Just Ramp-Up: Debiasing Regression-Based Causal Estimation under Network Interference

Qianyi Chen, Bo Li*
School of Economics and Management, Tsinghua University

Abstract

Network interference complicates A/B testing on online platforms, such as social networks and marketplaces, where causal methods based on single experiments often suffer from significant bias due to complex interference patterns. This paper demonstrates the statistical benefits of merging data from multiple experiments with varying treatment proportions. Sequential experimentation with increasing traffic, or ramp-up, is widely used in tech companies for risk management and cost control. Beyond operational benefits, we show that regression-based estimators trained on merged data achieve substantial bias reduction, even under simple randomization schemes and regression models. We focus on the global average treatment effect (GATE), a key estimand in the tech industry, and consider a general interference pattern that extends beyond the 1-hop setting. We present a closed-form bias–variance analysis of the linear regression estimator and show that, in practical settings, the bias term dominates. Moreover, we characterize how merging data across ramp-up stages improves regression training and reduces bias. We also offer an intuitive explanation for this reduction and highlight the synergy between cluster-level randomization and our approach. Furthermore, we consider a refined estimator based on graph neural networks (GNN). Extensive simulations across challenging scenarios confirm that our methodology significantly improves the accuracy of regression-based estimators. Code is available at github.com/Cqyiiii/Just-Ramp-up.

1 Introduction

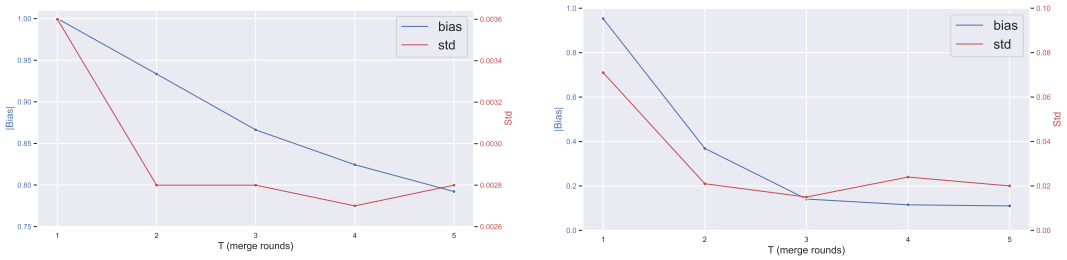
A/B testing has become a core paradigm for data-driven decision-making on modern online platforms. In many applications, the target causal estimand is the global average treatment effect (GATE), which measures the difference in average outcomes between globally deploying a new feature and maintaining the status quo. In complex systems such as social networks and marketplaces, however, this task is complicated by network interference: a unit's outcome may depend not only on its own treatment, but also on the treatments assigned to other connected units [25, 22, 13].

Estimating GATE under network interference is particularly challenging in practice because platforms rarely expose all users to treatment during experimentation. Instead, new features are typically rolled out gradually through a ramp-up process, starting from a very small treatment proportion and increasing traffic only for promising variants [31, 17]. As a result, the available data are collected under limited treatment proportions, whereas GATE concerns two extreme counterfactual regimes: global treatment and global control. This makes GATE estimation fundamentally an extrapolation problem, and interference further amplifies its difficulty [21].

The ramp-up process also provides a statistical opportunity: data from different stages can be pooled to train a regression model for counterfactual prediction. Pooling enriches variation in units' treatment

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) linear regression

(b) graph neural network

Figure 1: Absolute bias (blue curve, left axis) and standard deviation (red curve, right axis) of (a) linear regression and (b) GNN estimators trained on **merged** ramp-up data. The horizontal axis shows the number T of merged stages. At each T, training uses the final T stages, $(c_{6-T}, \ldots, c_5)$, where $(c_1, \ldots, c_5) = (2\%, 5\%, 10\%, 25\%, 50\%)$. Thus, $T = 1$ uses only the final experiment at 50% treatment, while $T = 5$ pools all five stages.

exposures—summaries of interference such as the fraction of treated 1-hop neighbors—and helps the model extrapolate to global treatment.

Our focus is on regression-based estimators for static GATE. Beyond their practical advantages, such as ease of implementation and the ability to incorporate covariates [11], regression-based estimators are especially attractive here because they can accommodate substantially richer interference structures than traditional estimators without a regression component. Much of the existing literature relies on neighborhood interference or related low-dimensional assumptions, which reduce the effect of the full treatment vector on a unit to the treatments within its 1-hop neighborhood [10, 7, 27, 20, 15, 5]. While such assumptions offer important technical simplifications, spillover beyond the 1-hop neighborhood remains underexplored despite its practical relevance. To address this gap, we study a novel interference pattern, *general linear interference*, which captures multihop spillover while still permitting *closed-form* bias and variance analysis, in contrast to the classical approximate neighborhood interference framework [19], which puts bounded-difference control on the impact of long-distant neighbors. Our analysis also incorporates the network topology explicitly, rather than relying only on coarse graph summaries such as the maximum degree, as is common in the literature.

We also emphasize the role of multiple experiments. While ramp-up experiments are standard in industry, most existing methods for network interference focus on a single experiment. There are some notable exceptions. For example, [7] models the mean outcome as a one-dimensional function of the treatment proportion and extrapolates to global treatment and control, while [3] uses multiple experiments for model selection over candidate interference structures. By contrast, our contribution is to show that multiple experiments can be pooled to improve regression-based GATE estimation itself. We interpret GATE estimation as a counterfactual prediction problem between two extreme exposure regimes, and the central benefit of merging lies in improving the training of the regression function through **richer network exposure variation** among units.

Our theoretical analysis starts with a simple network-agnostic linear regression estimator under general linear interference. Despite the complexity of long-range spillovers, this model admits a *closed-form* bias–variance decomposition and yields insights into more expressive regression-based estimators. We show that bias dominates estimation error in practically relevant regimes with regular interference intensity. Merging data across ramp-up stages substantially reduces this bias by varying treatment proportions and enriching the resulting network exposures.

Motivated by this finding, we then study a more expressive estimator based on graph neural networks (GNNs). Extensive simulations show that the same qualitative pattern persists: merged data markedly improve estimation accuracy, and powerful models such as GNNs benefit especially strongly from the richer exposure variation induced by merging. In this paper, we intentionally examine the performance of lightweight models, in contrast to GNN-based approaches with carefully designed architectures [29, 12]. Our key message is that, with our data-centric ramp-up augmentation, even simple regression models can achieve strong performance in GATE estimation.

Figure 1 provides an illustrative example. For both linear regression and GNN estimators, merging data from multiple ramp-up stages leads to a clear reduction in bias, while the variance remains

comparatively small. The figure also highlights a key practical message: even expressive models such as GNNs perform unsatisfactorily when trained on a single experiment, but become much more effective once trained on merged datasets with distinct treatment proportions. This comparison changes both sample size and treatment-proportion diversity; the repeated-experiment control in Section 5 examines whether additional observations at a fixed proportion offer the same benefit.

Our main takeaway is therefore simple: by pooling ramp-up experiments with different treatment proportions, one can train a regression model that more accurately predicts potential outcomes under global treatment and global control. Since ramp-up is already standard practice on many online platforms, this methodology requires no additional experimentation cost and can be deployed as a plug-and-play improvement for GATE estimation under network interference.

2 Basic Setting

2.1 Preliminary

We adopt the Neyman-Rubin potential outcome framework [24] and consider a finite population of n units. The treatment vector is represented as $\mathbf{z} = (z_1, z_2, \ldots, z_n) \in \{0,1\}^n$. Without assuming the Stable Unit Treatment Value Assumption (SUTVA), we define the potential outcome of unit i as $Y_i = Y_i(\mathbf{z})$ to account for its possible dependency on the treatments of other units. The parameter of interest is the global average treatment effect (GATE), which is defined as:

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

Here, $[n] = \{1, 2, \ldots, n\}$ is the unit set, and $\mathbf{1}$ and $\mathbf{0}$ are the n-dimensional all-ones and all-zeros vectors, representing global treatment and global control. We take the network as a given undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = [n]$ and adjacency matrix A. Let $\deg_i$ denote the degree of node i and write $D = \mathrm{diag}(\deg_1, \ldots, \deg_n)$. We assume every node has positive degree.

2.2 General Linear Interference Model

We consider the following data generation mechanism, called *general linear interference*:

$$Y(\mathbf{z}) = \beta_0 \mathbf{1} + \beta_1 \mathbf{z} + B\mathbf{z} + \boldsymbol{\epsilon}, \tag{2}$$

where (β_0, β_1) and B are all unknown parameters. With these, the true GATE is given by:

$$\tau = \beta_1 + \frac{1}{n} \mathbf{1}^\top B \mathbf{1}, \tag{3}$$

For convenience, we further assume an i.i.d. mean-zero noise structure.

$$\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma_e^2 I_n, \qquad \mathbf{z} \perp\!\!\!\perp \boldsymbol{\epsilon}. \tag{4}$$

The matrix B in Equation (2) characterizes interference effects. In this paper, we impose $B_{ii} = 0$ for all $i \in [n]$ to separate the direct treatment effect from the interference effect. For simplicity, we do not model heterogeneity in the baseline effect β_0 or direct effect β_1 across units. Instead, we place the modeling flexibility on a general interference matrix B, which can capture complex and potentially long-range dependencies on the network. We provide two examples of such matrices B below.

Example 2.1 (1-hop linear potential outcome model)**.** If we take $B = rD^{-1}A$, where r is the scale parameter, then our potential outcome model is reduced to the classic linear-in-means model [1]:

$$Y_i(\mathbf{z}) = \beta_0 + \beta_1 z_i + r \frac{\sum_{j \in \mathcal{N}(i)} z_j}{\deg_i} + \epsilon_i, \tag{5}$$

where $\mathcal{N}(i)$ denotes the 1-hop neighborhood of unit i in the graph.

Example 2.2 (Multihop linear potential outcome model)**.** We consider

$$B = (\mathbf{1}\mathbf{1}^\top - I_n) \odot \left(\sum_{m=1}^{M} r_m (D^{-1}A)^m \right), \tag{6}$$

where $\odot$ denotes the Hadamard product, M is the interference range, and $r_1 \geq r_2 \geq \cdots \geq r_M > 0$.

Notice that A^m characterizes the m-hop connections on the graph, where A^m_{ij} represents the number of walks of length m connecting units i and j. The term D^{-1} is included here as a normalization factor. The sequence of scale parameters, $\{r_i\}_{i=1}^M$, describes how interference decreases with increasing distance on the graph. Additionally, we exclude the diagonal elements of matrix B. This exclusion addresses the identification issue and can also be interpreted as eliminating the circular effects associated with closed walks that begin and end at the same unit. Since $D = \mathrm{diag}(A\mathbf{1})$, we can summarize the examples above as $B = g(A)$ for certain function g, namely, B models the interference that is conducted through the network topology. We consider the proposed general linear model for carrying out *closed-form* bias and variance analysis while maintaining sufficient generality.

2.3 Experimental Design

In this section, we discuss our default experimental design. Specifically, we adopt unit-level *complete randomization* as the basic randomization scheme, which randomly chooses exactly d units from a population of n units to receive treatment. To formalize it, we uniformly sample the binary treatment vector $\mathbf{z}$ from $\{\mathbf{z} \in \{0,1\}^n : \sum_{i=1}^n z_i = d\}$. In particular, $\mathbb{E}[z_i] = d/n$.

We note that the performance of unit-level complete randomization is often very similar to the independent Bernoulli randomization, which assigns treatment to each unit with probability $p = d/n$ independently. However, unit-level complete randomization ensures exactly d treated units, which brings considerable convenience in our derivation for closed-form expressions of bias and variance.

We next discuss the setting of multiple experiments. We consider a ramp-up process that includes T experiments that are subject to complete randomization with treated units $d_1 < d_2 < \cdots < d_T$. For convenience, we introduce another notation $c_t = d_t/n$. For instance, $T = 5$ and $(c_t)_{t=1}^T = (2\%, 5\%, 10\%, 25\%, 50\%)$, as also illustrated in [31]. Since our core methodology is merging multiple experimental data, we will also discuss the selection of T and $(c_t)_{t=1}^T$ later in the analysis.

2.4 Regression-based Estimator

We consider a regression-based estimator based on a function $f \in \mathscr{F}$ for predicting the counterfactual $Y(\mathbf{z})$, where $f : \{0,1\}^n \times \mathscr{A} \to \mathbb{R}^n$ and $\mathscr{A}$ is the space of adjacency matrices of simple graphs on n nodes. Given a prediction function f trained on experimental data, we define the estimator as the difference between the predicted mean outcomes under global treatment and global control:

$$\hat{\tau}(f) = \frac{1}{n}\mathbf{1}^\top \left(f(\mathbf{1}, A) - f(\mathbf{0}, A)\right). \tag{7}$$

In classical causal inference, the augmented inverse propensity weighting (AIPW; [6]) estimator is generally preferred to purely regression-based estimators. Under network interference, however, inverse propensity weights often lead to explosive variance [27, 5] and are also vulnerable to exposure misspecification. Regression-based estimators therefore occupy a distinctive and important role. In this paper, we focus on two regression model classes, linear regression $\mathscr{F}_{\text{LR}}$ and graph neural networks $\mathscr{F}_{\text{GNN}}$, and develop a data-centric approach for bias reduction in regression-based estimation.

3 Analysis on Linear Regression Estimator

We analyze a graph-agnostic, deliberately misspecified linear regression model under general linear interference. Its tractable bias and variance illustrate how merging data affects training. Our simulations show that these qualitative insights also apply to more expressive GNN models.

We fit $f(\mathbf{z}, A) = X(\mathbf{z})\hat{\theta}$, with $X(\mathbf{z}) = (\mathbf{1}, \mathbf{z})$ and ordinary least squares (OLS) coefficient

$$\hat{\theta} = (X(\mathbf{z})^\top X(\mathbf{z}))^{-1} X(\mathbf{z})^\top Y. \tag{8}$$

We retain the notation $X(\mathbf{z})$ to emphasize its dependence on the experimental design.

Linear regression is technically convenient because it admits a closed-form estimator, unlike many regression-based methods defined via variational formulations. The main difficulty is that the design matrix $X(\mathbf{z})$ is itself random under treatment assignment, so $(X(\mathbf{z})^\top X(\mathbf{z}))^{-1}$ contains a random determinant in the denominator, severely hindering exact analysis. However, we find that *complete randomization* offers a crucial source of tractability by fixing the number of treated units.

3.1 One-step Experiment

Throughout the section, we use $\hat{\tau}_{(T)}$ to denote the linear regression estimator based on merged T-step experiments. We now examine the case of $T = 1$ as a preparation. We specify the number of treated units in the experimental data as $d = cn$ and derive the bias and variance for $\hat{\tau}_{(1)}$.

Theorem 3.1. *The bias and variance of $\hat{\tau}_{(1)}$ under general linear interference are given by:*

$$\text{Bias}(\hat{\tau}_{(1)}) = -\frac{(\sum_{i,j} B_{ij})}{n-1}, \tag{9}$$

and

$$\begin{aligned}\text{Var}(\hat{\tau}_{(1)}) = &\left(\frac{1}{nc(1-c)}\right)^2 \Bigg((\sum_{i,j} B_{ij})^2 \left(\frac{c^3(c-1)}{n} + O\left(\frac{1}{n^2}\right)\right) \\ &+ (\sum_{i,j,k} B_{ij}B_{kj}) \left(\frac{c(c-1)}{n} + O\left(\frac{1}{n^2}\right)\right) + (\sum_{i,j,l} B_{ij}B_{il}) \left(c^3(1-c) + O\left(\frac{1}{n}\right)\right) \\ &+ (\sum_{i,j} B_{ij}^2) \left(c^2(1-c) + O\left(\frac{1}{n}\right)\right) + (\sum_{i,j} B_{ij}B_{ji}) \left((1-c)^2c^2 + O\left(\frac{1}{n}\right)\right) \Bigg) \\ &- (\sum_{i,j} B_{ij})^2 \frac{1}{(n(n-1))^2} + \frac{\sigma_e^2}{nc(1-c)}.\end{aligned} \tag{10}$$

Here, notations like $\sum_{i,j}$ denote summation over $i \in [n], j \in [n]$ without additional constraints. We suppress lower-order terms for readability; the complete expansion is given in Appendix C.1, and the hidden constants are indeed small. We now proceed to elaborate on the results presented above. First, we notice that the intensity of direct effect, β_1, does not influence either bias or variance. In total, the primary source of bias is interference, which corresponds to the second term in the expression.

Although the variance expression involves several quadratic functionals of B, the main takeaway is simple: under a regular interference regime (the interference is comparable to the direct effect), all of these terms are controlled at a lower order than the squared bias. This motivates the next assumption, which compresses the relevant requirements into three interpretable conditions.

Assumption 3.2 (Regular interference)**.** With general linear interference, we further assume that

$$\sum_{j=1}^{n} |B_{ij}| = O(1) \qquad \forall i \in [n]. \tag{11}$$

Moreover,

$$\sum_{i,j} B_{ij} = \Theta(n) \qquad \sum_{j} (\sum_{i} B_{ij})^2 = O(n). \tag{12}$$

The three conditions serve distinct purposes: the first bounds the total spillover received by each unit, the second ensures that average interference remains non-negligible, and the third rules out a vanishing fraction of highly influential nodes from dominating the variance. Under this regime, the variance is of lower order than the squared bias. We further provide a detailed discussion on this regime in Appendix A.1. Under Assumption 3.2, we obtain the following corollary:

Corollary 3.3. *Based on Assumption 3.2 and Theorem 3.1, we further conclude that*

$$\text{Bias}(\hat{\tau}_{(1)}) = \Theta(1) \qquad \text{Var}(\hat{\tau}_{(1)}) = \Theta(1/n). \tag{13}$$

This corollary shows that, for the estimator and regime considered here, the main challenge introduced by interference lies in the bias rather than the variance, especially since social networks on online platforms typically involve a large number of units. Therefore, to reduce the mean-squared error (MSE) of the estimator effectively, our primary focus should be on addressing the bias term.

3.2 Two-step Experiment

Although the derivation for $T > 2$-step experiments does not introduce substantial additional complexity compared with the case of $T = 2$, we discuss the latter separately because the benefit of merging experimental data from two ramp-up stages with distinct treatment proportions is fundamental, compared to the single-experiment case. First, we outline the changes induced by merging two experimental datasets. For merged data, we stack the observations from different ramp-up stages and fit one OLS regression on the resulting $2n$ rows. Thus each stage contributes n observations, while the treatment proportions differ across stages. In the baseline analysis, the stage-specific noises are independent across stages and units, with common variance σ_e^2. The treatment proportions in two experiments are c_1 and c_2. Moreover, there are two additional changes:

$$B \xrightarrow{\text{expand}} \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix} \qquad \mathbf{z} \xrightarrow{\text{expand}} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}. \tag{14}$$

Remark 3.4. The block-diagonal expansion in (14) rules out cross-round spillovers, as our focus is on the cross-sectional interference (see also [3]) and on estimating the static GATE. In principle, temporal interference can be incorporated by replacing $\mathrm{diag}(B, B)$ with $\begin{pmatrix} B_1 & 0 \\ B^{(1\to 2)} & B_2 \end{pmatrix}$, where $B^{(1\to 2)}$ captures the linear effect of first-round treatments on second-round outcomes. Dynamic graphs can likewise be accommodated by allowing $B_1 \neq B_2$. We focus on the block-diagonal case because temporal interference changes the estimand and would blur the cross-sectional bias structure of interest. Nevertheless, our framework is flexible enough to accommodate both temporal extensions.

We then characterize the bias of the 2-step estimator $\hat{\tau}_{(2)}$ in the following theorem. In the next subsection, we will show that the variance for general $\hat{\tau}_{(T)}$ remains negligible compared to the bias under Assumption 3.2. We maintain the setting $c_1, c_2 \in (0, 1)$ and $c_1 < c_2$ as introduced in Section 2.

Theorem 3.5. *The bias of the $\hat{\tau}_{(2)}$ under general linear interference is given by:*

$$\mathrm{Bias}(\hat{\tau}_{(2)}) = -\frac{\sum_{ij} B_{ij}}{n}\left(1 - \frac{(n-1)(c_1-c_2)^2 + 2(c_1^2+c_2^2) - 2(c_1+c_2)}{(n-1)(c_1+c_2)(2-(c_1+c_2))}\right). \tag{15}$$

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

$$0 < \frac{(n-1)(c_1-c_2)^2 + 2(c_1^2+c_2^2) - 2(c_1+c_2)}{(n-1)(c_1+c_2)(2-(c_1+c_2))} < 1. \tag{16}$$

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

$$|\,\mathrm{Bias}(\hat{\tau}_{(1)})| - |\,\mathrm{Bias}(\hat{\tau}_{(2)})| = \frac{|\sum_{ij} B_{ij}|}{n}\frac{(c_1-c_2)^2}{(c_1+c_2)(2-(c_1+c_2))} + O\left(\frac{1}{n}\right). \tag{17}$$

From the perspective of the online platform, $c_1 + c_2$ can be viewed as the total budget to be allocated in the experiment. If we fix $c_1 + c_2$, increasing $|c_1 - c_2|$ corresponds to an increase in exposure variation and a decrease in bias. This bias reduction is intuitive, since estimating the GATE inherently requires estimating the two mean outcomes under treatment proportions 0 and 1, and lowering c_1 while increasing c_2 brings us closer to these target regimes, reflecting a similar intuition in [9], namely that 'moving closer to the counterfactual situation of interest' induces bias reduction.

A complementary constraint caps the treatment proportion at $\max\{c_1, c_2\} \leq c$ to limit risk to revenue and user experience. Setting $c_1 = 0$ and $c_2 = c$ maximizes bias reduction under this cap, yielding

$$-\frac{\sum_{ij} B_{ij}}{n}\left(1 - \frac{c}{2-c}\right) + O\left(\frac{1}{n}\right), \tag{18}$$

which is still numerically unsatisfactory, particularly since $c \leq 0.5$ is common in practice.

Moreover, somewhat counterintuitively, merging *repeated* experiments with the same treatment proportion c does **not** substantially reduce bias, even though it increases the amount of data drawn from the same distribution. We discuss the implications of Theorem 3.5 further in Appendix A.2. Finally, since the variance analysis and conclusions for $T > 2$ closely parallel those for $T = 2$, we defer the detailed results to the next section. In particular, the variance remains of order $\Theta(1/n)$ and is therefore still negligible compared to the bias when merging T ramp-up stages.

3.3 T-step Experiment

In this section, we show that merging T ramp-up stages can still substantially reduce bias, although adding more intermediate steps may even be harmful compared to merging only the lowest and largest treatment proportions. We then establish that the variance remains secondary relative to the bias. Here, we emphasize that our regime is one in which n is large while T remains a small fixed constant (e.g., $n = 10000$ and $T = 4$), in line with practical settings on technology platforms.

Theorem 3.6. *Assume T is fixed. The bias of $\hat{\tau}_{(T)}$ under general linear interference is given by:*

$$\mathrm{Bias}(\hat{\tau}_{(T)}) = -\frac{\sum_{ij} B_{ij}}{n}\left(1 - \frac{T\sum_{t=1}^{T} c_t^2 - (\sum_{t=1}^{T} c_t)^2}{(\sum_{t=1}^{T} c_t)(T - \sum_{t=1}^{T} c_t)}\right) + O\left(\frac{1}{n}\right). \tag{19}$$

Thus, bias reduction depends on the dispersion of treatment proportions across stages. Repeated experiments with the same c have no such dispersion and do not reduce the leading bias. Moreover, simple algebra shows that, for fixed $c_1, \ldots, c_T$, merging more experimental stages does **not** necessarily lead to further bias reduction, and merging only the two extreme steps c_1 and c_T is already sufficiently effective. We defer the detailed discussion to Appendix A.3 and Appendix B.2.

Next, we analyze the variance of the regression-based estimator trained on merged data. Merging data across time introduces an additional temporal dimension in the design and may in principle create new covariance components. Nevertheless, we show that the variance remains of order $\Theta(1/n)$. We focus on two representative settings: temporally independent experiments and staggered rollout experiments [30], where treatment levels of units are non-decreasing over time.

Theorem 3.7. *For the linear regression estimator trained on merged T-step data from either **temporally independent** or **staggered rollout** experiments, the order of variance is given by:*

$$\mathrm{Var}(\hat{\tau}_{(T)}) = \Theta\left(1/n\right). \tag{20}$$

The final part of Assumption 3.2 restricts the order of the sum of the total impact that unit i imposes on other units. Intuitively, it restricts the number and influence of highly influential opinion leaders within the network. In Example 2.1, this assumption does not always hold and thus imposes substantial constraints on the adjacency matrix A, specifically, $\mathbf{1}^\top D^{-1}A^2D^{-1}\mathbf{1} = O(n)$.

For instance, in a ring graph, we have $\sum_j(\sum_i B_{ij})^2 = \Theta(n)$, while in a star graph—where one opinion leader connects to all other units—we have $\sum_j(\sum_i B_{ij})^2 = \Theta(n^2)$. In the context of Example 2.1, a sufficient condition for satisfying this assumption is the restricted growth rate that is commonly employed in related works [26, 27, 14]. However, a restricted growth rate is not a necessary condition; e.g. $\mathbf{1}^\top D^{-1}A^2D^{-1}\mathbf{1} = n$ in the complete graph; despite being the densest graph, it lacks any prominent opinion leaders.

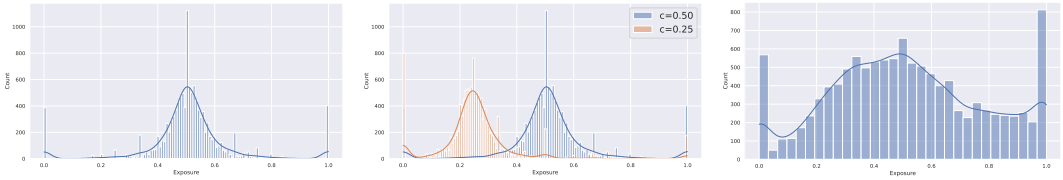
Lastly, the final part of Assumption 3.2 is not required for deriving the Corollary 3.3 but is instead intended to support the conclusion in the case of merging data, specifically that the variance of the linear regression estimator remains at the order of $\Theta(1/n)$.

4 An Intuitive Perspective on Exposure Variation

Although we study the graph-agnostic linear regression estimator for theoretical tractability, the conclusion of the previous section should extend more broadly. The key reason is that merging data across stages increases variation in macro-level treatment proportions, which **translates to variation in micro-level network exposures**. This added exposure diversity can likewise facilitate learning the interference pattern for more flexible models, such as GNN-based regressors.

Specifically, we use the case $T = 2$ and Example 2.1 for illustration. In this classical setting, the treatment exposure of unit i is defined as the fraction of its treated neighbors, $e_i = \sum_{j\in\mathcal{N}(i)} z_j/\deg_i$. We argue that the main challenge in modeling interference lies in obtaining sufficient variation in the treatment exposures $\{e_i\}_{i=1}^n$. Fortunately, because each e_i is induced by treatment randomization, the exposures naturally exhibit variation. We visualize their distribution in Figure 2.

In a single experiment under complete randomization, $\deg_i e_i$ follows a hypergeometric distribution and is well-approximated by $\mathrm{Binom}(\deg_i, c)$ when n is large. The overall exposure distribution is



(a) (b) (c)

Figure 2: Fig. 2a shows the exposure distribution under treatment proportion $c = 0.5$. Fig. 2b shows the corresponding distribution with $c = 0.25, 0.5$, and Fig. 2c shows that under *cluster-level randomization* with $c = 0.5$. The exposure variances are 0.0243, 0.0377, and 0.0712, respectively.

therefore a mixture of these unit-level laws across i, as shown in Figure 2a. Next, Figure 2b presents two histograms for the case $c_1 = 0.25, c_2 = 0.5$. The variation in treatment exposures within the merged data increases substantially. Beyond the variation induced by randomization, merging data from experiments with different treatment proportions naturally further enriches exposure variation.

Moreover, we note that the possible harmfulness of merging intermediate steps besides the first and T-th step, as detailed in Appendix A.3, can also be explained by this principle, i.e., merging intermediate steps would in turn decrease exposure variation. For example, merging stages with treatment proportions $(10\%, 50\%)$ yields an exposure variance of 0.0587. Adding a stage with $c = 0.25$ reduces this variance to 0.0459, illustrating how an intermediate stage can increase bias. Exposure variance is maximized when half the units have exposure 0 and half have exposure 1. This ideal, generally impractical configuration directly covers both counterfactual regimes defining GATE.

Figure 2c illustrates cluster-level randomization [13, 26, 9], which assigns a common treatment to units within each network community. These correlated assignments increase exposure variation relative to unit-level randomization: here, the exposure variance rises to 0.0712. This offers a learning-based explanation for the benefits of cluster randomization. By inducing covariance among treatments, it enriches the exposures available to the regression model. Merging and cluster randomization can therefore act jointly to improve estimation, as the simulations in Figure 3 demonstrate.

Finally, although the illustration is based on a particular exposure definition, we believe this intuition extends naturally to the most general setting, where $e_i = h_i(\mathbf{z}, A)$ with a unit-specific function h_i.

5 Simulation Study

5.1 Basic Setting

Data. We use the Facebook social network FB-Stanford3 [23], which has also been adopted in the experiments of [27, 4]. The network topology $\mathcal{G}$ contains $|\mathcal{V}| = 11586$ nodes and $|\mathcal{E}| = 568309$ edges. Based on this real network, we generate potential outcomes under the following models: linear-in-means [1], multiplicative [27], quadratic [7], multihop, and linear-in-means with dynamic graph structure. The linear-in-means and multihop models were introduced in Examples 2.1 and 2.2, respectively, and we take $M = 2$ (i.e., 2-hop) here. We defer the details of the dynamic-graph setting to Appendix B.1. We introduce the multiplicative and quadratic interference models below:

$$Y_i(\mathbf{z}) = (\beta_0 + \epsilon_i) \cdot \frac{\deg_i}{\overline{\deg}} \cdot (1 + \beta_1 z_i + r \frac{\sum_{j \in \mathcal{N}_i} z_j}{\deg_i}), \tag{21}$$

$$Y_i(\mathbf{z}) = \beta_0 + \beta_1 z_i + r \left(\frac{\sum_{j \in \mathcal{N}(i)} z_j}{\deg_i} \right)^2 + \epsilon_i. \tag{22}$$

Here, $\epsilon_i \sim N(0, \sigma_e^2)$ are i.i.d. normal noises. We set the parameters as $(\beta_0, \beta_1, r, \sigma_e) = (1, 1, 1, 0.1)$.

Regression model. Throughout this section, we intentionally use a lightweight 3-layer GNN as the regression model. The details of the model architecture are reported in Appendix B.1. We choose the Chebyshev Spectral Graph Convolution (ChebConv) [8] as our regression model, implemented through the Deep Graph Library [28]. Our GNN model consists of three ChebConv layers, with input dimension, output dimension, and filter size $(2, 16, 2), (16, 16, 1), (16, 1, 1)$, respectively.

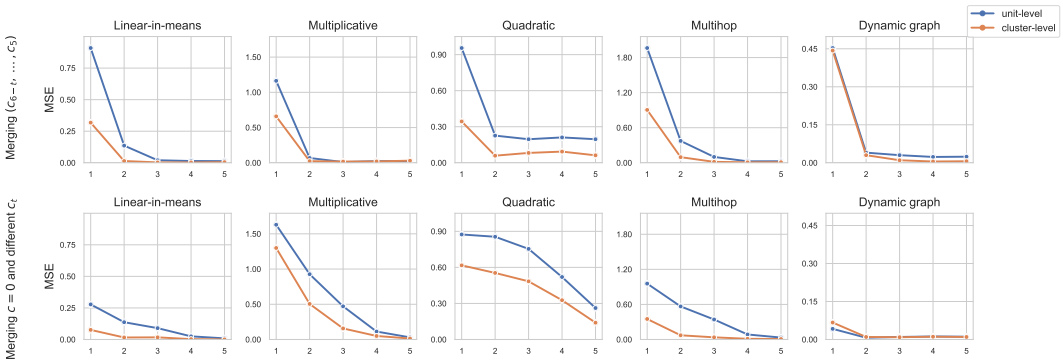


Figure 3: MSE of the GNN-based estimator under unit-level and cluster-level randomization. The top row merges stages $(c_{6-t}, \ldots, c_5)$; the bottom row combines each c_t with an initial stage treating one unit or cluster ($c \approx 0$). Both rows use the same vertical scale within each interference setting.

Table 1: Performance under **2-hop** interference for the two merging schemes.

	Unit-level			Cluster-level		
Stage	Bias	Std	MSE	Bias	Std	MSE
(a) Merging t stages: $(c_{6-t}, \ldots, c_5)$						
$t = 1$	-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
(b) Merging the initial stage ($c \approx 0$) with c_t						
$c_t = 0.02$	-0.977	0.032	0.956	-0.537	0.251	0.351
$c_t = 0.05$	-0.751	0.043	0.566	-0.217	0.159	0.072
$c_t = 0.10$	-0.580	0.078	0.342	-0.129	0.141	0.037
$c_t = 0.25$	-0.293	0.020	0.086	-0.094	0.061	0.012
$c_t = 0.50$	-0.183	0.012	0.033	-0.083	0.037	0.008

We train the GNN with Adam [16] to minimize MSE, using only treatment z_i and degree $\deg_i$ as initial node features. In most settings, the training loss approaches the irreducible noise variance σ_e^2, indicating adequate optimization of the regression model.

Merging scheme. We use two schemes with $(c_1, \ldots, c_5) = (2\%, 5\%, 10\%, 25\%, 50\%)$. The first fixes the final stage at $c_5 = 0.5$ and merges $\{c_{6-t}, \ldots, c_5\}$ for $t = 1, \ldots, 5$, measuring the benefit of incorporating earlier ramp-up data. The second combines each c_t with an initial stage that treats exactly one unit or cluster, approximating a zero-treatment baseline. This compares the current stage with an early stage. It is motivated by Theorem 3.6: under linear interference, merging the two extreme stages is already effective, while adding intermediate stages may increase bias.

Randomization scheme. We compare unit-level and cluster-level complete randomization under staggered rollout; Appendix B.1 describes cluster construction. Appendix B.4 also compares Bernoulli and complete randomization under both staggered rollout and temporally independent experiments, with similar performance across these alternatives.

5.2 Results and Analysis

We report MSE in Figure 3 and detailed 2-hop results in Table 1. Appendices B.2 and B.4 provide full numerical results and comparisons across randomization schemes, interference patterns, estimators without regression components, and network topologies.

Bias and standard deviation (Std) are estimated from 100 Monte Carlo repetitions. Std summarizes variability of the GATE estimate. The one-stage experiment at $c = 0.5$ serves as the baseline, shown at $t = 1$ in the top row of Figure 3 and in Table 1(a).

Table 2: Performance under 2-hop interference with **merged repeated experiments over t steps**.

	Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE
$t = 1$	-1.402	0.034	1.968	-0.954	0.049	0.912
$t = 2$	-1.460	0.015	2.132	-0.981	0.028	0.963
$t = 3$	-1.480	0.012	2.189	-0.992	0.025	0.984
$t = 4$	-1.486	0.012	2.208	-0.993	0.025	0.987
$t = 5$	-1.488	0.012	2.214	-0.994	0.026	0.989

The results support five conclusions: (i) **two-step merging already provides substantial benefits**; (ii) **merging and cluster-level randomization act synergistically** by jointly enriching exposure variation, as anticipated in Section 4; (iii) **bias dominates GNN estimation error in most settings**, so reducing it substantially improves MSE (Table 1(a) and Appendix B.2); (iv) **intermediate stages can help under complex interference**, including nonlinear models and dynamic graphs (Tables 5–7); (v) repeating experiments at a fixed proportion provides little bias reduction (Table 2).

Table 1(a) also shows that the gains need not be monotone. Under unit-level randomization, MSE falls from 1.963 at $t = 1$ to 0.374 at $t = 2$ and 0.025 at $t = 4$, while the fifth stage leaves it at 0.026. Cluster-level randomization reaches an MSE of 0.008 at $t = 4$ and 0.010 at $t = 5$. Thus, additional stages can help, but their value depends on the exposure information they contribute. These patterns support the interpretation in Sections 3 and 4; the GNN evidence complements the exact linear-regression analysis without extending its guarantees to arbitrary regressors.

Randomization schemes. Independent Bernoulli randomization is common in practice because of its simplicity. Complete randomization induces only weak correlations among unit treatments and yields similar performance in our simulations. We also compare temporally independent experiments with staggered rollout, forming a 2×2 comparison reported in Tables 8 and 9.

Merging repeated experiments. To distinguish the effect of exposure variation from that of training-set size, we merge repeated experiments at a fixed treatment proportion.

Table 2 reports results under 2-hop interference when all repeated experiments treat 50% of units. Repetition reduces standard deviation relative to one stage, but does not reduce bias or MSE. Increasing training-set size alone therefore does not reproduce the benefits of merging stages with **different** treatment proportions, consistent with Theorems 3.5 and 3.6.

Together, the experiments support combining cluster-level randomization with at least two stages having distinct treatment proportions when training a regression-based GATE estimator.

6 Conclusion

Ramp-up provides a statistical resource for GATE estimation in A/B tests under network interference. Our closed-form analysis shows that, under regular general linear interference, bias dominates the error of the linear regression estimator. Merging stages with different treatment proportions reduces this bias by increasing exposure variation. This perspective also explains the synergy with cluster-level randomization observed in our simulations. Across the studied interference settings, the benefits extend to lightweight GNN estimators, providing a practical way to improve counterfactual prediction using existing ramp-up data.

The conclusions concern static GATE and do not establish unbiasedness under arbitrary interference. Under a fixed traffic cap, merging can still leave substantial bias, as Equation (18) illustrates. Our closed-form analysis holds potential-outcome parameters fixed across stages and excludes cross-round spillovers; the nonlinear and dynamic-graph results provide simulation evidence rather than a general guarantee. In deployment, the available stages must also satisfy operational risk and traffic constraints. Stage selection should therefore consider exposure diversity as well as sample size, with the earliest and latest usable stages providing a natural starting point. Extending the analysis to carryover effects and adaptive rollout decisions remains an important direction.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.72171131).

References

[1] Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947, 2017.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[3] Ariel Boyarsky, Hongseok Namkoong, and Jean Pouget-Abadie. Modeling interference using experiment roll-out. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, 2023.

[4] Qianyi Chen, Bo Li, Lu Deng, and Yong Wang. Optimized covariance design for ab test on social network under interference. In *Advances in Neural Information Processing Systems*, 2023.

[5] Qianyi Chen, Anpeng Wu, Bo Li, Lu Deng, and Yong Wang. Journey to the centre of cluster: Harnessing interior nodes for a/b testing under network interference. In *The Fourteenth International Conference on Learning Representations*, 2026.

[6] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[7] Mayleen Cortez, Matthew Eichhorn, and Christina Yu. Staggered rollout designs enable causal inference under interference without network knowledge. In *Advances in Neural Information Processing Systems*, 2022.

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[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):20150021, 2016.

[10] Laura Forastiere, Fabrizia Mealli, Albert Wu, and Edoardo M Airoldi. Estimating causal effects under network interference with bayesian generalized propensity scores. *Journal of Machine Learning Research*, 23(289):1–61, 2022.

[11] Mengsi Gao and Peng Ding. Causal inference in network experiments: regression-based analysis and design-based properties. *Journal of Econometrics*, 252:106119, 2025.

[12] Qiang Huang and Jin Tian. Modeling interference for treatment effect estimation in network dynamic environment. In *The Fourteenth International Conference on Learning Representations*, 2026.

[13] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[14] Su Jia, Nathan Kallus, and Christina Lee Yu. Clustered switchback experiments: Near-optimal rates under spatiotemporal interference. *arXiv preprint arXiv:2312.15574*, 2025.

[15] Vardis Kandiros, Charilaos Pipis, Constantinos Daskalakis, and Christopher Harshaw. The conflict graph design: Estimating causal effects under arbitrary neighborhood interference. *arXiv preprint arXiv:2411.10908*, 2026.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

[17] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

[18] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, 2005.

[19] Michael P Leung. Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293, 2022.

[20] Yang Liu, Yifan Zhou, Ping Li, and Feifang Hu. Cluster-adaptive network a/b testing: From randomization to estimation. *Journal of Machine Learning Research*, 25(170):1–48, 2024.

[21] Ben M Parker, Steven G Gilmour, and John Schormans. Optimal design of experiments on connected units with application to social networks. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, pages 455–480, 2017.

[22] Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007.

[23] Ryan Rossi and Nesreen Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the AAAI conference on artificial intelligence*, 2015.

[24] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[25] Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[26] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.

[27] Johan Ugander and Hao Yin. Randomized graph cluster randomization. *Journal of Causal Inference*, 11(1):20220014, 2023.

[28] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2020.

[29] Anpeng Wu, Haiyi Qiu, Zhengming Chen, Zijian Li, Ruoxuan Xiong, Fei Wu, and Kun Zhang. Causal graph transformer for treatment effect estimation under unknown interference. In *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Ruoxuan Xiong, Susan Athey, Mohsen Bayati, and Guido Imbens. Optimal experimental design for staggered rollouts. *Management Science*, 70(8):5317–5336, 2023.

[31] Ya Xu, Weitao Duan, and Shaochen Huang. Sqr: Balancing speed, quality and risk in online experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 895–904, 2018.

A Additional Discussion on Theoretical Results

A.1 Discussion on Assumption 3.2

In this subsection, we revisit Assumption 3.2 and discuss its implications and practicality.

Scenarios where interference is excessively strong are relatively rare. Moreover, to assess the relative order of variance compared to bias in such cases, we will inevitably encounter the order of $\sum_{(i,j,k,l)\in\mathcal{I}} B_{ij}B_{kl}/\sum_{i,j} B_{ij}$ for various index sets $\mathcal{I}$ as outlined in Theorem 3.1. This complexity makes it challenging to derive succinct conclusions without specifying the interference matrix B.

For a specific unit i, the interference on its outcome is represented by $\sum_j B_{ij}z_j$, and the treatment effect is β_1. The first part of Assumption 3.2 restricts the interference effect received by unit i to be at most of the constant order, ensuring that interference does not overshadow the outcome. Additionally, this assumption allows for some units to be minimally affected by interference. Finally, through basic algebra, one can verify that $\sum_j |B_{ij}|$ is equal to 1 in Example 2.1 and bounded above by $\sum_{m=1}^{M} r_m$ in Example 2.2.

The second part of Assumption 3.2 restricts the order of total interference effect to be regular. While we assume that $\sum_{i,j} B_{ij} = \Theta(n)$, the bias reduction results discussed in the subsequent subsections also apply when $\sum_{i,j} B_{ij} = o(n)$, which indicates very weak interference. We do not incorporate this case here because the bias introduced by interference diminishes as n increases, given that $\sum_{i,j} B_{ij}/n = o(1)$. As illustrated in the expression of bias, it is unnecessary to use specific techniques designed to address interference, in this case. Therefore, we focus on the meaningful and challenging scenario, where $\sum_{i,j} B_{ij} = \Theta(n)$. For instance, in Examples 2.1 and 2.2, the order of $\sum_{ij} B_{ij}$ are both $\Theta(n)$.

The graph-structural interpretation of this condition is given in Section 3. This assumption is particularly reasonable for typical social networks, where, even with over a billion nodes, it is unlikely for any single unit to have thousands of friends while maintaining significant influence over each one of them.

A.2 Discussion on Theorem 3.5

Lowering c_1 and increasing c_2 brings us closer to these target mean outcomes. The traffic-capped bias expression is now given in Equation (18). Moreover, as demonstrated in the next subsection, merging multistep experimental data does not lead to further bias reduction compared to the two-step case. Therefore, we choose to explore more powerful regression models, specifically GNNs, in our simulation study.

A.3 Discussion on Theorem 3.6

In machine learning practice, it is common to see that merging more related data enhances the performance of the predictive models. However, through basic algebra, one can verify that merging more experimental data does **not** guarantee further bias reduction. This raises a natural question: Given the completed experiments with treatment proportions $c_1 < c_2 < \cdots < c_T$, which new experiment with proportion x will further reduce bias? We address this question through the following corollary.

Corollary A.1. *To further reduce the bias, x should satisfy:*

$$\left(T^2C_1 - (T+1)C_1^2 + TC_2\right)x^2 - \left((T-1)C_1^2 + 2TC_1C_2 + T(T+1)C_2\right)x + C_1^3 - C_1^2C_2 > 0, \tag{23}$$

here $C_1 = \sum_{t=1}^{T} c_t,\ C_2 = \sum_{t=1}^{T} c_t^2$

For this quadratic function with respect to x, we notice the following two facts about the coefficient of the quadratic term and the constant term:

$$\begin{aligned} T^2C_1 - (T+1)C_1^2 + TC_2 &= (TC_1(T-C_1)) + \left(TC_2 - C_1^2\right) > 0 \\ C_1^3 - C_1^2C_2 &= C_1^2(C_1 - C_2) > 0. \end{aligned} \tag{24}$$

We then draw several conclusions about the quadratic function in Corollary A.1. First, when $c_1 > 0$, merging data with $x = 0$—representing data under global control—can always reduce bias. In practice, this type of data is almost always accessible on online platforms. Second, intuitively, a beneficial proportion x should either be lower than c_1 or higher than c_T, rather than lying within the middle section of $[c_1, c_T]$. This intuition is only for illustration and not rigorous, since x around c_1 or c_T may also bring slight bias reduction, which depends on the specific values of $(c_t)_{t=1}^T$. Considering the first point discussed above, an effective way to further reduce bias is to conduct a larger-scale experiment, i.e., with $x > c_T$.

Additionally, we can apply Corollary A.1 in reverse: given the treatment proportions $\{c_i\}_{i=1}^T / \{c_t\}$, removing data with the treatment proportion c_t from the merged dataset may further reduce bias. Repeating this procedure suggests that the optimal solution may involve merging only the datasets with proportions c_1 and c_T. This finding highlights the unique importance of the $T = 2$ case, which we will further validate in the simulation study. Nonetheless, with nonlinear interference terms, merging multistep experimental data can offer advantages, as will be discussed in Section 5.

While it is theoretically possible to solve an optimization problem to select the treatment proportions $(c_t)_{t=1}^T$ for maximizing bias reduction, we do not recommend this approach. First, the optimization problem is dependent on the specific potential outcome model and the regression model used. Although we provide a detailed analysis of the simple regression estimator, it is not typically used in practice. Second, when selecting ramp-up steps in real-world experiments, the experimenter must consider factors such as risk and budget in advance, while such an optimization problem does not capture these practical considerations.

Finally, we discuss our proposed intuition of increasing exposure variation, as introduced in Section 4. The corresponding numerical illustration is reported in Section 4.

A.4 Discussion on Temporal Interference

The focus of this paper is network interference, namely the cross-sectional component. Although temporal interference may arise across multiple experiments, it is not our primary concern, and we abstract it away in our analysis. In principle, our analytical framework can be extended to accommodate both temporal interference and dynamic graph structures; however, doing so would alter the target estimand—the static GATE—and complicate the main narrative of this paper. In short, we study the regime in which cross-sectional interference is more substantial than temporal interference and therefore deserves primary attention.

B Detailed Experiments Results

B.1 Details of Basic Setting

Data. The main network topology is a Facebook social network used in many related works, available at https://networkrepository.com/socfb-Stanford3.php.

GNN. The model architecture and optimization objective are described in Section 5.

Other miscellaneous parameters are given below:

1. Resolution of clustering: 10
2. Number of clusters: 192
3. Activation function: ReLU
4. Number of training epochs: 400
5. Optimizer: Adam with learning rate 0.004 and momentum coefficient (betas) 0.9,0.999.

Potential outcome models. Beyond the results of the standard linear-in-means model (Example 2.1), we examine three non-linear interference patterns to demonstrate the generality of our theoretical

Table 3: **Linear-in-means** model: (left) merge t stages; (right) merge $c = 0$ and c_t.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Proportion**	Bias	Std	MSE	Bias	Std	MSE
$t = 1$	-0.951	0.071	0.909	-0.561	0.059	0.318	$c_t = 0.02$	-0.526	0.034	0.278	-0.216	0.173	0.076
$t = 2$	-0.369	0.021	0.136	-0.112	0.037	0.014	$c_t = 0.05$	-0.369	0.025	0.137	-0.032	0.121	0.016
$t = 3$	-0.142	0.014	0.020	-0.034	0.044	0.003	$c_t = 0.10$	-0.299	0.028	0.090	-0.010	0.130	0.017
$t = 4$	-0.116	0.023	0.014	-0.030	0.048	0.003	$c_t = 0.25$	-0.155	0.020	0.025	-0.037	0.045	0.003
$t = 5$	-0.110	0.020	0.013	-0.042	0.054	0.005	$c_t = 0.50$	-0.092	0.015	0.009	-0.015	0.033	0.001

Table 4: **Multiplicative** model: (left) merge t stages; (right) merge $c = 0$ and c_t.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Proportion**	Bias	Std	MSE	Bias	Std	MSE
$t = 1$	-1.078	0.039	1.163	-0.811	0.046	0.659	$c_t = 0.02$	-1.275	0.055	1.629	-1.103	0.288	1.299
$t = 2$	-0.257	0.037	0.068	-0.122	0.089	0.023	$c_t = 0.05$	-0.959	0.082	0.927	-0.666	0.249	0.505
$t = 3$	-0.061	0.077	0.010	-0.058	0.115	0.017	$c_t = 0.10$	-0.681	0.082	0.471	-0.340	0.205	0.157
$t = 4$	-0.110	0.106	0.023	-0.003	0.137	0.019	$c_t = 0.25$	-0.321	0.098	0.113	-0.193	0.111	0.050
$t = 5$	-0.120	0.104	0.025	-0.118	0.125	0.030	$c_t = 0.50$	-0.164	0.058	0.030	-0.101	0.063	0.014

insights, including **multiplicative**, **quadratic**, and **square-root interference**.

$$Y_i(\mathbf{z}) = (\beta_0 + \epsilon_i) \cdot \frac{\deg_i}{\overline{\deg}} \cdot (1 + \beta_1 z_i + r \frac{\sum_{j \in \mathcal{N}_i} z_j}{\deg_i}) \tag{25}$$

$$Y_i(\mathbf{z}) = \beta_0 + \beta_1 z_i + r \left(\frac{\sum_{j \in \mathcal{N}(i)} z_j}{\deg_i} \right)^2 + \epsilon_i \tag{26}$$

$$Y_i(\mathbf{z}) = \beta_0 + \beta_1 z_i + r \sqrt{\left(\frac{\sum_{j \in \mathcal{N}(i)} z_j}{\deg_i} \right)} + \epsilon_i, \tag{27}$$

Here, $\epsilon_i \sim N(0, \sigma_e^2)$ are i.i.d. normal noises. We set the parameters as $(\beta_0, \beta_1, r, \sigma_e) = (1, 1, 1, 0.1)$. We defer the introduction of our dynamic graph structure to Appendix B.3.

B.2 Detailed Results

We present the main numerical results below. Table 3 shows that the theoretical conclusions derived in Section 3 and Appendix A continue to hold for the GNN regressor. In particular, merging the first and last ramp-up stages (c_1, c_T) is already highly effective, and can even outperform merging intermediate stages. Under more complex interference patterns, however, incorporating additional intermediate stages can be substantially more beneficial, much like two points suffice to determine a line, whereas a cubic curve requires four.

B.3 Dynamic Graph Structure

We introduce our setting of a dynamic graph structure, which is novel in the literature but common in practice. For instance, if each experiment is conducted over a period of one week, the edges of the graph can change within that week, even in social networks. New friendships may form, and existing ones may be dissolved. The evolution of network structure is well-documented in network science, with models such as the Barabási-Albert Model [2] and the Forest Fire Model [18]. Although this aspect is still underexplored in the context of network experiments, it is valuable to consider this practical factor in sequential multiple experiments, such as in the ramp-up process.

Table 5: **Quadratic** interference: (left) merge t stages; (right) merge $c = 0$ and c_t.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Proportion**	Bias	Std	MSE	Bias	Std	MSE
$t = 1$	-0.976	0.044	0.955	-0.583	0.063	0.344	$c_t = 0.02$	-0.932	0.067	0.874	-0.776	0.120	0.617
$t = 2$	-0.475	0.020	0.226	-0.240	0.040	0.059	$c_t = 0.05$	-0.924	0.032	0.855	-0.739	0.084	0.554
$t = 3$	-0.443	0.010	0.196	-0.284	0.038	0.082	$c_t = 0.10$	-0.868	0.021	0.754	-0.692	0.074	0.484
$t = 4$	-0.460	0.015	0.211	-0.301	0.044	0.093	$c_t = 0.25$	-0.721	0.023	0.520	-0.569	0.058	0.327
$t = 5$	-0.442	0.032	0.196	-0.241	0.064	0.062	$c_t = 0.50$	-0.512	0.010	0.262	-0.373	0.035	0.140

Table 6: **Square-root** interference: (left) merge t stages; (right) merge $c = 0$ and c_t.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Proportion**	Bias	Std	MSE	Bias	Std	MSE
$t = 1$	-0.905	0.042	0.821	-0.623	0.051	0.391	$c_t = 0.02$	-0.324	0.027	0.106	0.218	0.304	0.140
$t = 2$	-0.387	0.014	0.150	-0.180	0.046	0.035	$c_t = 0.05$	-0.090	0.109	0.020	0.696	0.241	0.543
$t = 3$	-0.185	0.012	0.034	-0.063	0.052	0.007	$c_t = 0.10$	0.519	0.050	0.272	0.895	0.145	0.822
$t = 4$	-0.086	0.010	0.008	0.021	0.057	0.004	$c_t = 0.25$	0.528	0.030	0.280	0.579	0.091	0.344
$t = 5$	-0.006	0.012	0.000	0.030	0.081	0.007	$c_t = 0.50$	0.218	0.008	0.048	0.250	0.057	0.066

Table 7: **Dynamic graph structure**: (left) merge t stages; (right) merge $c = 0$ and c_t.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Proportion**	Bias	Std	MSE	Bias	Std	MSE
$t = 1$	-0.666	0.095	0.453	-0.665	0.025	0.443	$c_t = 0.02$	-0.196	0.061	0.042	-0.220	0.135	0.067
$t = 2$	-0.150	0.132	0.040	-0.161	0.061	0.030	$c_t = 0.05$	-0.069	0.048	0.007	-0.086	0.055	0.010
$t = 3$	-0.106	0.137	0.030	-0.094	0.028	0.010	$c_t = 0.10$	-0.092	0.037	0.010	-0.084	0.041	0.009
$t = 4$	-0.060	0.141	0.023	-0.058	0.044	0.005	$c_t = 0.25$	-0.107	0.021	0.012	-0.100	0.025	0.011
$t = 5$	-0.045	0.149	0.024	-0.034	0.070	0.006	$c_t = 0.50$	-0.099	0.031	0.011	-0.098	0.028	0.010

Remark B.1. Here, for simplicity, we consider the exogenous network evolution that is independent of treatment assignment. This is because the treatments typically considered by online platforms, such as new product features, generally do not have a significant impact on social relationships. However, there is existing literature in economics that examines cases where treatment does endogenously influence network formation, usually when the treatment itself is more consequential.

For simplicity, we keep the node set unchanged throughout the ramp-up process and focus on modeling the addition of new edges over time. Inspired by classic preferential attachment, we generate edges for nodes with degrees in the top 50%. For each such node i, we connect it to node j with the following probability:

$$p_{ij} = \frac{\deg_j}{\sum_k \deg_k}, \tag{28}$$

In each evolution, we repeat this sampling procedure 5 times with replacement for each node with degrees in the top 50%. If a unit i is sampled when generating edges for itself, we discard this sampling to avoid self-loops. Additionally, if the sampled unit j is already connected to unit i at the current step, we also discard this sampling to avoid creating multiple edges.

Given the ramp-up steps $(c_1, c_2, \ldots, c_5) = (2\%, 5\%, 10\%, 25\%, 50\%)$, we take the initial network topology $\mathcal{G}$ as baseline and repeat such evolution 5 times to generate evolved graphs with the same node set. The number of edges:

$$|\mathcal{E}_1|, \ldots, |\mathcal{E}_5| = 596414, 624465, 652391, 680313, 708274, \tag{29}$$

where $\mathcal{E}_t$ is the edge set at the t-th step.

Moreover, we go back to use the classic linear-in-means model and examine the influence of dynamic graph structure. The main parameter is set as $(\beta_0, \beta_1, r, \sigma_e) = (1, 1, 1, 0.1)$, as in the previous settings. Even if the graph structure evolves, the GATE remains $\beta_1 + r = 2$. We present the result in Table 7.

The impact of evolutionary changes is quite complex, with classic models typically only characterizing the limiting behavior of preferential attachment. Nonetheless, our methodology consistently enhances the GNN model's performance. During the evolution process, we preserve the clustering of the original topology $\mathcal{G}$ to implement the staggered rollout design[1], although this clustering may become less effective as edges are continuously added. As a result, our suggested practice of merging data with global control and the highest treatment proportion faces challenges due to the increasing discrepancy between $\mathcal{G}_t$ and $\mathcal{G}$ as step t progresses. This provides an intuitive explanation for why the monotonic improvement observed with increasing c_t, as shown in our basic experiments, no longer persists in Table 7. This reveals a potential limitation of this merging strategy in practice.

Remark B.2. The rationale for generating edges solely for high-degree nodes stems from the distribution of treatment exposure, as detailed in the main paper. Low-degree nodes typically have exposure values concentrated at the boundaries of the interval $(0, 1)$, which significantly contributes

[1]The time complexity of the Louvain algorithm supports weekly updates for clustering on large online platforms. However, when clustering changes, monotonic treatment assignment may no longer be guaranteed under cluster-level randomization.

Table 8: Performance under **independent Bernoulli randomization**: (Left) **temporal independence**, (Right) **staggered rollout**.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Rounds**	Bias	Std	MSE	Bias	Std	MSE
$t=1$	-0.945	0.071	0.899	-0.562	0.058	0.320	$t=1$	-0.944	0.072	0.896	-0.560	0.058	0.317
$t=2$	-0.365	0.023	0.133	-0.118	0.046	0.016	$t=2$	-0.365	0.022	0.134	-0.117	0.047	0.016
$t=3$	-0.142	0.015	0.020	-0.038	0.041	0.003	$t=3$	-0.142	0.015	0.020	-0.036	0.044	0.003
$t=4$	-0.115	0.024	0.014	-0.032	0.050	0.004	$t=4$	-0.116	0.023	0.014	-0.033	0.051	0.004
$t=5$	-0.110	0.019	0.012	-0.042	0.056	0.005	$t=5$	-0.110	0.019	0.012	-0.042	0.056	0.005

Table 9: Performance under **complete randomization**: (Left) **temporal independence**, (Right) **staggered rollout**.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Rounds**	Bias	Std	MSE	Bias	Std	MSE
$t=1$	-0.953	0.071	0.914	-0.560	0.059	0.317	$t=1$	-0.951	0.071	0.909	-0.561	0.059	0.318
$t=2$	-0.369	0.021	0.136	-0.114	0.037	0.014	$t=2$	-0.369	0.021	0.136	-0.112	0.037	0.014
$t=3$	-0.141	0.015	0.020	-0.033	0.045	0.003	$t=3$	-0.142	0.014	0.020	-0.034	0.044	0.003
$t=4$	-0.115	0.024	0.014	-0.031	0.047	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	$t=5$	-0.110	0.020	0.013	-0.042	0.054	0.005

to the variation in exposures. To preserve this distinctive structure, we intentionally avoid generating new edges for all nodes. However, we do allow low-degree nodes to receive new edges initiated by high-degree nodes.

B.4 Supplementary Experiments

The randomization comparison is discussed in Section 5; the complete numerical results are reported below.

Merging repeated experiments. In Theorem 3.5 and 3.6, we conclude that the potential bias reduction of merging experiments with the same treatment proportion should be very marginal, i.e., at most of order $O(1/n)$, and we examine this for the linear regression estimator here.

The empirical comparison is reported in Table 2 in the main text.

Linear regression estimator We then supplement the statistical performance of the linear regression estimator under the basic setting, as shown in Figure 1. As in the main paper, we fix the randomization scheme as complete randomization with staggered rollout. We consider treatment proportions $(c_1, c_2, \ldots, c_5) = (2\%, 5\%, 10\%, 25\%, 50\%)$ in the ramp-up process and train on merged data from $t = 1, 2, 3, 4, 5$ steps, corresponding to treatment proportions $(c_{6-t}, \ldots, c_5)$. The detailed results are reported in Table 10. By basic algebra, these numerical results validate Theorem 3.5.

Lagrange interpolation estimator We next supplement the performance of the Lagrange interpolation estimator proposed in [7]. Unlike our regression-based approach, this is a system-level estimator that is agnostic to the graph microstructure: it approximates the one-dimensional mean outcome curve $\bar{Y}(p)$ as a function of the treatment proportion p under unit-level independent Bernoulli randomization. Its polynomial basis effectively assumes a low-order interaction structure of the form

$$\sum_{\ell=1}^{\beta} w_l \left(\frac{\sum_{j \in \mathcal{N}_i} \tilde{c}_{ij} z_j}{\sum_{j \in \mathcal{N}_i} \tilde{c}_{ij}} \right)^{\ell}, \tag{30}$$

Table 10: Performance with **linear regression estimator** when merging t steps.

	Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE
$t=1$	-1.000	0.004	1.000	-0.595	0.010	0.354
$t=2$	-0.933	0.003	0.871	-0.554	0.019	0.307
$t=3$	-0.866	0.003	0.750	-0.518	0.027	0.269
$t=4$	-0.824	0.003	0.680	-0.493	0.028	0.244
$t=5$	-0.792	0.003	0.628	-0.471	0.031	0.223

Table 11: **Lagrange interpolation** under **multiplicative** interference with t stages.

	Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE
$t=1$	N/A	N/A	N/A	N/A	N/A	N/A
$t=2$	-0.002	0.036	0.001	0.010	0.237	0.056
$t=3$	-0.005	0.074	0.005	0.045	0.479	0.231
$t=4$	0.013	0.561	0.315	0.740	3.501	12.806
$t=5$	0.045	8.437	71.192	4.802	49.852	2508.301

Table 12: **Lagrange interpolation** under **square-root** interference with t stages.

	Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE
$t=1$	N/A	N/A	N/A	N/A	N/A	N/A
$t=2$	-0.131	0.015	0.017	-0.079	0.331	0.116
$t=3$	-0.387	0.032	0.150	-0.282	0.685	0.549
$t=4$	1.276	0.366	1.761	1.787	5.708	35.769
$t=5$	-18.021	7.106	375.254	-8.481	88.365	7880.220

where $\tilde{c}_{ij}$ characterizes the influence of unit j on unit i, and w_ℓ is the coefficient of the ℓ-th order term.

We note that Lagrange interpolation estimator is a **graph-agnostic** estimator that summarizes the experimental data at each timestep into a single value—mean outcomes. Thus, it heavily relies on the exact knowledge of β and incurs misspecification risk for both potential outcome model and **number of timesteps** (note that the order of polynomial in Lagrange interpolation is often exactly $T-1$ due to random error). We report its performance on two settings: multiplicative interference in Table 11 and square-root interference in Table 12.

In the multiplicative setting, we observe that Lagrange interpolation performs very well for the case of $t=2$, while it rapidly deteriorates with the increase of t, which is counterintuitive but can be explained by our summary: **it heavily relies on the exact knowledge on** β, which is usually inaccessible in practice. Moreover, although the theoretical derivation of both our paper and [7] are based on unit-level randomization, the functional form of Lagrange interpolation estimator there substantially relies on the randomization, while our idea is generalizable (through the lens of increasing exposure variation), i.e., improving the data for learning representation. Therefore, we observe a general synergistic effect with cluster-level randomization and our merging methodology, while Lagrange interpolation estimator performs even much worse on experimental data generated with cluster-level randomization.

In the square-root setting (the performance of GNN-based estimator with our methodology is provided in Table 6), which corresponds a scenario of diminishing marginal utility, our methodology benefits a lot from merging more steps (left part of Table 6), while

We summarize that our GNN-based approach learns a flexible node-level representation from all units directly, and can naturally accommodate richer microstructure, nodal covariates, and more complex interaction patterns.

Network-aware estimators without regression component. We next provide the performance of two network-aware estimators without regression component for comparison, including MII [5] and Hájek estimator [27], in Table 13. We empirically find that naively averaging the estimates acquired on different timesteps t brings no improvement, so we only report the performance with treatment proportion $t=50\%$. This table is provided only for comparison.

Table 13: Performance of MII and Hájek estimators under 2-hop interference.

	Unit			Cluster		
Estimator	Bias	Std	MSE	Bias	Std	MSE
MII	N/A	N/A	N/A	-0.276	0.010	0.077
Hájek	-0.471	0.043	0.223	-0.263	0.039	0.071

Table 14: **2-hop** interference on **MIT8**: (left) merge t stages; (right) merge $c = 0$ and c_t.

	Unit			Cluster				Unit			Cluster		
Rounds	Bias	Std	MSE	Bias	Std	MSE	**Proportion**	Bias	Std	MSE	Bias	Std	MSE
$t = 1$	-1.439	0.023	2.072	-1.022	0.025	1.044	$c_t = 0.02$	-0.880	0.052	0.776	-0.467	0.256	0.284
$t = 2$	-0.729	0.018	0.531	-0.411	0.052	0.172	$c_t = 0.05$	-0.633	0.052	0.403	-0.246	0.119	0.075
$t = 3$	-0.351	0.018	0.123	-0.170	0.033	0.030	$c_t = 0.10$	-0.545	0.032	0.298	-0.192	0.055	0.040
$t = 4$	-0.208	0.019	0.044	-0.101	0.031	0.011	$c_t = 0.25$	-0.292	0.031	0.086	-0.088	0.051	0.010
$t = 5$	-0.177	0.025	0.032	-0.086	0.039	0.009	$c_t = 0.50$	-0.184	0.019	0.034	-0.091	0.022	0.009

Another network topology. For completeness, we supplement the performance of GNN-based estimator on MIT8[2] [23]. We examine the case of 2-hop interference, since it is concerned with the concrete topology. We report the results in Table 14, and the results demonstrate that our methodology remains effective.

C Proofs

C.1 Proof of Theorem 3.1

Proof. First, our simple linear regression estimator gives:

$$\hat{\tau} = (\hat{\theta}_1 + \hat{\theta}_0) - \hat{\theta}_0 = \hat{\theta}_1, \tag{31}$$

and

$$\hat{\theta} = (X(\mathbf{z})^\top X(\mathbf{z}))^{-1} X(z)^\top Y, \tag{32}$$

For the simplicity in the derivation, we write $X = X(\mathbf{z})$ in the following. We recap that $X(\mathbf{z}) = (\mathbf{1} \quad \mathbf{z})$. Since we consider complete randomization, there are n units and we assign d treatments exactly. We have:

$$X^\top X = \begin{pmatrix} n & \mathbf{z}^\top \mathbf{1} \\ \mathbf{z}^\top \mathbf{1} & \|\mathbf{z}\|_2^2 \end{pmatrix} = \begin{pmatrix} n & d \\ d & d \end{pmatrix}, \tag{33}$$

We remark that, for precise analysis on the bias and variance of linear regression estimator, it is critical to remove the randomness in $\det\left(X(\mathbf{z})^T X(\mathbf{z})\right)$.

Naturally,

$$(X^\top X)^{-1} X^\top = \frac{1}{d(n-d)} \begin{pmatrix} d(\mathbf{1} - \mathbf{z})^\top \\ (n\mathbf{z} - d\mathbf{1})^\top \end{pmatrix}, \tag{34}$$

Notice that Y can be written as

$$Y = (X \quad B\mathbf{z})\, \beta, \tag{35}$$

Here, $\beta = (\beta_0 \quad \beta_1 \quad 1)$. The shape of X is $n \times 2$, $B\mathbf{z}$ is $n \times 1$, and β is 3×1.

The expression of $\hat{\theta}$ is given by

$$\begin{aligned} \hat{\theta} &= (X^\top X)^{-1} X^\top Y \\ &= \left(I_2 \quad (X^\top X)^{-1} X^\top B\mathbf{z}\right) \beta + (X^\top X)^{-1} X^\top \epsilon \\ &= \left(I_2 \quad \tfrac{1}{d(n-d)} \begin{pmatrix} d(\mathbf{1} - \mathbf{z})^\top B\mathbf{z} \\ (n\mathbf{z} - d\mathbf{1})^\top B\mathbf{z} \end{pmatrix}\right) + (X^\top X)^{-1} X^\top \epsilon. \end{aligned} \tag{36}$$

Since we're only concerned with $\hat{\theta}_1$, we directly take the second term of $\hat{\theta}$:

$$\hat{\theta}_1 = \beta_1 + \frac{1}{d(n-d)} (n\mathbf{z} - d\mathbf{1})^\top B\mathbf{z} + \frac{1}{d(n-d)} (n\mathbf{z} - d\mathbf{1})^\top \epsilon, \tag{37}$$

Notice that $\mathbf{z} \perp \epsilon$. The **bias** is given by

$$\mathbb{E}[\hat{\theta}_1] - (\beta_1 + \frac{1}{n} \operatorname{trace}(\mathbf{1}\mathbf{1}^\top B)) = \frac{1}{nc(1-c)} \mathbb{E}[(\mathbf{z} - c\mathbf{1})^\top B\mathbf{z}] - \frac{1}{n} \operatorname{trace}(\mathbf{1}\mathbf{1}^\top B), \tag{38}$$

[2]The network topology is available at https://networkrepository.com/socfb-MIT8.php.

Next, we evaluate the term $\mathbb{E}[(\mathbf{z}-c\mathbf{1})^\top B\mathbf{z}]$, which is given by:

$$\begin{aligned}\mathbb{E}[(\mathbf{z}-c\mathbf{1})^\top B\mathbf{z}] &= \mathbb{E}\left[\operatorname{trace}\left((\mathbf{z}-c\mathbf{1})^\top B\mathbf{z}\right)\right] \\ &= \mathbb{E}\left[\operatorname{trace}\left(\mathbf{z}(\mathbf{z}-c\mathbf{1})^\top B\right)\right] \\ &= \operatorname{trace}\left(\mathbb{E}\left[\mathbf{z}(\mathbf{z}-c\mathbf{1})^\top\right] B\right) \\ &= \operatorname{trace}\left((c\frac{d-1}{n-1}-c^2)\mathbf{1}\mathbf{1}^\top B\right) \\ &= -\frac{c(1-c)}{n-1}\operatorname{trace}(\mathbf{1}\mathbf{1}^\top B).\end{aligned} \tag{39}$$

Plugging it into the expression of bias, we get:

$$\text{Bias}(\hat{\tau}_{(1)}) = -\frac{\sum_{i,j} B_{ij}}{n}\left(1+\frac{1}{n-1}\right) = -\frac{\sum_{i,j} B_{ij}}{n-1}. \tag{40}$$

Subsequently, we derive the expression of variance. Noting that $\mathbf{z} \perp \epsilon$ and $\mathbb{E}[\epsilon] = \mathbf{0}$, we have:

$$\operatorname{Cov}[(n\mathbf{z}-d\mathbf{1})^\top B\mathbf{z}, (n\mathbf{z}-d\mathbf{1})^\top \epsilon] = 0, \tag{41}$$

Thus, we can omit all the covariance terms in the expression of variance, yielding:

$$\operatorname{Var}(\hat{\tau}) = (\frac{1}{nc(1-c)})^2 \operatorname{Var}\left((\mathbf{z}-c\mathbf{1})^\top B\mathbf{z}\right) + \frac{1}{(nc(1-c))^2}\mathbb{E}[(\mathbf{z}-c\mathbf{1})^\top(\mathbf{z}-c\mathbf{1})]\sigma_e^2, \tag{42}$$

First, we consider the second term above, which represents the variance induced by ϵ:

$$\begin{aligned}\operatorname{Var}_e &= \frac{1}{(d(n-d))^2}\mathbb{E}[(n\mathbf{z}-d\mathbf{1})^\top(n\mathbf{z}-d\mathbf{1})]\sigma_e^2 \\ &= \left(\frac{n}{d(n-d)}\right)^2 \mathbb{E}[(\mathbf{z}-\frac{d}{n}\mathbf{1})^\top(\mathbf{z}-\frac{d}{n}\mathbf{1})]\sigma_e^2 \\ &= \left(\frac{n\sigma_e}{d(n-d)}\right)^2 \mathbb{E}[d-\frac{2d}{n}\times d+\frac{d^2}{n}] \\ &= \left(\frac{n\sigma_e}{d(n-d)}\right)^2 (d-\frac{d^2}{n}) \\ &= \frac{\sigma_e^2}{nc(1-c)}.\end{aligned} \tag{43}$$

Next, we discuss the first term:

$$\left(\frac{1}{nc(1-c)}\right)^2 \operatorname{Var}\left((\mathbf{z}-c\mathbf{1})^\top B\mathbf{z}\right), \tag{44}$$

We use the expansion $\operatorname{Var}(X) = \mathbb{E}[X^2] - E[X]^2$, and the square of expectation term has been simplified in equation (39). Thus, our focus is the expectation of square term, which can be expanded with brute-force:

$$\left(\frac{1}{c(1-c)n}\right)^2 \sum_{i,j,k,l} B_{ij}B_{kl}\mathbb{E}[(z_i-c)(z_k-c)z_jz_l], \tag{45}$$

As mentioned in the main paper, our technique involves classifying the index set $I \subset [n]^4$ that restricts the quadruples (i,j,k,l) and calculating $\mathbb{E}[(z_i-c)(z_k-c)z_jz_l]$ respectively. Since $B_{ii} = 0$ for all $i \in [n]$, we do not need to consider the cases where $i = j$ or $k = l$. We present the result as follows:

1. $i = k, j = l, i \neq k$,

$$\begin{aligned}\mathbb{E}[(z_i - c)(z_k - c)z_j z_l] &= \mathbb{E}[(z_i - c)^2 z_j^2] \\ &= \mathbb{E}[(z_i - 2cz_i + c^2)z_j] \\ &= (1 - 2c)\mathbb{E}[z_i z_j] + c^2\mathbb{E}[z_j] \\ &= (1 - 2c)c\frac{d-1}{n-1} + c^3 \\ &= c^2(1 - c) + O\left(\frac{1}{n}\right).\end{aligned} \tag{46}$$

2. $i = l, k = j, i \neq k$

$$\begin{aligned}\mathbb{E}[(z_i - c)(z_k - c)z_j z_l] &= \mathbb{E}[(z_i - c)(z_k - c)z_i z_k] \\ &= \mathbb{E}[(1 - c)^2 z_i z_k] \\ &= (1 - c)^2 c\frac{d-1}{n-1} \\ &= (1 - c)^2 c^2 + O\left(\frac{1}{n}\right).\end{aligned} \tag{47}$$

3. Only $i = k$ holds, with i, j, and l being distinct.

$$\begin{aligned}\mathbb{E}[(z_i - c)(z_k - c)z_j z_l] &= \mathbb{E}[(z_i - c)^2 z_k z_l] \\ &= (1 - 2c)\mathbb{E}[z_i z_k z_l] + c^2\mathbb{E}[z_k z_l] \\ &= (1 - 2c)c\frac{(d-1)(d-2)}{(n-1)(n-2)} + c^3\frac{d-1}{n-1} \\ &= c^3(1 - c) + O\left(\frac{1}{n}\right).\end{aligned} \tag{48}$$

4. Only $j = l$ holds, with i, j, and k being distinct.

$$\begin{aligned}\mathbb{E}[(z_i - c)(z_k - c)z_j z_l] &= \mathbb{E}[(z_i - c)(z_k - c)z_j] \\ &= \frac{nd(d-n) + 2d^2 + 2n^2 - 4nd}{n^2(n-1)(n-2)} \\ &= \frac{c(d-n) + 2c^2 + 2 - 4c}{(n-1)(n-2)} \\ &= \frac{c(c-1)}{n} + O(\frac{1}{n^2}).\end{aligned} \tag{49}$$

5. i, j, k, l are all distinct.

$$\begin{aligned}&\mathbb{E}[(z_i - c)(z_k - c)z_j z_l] \\ &= \mathbb{E}[(z_i z_k - c(z_i + z_k) + c^2)z_j z_l] \\ &= \frac{d(d-1)(d-2)(d-3)}{n(n-1)(n-2)(n-3)} - 2c\frac{d(d-1)(d-2)}{n(n-1)(n-2)} + c^2\frac{d(d-1)}{n(n-1)} \\ &= \frac{d(d-1)}{n(n-1)}\left(\frac{(d-2)(d-3)}{(n-2)(n-3)} + \frac{d^2}{n^2} - 2\frac{d(d-2)}{n(n-2)}\right) \\ &= \frac{d(d-1)}{n(n-1)}\left(\frac{(d^2-5d+6)n^2 + d^2(n^2-5d+6) - 2dn(dn-2n-3d+6)}{n^2(n-2)(n-3)}\right) \\ &= \frac{d(d-1)}{n(n-1)}\left(\frac{(d^2-5d+6)n^2 + d^2(n^2-5d+6) - 2dn(dn-2n-3d+6)}{n^2(n-2)(n-3)}\right) \\ &= \frac{c(d-1)}{(n-1)}\left(\frac{c(d-n) + 6 + 6c^2 - 12c}{(n-2)(n-3)}\right) \\ &= \frac{c^3(c-1)}{n} + O(\frac{1}{n^2}).\end{aligned} \tag{50}$$

Combine these five cases and corresponding index sets, we get the expression of variance as presented in Theorem 1. For the convenience of proof of Corollary 3.3 later, we also annotate the order of each term at last line. □

C.2 Proof of Corollary 3.3

Proof. Since Assumption 3.2 restricts

$$\sum_{i,j} B_{ij} = \Theta(n), \tag{51}$$

We get the order of bias instantly:

$$\text{Bias}(\hat{\tau}_{(1)}) = -\frac{\sum_{i,j} B_{ij}}{n-1} = \Theta(1). \tag{52}$$

Next, we discuss the order of variance. It concerns bounding the order of the summations over five index sets mentioned in Appendix C.1. To begin with, we notice that the variance is non-negative and the variance induced by ϵ is:

$$\frac{\sigma_e^2}{nc(1-c)} = \Theta(\frac{1}{n}), \tag{53}$$

Thus, we can first claim that

$$\text{Var}(\hat{\tau}) = \Omega(\frac{1}{n}), \tag{54}$$

We then prove the first term in the expression of variance has a order of $O(1/n)$. Specifically, we bound those five summations respectively. According to the Assumption 3.2, we get a naive bound that will be used repeatedly:

$$|B_{ij}| \leq \sum_j |B_{ij}| \leq b. \tag{55}$$

1. $i = k, j = l, i \neq k$.

$$\sum_{ij} B_{ij}^2 \leq \sum_i b \sum_j |B_{ij}| \leq b^2 n, \tag{56}$$

   Thus,

$$\sum_{ij} B_{ij}^2 = O(n). \tag{57}$$

2. $i = l, k = j, i \neq k$

$$|\sum_{i,j} B_{ij} B_{ji}| \leq \sum_{i,j} |B_{ij} B_{ji}| \leq b \sum_{i,j} B_{ij} \leq b^2 n, \tag{58}$$

   Thus,

$$\sum_{ij} B_{ij} B_{ji} = O(n). \tag{59}$$

3. Only $i = k$ holds, with i, j, and l being distinct.

$$|\sum_{i,j,l} B_{ij} B_{il}| \leq \sum_{i,j,l} |B_{ij} B_{il}| = \sum_{i,j} |B_{ij}| \sum_l |B_{il}| \leq b \sum_{i,j} |B_{ij}| \leq b^2 n, \tag{60}$$

   Thus,

$$\sum_{i,j,l} B_{ij} B_{il} = O(n). \tag{61}$$

4. Only $j = l$ holds, with i, j, and k being distinct.

$$|\sum_{i,j,k} B_{ij} B_{kj}| \leq \sum_{i,j,k} |B_{ij} B_{kj}| \leq b \sum_{i,j,k} |B_{ij}| \leq b^2 n^2, \tag{62}$$

   Thus,

$$\sum_{i,j,k} B_{ij} B_{kj} = O(n^2). \tag{63}$$

5. i, j, k, l are all distinct.

$$\sum_{i,j,k,l} B_{ij}B_{kl} = (\sum_{ij} B_{ij})^2 = \Theta(n^2). \tag{64}$$

Combine the orders of five summations above and the order of corresponding $\mathbb{E}[(z_i - c)(z_k - c)z_j z_l]$, we get:

$$\mathrm{Var}(\hat{\tau}) = \Theta(\frac{1}{n}), \tag{65}$$

Finally, we comment that the bound on $|\sum_{i,j,k} B_{ij}B_{kj}|$ is rough, while we do not apply the third part of Assumption 3.2, i.e. restricting the order $\sum_{i,j,k} B_{ij}B_{kj} = \sum_j (\sum_i B_{ij})^2$ directly.

Such a rough bound is still acceptable because of the equation (49), i.e.

$$\mathbb{E}[(z_i - c)(z_k - c)z_j z_l] = \Theta(\frac{1}{n}), \tag{66}$$

For the situation of merging $T \geq 2$ experimental data, if $\frac{\sum_{t=1}^T c_t}{T} \neq \mathbb{E}[z_i]$, we would have:

$$\mathbb{E}[(z_i - \frac{\sum_{t=1}^T c_t}{T})(z_k - \frac{\sum_{t=1}^T c_t}{T})z_j z_l] = \Theta(1), \tag{67}$$

This explains why we need a further assumption to control the order of variance there. □

C.3 Proof of Theorem 3.5

Proof. In the situation of $T = 2$, the sample size increases from n to $2n$, and the number of treated units changes from d to $d_1 + d_2$. Correspondingly, the treatment proportions in two experiments are c_1 and c_2. In linear regression, this additionally leads to the following two changes:

$$B \to \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix} \qquad \mathbf{z} \to \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}. \tag{68}$$

We make the changes above and get

$$\begin{aligned}
&\mathrm{Bias}(\hat{\tau}_{(2)})/(\sum_{i,j} B_{ij}) + 1 \\
=& \frac{2n}{(d_1+d_2)(2n-(d_1+d_2))} \mathrm{trace}\left(\mathbb{E}\left[\mathbf{z}(\mathbf{z} - \frac{d_1+d_2}{2n}\mathbf{1})^\top\right]\begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix}\right) \\
=& \frac{2n}{(d_1+d_2)(2n-(d_1+d_2))} \mathrm{trace}\left(\left(\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right] - \frac{d_1+d_2}{2n}\mathbb{E}\left[\mathbf{z}\mathbf{1}^\top\right]\right)\begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix}\right) \\
=& \frac{2n}{(d_1+d_2)(2n-(d_1+d_2))} \begin{pmatrix} \mathrm{trace}\left((\mathbb{E}[\mathbf{z}_1\mathbf{z}_1^\top] + \mathbb{E}[\mathbf{z}_2\mathbf{z}_2^\top])B\right) \\ -\frac{d_1+d_2}{2n}(\frac{d_1}{n} + \frac{d_2}{n})\,\mathrm{trace}\left(\mathbf{1}\mathbf{1}^\top B\right) \end{pmatrix} \\
=& \frac{2n}{(d_1+d_2)(2n-(d_1+d_2))} \left(\frac{d_1(d_1-1)+d_2(d_2-1)}{n(n-1)} - \frac{(d_1+d_2)^2}{2n^2}\right)\mathrm{trace}\left(\mathbf{1}\mathbf{1}^\top B\right) \\
=& \frac{2n^2}{(d_1+d_2)(2n-(d_1+d_2))} \frac{(n-1)(d_1-d_2)^2 + 2(d_1^2+d_2^2) - 2n(d_1+d_2)}{2n^2(n-1)} \\
=& \frac{(n-1)(d_1-d_2)^2 + 2(d_1^2+d_2^2) - 2n(d_1+d_2)}{(n-1)(d_1+d_2)(2n-(d_1+d_2))} \\
=& \frac{(n-1)\left(c_1-c_2\right)^2 + 2\left(c_1^2+c_2^2\right) - 2\left(c_1+c_2\right)}{(n-1)\left(c_1+c_2\right)\left(2-(c_1+c_2)\right)}.
\end{aligned} \tag{69}$$

Thus, we get:

$$\mathrm{Bias}(\hat{\tau}_{(2)}) = -\frac{\sum_{ij} B_{ij}}{n}\left(1 - \frac{(n-1)\left(c_1-c_2\right)^2 + 2\left(c_1^2+c_2^2\right) - 2\left(c_1+c_2\right)}{(n-1)\left(c_1+c_2\right)\left(2-(c_1+c_2)\right)}\right). \tag{70}$$

C.4 Proof of Theorem 3.6

Proof. As mentioned in the main paper, there is no additional challenge in deriving the result of $T > 2$ case. Refer to the fourth equation in equation (69), we now have:

$$\begin{aligned}\text{Bias}(\hat{\tau}_{(T)})/(\sum_{i,j} B_{ij}) + 1 &= \frac{Tn}{\sum_{t=1}^{T} d_t(Tn - \sum_{t=1}^{T} d_t)} \left(\frac{\sum_{t=1}^{T} d_t(d_t-1)}{n(n-1)} - \frac{(\sum_{t=1}^{T} d_t)^2}{Tn^2} \right) \\ &= \frac{T}{n(\sum_{t=1}^{T} c_t)(T - \sum_{t=1}^{T} c_t)} \left(\frac{T\sum_{t=1}^{T} c_t^2 - (\sum_{t=1}^{T} c_t)^2}{T} + O\left(\frac{1}{n}\right) \right).\end{aligned} \tag{71}$$

Thus, since $\sum_{i,j} B_{ij} = \Theta(n)$, we have:

$$\text{Bias}(\hat{\tau}_{(T)}) = -\frac{\sum_{ij} B_{ij}}{n} \left(1 - \frac{T\sum_{t=1}^{T} c_t^2 - (\sum_{t=1}^{T} c_t)^2}{(\sum_{t=1}^{T} c_t)(T - \sum_{t=1}^{T} c_t)} \right) + O\left(\frac{1}{n}\right). \tag{72}$$

C.5 Proof of Corollary A.1

Proof. We continue to use the notation in the main paper, and further bias reduction corresponds to:

$$\frac{kC_2 - C_1^2}{C_1(k - C_1)} < \frac{(k+1)(C_2 + x^2) - (C_1 + x)^2}{(C_1 + x)(k + 1 - C_1 - x)}, \tag{73}$$

We expand the two sides of this inequality with brute-force, and get following condition:

$$\begin{aligned}&k(k+1)C_1C_2 + k(k+1)C_1x^2 - (k+1)C_1^2C_2 - (k+1)C_1^2x^2 \\ &- kC_1^3 - 2kC_1^2x - kC_1x^2 + C_1^4 + 2C_1^3x + C_1^2x^2 \\ &> k(k+1)C_1C_2 + k(k+1)C_2x - kC_1^2C_2 - 2kC_1C_2x \\ &- kC_2x^2 - (k+1)C_1^3 + C_1^4 + C_1^3x - (k+1)C_1^2x + C_1^3x + C_1^2x^2,\end{aligned} \tag{74}$$

which is equivalent to:

$$\left(k^2C_1 - (k+1)C_1^2 + kC_2\right)x^2 - \left((k-1)C_1^2 + 2kC_1C_2 + k(k+1)C_2\right)x + C_1^3 - C_1^2C_2 > 0. \tag{75}$$

□

C.6 Lemma C.1 for Variance Analysis with Proof

We construct the following lemma for the convenience of variance analysis:

Lemma C.1. *Given Assumption 3.2, for a treatment vector* $\mathbf{z}$ *with treatment proportion* $c \in (0,1)$ *and any constant* $w \in (0,1)$*, we have:*

$$\text{Var}\left((\mathbf{z} - w\mathbf{1})^\top B\mathbf{z}\right) = O\left(\frac{1}{n}\right). \tag{76}$$

Proof. First, we will continue to apply the idea of expansion in the proof of variance part of Theorem 1. However, the proof in the Theorem 1 can't apply directly because the order of $\mathbb{E}[(z_i - c)(z_k - c)z_jz_l]$ increases from $\Theta(1/n)$ to $\Theta(1)$, when $w \neq c$. These two cases are:

1. Only $j = l$ holds, with i, j, and k being distinct.
2. i, j, k, l are all distinct.

For the first case, we resort to the third part of Assumption 3.2:

$$\sum_{i,j,k} B_{ij}B_{kj} = \sum_{j}(\sum_{i} B_{ij})^2 = O(n), \tag{77}$$

For the second case, we will prove that the two coefficients of term $(\sum_{ij} B_{ij})^2$ in the expression of $\mathrm{Var}((\mathbf{z}-w\mathbf{1})^\top B\mathbf{z})$ can cancel out. Namely, we will prove:

$$\mathbb{E}[(z_i-w)(z_k-w)z_jz_l] - \left(\mathbb{E}[(\mathbf{z}-w\mathbf{1})^\top B\mathbf{z}]/\operatorname{trace}(\mathbf{1}\mathbf{1}^\top B)\right)^2 = O\left(\frac{1}{n}\right), \tag{78}$$

for the case i, j, k, l are not equal to each other.

Notice that, for $s = 1, 2, 3$, we have:

$$\frac{d-s}{n-s} = c + O\left(\frac{1}{n}\right), \tag{79}$$

On one hand,

$$\begin{aligned}\mathbb{E}[(z_i-w)(z_k-w)z_jz_l] &= \mathbb{E}[(z_i-w)(z_k-w) \mid z_j = z_l = 1]\left(c^2 + O\left(\frac{1}{n}\right)\right)\\ &= (c-w)^2c^2 + O\left(\frac{1}{n}\right).\end{aligned} \tag{80}$$

On the other hand,

$$\begin{aligned}\frac{\mathbb{E}[(\mathbf{z}-w\mathbf{1})^\top B\mathbf{z}]}{\operatorname{trace}(\mathbf{1}\mathbf{1}^\top B)} &= \frac{\operatorname{trace}\left(\mathbb{E}\left[\mathbf{z}(\mathbf{z}-w\mathbf{1})^\top\right]B\right)}{\operatorname{trace}(\mathbf{1}\mathbf{1}^\top B)}\\ &= \frac{\operatorname{trace}\left((\mathbb{E}[\mathbf{z}\mathbf{z}^\top] - cw\mathbf{1}\mathbf{1}^\top)B\right)}{\operatorname{trace}(\mathbf{1}\mathbf{1}^\top B)}\\ &= c(c-w) + O\left(\frac{1}{n}\right).\end{aligned} \tag{81}$$

In the last equation, we again utilize $B_{ii} = 0$ for all unit $i \in [n]$. Combine two equations above, we successfully control the order of coefficient of term $(\sum_{ij} B_{ij})^2$, in the expression of concerned variance $\mathrm{Var}((\mathbf{z}-w\mathbf{1})^\top B\mathbf{z})$.

For other three cases of $\mathbb{E}[(z_i-c)(z_k-c)z_jz_l]$, the proof of Theorem 1 can be utilized directly here, because $\mathbb{E}[(z_i-c)(z_k-c)z_jz_l] = \Theta(1)$ there, and the order of this term can't exceed constant no matter what value w takes. Therefore, we finish the proof. □

C.7 Proof of Theorem 3.7

Proof. We first present the proof for the case of **temporally independent** T experiments is a direct extension of Lemma C.1. In fact, with Lemma C.1, we can conclude that

$$\mathrm{Var}\left(\sum_{t=1}^{T}\left(\mathbf{z}_t - \frac{\sum_{k=1}^{\top} c_k}{T}\mathbf{1}\right)^\top B\mathbf{z}_t\right) = \sum_{t=1}^{T}\mathrm{Var}\left(\left(\mathbf{z}_t - \frac{\sum_{k=1}^{\top} c_k}{T}\mathbf{1}\right)^\top B\mathbf{z}_t\right) = O\left(\frac{1}{n}\right), \tag{82}$$

On the other hand, the order of variance introduced by ϵ is always $\Theta(1/n)$, and the concerned variance is the sum of these non-negative variance terms. Therefore, we can conclude that

$$\mathrm{Var}(\hat{\tau}) = \Theta\left(\frac{1}{n}\right), \tag{83}$$

Next, we study the case of **staggered rollout** experiments. The complexity here is strong temporal covariance of treatments.

Based on the Lemma C.1, the remaining work is bounding the sum of covariance terms. We have:

$$\begin{aligned}\mathrm{Var}\left(\sum_{t=1}^{T}(\mathbf{z}_t - \frac{\sum_{k=1}^{\top} c_k}{T}\mathbf{1})^\top B\mathbf{z}_t\right) &= \sum_{t=1}^{T}\mathrm{Var}\left((\mathbf{z}_t - \frac{\sum_{k=1}^{\top} c_k}{T}\mathbf{1})^\top B\mathbf{z}_t\right)\\ &+2\sum_{t_1<t_2}\mathrm{Cov}[(\mathbf{z}_{t_1} - \frac{\sum_{k=1}^{\top} c_k}{T}\mathbf{1})^\top B\mathbf{z}_{t_1}, (\mathbf{z}_{t_2} - \frac{\sum_{k=1}^{\top} c_k}{T}\mathbf{1})^\top B\mathbf{z}_{t_2}].\end{aligned} \tag{84}$$

For simplicity, we use $\mathbf{u}, \mathbf{v}$ to denote $\mathbf{z}_{t_1}, \mathbf{z}_{t_2}$, respectively, and use $w = \sum_{t=1}^{T} c_t/T$. We continue to expand the covariance term, and get:

$$\sum_{i,j,k,l} B_{ij}B_{kl} \left(\mathbb{E}\left[(u_i - w)\, u_j\, (v_k - w)\, v_l \right] - \mathbb{E}\left[(u_i - w)\, u_j \right] \mathbb{E}\left[(v_k - w)\, v_l \right] \right), \tag{85}$$

Correspondingly, we need to discuss the order of the this term in five cases, as well as that in the Appendix C.1. Similarly, for the first four cases (except for the case where i, j, k, l are not equal to each other), we allow for:

$$\mathbb{E}\left[(u_i - w)\, u_j\, (v_k - w)\, v_l \right] - \mathbb{E}\left[(u_i - w)\, u_j \right] \mathbb{E}\left[(v_k - w)\, v_l \right] = \Theta(1), \tag{86}$$

Therefore, we only need to prove:

$$\mathbb{E}\left[(u_i - w)\, u_j\, (v_k - w)\, v_l \right] - \mathbb{E}\left[(u_i - w)\, u_j \right] \mathbb{E}\left[(v_k - w)\, v_l \right] = O\left(\frac{1}{n} \right), \tag{87}$$

when i, j, k, l are not equal to each other.

Since $\mathbf{u}$ is the treatment vector that is assigned before $\mathbf{v}$, we should first condition on u_i, u_j .

$$\begin{aligned}
\mathbb{E}\left[(u_i - w) u_j (v_k - w) v_l \right] &= \left(c_{t_1}^2 + O\left(\frac{1}{n} \right) \right) (1 - w) \mathbb{E}[(v_k - w) v_l \mid u_i = u_j = 1] \\
&\quad - \left(c_{t_1}(1 - c_{t_1}) + O\left(\frac{1}{n} \right) \right) w \mathbb{E}[(v_k - w) v_l \mid u_j = 1] \\
&= (1 - w) c_{t_1}^2 (c_{t_2} - w) c_{t_2} - w c_{t_1}(1 - c_{t_1})(c_{t_2} - w) c_{t_2} + O\left(\frac{1}{n} \right) \\
&= (c_{t_1} - w) c_{t_1} (c_{t_2} - w) c_{t_2} + O\left(\frac{1}{n} \right).
\end{aligned} \tag{88}$$

On the other hand, it is easy to check that

$$\mathbb{E}\left[(u_i - w)\, u_j \right] \mathbb{E}\left[(v_k - w)\, v_l \right] = (c_{t_1} - w) c_{t_1} (c_{t_2} - w) c_{t_2} + O\left(\frac{1}{n} \right), \tag{89}$$

Thus, we finish the proof. □

Remark C.2. In fact, the covariance term in other four cases are likely of constant order. For example, we check the case where only $j = l$ holds:

$$\begin{aligned}
\mathbb{E}\left[(u_i - w) u_j (v_k - w) v_l \right] &= \mathbb{E}\left[(u_i - w) u_j (v_k - w) v_j \right] \\
&= c_{t_1} \mathbb{E}\left[(u_i - w)(v_k - w) v_j \mid u_j = 1 \right] \\
&= c_{t_1} \left(c_{t_1} - w + O\left(\frac{1}{n} \right) \right) \left(c_{t_2} - w + O\left(\frac{1}{n} \right) \right) \\
&= c_{t_1} (c_{t_1} - w)(c_{t_2} - w) + O\left(\frac{1}{n} \right).
\end{aligned} \tag{90}$$

Therefore, in this case, the value of covariance term

$$\begin{aligned}
&\mathbb{E}\left[(u_i - w)\, u_j\, (v_k - w)\, v_l \right] - \mathbb{E}\left[(u_i - w)\, u_j \right] \mathbb{E}\left[(v_k - w)\, v_l \right] \\
&= c_{t_1}(c_{t_1} - w)(c_{t_2} - w) - c_{t_1} c_{t_2} (c_{t_1} - w)(c_{t_2} - w) + O\left(\frac{1}{n} \right) \\
&= \Theta(1),
\end{aligned} \tag{91}$$

if $c_{t_1} > 0$ and $c_{t_2} < 1$. Therefore, the third part of Assumption 3.2 is still necessary here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions, assumptions, and scope, and these are supported by the theory and experiments in Sections 1–5 and the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses key limitations and scope restrictions, including the focus on static GATE, suppression of temporal interference, and the simplifying assumptions behind the exact theory; see Sections 3–6 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results are stated with explicit assumptions and cross-referenced proofs, with detailed derivations deferred to the appendix and supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the experimental design, datasets, models, and evaluation protocol, and we will further ensure reproducibility by providing runnable code in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will provide anonymized code and instructions in the supplementary material so that the main experimental results can be faithfully reproduced.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the network data, interference models, merging schemes, randomization schemes, and main model choices in Section 5 and Appendix B, and the supplementary code will include full training details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments report variability information through standard deviations, and the paper explicitly states that the bias and standard deviation are computed from Monte Carlo repetitions; see Section 5.2 and the result tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Computational cost is not a central concern of this paper, whose main focus is the statistical properties of regression-based estimators under network interference. The experiments use lightweight models and standard simulation pipelines, and we will provide implementation details in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the research complies with the NeurIPS Code of Ethics and does not involve unethical data collection or human-subject procedures.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The current submission does not include a dedicated discussion of broader societal impacts. This is a methodological paper on experimental design and estimation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk models or scraped datasets, so special safeguards for responsible release are not applicable here.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: Existing assets such as the FB-Stanford3 network and software packages are credited by citation, but the manuscript does not yet explicitly document licenses and terms of use. We will add this information in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release code for the proposed method and experiments, together with documentation and instructions in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve human subjects, so IRB approval or an equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important or non-standard component of the core methodology in this research; any LLM use was limited to writing assistance only, if any.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.