

- Star graph ×
- Complete graph √

One-step Experiment

- In the regime of regular interference, we arrange the results and conclude that the bias is the dominant factor:

Corollary 1 *Based on Assumption 1 and Theorem 1, we further conclude that:*

$$\mathrm{Bias}\left(\hat{\tau}_{(1)}\right) = \Theta(1)$$
$$\mathrm{Var}\left(\hat{\tau}_{(1)}\right) = \Theta(1/n)$$

Two-step Experiment

- What's new:

$$B \xrightarrow{\text{expand}} \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix} \qquad \mathbf{z} \xrightarrow{\text{expand}} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}$$

- Sample size: $n \to 2n$

- Remark: For incorporating temporal interference, just add off-diagonal elements into $diag(B, B)$. Even if the temporal effect exists, we can still focus on the cross-sectional part by imposing appropriate assumptions.

Bias Reduction Brought by Merging

Theorem 2 *The bias of the linear regression estimator trained on merged data ($T = 2$) under general linear interference is given by:*

$$\mathrm{Bias}\left(\hat{\tau}_{(2)}\right) = -\frac{\sum_{ij} B_{ij}}{n}\left(1 - \frac{(n-1)\left(c_1 - c_2\right)^2 + 2\left(c_1^2 + c_2^2\right) - 2\left(c_1 + c_2\right)}{(n-1)\left(c_1 + c_2\right)\left(2 - \left(c_1 + c_2\right)\right)}\right)$$

When $n > 2\left(c_1 + c_2 - c_1^2 - c_2^2\right) / \left(c_1 - c_2\right)^2 + 1,$ *we have:*

$$\frac{(n-1)\left(c_1 - c_2\right)^2 + 2\left(c_1^2 + c_2^2\right) - 2\left(c_1 + c_2\right)}{(n-1)\left(c_1 + c_2\right)\left(2 - \left(c_1 + c_2\right)\right)} \in (0, 1)$$

This directly implies that $\mathrm{Bias}\left(\hat{\tau}_{(1)}\right)$ *and* $\mathrm{Bias}\left(\hat{\tau}_{(2)}\right)$ *shares the same sign. A substantial reduction in the magnitude of bias from* $\hat{\tau}_{(1)}$ *to* $\hat{\tau}_{(2)}$ *is given by:*

$$\boxed{\left|\mathrm{Bias}\left(\hat{\tau}_{(1)}\right)\right| - \left|\mathrm{Bias}\left(\hat{\tau}_{(2)}\right)\right| = \frac{\left|\sum_{ij} B_{ij}\right|}{n} \frac{\left(c_1 - c_2\right)^2}{\left(c_1 + c_2\right)\left(2 - \left(c_1 + c_2\right)\right)} + O\left(\frac{1}{n}\right)}$$

Bias Reduction Brought by Merging

- Bias reduction:

$$\left|\text{Bias}\left(\hat{\tau}_{(1)}\right)\right| - \left|\text{Bias}\left(\hat{\tau}_{(2)}\right)\right| = \frac{\left|\sum_{ij} B_{ij}\right|}{n} \frac{\left(c_1 - c_2\right)^2}{\left(c_1 + c_2\right)\left(2 - \left(c_1 + c_2\right)\right)} + O\left(\frac{1}{n}\right)$$

- Intuition: lower c_1 and larger c_2 will bring us closer to the desired scenario of global control and global treatment.

- Given a budget $\max\{c_1, c_2\} \leq c$, the best we can do: $c_1 = 0, c_2 = c$

$$\text{Bias}\left(\hat{\tau}_{(2)}\right) = -\frac{\sum_{ij} B_{ij}}{n}\left(1 - \frac{c}{2 - c}\right) + O\left(\frac{1}{n}\right)$$

- This is still unsatisfactory, which motivates refined regression functions.

Bias Reduction Brought by Merging (T-step)

Theorem 3 *For linear regression estimator trained on merged T-step data, the relative bias is given by:*

$$\text{Bias}\left(\hat{\tau}_{(T)}\right) = -\frac{\sum_{ij} B_{ij}}{n}\left(1 - \frac{T\sum_{t=1}^{T} c_t^2 - \left(\sum_{t=1}^{T} c_t\right)^2}{\left(\sum_{t=1}^{T} c_t\right)\left(T - \sum_{t=1}^{T} c_t\right)}\right) + O\left(\frac{1}{n}\right)$$

- Is merging still effective? Yes.
- Does T-step necessarily bring further bias reduction (v.s. fewer steps)? No. It improves only when an experiment with more extreme proportion is merged in.
- The benefit of merging 2-step experimental data is intrinsic.

Variance Remains Negligible

- Besides complex cross-units variance, there can also be correlations of treatments in the temporal dimension.
- We consider two cases, which is *temporally independent experiments* and *staggered rollout experiments* (non-decreasing treatments on the fly)

Theorem 4 *For the linear regression estimator trained on merged T-step temporally independent and staggered rollout experimental data, the order of variance is given by*

$$\mathrm{Var}\left(\hat{\tau}_{(T)}\right) = \Theta\left(\frac{1}{n}\right)$$

Further Intuition: Variation of Exposures

- The network exposure can be viewed as a representation of treatment vector in the interference term.

- For the linear-in-means model, the exposure can be specified as:

$$Y_i(\mathbf{z}) = \beta_0 + \beta_1 z_i + r \frac{\sum_{j \in \mathcal{N}(i)} z_j}{\deg_i} + \epsilon \qquad e_i = \frac{\sum_{j \in \mathcal{N}(i)} z_j}{\deg_i}$$

- We claim that the key challenge in learning the interference effect is ensuring sufficient variation in treatment exposures $\{e_i\}_{i=1}^{n}$.

Further Intuition: Variation of Exposures

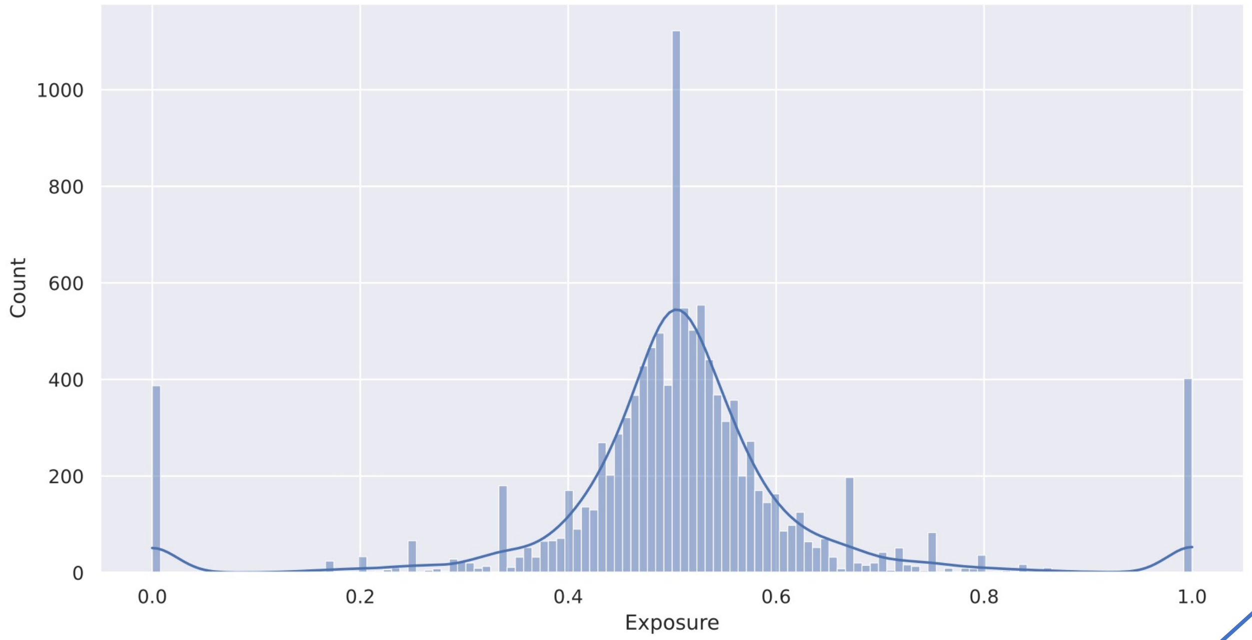


Figure 2: The distribution of treatment exposures under complete randomization with a treatment proportion of $c = 0.5$ is presented. The variance of treatment exposures is $\boxed{0.0243.}$ The network topology is sourced from the FB-Stanford3 dataset in [51], which represents a Facebook social network comprising $|\mathcal{V}| = 11586$ nodes and $|\mathcal{E}| = 568309$ edges. This network will be used in our simulation study.

Merging Increases the Variation of Exposures

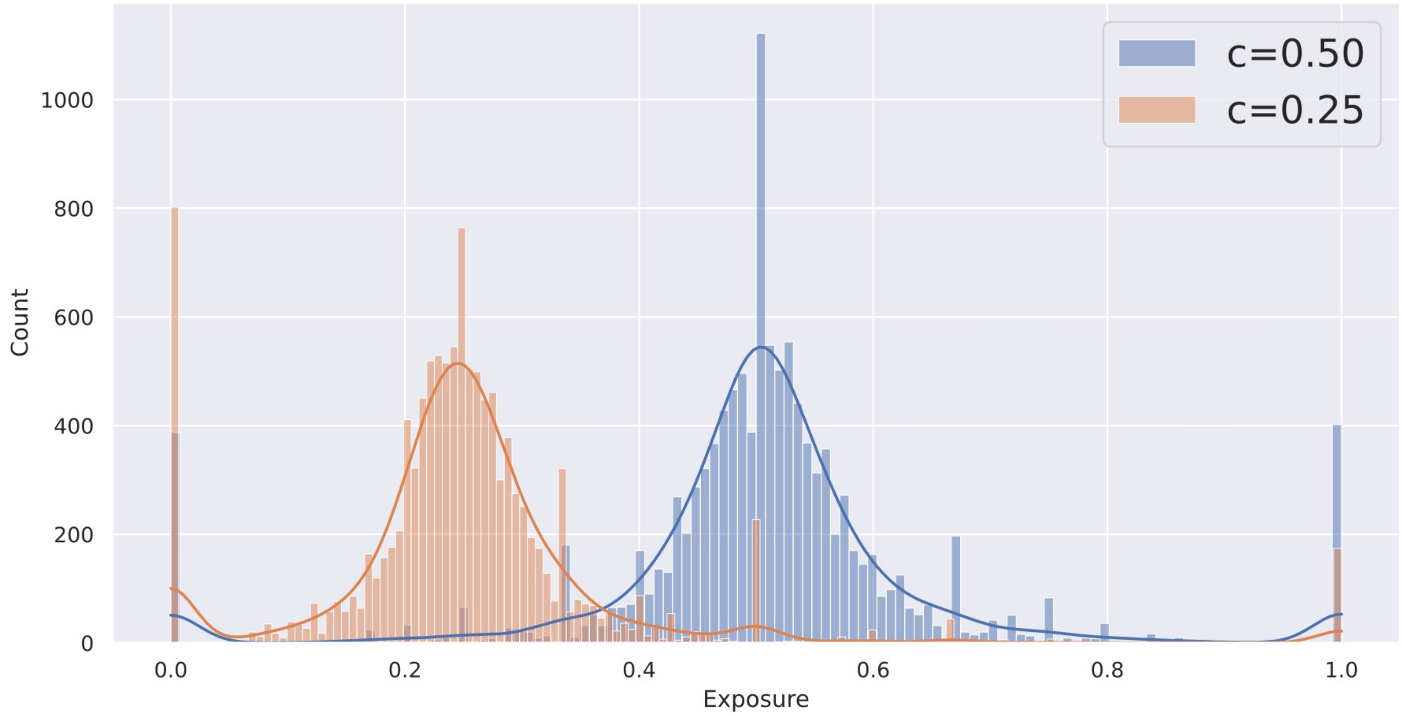


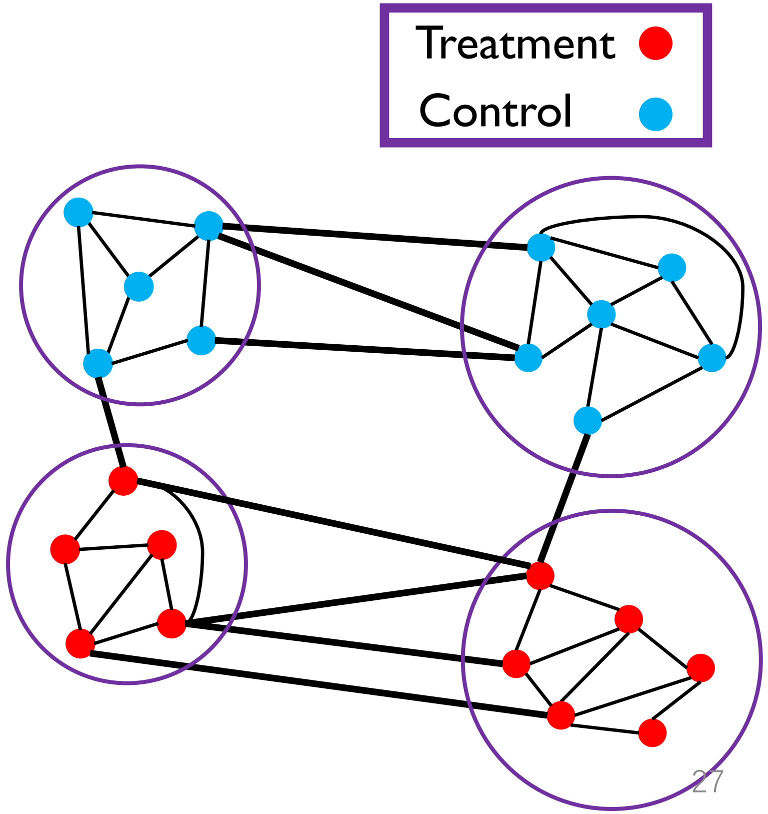
Figure 3: Distributions of treatment exposures under complete randomization with treatment proportion $c = 0.25$ and $c = 0.5$. The variances of treatment exposures are 0.0188 and 0.0243, respectively, with a variance of $\boxed{0.0377}$ for the merged data.

Synergy with Cluster-level Randomization

- The well-established methodology for tackling with interference is cluster-level randomization.

- Def. Allocate treatments at cluster-level, which makes units within the same cluster share the treatment level.

- Our new idea: it introduces strong correlations among treatments of units and increases the variation of exposures.



Synergy with Cluster-level Randomization

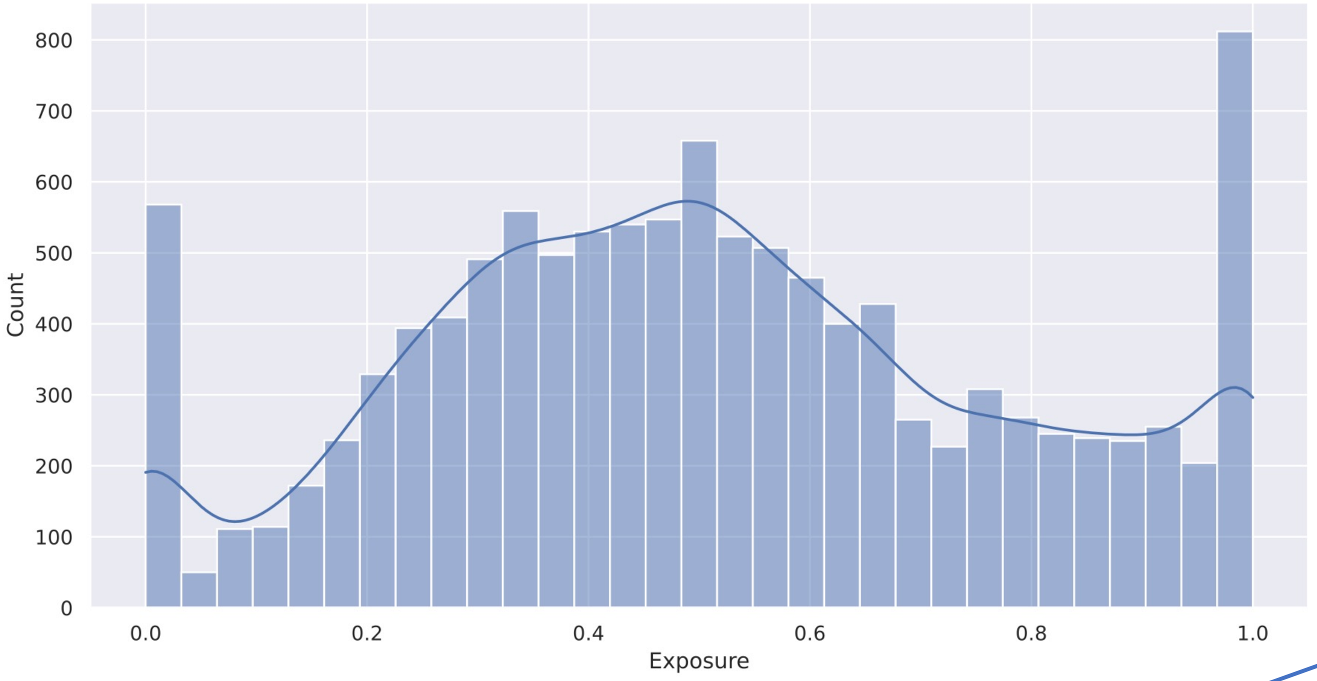
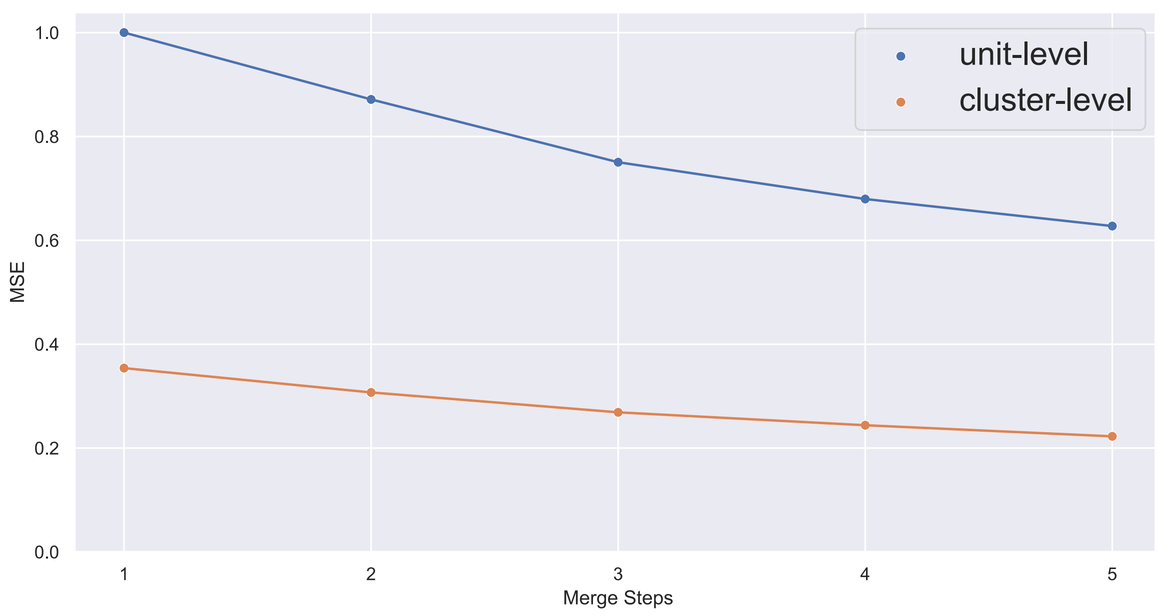
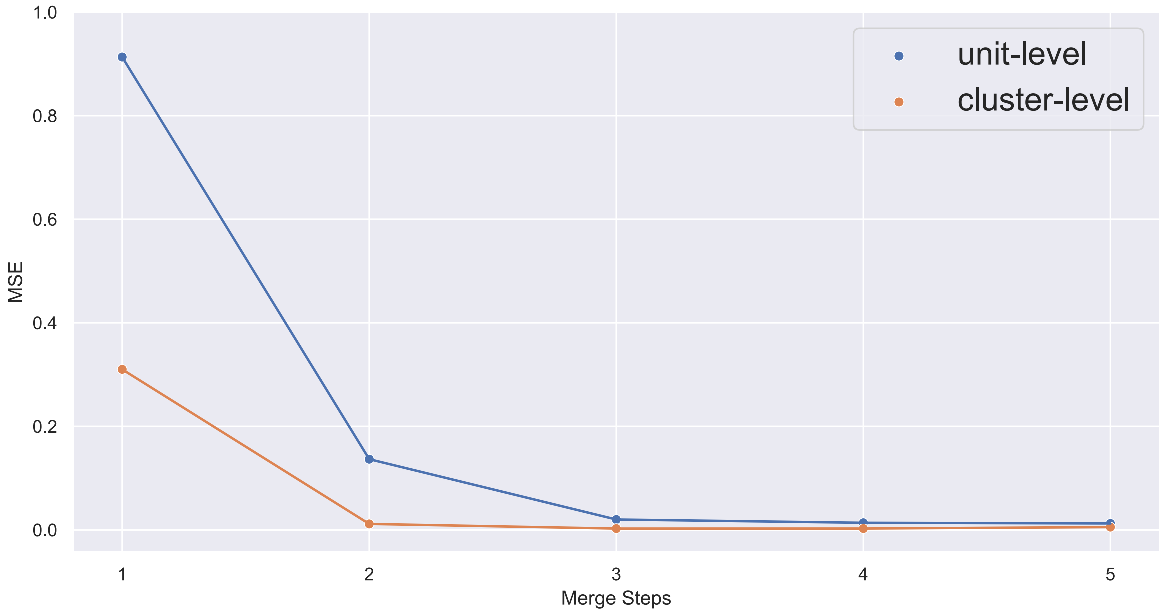


Figure 4: Distribution of treatment exposures under cluster-level complete randomization with treatment proportion $c = 0.5$. The variance of treatment exposures is $\boxed{0.0712.}$

Synergy with Cluster-level Randomization



Linear Regression



Graph Neural Network

Simulation Study TL;DR

- We examine a more refined estimator, graph neural network (GNN) with three layers of graph convolution.

- We further examine the following cases, beyond traditional settings
  - 2-hop interference (beyond 1-hop neighborhood interference)
  - Multiplicative, quadratic, square-root interference (beyond linear interference)
  - Dynamic graph structure (beyond static graph)

- Main takeaway:
  - Our findings generalize to GNN-based estimator.
  - In linear case, just merging two steps with lowest and largest proportion.
  - In complex cases, merging more intermediate steps can be beneficial.

QR code of paper

Simulation Study 1

randomization level

#merging rounds

level	Unit			Cluster		
metric	Bias	Std	MSE	Bias	Std	MSE
rounds						
$t = 1$	-0.951	0.071	0.909	-0.561	0.059	0.318
$t = 2$	-0.369	0.021	0.136	-0.112	0.037	0.014
$t = 3$	-0.142	0.014	0.020	-0.034	0.044	0.003
$t = 4$	-0.116	0.023	0.014	-0.030	0.048	0.003
$t = 5$	-0.110	0.020	0.013	-0.042	0.054	0.005

Performance of GNN estimator with complete randomization and staggered rollout

Merging setting:
$(c_1, c_2, \ldots, c_5) = (2\%, 5\%, 10\%, 25\%, 50\%)$. The t-th point means the result with merging the steps $(c_{6-t}, \ldots, c_5)$, namely, we stand at the point of ramp%=50% and consider merging previous steps.

Simulation Study I

randomization level → level; #merging rounds → rounds

level	Unit			Cluster		
metric	Bias	Std	MSE	Bias	Std	MSE
rounds						
$t = 1$	-0.951	0.071	0.909	-0.561	0.059	0.318
$t = 2$	-0.369	0.021	0.136	-0.112	0.037	0.014
$t = 3$	-0.142	0.014	0.020	-0.034	0.044	0.003
$t = 4$	-0.116	0.023	0.014	-0.030	0.048	0.003
$t = 5$	-0.110	0.020	0.013	-0.042	0.054	0.005

Performance of GNN estimator with complete randomization and staggered rollout

Messages:

- Merging remains effective for GNN estimator.
- Cluster-level randomization and merging methodology work synergistically.

Simulation Study 2

level		Unit			Cluster	
metric	Bias	Std	MSE	Bias	Std	MSE
rounds						
$t = 1$	-0.997	0.013	0.994	-0.600	0.018	0.360
$t = 2$	-0.995	0.012	0.990	-0.599	0.018	0.359
$t = 3$	-0.989	0.015	0.978	-0.595	0.020	0.354
$t = 4$	-0.975	0.025	0.952	-0.583	0.026	0.341
$t = 5$	-0.951	0.071	0.909	-0.561	0.059	0.318

Performance of GNN estimator with repeated experiments

Messages:

- The benefit is not simply from the increase of data volume.
- One should merge experimental data with the same population and different treatment proportions.

Simulation Study 3

level		Unit			Cluster	
metric	Bias	Std	MSE	Bias	Std	MSE
rounds						
$t = 1$	-1.401	0.034	1.963	-0.950	0.045	0.904
$t = 2$	-0.611	0.024	0.374	-0.303	0.064	0.096
$t = 3$	-0.316	0.013	0.100	-0.117	0.046	0.016
$t = 4$	-0.159	0.015	0.025	-0.061	0.064	0.008
$t = 5$	-0.160	0.024	0.026	-0.074	0.070	0.010

Performance of GNN estimator with square-root interference term

Messages:

- When interference becomes non-linear, merging more intermediate steps can be beneficial.

Simulation Study 4

level		Unit			Cluster	
metric	Bias	Std	MSE	Bias	Std	MSE
rounds						
$t = 1$	-1.401	0.034	1.963	-0.950	0.045	0.904
$t = 2$	-0.611	0.024	0.374	-0.303	0.064	0.096
$t = 3$	-0.316	0.013	0.100	-0.117	0.046	0.016
$t = 4$	-0.159	0.015	0.025	-0.061	0.064	0.008
$t = 5$	-0.160	0.024	0.026	-0.074	0.070	0.010

Performance of GNN estimator with multi-hop interference

Messages:

- Our findings remain valid when it comes to the case of multi-hop interference.

Simulation Study 5

level		Unit			Cluster	
metric	Bias	Std	MSE	Bias	Std	MSE
rounds						
$t = 1$	-0.666	0.095	0.453	-0.665	0.025	0.443
$t = 2$	-0.150	0.132	0.040	-0.161	0.061	0.030
$t = 3$	-0.106	0.137	0.030	-0.094	0.028	0.010
$t = 4$	-0.060	0.141	0.023	-0.058	0.044	0.005
$t = 5$	-0.045	0.149	0.024	-0.034	0.070	0.006

Performance of GNN estimator with dynamic graph structure (preferential attachment)

Messages:

- Our findings remains valid when graph structure is dynamic
- Merging intermediate steps are beneficial when graph structure becomes dynamic.

Thanks for Your Attention

Qianyi Chen, Bo Li

Tsinghua University, School of Economics and Management