Analysis of Two-Stage Rollout Designs with Clustering for Causal Inference under Network Interference

Mayleen Cortez-Rodriguez
Cornell University

Matthew Eichhorn
Cornell University

Christina Lee Yu
Cornell University

Abstract

Estimating causal effects under interference is pertinent to many real-world settings. Recent work with low-order potential outcomes models uses a rollout design to obtain unbiased estimators that require no interference network information. However, the required extrapolation can lead to prohibitively high variance. To address this, we propose a two-stage experiment that selects a subpopulation in the first stage and restricts treatment rollout to this sub-population in the second stage. We explore the role of clustering in the first stage by analyzing the bias and variance of a polynomial interpolation-style estimator under this experimental design. Bias increases with the number of edges cut in the clustering of the interference network, but variance depends on qualities of the clustering that relate to homophily and covariate balance. There is a tension between clustering objectives that minimize the number of cut edges versus those that maximize covariate balance across clusters. Through simulations, we explore a bias-variance tradeoff and compare the performance of the estimator under different clustering strategies.

1 Introduction

The stable unit treatment value assumption (SUTVA) is critical for many classic causal inference methods, but is violated in settings with interference, where the outcome of an individual can be affected by the treatment assignment of another. Interference introduces bias into estimators, potentially leading to inaccurate conclusions about causal effects when ignored (Sobel, 2006). Many domains experience interference. In evaluating the effect of a public health intervention such as a vaccine, peer effects from herd immunity play a role (Hudgens and Halloran, 2008). In evaluating the effect of a new feature on a social media platform, a user's engagement is affected by the behaviors of their social connections, whether they are directly exposed to the new feature or not (Biswas and Airoldi, 2018; Aral and Walker, 2012).

We exploit both a rich two-stage randomized experimental design and a flexible potential outcomes model to estimate the *total treatment effect* (TTE), the difference in average outcomes of the population when everyone is treated versus untreated. The TTE estimand, sometimes called the *global average treatment effect*, is a natural choice in applications where the decision-maker needs to decide between adopting the new intervention for everyone or sticking with the status quo, such as a tech company choosing a single user-feed content recommendation algorithm for their social media platform. The class of experimental designs we consider in this paper are called *staggered rollout designs*, where treatment is assigned over different time periods to increasing subsets of participants until it has been rolled out to all subjects designated for treatment. This style of experiment is common on online platforms (Xu et al., 2018) to mitigate the possible, unknown risks related to introducing a new feature, and in medicine (Brown and Lilford, 2006), where for logistical or financial reasons it may be impossible to deliver treatment to all participants at once.

Related Work. Many prior approaches for causal inference under interference consider cluster randomized designs (Sobel, 2006; Hudgens and Halloran, 2008; Liu and Hudgens, 2014; Ugander et al., 2013; Gui et al., 2015; Eckles et al., 2017; Auerbach and Tabord-Meehan, 2021; Brennan et al., 2022; Ugander and Yin, 2023). These works exploit structural assumptions on the underlying interference network to reduce bias in the difference in means estimator or reduce variance in the Horvitz-Thompson estimator via cluster randomized designs. Some of these works rely on the assumption of *partial interference*, which posits that the underlying network is made up of disjoint groups

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

and interference only occurs within, not across, groups (Sobel, 2006; Hudgens and Halloran, 2008; Liu and Hudgens, 2014; Bhattacharya et al., 2020; Auerbach and Tabord-Meehan, 2021). Other works are devoted to proposing cluster randomized designs that exploit knowledge about the network and its structure to minimize the number of edges that cross clusters in settings where partial interference may not hold (Ugander et al., 2013; Gui et al., 2015; Eckles et al., 2017; Brennan et al., 2022; Ugander and Yin, 2023).

Another strand of literature exploits assumptions on the potential outcomes in their methodology while considering simpler, unit-randomized designs instead of, or in addition to, cluster-randomized designs (Toulis and Kao, 2013; Cai et al., 2015; Gui et al., 2015; Parker et al., 2017; Chin, 2019). These approaches assume linear or generalized linear potential outcomes models, reducing the estimation task to regression. A drawback of these approaches is their assumption of *anonymous interference*, which posits that only the number, not identity, of treated units affects an individual's outcome (Hudgens and Halloran, 2008; Liu and Hudgens, 2014; Li and Wager, 2022), and the entire population shares the same outcomes model.

To address this, Cortez et al. (2022) introduced a flexible class of potential outcomes models that allow for heterogeneous treatment effects by relaxing the anonymous interference assumption and instead imposing the *β-order interactions* assumption, where interference effects are constrained to small subsets of the population. Other recent work has also adopted this model (Cortez-Rodriguez et al., 2023; Eichhorn et al., 2024). In both our work and Cortez et al. (2022), the class of estimators under consideration are based on polynomial interpolation, where a key insight is that under a β-order potential outcomes model, the expected average outcome of the population is a β-degree polynomial in the treatment level. This style of estimator was first introduced by Yu et al. (2022) for the case when the polynomial has degree 1, and generalized by Cortez et al. (2022) to polynomials of higher degree. A drawback is that the estimator has prohibitively high variance when the polynomial degree is greater than 1, especially when the treatment probability is small. The present work addresses this by using a two-stage experimental design to reduce variance in polynomial interpolation estimators for the TTE. A key contribution of Yu et al. (2022) and Cortez et al. (2022) is the unbiased estimation of causal effects *without any knowledge of the underlying interference network*, which the majority of prior approaches require. Our approach does not require network knowledge, but we show how using graph knowledge to select a good graph clustering with which to correlate treatments may improve our estimator's performance.

A few recent works also study rollout designs for causal inference under interference. Han et al. (2022) present statistical tests to detect the presence of interference using rollout designs. Our work differs from theirs in that our focus is estimating treatment effects under interference, not detecting its presence. Boyarsky et al. (2023) leverage rollout designs as part of a model selection mechanism to select a "best" model for interference. While they also study the TTE, their focus is on its identification conditions and how rollout designs aid in satisfying them. Viviano (2020) does not leverage rollout designs, instead considering a two-wave experiment that uses a pilot study in the first wave to minimize variance in causal effect estimation from a main experiment in the second wave. Unlike our two-stage approach, their two-wave design does not utilize a staggered rollout. Furthermore, Viviano (2020) requires anonymous interference, whereas our approach does not.

Contributions We propose a two-stage experiment design to address the high variance of polynomial interpolation estimators under β-order interactions with large β. Given an overall treatment budget p, for a chosen parameter $q \in [p, 1]$, the first stage samples a p/q fraction subset of the population, and the second stage runs a staggered rollout on the selected subset with an effective budget of q. We propose a polynomial interpolation estimator that uses the higher effective budget q, and scales the final outcome to account for the fact that only p/q fraction of units are eligible for treatment in stage two. The increased effective budget reduces the variance from polynomial interpolation. We show the following insights:

- This two-stage estimator has less variance than a one-stage rollout interpolation estimator, but the sub-sampling in the first stage introduces bias.
- Larger values of q lead to higher bias due to edges cut in stage one of the design (that is, edges crossing between selected units and unselected units).
- When clustering is used in the first stage, the bias and variance of the estimator are affected by the edges between clusters and the variance of the average treatment effect across clusters.
- Since the variance of average cluster effects relates to homophily, we see a tension between two clustering objectives: minimizing cut edges versus maximizing covariate balance.
- Even without network or covariate information, the two-stage approach improves for large values of β (i.e. richer models); a good clustering can help improve further.

2 Preliminaries

We estimate the effect of a treatment on a population of n individuals, denoted $[n] := \{1, \ldots, n\}$, via a randomized experiment. Their treatment assignments are collected in a binary vector $\mathbf{z} \in \{0,1\}^n$, where $z_i = 1$ (resp. $z_i = 0$) indicates that unit i is assigned to treatment (resp. control). We allow for *interference*, so the potential outcome of individual i may be a function of the entire treatment vector $Y_i : \{0,1\}^n \to \mathbb{R}$. We frame our analysis around potential outcomes with the following two features.

First, individual i's outcome is a function of a small subset of the population. We visualize this subset as i's in-neighborhood $\mathcal{N}_i$ in a directed interference graph, where an edge from j to i indicates that j's treatment affects i's outcome; we call j an *in-neighbor* of i. We assume the graph does not change over the timescale of the experiment.

Assumption 1 (Neighborhood Interference):

If $\mathbf{z}, \mathbf{z}'$ have $z_j = z'_j \ \forall\, j \in \mathcal{N}_i$, then $Y_i(\mathbf{z}) = Y_i(\mathbf{z}') \qquad \forall\, i.$

We use $d := \max_i |\mathcal{N}_i|$ to denote the maximum in-degree of the network. Differing from most prior work with neighborhood interference, the underlying interference graph may be *unknown*. In some cases, we leverage various levels of network knowledge such as covariate information or full edge information.

Second, following Cortez et al. (2022), we use the binary nature of the treatments $z_i \in \{0,1\}$ to represent *any* potential outcomes under neighborhood interference as a polynomial in $\mathbf{z}$:

$$Y_i(\mathbf{z}) = \sum_{\mathcal{S} \subseteq \mathcal{N}_i} c_{i,\mathcal{S}} \prod_{j \in \mathcal{S}} z_j, \tag{2.1}$$

where the coefficients $c_{i,\mathcal{S}}$ represent the additive effect to individual i's outcome if everyone in $\mathcal{S}$ is treated. Our second assumption posits that each outcome is affected only by *small* treated subsets.

Assumption 2 (β-Order Interactions): $c_{i,\mathcal{S}} = 0$ for all $|\mathcal{S}| > \beta$.

Under this assumption, $Y_i(\mathbf{z})$ is a polynomial with degree at most β. This is motivated by settings where an individual is separately affected by smaller subcommunities of their neighbors (e.g. colleagues, family members, close friends) rather than the neighborhood as a whole. The case $\beta = 1$ is the heterogeneous linear outcomes model explored in Yu et al. (2022), which generalizes the linear models commonly used in applied settings. When $\beta = d$, we return to the unrestricted neighborhood interference setting.

Our estimand of interest is the *total treatment effect* (TTE), the average difference in outcomes when everyone versus no one is treated:

$$\tfrac{1}{n}\sum_{i=1}^{n}\big(Y_i(\mathbf{1}) - Y_i(\mathbf{0})\big) = \tfrac{1}{n}\sum_{i\in[n]}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}, \tag{2.2}$$

where $\mathcal{S}_i^{\beta} := \{\mathcal{S} \subseteq \mathcal{N}_i \ : \ |\mathcal{S}| \le \beta\}$.

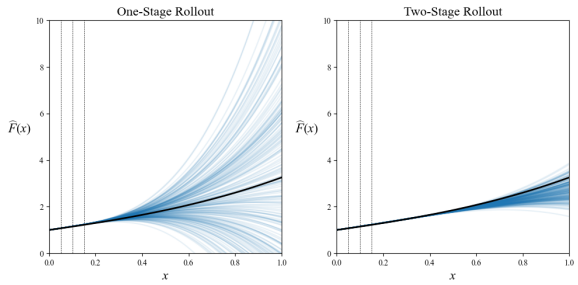


Figure 1: Visualization of extrapolated polynomials used to estimate TTE across 200 runs of a rollout experiment on a 20×20 lattice with $\beta = 3$. The left plot uses a one-stage rollout ($p = 0.15$), as in Cortez et al. (2022), while the right plot uses a two-stage rollout ($q = 0.375$). The two-stage design incurs bias, but extrapolation in the one-stage design leads to higher variance.

Throughout, we consider *completely randomized* experimental designs, $\mathbf{z} \sim \mathrm{CRD}(xn, n)$, wherein a uniform random subset of xn entries of $\mathbf{z}$ are assigned 1. We use the notation $\left[\frac{xn}{n}\right]_m = \prod_{i=0}^{m-1} \frac{xn-i}{n-i}$ to denote the probability that a subset of m individuals is fully treated under such a design.

3 Two-Stage Rollout Designs

Under the β-order interactions model described in Section 2, the quantity

$$\mathbb{E}\big[\tfrac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z})\big] = \tfrac{1}{n}\sum_{i\in[n]}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}} c_{i,\mathcal{S}} \cdot \left[\tfrac{xn}{n}\right]_{|\mathcal{S}|} =: F(x),$$

with the expectation taken over $\mathbf{z} \sim \mathrm{CRD}(xn, n)$ is a polynomial F with degree at most β. Here, x is a variable that represents the proportion of individuals assigned to treatment. We can recover this polynomial (along with $\mathrm{TTE} = F(1) - F(0)$) by interpolating through $\beta + 1$ evaluation points. Cortez et al. (2022) use this observation to describe a TTE estimator that leverages a *staggered rollout* experimental design, wherein $\beta+1$ observations are taken as treatment is rolled out to increasing proportions of the population. This estimator, equation 4 of Cortez et al. (2022), is unbiased with variance $O\big(\frac{d^2\beta^{2\beta+2}}{np^{2\beta}}\big)$, where p represents the *treatment budget*, i.e. proportion of the population assigned to treatment in the final stage of the experiment. In each time step $t \in \{0, \ldots, \beta\}$, where tpn/β individuals are assigned to treatment, the estimator accrues sampling variance by using the observed

mean outcome $\widehat{F}(tp/\beta)$ as a proxy for $F(tp/\beta)$. Then, the extrapolation of these observations to estimate $\widehat{F}(1)$ magnifies this variance by a factor of $(\beta/p)^{2\beta}$ (see Figure 1). To lessen the effects of extrapolation, we would like to sample points from $F(x)$ closer to 1. However, to adhere to our treatment budget p, we must restrict the rollout to a subset of the population. This motivates the following two-stage rollout design.

Definition 3.1 (Two-Stage Rollout Design)**.** *Given a population $[n]$, model degree β, treatment budget p, and parameter $q \in [p, 1]$, we consider experimental designs with the following two stages.*

Stage 1: *Select a subset of the population $\mathcal{U} \subseteq [n]$ with $|\mathcal{U}| = pn/q$ and marginals $\Pr(i \in \mathcal{U}) = p/q$.*

Stage 2: *Run a $(\beta+1)$-stage staggered CRD rollout experiment on the units in $\mathcal{U}$, leaving all other units untreated. Such an experiment satisfies:*

Treatment Restriction: $z_i^t = 0 \qquad \forall\, t, i \notin \mathcal{U}.$

Per-Round Treatment: $\mathbf{z}_{\mathcal{U}}^t \sim \mathrm{CRD}\big(\frac{tpn}{\beta}, |\mathcal{U}|\big) \quad \forall\, t.$

Monotonicity: $z_i^t \geq z_i^{t-1} \qquad \forall\, i, t \geq 1.$

We interpret the parameter q as the *effective treatment budget* of the units within set $\mathcal{U}$. The treatment restriction condition emphasizes that only indivudals chosen in the first stage are eligible for treatment in the second stage. The per-round treatment condition ensures that at each time step t of the rollout, tpn/β units are selected with a completely randomized design from the set of individuals chosen in the first stage ($\mathcal{U}$). When $q = p$, $\mathcal{U} = [n]$ and this design reduces to the CRD (one-stage) staggered rollout designs from (Cortez et al., 2022). The monotonicity assumption is natural in settings where once a unit is treated, they are always treated because treatment cannot be "taken back." We consider two variants of the two-stage rollout design.

Example 3.2 (Unit CRD Rollout Design)**.** *Select $\mathcal{U}$ according to a $\mathrm{CRD}\big(np/q, n\big)$ design. To realize this design, one can sample $U_i \sim \mathrm{Unif}(0,1)$ i.i.d. for each $i \in [n]$, let $\mathcal{U}$ comprise the np/q individuals with highest U_i, and let each $\mathbf{z}^t$ indicate the tnp/β individuals with highest U_i.*

Example 3.3 (Clustered CRD Rollout Design)**.**

Partition the individuals into n_c equal-sized clusters. In Stage 1, use a $\mathrm{CRD}\big(n_c p/q, n_c\big)$ design to select a subset of clusters, and include all individuals from these clusters in $\mathcal{U}$.

We assume throughout that the parameters are appropriately chosen to make all treatment sizes whole numbers. Now, following the presentation of Cortez et al. (2022), we develop a TTE estimator for data collected under such an experiment that leverages the connection to Lagrange polynomial interpolation.

$$\widehat{\mathrm{TTE}} := \tfrac{q}{np} \sum_{t=0}^{\beta} h_{t,q} \sum_{i=1}^{n} Y_i(\mathbf{z}^t), \tag{3.1}$$

where $h_{t,q} = \prod_{\substack{s=0 \\ s\neq t}}^{\beta} \frac{\beta/q - s}{t-s} - \prod_{\substack{s=0 \\ s\neq t}}^{\beta} \frac{-s}{t-s}$ come from the Lagrange coefficients. When $q = p$, this estimator coincides with the polynomial interpolation estimator of Cortez et al. (2022). The estimator is equivalent to applying the polynomial interpolation estimator with the Stage 2 budget q and then scaling the result by q/p, since only a p/q fraction of units are selected in Stage 1 to be eligible for treatment.

The estimate can be evaluated in $O(n\beta)$ time and requires no information about the edges in the interference network. While β is a parameter of the potential outcomes model, it also appears in the estimator due to its use of polynomial interpolation: fitting a β-degree polynomial requires $\beta+1$ points. Note that this estimator requires knowledge of β, as does the estimator in Cortez et al. (2022). Determining or choosing β, while an interesting and practical research direction, is beyond the scope of this paper.

Remark 3.4. *The estimator defined in* (3.1) *does not require knowledge of the interference network. However, the design described in Definition 3.1 may or may not require knowledge of the network depending on how the subset is selected in Stage 1. For example, the design described in Example 3.2 does not require knowledge of the interference network, but the design described in Example 3.3 will require graph knowledge if the clustering method requires it.*

To analyze a Clustered CRD Rollout Design, we introduce the following notation. A clustering Π of the interference network is a partition of $[n]$ into n_c disjoint sets; $\pi \in \Pi$ is a subset $\pi \subseteq [n]$ of units assigned to the same cluster. Given a unit $i \in [n]$, we define $\pi(i)$ as the cluster containing i. Given $\mathcal{S} \subseteq [n]$, we define $\Pi(\mathcal{S}) := \{\pi \in \Pi \mid \exists\, i \in \mathcal{S} \colon \pi = \pi(i)\}$. We define the average treatment effect of a particular cluster $\pi \in \Pi$ as

$$\bar{L}_\pi := \frac{n_c}{n} \sum_{i\in[n]} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta \setminus \varnothing} c_{i,\mathcal{S}} \cdot \mathbb{I}\big(\mathcal{S} \subseteq \pi\big),$$

which is equivalent to the portion of the TTE contained in a single cluster π.

Throughout our analysis, it will be useful to consider the *cut effect*, defined as:

$$C(\delta(\Pi)) := \tfrac{1}{n} \sum_{i\in[n]} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}} \cdot \mathbb{I}\big(|\Pi(\mathcal{S})| \geq 2\big).$$

The cut effect $C(\delta(\Pi))$ denotes the average treatment effect attributable to subsets $\mathcal{S}$ that span across multiple clusters. Since each $\mathcal{S} \subseteq \mathcal{N}_i$ for some $i \in [n]$, $C(\delta(\Pi)) = 0$ when there are no edges cut by the clustering, and it increases with the number of cut edges. Note that the cut effect can be equivalently expressed as $C(\delta(\Pi)) = \text{TTE} - \frac{1}{n_c}\sum_{\pi\in\Pi}\bar{L}_\pi$.

4 Theoretical Results

In this section, we consider the bias and variance of estimator (3.1), with a focus on the Clustered CRD Rollout Design from Example 3.3. In our results, we notationally distinguish between theoretical variance (Var) and empirical variance ($\widehat{\text{Var}}$) with the hat notation. Our first theorem gives a general expression for the bias of estimator (3.1) under two-stage rollout designs.

Theorem 4.1. *Under a β-order potential outcomes model and a Two-Stage Rollout Design, estimator* (3.1) *has bias*

$$\tfrac{1}{n}\sum_{i\in[n]}\sum_{\substack{\mathcal{S}\in\mathcal{S}_i^\beta\\ \mathcal{S}\neq\varnothing}} c_{i,\mathcal{S}}\Big[\tfrac{q}{p}\cdot\Pr\left(\mathcal{S}\subseteq\mathcal{U}\right)-1\Big].$$

A proof appears in Appendix A and uses the law of total expectation to derive an expression for $\mathbb{E}\left[\widehat{\text{TTE}}\right]$ by first conditioning on $\mathcal{U}$. When $|\mathcal{S}| = 1$, i.e. $\mathcal{S} = \{j\}$, the marginal condition $\Pr\left(j \in \mathcal{U}\right) = p/q$ in Stage 1 of the experiment ensures that these terms will not contribute any bias. Rather, all of the bias comes from larger subsets $\mathcal{S}$. Intuitively, bias arises from the possibility that an individual's neighborhood can be partially within and partially outside of $\mathcal{U}$; the estimation of such an individual i's treatment effect via interpolation will be biased as the rollout proportion in each time step will not match the expected proportion of $\mathcal{N}_i$ that is treated. However, when $q = p, \mathcal{U} = [n]$ such that $\Pr(\mathcal{S}\subseteq\mathcal{U}) = 1$ and the estimator is unbiased.

The next two corollaries specialize Theorem 4.1 to the two-stage designs described in Examples 3.2 and 3.3.

Corollary 4.2. *Under a β-order potential outcomes model, the bias of* (3.1) *under a two-stage Unit CRD Rollout Design is*

$$\tfrac{1}{n}\sum_{i\in[n]}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\Big[\tfrac{q}{p}\cdot\big[\tfrac{(p/q)n}{n}\big]_{|\mathcal{S}|}-1\Big].$$

Corollary 4.3. *Under a β-order potential outcomes model and a clustering Π of the interference network into n_c equal-sized clusters, the bias of* (3.1) *under a two-stage Clustered CRD Rollout Design is*

$$\tfrac{1}{n}\sum_{i\in[n]}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\Big[\tfrac{q}{p}\cdot\big[\tfrac{(p/q)n_c}{n_c}\big]_{|\Pi(\mathcal{S})|}-1\Big].$$

When each $c_{i,\mathcal{S}} \geq 0$, the bias is always negative, as we are omitting some of the effects corresponding to sets $\mathcal{S}$ for which $|\Pi(\mathcal{S})| \geq 2$. The magnitude of the bias can be upper-bounded by $\frac{q-p}{q}\cdot C(\delta(\Pi))$. Next, we present a general bound on the variance of estimator (3.1).

Theorem 4.4. *Under a β-order potential outcomes model with each $c_{i\mathcal{S}} \geq 0$ and a clustering Π of the interference network into n_c equal-sized clusters, we bound the variance of* (3.1) *under a two-stage Clustered CRD Rollout Design by:*

$$\mathbb{I}(q<1)\cdot\tfrac{q^3\beta^2Y_{\max}^2}{p^2n}\left(\tfrac{\beta}{q}\right)^{2\beta}\left(d^2+4\beta^3\right) \;+\; \tfrac{q-p}{p(n_c-1)}\widehat{\text{Var}}(\bar{L}_\pi)$$
$$+\;\mathbb{I}(q>p)\cdot\tfrac{2d^2Y_{\max}}{n_c}\cdot C(\delta(\Pi)).$$

where $Y_{\max}$ is a bound on the outcomes.

The first term, which has an exponential dependence on β, is from upper bounding the $\mathbb{E}_{\mathcal{U}}[\text{Var}_{\mathbf{z}}(\widehat{\text{TTE}}_{\text{PI}}\mid\mathcal{U})]$ term from the law of total variance. As this term grows exponentially with β, it is large when β is not small. Since it also does not depend on the clustering, we can only make it small by choosing q close to 1. When $q = 1$ all selected units are assigned treatment, so $\mathbb{E}_{\mathcal{U}}[\text{Var}_{\mathbf{z}}(\widehat{\text{TTE}}_{\text{PI}}\mid\mathcal{U})] = 0$ because there is no randomness in the second stage conditioned on $\mathcal{U}$.

The second and third terms are from upper bounding $\text{Var}_{\mathcal{U}}[\mathbb{E}_{\mathbf{z}}(\widehat{\text{TTE}}_{\text{PI}}\mid\mathcal{U})]$. When $q = p$, $\mathcal{U} = [n]$ is deterministic, such that $\text{Var}_{\mathcal{U}}[\mathbb{E}_{\mathbf{z}}(\widehat{\text{TTE}}_{\text{PI}}\mid\mathcal{U})] = 0$. When $q > p$, these terms reflect the impact of the clustering on the performance of the estimator. The second term is small if $\widehat{\text{Var}}(\bar{L}_\pi)$ is small while n_c is still large. Intuitively $\widehat{\text{Var}}(\bar{L}_\pi)$ is small if the average effects within clusters is well-balanced across clusters, i.e. clusters are similar to each other with respect to their average treatment effects. If covariates are positively correlated with the treatment effects, this encourages clusters that have good covariate balance. The third term depends on the magnitude of the effects from sets $\mathcal{S}$ with membership in more than one cluster, which is the same expression that showed up in the upper bound on the magnitude of the bias. This encourages clusters that minimize cut edges.

When the interference network exhibits strong homophily, these two clustering objectives are in tension: minimizing cut edges may decrease the cut effect, but increase $\widehat{\text{Var}}(\bar{L}_\pi)$ by lowering covariate balance. This suggests that traditional clustering algorithms that focus only on graph-based objectives like minimizing the cut may not be optimal. This is important because the literature on graph clustering often focuses on clustering objectives related to graph structure (e.g. edges). Meanwhile, covariate balance is important in causal inference settings where potential outcomes may be correlated with covariates. At the intersection of these

two research areas is causal inference under network interference, where an effective clustering should capture more than just graph structure when there is homophily. This is an important reminder that with cluster-randomized designs for causal inference under network interference, considering *both* graph structure and covariate balance may be crucial.

Remark 4.5.

When we plug $q = p$ into the bound from Theorem 4.4, we obtain variance bound $\frac{\beta^{2\beta+2} Y_{\max}^2}{p^{2\beta-1} n}\big(d^2 + 4\beta^3\big)$, which we can compare to the asymptotic variance bound $O\Big(\frac{d^2 \beta^{2\beta+2} Y_{\max}^2}{p^{2\beta} n}\Big)$ from Cortez et al. (2022). If we assume we are in the $\beta \ll d$ regime, then these bounds differ by a factor of $\frac{1}{p}$.

4.1 The $\beta = 1$ Setting

Here, we strengthen our variance bounds for the setting of linear ($\beta = 1$) heterogeneous outcomes models. As noted above, estimator (3.1) is unbiased in this setting. For each $j \in [n]$, let us introduce the quantity $L_j = \sum_{i:\, j \in \mathcal{N}_i} c_{i,\{j\}}$ to represent the total *outgoing* effect that treating j has on the population. We use Lemma 4.6, restated from Yu et al. (2022) to understand the variance of $\widehat{\mathrm{TTE}}$.

Lemma 4.6. *Suppose that* $\mathbf{z} \sim \mathrm{CRD}(p \cdot |\mathbf{z}|, |\mathbf{z}|)$. *Then,*

$$\mathrm{Var}\big(\tfrac{1}{p \cdot |\mathbf{z}|} \textstyle\sum_i L_i z_i\big) = \tfrac{1-p}{p \cdot (|\mathbf{z}|-1)} \cdot \widehat{\mathrm{Var}}\big(L_i\big),$$

where $\widehat{\mathrm{Var}}\big(L_i\big) = \frac{1}{|\mathbf{z}|} \sum_j (L_i)^2 - \big(\frac{1}{|\mathbf{z}|} \sum_j (L_i)\big)^2$ *and* $|\mathbf{z}|$ *is the total number of entries in* $\mathbf{z}$.

We use this lemma to derive the variance expression in the following theorem.

Theorem 4.7. *Under a potential outcomes model with* $\beta = 1$ *and a two-stage Clustered CRD rollout design with clustering* Π, *estimator* (3.1) *has variance*

$$\tfrac{1-q}{pn-q} \cdot \widehat{\underset{j \in [n]}{\mathrm{Var}}}(L_j) \;+\; \tfrac{(q-p)(pn-1)}{p(n_c-1)(pn-q)} \cdot \widehat{\underset{\pi \in \Pi}{\mathrm{Var}}}\big(\bar{L}_\pi\big).$$

The proof is in Appendix A. The $\widehat{\mathrm{Var}}_{j \in [n]}(L_j)$ term is the empirical variance of treatment effects across the population and comes from applying Lemma 4.6 where the outer sum is over units $i \in [n]$. The $\widehat{\mathrm{Var}}(\bar{L}_\pi)$ term is the across-cluster variance of average cluster treatment effects and comes from applying the lemma where the outer sum is indexed over clusters $\pi \in \Pi$. When $q = 1$, the design is a simple cluster randomized design where every selected cluster in the first stage is treated in the second stage, and the variance expression simplifies to $\frac{1-p}{p(n_c-1)} \widehat{\mathrm{Var}}_{\pi \in \Pi}\big(\widehat{\mathbb{E}}_{j \in \pi}[L_j]\big)$. This aligns exactly with the cluster-randomized result from Yu et al. (2022). When $q = p$, our variance expression is $\frac{1-p}{p(n-1)} \widehat{\mathrm{Var}}_{j \in [n]}(L_j)$ and aligns exactly with the completely randomized design result from Yu et al. (2022). When $q \in (p, 1)$, some of the variance is attributable to the population-wide variance of influences (which we have no control over) and some of the variance is attributable to variance of average influences across clusters, which can be controlled with a clustering that enforces covariate balance in settings where covariates positively correlate with treatment effects. Note that cut edges do not play a role in the bias or variance when $\beta = 1$.

5 Experiments

In this section, we use experiments on both synthetic and real-world networks to analyze the performance of the two-stage estimator.

Potential Outcomes Model. We use synthetic potential outcomes that generalize the response model of Ugander and Yin (2023) to incorporate β-order interactions; refer to Section 6.2 in their paper for an in-depth description of the design choices of this model. The model incorporates homophily, degree-correlated effects, and β-order interference. Unless otherwise noted, our choices for the parameter values agree with Ugander and Yin (2023). Refer to Appendix B.2 for further details about the model and parameters.

Networks. The synthetic networks that we consider are $\sqrt{n} \times \sqrt{n}$ lattice graphs. We include all self-loops in these networks.

In addition, we consider three real-world networks: an email communication network (Leskovec et al., 2007b), a social network (Rossi and Ahmed, 2015), and a co-purchase network in an online marketplace (Leskovec et al., 2007a); details of these networks can be found in Appendix B.2. Each dataset includes a network and a set of feature labels F assigned to its vertices. In our experiments, we sometimes use these features to cluster the networks.

Running the Experiments. The source code for our experiments and plots found within our manuscript is available in the supplementary materials. All of our experiments were run on a MacBook Pro with an M3 chip and 16GB of memory and ran in under two hours parallelized across its 8 cores.

5.1 Comparison with other estimators

We first empirically explore the performance of our two-stage approach without clustering. We compare the bias and variance of the following estimators:

- 2-Stage, the Polynomial Interpolation (PI) estimator under a unit two stage rollout (with no clustering) and $q = 0.5$

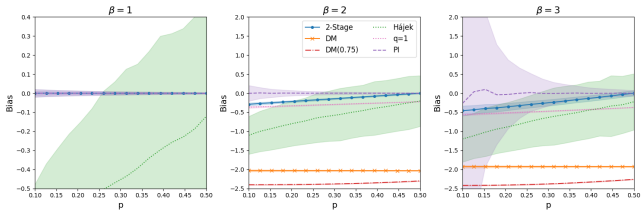


Figure 2: Performance of different estimators on the AMAZON network for various values of p. The bold line indicates the mean over 1000 replications. The shading indicates the experimental standard deviation, calculated by taking the square root of the experimental variance over all replications. The 2-Stage estimator uses $q = 0.5$ and does not utilize clustering. Note the scaling of the y-axes are not the same across β.

- Two difference-in-means style estimators. The classical DM estimator, and a thresholded version, DM(λ) that only considers individuals for which a λ-proportion of their neighborhood shares their treatment assignment.
- The Hájek estimator, an inverse probability weighted-style estimator.
- The PI estimator under a one-stage CRD(pn, n) rollout over $\beta + 1$ time steps from Cortez et al. (2022).
- The two-stage PI estimator with q=1, equivalent to the PI estimator under a one-stage CRD(pn, n) rollout over 2 time steps (with no clustering).

The exact formulas for these estimators can be found in Appendix B.3. Although the Horvitz-Thompson estimator is also considered a baseline, due to its high variance, it consistently performs worse than all the other estimators considered so it is omitted. Of the non-PI estimators, only the (unthresholded) difference in means does not require knowledge of the underlying interference network.

Figure 2 shows the bias and standard deviation of these estimators as we vary the treatment budget p. The column faceting distinguishes between the cases of $\beta = 1$, $\beta = 2$, and $\beta = 3$. While the difference in means estimators have low variance, their bias leads to higher mean squared error (MSE) than the two-stage estimator. In the $\beta = 1$ case, we have zoomed in on the scaling since the variance of the polynomial interpolation estimators is very small. In this setting they are unbiased. PI is equivalent to q=1 in this case so they perfectly overlap and have smaller variance than 2-Stage. This suggests that when you have a truly linear model, the two-stage approach may not improve over the one-stage approach. The remaining estimators either have much worse variance or much worse bias, and thus they do not show up on the plot.

In the $\beta = 2$ case, the difference in means estimators are biased with very low variance. The Hájek estimator has bias that decreases as p increases, but significantly higher variance than all other approaches. Between the three polynomial interpolation-based estimators, PI is unbiased with slightly larger variance for small values of p, while 2-Stage has bias that decreases as p increases and q=1 has bias that remains about the same regardless of p. In this case, the MSE of these estimators is relatively similar, with PI doing slightly better than 2-Stage for smaller values of p, again suggesting this is not a setting where the two-stage approach necessarily does better.

In the $\beta = 3$ case, we can see the significant variance reduction of the two-stage estimator over PI, which comes at the expense of a smaller introduction of bias relative to the remaining estimators. Due to the richer model, we see that the one-stage approach (PI) has an extremely high variance for smaller values of p, much larger than the bias incurred by the two-stage approach.

Overall, the performance of the 2-Stage and PI estimators is similar for most of the parameter landscape when $\beta = 1$ or $\beta = 2$, but the variance reduction of the 2-Stage estimator for small p and $\beta = 3$ results in a lower MSE despite the additional bias. This makes sense because we only expect a large error reduction for richer models and small treatment probabilities.

These results highlight a setting where the two-stage approach improves over the one-stage approach, even without network information, as these experiments did not use any clustering. Plots of the MSE and the two other network datasets can be found in Appendix B.

5.2 Clustering effect in two-stage estimator

In this section we conduct experiments to empirically explore the impact of clustering.

Lattice. In Figure 3, we compare the MSE of the 2-Stage estimator under two clusterings and no clustering on a 100×100 lattice. The Coarse clustering is a 10×10 grid on top of the lattice; there are 100 clusters with 100 people in each cluster. The Fine clustering is a 2×2 grid on top of the lattice; there are 2500 clusters with 4 people in each cluster. Table 1 displays some metrics for these clusterings: $\widehat{\mathrm{Var}}(\bar{L}_\pi)$, $C(\delta(\Pi))$, and the number of cut edges.

Table 1: Clustering Metrics for Figure 3.

Clustering	$\widehat{\mathrm{Var}}(\bar{L}_\pi)$	$C(\delta(\Pi))$	Cut Edges
Coarse	0.0002	0.1229	3600
Fine	0.002	0.5703	19600

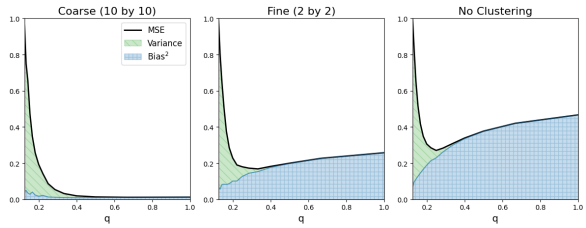


Figure 3: Mean Squared Error of the Two-Stage TTE estimator for two clusterings of a 100×100 lattice graph, compared with no clustering, for a β-degree potential outcomes model with $\beta = 3$. Even with no network knowledge, we see a drastic decrease in MSE even at the cost of incurring bias.

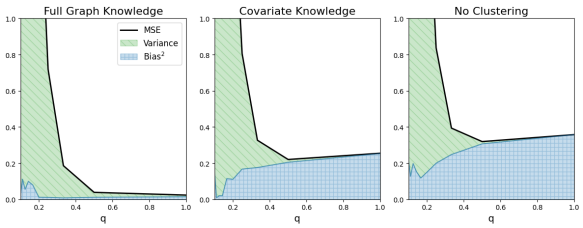


Figure 4: Mean Squared Error of the Two-Stage TTE estimator for two clusterings (with 250 clusters) of the Amazon network, compared with no clustering, for a β-degree potential outcomes model with $\beta = 3$. Even with no network knowledge, we see a drastic decrease in MSE even at the cost of incurring bias.

In Figure 3, we vary q (the treatment probability for individuals selected in the first stage) on the y-axis and plot the MSE, with the different shading corresponding to the variance and squared bias components. The leftmost endpoint corresponds to $q = p = 0.15$, equivalently the one-stage setting from Cortez et al. (2022). Since the treatment budget p is small and $\beta = 3$ (indicating a richer model), the left side of each plot exhibits high variance and low bias. As q increases, the variance decreases but the bias increases due to cut edges. Clustering reduces bias by reducing the number of cut edges. The coarse clustering in the left plot drastically decreases the error, especially as q approaches 1. The middle plot is a finer clustering, and results in much more bias as q approaches 1. Table 1 helps elucidate the difference in performance. The fine clustering cuts five times more edges than the coarse clustering, resulting in a cut effect that is about five times larger. Finally, the rightmost plot shows the MSE of the two-stage estimator under no clustering, i.e. a unit CRD 2-stage rollout, and incurs the largest amount of bias. Overall, the error for $q > p$ is smaller than at $q = p$ across all plots, showing settings where a two-stage design leads to improvement over a one-stage design.

Real-world Networks. We compare two methods of clustering the real-world networks. In the clustering with Full Graph Knowledge, we cluster the true underlying graph using the METIS clustering library by Karypis and Kumar (1998). In the clustering with Covariate Knowledge, clusters are based on features. When each vertex is assigned to one feature, we use these assignments as the clustering. When vertices may have multiple features we form a feature graph — a weighted graph, where the weight of edge (i, j) is the number of feature labels shared by i and j — and cluster this feature graph using METIS.

We highlight the Amazon network here, but additional experiments with the other networks are in Appendix B. Figure 4 depicts the results of such an experiment run on a co-purchasing network of Amazon products. We compare the two clusterings of this graph as described above, each time partitioning the network into 250 parts, against no clustering. To generate these plots, we compute the experimental bias, sampling variance, and total variance, over 1000 replications. Here, we vary q from $q = p = 0.1$ to $q = 1$. In the first plot, showing the MSE of the estimator when the clustering uses full network knowledge, the MSE is minimized at $q = 1$ with value 0.024. In the second plot, showing the MSE of the estimator when the clustering only uses covariate knowledge, the MSE is minimized around $q = 0.5$ with value 0.22. Table 2 gives insight into the difference in performance under these clusterings. The clustering with covariate knowledge cuts about five times as many edges as the clustering with full graph knowledge, resulting in a cut effect that is about 5 times larger.

Table 2: Clustering Metrics for Figure 4.

Cluster	$\widehat{\mathrm{Var}}(\bar{L}_\pi)$	$C(\delta(\Pi))$	Cuts
Full	0.2488	0.1258	7670
Covariate	0.0426	0.5436	41243

In the final plot, showing the MSE of the estimator under a two-stage unit CRD design, the MSE is minimized around $q = 0.5$ with value 0.32. Recall that the leftmost endpoint of each plot corresponds to the error when $q = p$, i.e. under the one-stage rollout. Although a clustering with full network knowledge achieves the best overall performance, we see a significant error reduction over a one-stage even for a two-stage unit CRD design. Thus, using the two-stage estimator may reduce MSE (versus a single-stage rollout) even without a clustering or network knowledge.

In Table 3, we record some metrics of clusterings of different sizes computed with full network or covariate knowledge. These experiments use a β-degree potential outcomes model with $\beta = 3$. The parameter n_c

Table 3: Clustering Metrics for Amazon Network

Cluster	n_c	$\widehat{\mathrm{Var}}(\bar{L}_\pi)$	$C(\delta(\Pi))$	$q_{\min}$	MSE
Full	50	0.156059	0.088226	1	0.035
Full	100	0.187048	0.102260	1	0.028
Full	250	0.248759	0.125772	1	0.024
Covariate	50	0.019543	0.517855	0.5	0.211
Covariate	100	0.025207	0.536876	0.5	0.228
Covariate	250	0.042644	0.543623	0.5	0.220

indicates the number of clusters. In each row, $q_{\min}$ is the value of q that minimizes the MSE and the column MSE contains that value. We computed the minimum empirically. Generally, having more clusters corresponds to a higher across-cluster variance of average cluster influences and a higher cut effect. However, regardless of cluster size, the MSE is still drastically decreased from 38 (at $q = p$) to about 0.2 under a covariate-based clustering (at $q = 0.5$) and to about 0.02 with a full graph knowledge-based clustering (at $q = 1$). Clustering with full knowledge has higher across-cluster variance of average cluster influences but smaller cut effect compared with clustering with covariate knowledge. This reminds us that there is a tension between cut edges and covariate balance. While a common clustering objective is to minimize cut edges, there may be settings where enforcing some covariate balance may be wise if there is homophily. This is because if there is strong homophily, edges are correlated with covariates. If there is reason to believe these covariates are highly correlated with potential outcomes, then minimizing cut edges might minimize the cut effect but maximize the variance of cluster influences.

Our theoretical and experimental results explore settings with equal-sized clusters, such as in Eckles et al. (2017); Brennan et al. (2022); Candogan et al. (2024), which may be too difficult a constraint to meet in many practical settings. In theory, unequal-size clusters should not affect bias but will affect the variance of the estimates. For the experiments on real-world networks, we use the METIS clustering library, which can only do equal-size clusters. Exploring the performance under unequal-sized clusters is a practical direction for future work.

Acknowledgements

We gratefully acknowledge financial support from the National Science Foundation grants CCF-2337796 and CNS-1955997, the National Science Foundation Graduate Research Fellowship grant DGE-1650441, and AFOSR grant FA9550-23-1-0301.

References

Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

Auerbach, E. and Tabord-Meehan, M. (2021). The local approach to causal inference under network interference. *arXiv preprint arXiv:2105.03810*.

Basu, D. (2011). *An essay on the logical foundations of survey sampling, part one*. Springer.

Bhattacharya, R., Malinsky, D., and Shpitser, I. (2020). Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, pages 1028–1038. PMLR.

Biswas, N. and Airoldi, E. M. (2018). Estimating peer-influence effects under homophily: Randomized treatments and insights. In *Complex Networks IX: Proceedings of the 9th Conference on Complex Networks CompleNet 2018 9*, pages 323–347. Springer.

Boyarsky, A., Namkoong, H., and Pouget-Abadie, J. (2023). Modeling interference using experiment rollout. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, page 298, New York, NY, USA. Association for Computing Machinery.

Brennan, J., Mirrokni, V., and Pouget-Abadie, J. (2022). Cluster randomized designs for one-sided bipartite experiments. *Advances in Neural Information Processing Systems*, 35:37962–37974.

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC medical research methodology*, 6(1):1–9.

Cai, J., Janvry, A. D., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Candogan, O., Chen, C., and Niazadeh, R. (2024). Correlated cluster-based randomized experiments: Robust variance minimization. *Management Science*, 70(6):4069–4086.

Chin, A. (2019). Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2):20180026.

Cortez, M., Eichhorn, M., and Yu, C. L. (2022). Staggered rollout designs enable causal inference under interference without network knowledge. *Advances in Neural Information Processing Systems*, 35:7437–7449.

Cortez-Rodriguez, M., Eichhorn, M., and Yu, C. L. (2023). Exploiting neighborhood interference with low-order interactions under unit randomized design. *Journal of Causal Inference*, 11(1):20220051.

Eckles, D., Karrer, B., and Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021.

Eichhorn, M., Khan, S., Ugander, J., and Yu, C. L. (2024). Low-order outcomes and clustered designs: combining design and analysis for causal inference under network interference.

Gui, H., Xu, Y., Bhasin, A., and Han, J. (2015). Network a/b testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pages 399–409.

Han, K., Li, S., Mao, J., and Wu, H. (2022). Detecting interference in a/b testing with increasing allocation. *arXiv preprint arXiv:2211.03262*.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

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

Karypis, G. and Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1):359–392.

Khan, S. and Ugander, J. (2023). Adaptive normalization for ipw estimation. *Journal of Causal Inference*, 11(1):20220019.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007a). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5–es.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007b). Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es.

Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

Li, S. and Wager, S. (2022). Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334–2358.

Liu, L. and Hudgens, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the american statistical association*, 109(505):288–301.

Parker, B. M., Gilmour, S. G., and Schormans, J. (2017). Optimal design of experiments on connected units with application to social networks. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(3):455–480.

Rossi, R. A. and Ahmed, N. K. (2015). The network data repository with interactive graph analytics and visualization. In *AAAI*.

Sobel, M. E. (2006). 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.

Tang, L. and Liu, H. (2009a). Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 817–826.

Tang, L. and Liu, H. (2009b). Scalable learning of collective behavior based on sparse social dimensions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1107–1116.

Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR.

Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337.

Ugander, J. and Yin, H. (2023). Randomized graph cluster randomization. *Journal of Causal Inference*, 11(1):20220014.

Viviano, D. (2020). Experimental design under network interference. *arXiv preprint arXiv:2003.08421*.

Xu, Y., Duan, W., and Huang, S. (2018). 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.

Yin, H., Benson, A. R., Leskovec, J., and Gleich, D. F. (2017). Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 555–564.

Yu, C. L., Airoldi, E. M., Borgs, C., and Chayes, J. T. (2022). Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Section 2.]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, in supplementary material.]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes, in Appendix]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, in supplementary material and as URL in camera-ready version.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, in Appendix and in Section 5]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

(d) Information about consent from data providers/curators. [Yes]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A PROOFS

Theorem 4.1

Proof. We use the Law of Total Expectation, conditioning on the set of individuals $\mathcal{U}$ selected in the first stage and reasoning about the randomness from the treatment assignments $\mathbf{z}$. We have,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{z}}\left[\widehat{\mathrm{TTE}}_{\mathrm{PI}} \,\middle|\, \mathcal{U}\right] &= \tfrac{q}{np}\sum_{t=0}^{\beta} h_{t,q} \sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}} \cdot \mathop{\mathbb{E}}_{\mathbf{z}}\Big[\prod_{j\in\mathcal{S}} z_j^t \,\Big|\, \mathcal{U}\Big] \\
&= \tfrac{q}{np}\sum_{t=0}^{\beta} h_{t,q} \sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}} \cdot \mathbb{I}(\mathcal{S}\subseteq\mathcal{U}) \cdot \Big[\tfrac{tpn/\beta}{|\mathcal{U}|}\Big]_{|\mathcal{S}|} \\
&= \tfrac{q}{np}\sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}} \cdot \mathbb{I}(\mathcal{S}\subseteq\mathcal{U}) \sum_{t=0}^{\beta} h_{t,q}\Big[\tfrac{tpn/\beta}{|\mathcal{U}|}\Big]_{|\mathcal{S}|} \\
&= \tfrac{q}{np}\sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}} \cdot \mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}\big)\Big(1^{|\mathcal{S}|} - 0^{|\mathcal{S}|}\Big) \\
&= \tfrac{q}{np}\sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}} \cdot \mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}\big) \qquad \text{(A.1)}
\end{aligned}$$

Here, the fourth line follows from the properties of Lagrange interpolation. Note that $h_{t,q} = \ell_{t,\mathbf{x}}(1) - \ell_{t,\mathbf{x}}(0)$, where $\ell_{t,\mathbf{x}}$ is the t'th Lagrange basis polynomial with evaluation points $\mathbf{x} = \big(\frac{tq}{\beta}\big)_{t\in 0,\ldots,\beta} = \big(\frac{tpn/\beta}{|\mathcal{U}|}\big)_{t\in 0,\ldots,\beta}$. Thus, for any polynomial $f(x)$ with degree at most β,

$$\sum_{t=0}^{\beta} h_{t,q}\cdot f\big(\tfrac{tq}{\beta}\big) = f(1) - f(0).$$

In this case, we let $f(x) = \Big[\frac{x|\mathcal{U}|}{|\mathcal{U}|}\Big]_{|\mathcal{S}|}$, to find that

$$\sum_{t=0}^{\beta} h_{t,q}\Big[\tfrac{tpn/\beta}{|\mathcal{U}|}\Big]_{|\mathcal{S}|} = [1]_{|\mathcal{S}|} - [0]_{|\mathcal{S}|} = 1^{|\mathcal{S}|} - 0^{|\mathcal{S}|}.$$

Now, taking the expectation over the randomness in $\mathcal{U}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathrm{TTE}}_{\mathrm{PI}}\right] &= \mathop{\mathbb{E}}_{\mathcal{U}}\Big[\mathop{\mathbb{E}}_{\mathbf{z}}\left[\widehat{\mathrm{TTE}}_{\mathrm{PI}} \,\middle|\, \mathcal{U}\right]\Big] \\
&= \tfrac{1}{n}\sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}} \cdot \tfrac{q}{p} \cdot \Pr(\mathcal{S}\subseteq\mathcal{U}).
\end{aligned}$$

The bias expression in the theorem statement follows from the expression for TTE given in (2.2). □

Theorem 4.4

We also use the following algebraic lemma.

Lemma A.1. *For all* $0 < k \le n_{\mathcal{U}} \le 1$,

$$|h_{t,q}| = \prod_{\substack{s=0\\ s\neq t}}^{\beta} \tfrac{\beta/q-s}{t-s} - \prod_{\substack{s=0\\ s\neq t}}^{\beta} \tfrac{-s}{t-s} \;\le\; \left(\tfrac{\beta}{q}\right)^{\beta}.$$

Proof. When $\beta = 1$, $|h_{0,q}| = |h_{1,q}| = \frac{1}{q}$, so the inequality holds (with equality). Thus, we can restrict our attention to $\beta \ge 2$, for which we consider in two cases. First, if $t \ge 1$, we have

$$\left|h_{t,q}\right| = \Big| \prod_{\substack{s=0\\ s\neq t}}^{\beta} \tfrac{\beta/q-s}{t-s} \Big| \le \left(\tfrac{\beta}{q}\right)^{\beta}.$$

The equality uses the definition of $h_{t,q}$, and the inequality upper bounds the numerator of each factor with β/q and lower bounds the denominator of each factor by 1. When $t = 0$, we apply the triangle inequality to conclude that

$$\left|h_{0,q}\right| = \Big| \prod_{s=1}^{\beta} \tfrac{\beta/q-s}{-s} - 1 \Big| \le \prod_{s=1}^{\beta} \tfrac{\beta}{sq} + 1 = \tfrac{1}{\beta!}\left(\tfrac{\beta}{q}\right)^{\beta} + 1.$$

Since $\beta \ge 2$ and $q \le 1$, we must have $1 \le \frac{1}{2}\left(\frac{\beta}{q}\right)^{\beta}$. Thus we can upper-bound this last expression by

$$\tfrac{1}{\beta!}\left(\tfrac{\beta}{q}\right)^{\beta} + \tfrac{1}{2}\left(\tfrac{\beta}{q}\right)^{\beta} = \left(\tfrac{1}{\beta!} + \tfrac{1}{2}\right) \cdot \left(\tfrac{\beta}{q}\right)^{\beta} \le \left(\tfrac{\beta}{q}\right)^{\beta}.$$

□

We also prove a slightly stronger version of Theorem 3 from Cortez et al. (2022) with the constants specified. It first relies on a slightly modified version of Lemma 8 from Cortez et al. (2022).

Lemma A.2. *For any* $x \in (0, 1]$ *and any constants* $a, b \in \mathbb{N}$ *such that* $xn \ge \sqrt{2}ab + b - 1$,

$$\left| \frac{\left[\frac{xn-a}{n-a}\right]_b}{\left[\frac{xn}{n}\right]_b} - 1 \right| \le \tfrac{2ab}{xn-b+1},$$

Proof. First, let us note that when $a = 0$ or $b = 0$, both sides of this inequality simplify to 0, so it holds with equality. Thus, we assume throughout the rest of the proof that $a, b > 0$. Note that our assumption $xn \ge \sqrt{2}ab + b - 1$ with $x \le 1$ implies that $n \ge a + b - 1$.

Now, given any any $i \in \{0, \dots, b-1\}$,

$$\frac{xn-a-i}{n-a-i} \le \frac{xn-i}{n-i} \quad \Rightarrow \quad \frac{\left[\frac{xn-a}{n-a}\right]_b}{\left[\frac{xn}{n}\right]_b} \le 1.$$

As a result, expanding the bracket notation, we have,

$$\begin{aligned}
\left| \frac{\left[\frac{xn-a}{n-a}\right]_b}{\left[\frac{xn}{n}\right]_b} - 1 \right| &= 1 - \prod_{i=0}^{b-1} \left(\frac{xn-a-i}{xn-i}\right)\left(\frac{n-i}{n-a-i}\right) \\
&= 1 - \prod_{i=0}^{b-1} \left(1 - \frac{a}{xn-i}\right) \underbrace{\left(1 + \frac{a}{n-a-i}\right)}_{\ge 1}
\end{aligned}$$

$$
\begin{aligned}
&\leq 1 - \prod_{i=0}^{b-1} \Big(1 - \frac{a}{xn - b + 1}\Big) && (i \leq b-1)\\
&= -\sum_{j=1}^{b} \binom{b}{j} \Big(-\frac{a}{(xn-b+1)}\Big)^j && \text{(binomial expansion)}\\
&\leq \sum_{j=1}^{b} \binom{b}{j} \Big(\frac{a}{(xn-b+1)}\Big)^j \cdot \mathbb{I}(j \text{ is odd})\\
&\leq \Big(\frac{ab}{xn-b+1}\Big) \sum_{j=0}^{\lfloor (b-1)/2 \rfloor} \Big(\frac{ab}{xn-b+1}\Big)^{2j}\\
&\leq \Big(\frac{ab}{xn-b+1}\Big) \sum_{j=0}^{\lfloor (b-1)/2 \rfloor} \Big(\frac{1}{\sqrt{2}}\Big)^{2j} && (xn \geq \sqrt{2}ab + b - 1)\\
&\leq \frac{2ab}{xn-b+1}. && \text{(geometric series with factor } \tfrac{1}{2})
\end{aligned}
$$

□

We use this to lemma to give an upper bound on the covariance of two sets being treated under a CRD rollout design with $\frac{ptn}{\beta}$ individuals treated in round t for each $t \in \{0, \ldots, \beta\}$.

Lemma A.3. *If $\frac{pt'n}{\beta} \geq 2\beta^2 + \beta - 1$, then for $t \leq t'$ and $\mathcal{S} \cap \mathcal{S}' = \emptyset$ with $|\mathcal{S}|, |\mathcal{S}'| \geq 1$, it follows that*

$$\Big|\mathrm{Cov}\Big[\prod_{j\in\mathcal{S}} z_j^t, \prod_{j'\in\mathcal{S}'} z_{j'}^{t'}\Big]\Big| \leq \frac{4p\beta^3}{n}.$$

Proof. First, let us note that if $t = 0$, then the first argument of this covariance is not random, so the covariance simplifies to 0, trivially satisfying the bound. Thus, we may assume that $1 \leq t \leq t'$. We can rewrite the covariance expression:

$$
\begin{aligned}
\Big|\mathrm{Cov}\Big[\prod_{j\in\mathcal{S}} z_j^t, \prod_{j'\in\mathcal{S}'} z_{j'}^{t'}\Big]\Big| &= \Big|\mathbb{E}\Big[\prod_{j\in\mathcal{S}} z_j^t \prod_{j'\in\mathcal{S}'} z_{j'}^{t'}\Big] - \mathbb{E}\Big[\prod_{j\in\mathcal{S}} z_j^t\Big]\mathbb{E}\Big[\prod_{j'\in\mathcal{S}'} z_{j'}^{t'}\Big]\Big|\\
&= \Big[\tfrac{ptn/\beta}{n}\Big]_{|\mathcal{S}|}\Big[\tfrac{pt'n/\beta}{n}\Big]_{|\mathcal{S}'|} \cdot \left|\frac{\Big[\frac{pt'n/\beta - |\mathcal{S}|}{n-|\mathcal{S}|}\Big]_{|\mathcal{S}'|}}{\Big[\frac{pt'n/\beta}{n}\Big]_{|\mathcal{S}'|}} - 1\right|.
\end{aligned}
$$

We can bound this last absolute value expression using Lemma A.2, letting $x = pt'/\beta$, $a = |\mathcal{S}|$, and $b = |\mathcal{S}'|$. Note that $a, b \leq \beta$, so our assumption that $\frac{pt'n}{\beta} \geq 2\beta^2 + \beta - 1$ ensures that $xn \geq \sqrt{2}ab + b - 1$. We find that

$$\Big|\mathrm{Cov}\Big[\prod_{j\in\mathcal{S}} z_j^t, \prod_{j'\in\mathcal{S}'} z_{j'}^{t'}\Big]\Big| \leq \Big(\tfrac{pt}{\beta}\Big)^{|\mathcal{S}|}\Big(\tfrac{pt'}{\beta}\Big)^{|\mathcal{S}'|} \cdot \frac{2|\mathcal{S}||\mathcal{S}'|}{\frac{pt'n}{\beta} - |\mathcal{S}'| + 1} \leq \frac{2p^2\beta^3}{pn - \beta^2} \leq \frac{4p\beta^3}{n}$$

Here, the final equality uses the fact that $pn \geq pt'n/\beta \geq 2\beta^2$ to conclude that $\frac{p}{pn-\beta^2} \leq \frac{2}{n}$. □

When $\mathcal{S} \cap \mathcal{S}' \neq \emptyset$ for $|\mathcal{S}|, |\mathcal{S}'| \geq 1$, it follows that

$$\Big|\mathrm{Cov}\Big[\prod_{j\in\mathcal{S}} z_j^t, \prod_{j'\in\mathcal{S}'} z_{j'}^{t'}\Big]\Big| \leq p.$$

Plugging this into Lemma 6 of Cortez et al. (2022), (so, in their notation, $\alpha = \Big(\frac{\beta}{p}\Big)^\beta$, $B_1 = p$, and $B_2 = \frac{4p\beta^3}{n}$), we can upper bound the variance of the staggered rollout estimator under a CRD rollout design by

$$\mathrm{Var}\left(\widehat{\mathrm{TTE}}\right) \leq \frac{\beta^2 Y_{\max}^2 p}{n} \cdot \Big(\tfrac{\beta}{p}\Big)^{2\beta} \cdot \Big(d^2 + 4\beta^3\Big). \tag{A.2}$$

Proof of Theorem 4.4.

By the Law of Total Variance, we have

$$\operatorname*{Var}_{\mathbf{z}}\left(\widehat{\mathrm{TTE}}\right) = \mathop{\mathbb{E}}_{\mathcal{U}}\left[\operatorname*{Var}_{\mathbf{z}|\mathcal{U}}\left(\widehat{\mathrm{TTE}}\right)\right] + \operatorname*{Var}_{\mathcal{U}}\left(\mathop{\mathbb{E}}_{\mathbf{z}}\left[\widehat{\mathrm{TTE}} \mid \mathcal{U}\right]\right).$$

We separately bound each of these terms.

First Term:

First, let us note that when $q = 1$, $h_{0,q} = -1$, $h_{\beta,q} = 1$ and $h_{t,q} = 0$ for all $0 < t < \beta$. In this case, we may simplify the estimator to

$$\widehat{\mathrm{TTE}} = \tfrac{1}{np}\sum_{i=1}^{n} Y_i(\mathbf{z}^\beta) - Y_i(\mathbf{z}^0).$$

Conditioned on $\mathcal{U}$, this quantity is deterministic, since $z_j^\beta = \mathbb{I}(j \in \mathcal{U})$ and $z_j^0 = 0$. Thus, the variance of the estimator conditioned on $\mathcal{U}$ is 0, making the first term of our variance expression 0. Thus, we may restrict our attention to the case when $q < 1$ and multiply the resulting expression by the indicator $\mathbb{I}(q < 1)$ in our final bound.

Now, let $\tilde{\mathbf{z}} \sim \mathrm{CRD}(qn, n)$ be a random vector with $z_j^t = \tilde{z}_j^t \cdot \mathbb{I}(j \in \mathcal{U})$. Conditioned on $\mathcal{U}$, we may rewrite our estimator:

$$\begin{aligned}
\widehat{\mathrm{TTE}} &= \tfrac{q}{np}\sum_{t=0}^{\beta} h_{t,q} \sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}} \prod_{j\in\mathcal{S}} z_j^t \\
&= \sum_{t=0}^{\beta} h_{t,q} \cdot \left(\tfrac{1}{n}\sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} \tfrac{q}{p} \cdot c_{i,\mathcal{S}} \cdot \mathbb{I}(\mathcal{S} \subseteq \mathcal{U}) \prod_{j\in\mathcal{S}} \tilde{z}_j^t\right) \\
&= \sum_{t=0}^{\beta} h_{t,q} \cdot \left(\tfrac{1}{n}\sum_{i=1}^{n} \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} \tilde{c}_{i,\mathcal{S}} \prod_{j\in\mathcal{S}} \tilde{z}_j^t\right) \\
&= \sum_{t=0}^{\beta} h_{t,q} \cdot \left(\tfrac{1}{n}\sum_{i=1}^{n} \tilde{Y}_i(\tilde{\mathbf{z}}^t)\right),
\end{aligned}$$

where

$$\tilde{c}_{i,\mathcal{S}} = \tfrac{q}{p} c_{i,\mathcal{S}} \cdot \mathbb{I}(\mathcal{S} \subseteq \mathcal{U}), \qquad \tilde{Y}_i(\tilde{\mathbf{z}}) = \sum_{\mathcal{S}\in\mathcal{S}_i^\beta} \tilde{c}_{i,\mathcal{S}} \prod_{j\in\mathcal{S}} \tilde{z}_j^t.$$

Writing it in this way, we can see that the distribution of $\widehat{\mathrm{TTE}}$ conditioned on $\mathcal{U}$ is equivalent to the distribution of the polynomial interpolation estimator in Cortez et al. (2022) with $\tilde{\mathbf{z}} \sim \mathrm{CRD}(qn, n)$ for a modified potential outcomes model given by the coefficients $\tilde{c}_{i,\mathcal{S}}$.

Under the assumption that $c_{i,\mathcal{S}} \geq 0$, then $\tilde{Y}_i(\mathbf{z}) \leq \frac{q}{p} Y_i(\mathbf{z})$.

As a result, the variance of $\widehat{\mathrm{TTE}}$ conditioned on $\mathcal{U}$ can be upper-bounded from (A.2). As this expression does not depend on $\mathcal{U}$,

$$\mathop{\mathbb{E}}_{\mathcal{U}}\left[\operatorname*{Var}_{\mathbf{z}|\mathcal{U}}\left(\widehat{\mathrm{TTE}}\right)\right] \leq \tfrac{q^3\beta^2 Y_{\max}^2}{p^2 n} \cdot \left(\tfrac{\beta}{q}\right)^{2\beta} \cdot \left(d^2 + 4\beta^3\right).$$

Second Term:

First, let us note that when $q = p$, every individual is deterministically included in $\mathcal{U}$ during Stage 1 of the experiment. In this case, the second term, which concerns a variance over $\mathcal{U}$, is 0. Thus, we may restrict our attention to the case when $q > p$ and multiply the resulting expression by the indicator $\mathbb{I}(q > p)$ in our final bound.

We first split $\mathbb{E}_{\mathbf{z}}\left[\widehat{\mathrm{TTE}}_{\mathrm{PI}} \,\middle|\, \mathcal{U}\right]$ from (A.1) into the terms associated to sets $\mathcal{S}$ that are fully contained inside a cluster as opposed to sets $\mathcal{S}$ that contain members of more than one cluster.

$$\mathbb{E}_{\mathbf{z}}\left[\widehat{\mathrm{TTE}}_{\mathrm{PI}} \,\middle|\, \mathcal{U}\right] = \tfrac{q}{np}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}, |\Pi(\mathcal{S})|=1\big) + \tfrac{q}{np}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}, |\Pi(\mathcal{S})|\geq 2\big). \quad \text{(A.3)}$$

We may rewrite the first term of (A.3):

$$\tfrac{q}{np}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}, |\Pi(\mathcal{S})|=1\big) \quad = \quad \tfrac{q}{np}\sum_{i=1}^{n}\sum_{\pi\in\Pi} x_\pi \sum_{\substack{\mathcal{S}\in\mathcal{S}_i^{\beta}\\ \mathcal{S}\neq\varnothing}} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(S\subseteq\pi\big) \quad = \quad \tfrac{q}{pn_c}\sum_{\pi\in\Pi} x_\pi \bar{L}_\pi,$$

where $x_\pi = \mathbb{I}\big(\pi\subseteq\mathcal{U}\big)$ and $\bar{L}_\pi$ is defined as in the main text, with

$$\bar{L}_\pi = \tfrac{n_c}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\subseteq[n]} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(\mathcal{S}\subseteq\pi\big),$$

which represents the effects associated with sets fully contained inside cluster π. In Stage 1, we select clusters according to a CRD design. In particular, $\mathbf{x}\sim\mathrm{CRD}(pn_c/q, n_c)$. Applying Lemma 4.6, we find that the variance of the first term of (A.3) is equal to

$$\tfrac{q-p}{p(n_c-1)}\cdot\widehat{\mathrm{Var}}\big(\bar{L}_\pi\big).$$

To upper bound the terms of the variance associated to the second term of $\mathbb{E}_{\mathbf{z}}\left[\widehat{\mathrm{TTE}}_{\mathrm{PI}} \,\middle|\, \mathcal{U}\right]$ associated to all the sets $\mathcal{S}$ for which $|\Pi(\mathcal{S})|\geq 2$, we use the bound that for any $\mathcal{S}$ such that $|\Pi(\mathcal{S})|\geq 2$,

$$\mathrm{Cov}\Big(\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}\big), \mathbb{I}\big(\mathcal{S}'\subseteq\mathcal{U}\big)\Big) \leq \Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\Big(\Pi(\mathcal{S})\cap\Pi(\mathcal{S}')\neq\emptyset\Big).$$

In addition, we'll make use of our assumption that each $c_{i\mathcal{S}}\geq 0$. Plugging in these bounds, it follows that

$$\begin{aligned}
&\mathrm{Cov}\Big(\tfrac{q}{np}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}, |\Pi(\mathcal{S})|\geq 2\big), \tfrac{q}{pn_c}\sum_{\pi\in\Pi} x_\pi\bar{L}_\pi\Big)\\
&= \tfrac{q^2}{nn_cp^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{\pi\in\Pi}\bar{L}_\pi\,\mathrm{Cov}(\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}\big), x_\pi)\\
&\leq \tfrac{q^2}{nn_cp^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{\pi\in\Pi}\bar{L}_\pi\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}(c\in\Pi(\mathcal{S}))\\
&= \tfrac{q^2}{p^2nn_c}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{\pi\in\Pi(\mathcal{S})}\bar{L}_\pi\\
&= \tfrac{q^2}{p^2n^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{\pi\in\Pi(\mathcal{S})}\sum_{i'\in[n]}\sum_{\mathcal{S}'\in\mathcal{S}_i'^{\beta}} c_{i',\mathcal{S}'}\cdot\mathbb{I}\big(\mathcal{S}'\subseteq\pi\big)\\
&\leq \tfrac{q^2}{p^2n^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{\pi\in\Pi(\mathcal{S})}\sum_{i'\in[n]}\mathbb{I}\big(\pi\in\Pi(\mathcal{N}_{i'})\big)\sum_{\mathcal{S}'\in\mathcal{S}_i'^{\beta}} c_{i',\mathcal{S}'}\\
&\leq \tfrac{q^2Y_{\max}}{p^2n^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{\pi\in\Pi(\mathcal{S})}\sum_{i'\in[n]}\mathbb{I}\big(\pi\in\Pi(\mathcal{N}_{i'})\big)\\
&\leq \tfrac{q^2Y_{\max}}{p^2n^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{\pi\in\Pi(\mathcal{S})}\tfrac{nd}{n_c}
\end{aligned}$$

$$= \tfrac{q^2 d\beta Y_{\max}}{p^2 n_c}\left(\tfrac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot \Pr(\mathcal{S}\subseteq\mathcal{U})\cdot \mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\right).$$

In addition,

$$\begin{aligned}
&\mathrm{Var}\left(\tfrac{q}{np}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U},|\Pi(\mathcal{S})|\geq 2\big)\right)\\
&= \tfrac{q^2}{n^2p^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{i'=1}^{n}\sum_{\mathcal{S}'\in\mathcal{S}_{i'}^\beta\setminus\varnothing} c_{i',\mathcal{S}'}\cdot\mathbb{I}\big(|\Pi(\mathcal{S}')|\geq 2\big)\cdot\mathrm{Cov}\left(\mathbb{I}\big(\mathcal{S}\subseteq\mathcal{U}\big),\mathbb{I}\big(\mathcal{S}'\subseteq\mathcal{U}\big)\right)\\
&\leq \tfrac{q^2}{n^2p^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{i'=1}^{n}\sum_{\mathcal{S}'\in\mathcal{S}_{i'}^\beta\setminus\varnothing} c_{i',\mathcal{S}'}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\Big(\Pi(\mathcal{S})\cap\Pi(\mathcal{S}')\neq\emptyset\Big)\\
&\leq \tfrac{q^2}{p^2n^2}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{i'=1}^{n}\mathbb{I}\Big(\Pi(\mathcal{N}_i)\cap\Pi(\mathcal{N}_i')\neq\emptyset\Big)\sum_{\mathcal{S}'\in\mathcal{S}_{i'}^\beta\setminus\varnothing} c_{i',\mathcal{S}'}\\
&\leq \tfrac{q^2Y_{\max}}{p^2n}\left(\tfrac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\sum_{i'=1}^{n}\mathbb{I}\Big(\Pi(\mathcal{N}_i)\cap\Pi(\mathcal{N}_i')\neq\emptyset\Big)\right)\\
&\leq \tfrac{q^2d^2Y_{\max}}{p^2n_c}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\Pr(\mathcal{S}\subseteq\mathcal{U})\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\right).
\end{aligned}$$

Putting it all together, we get that

$$\begin{aligned}
\operatorname*{Var}_{\mathcal{U}}\left[\mathop{\mathbb{E}}_{\mathbf{z}}\left(\widehat{\mathrm{TTE}}_{\mathrm{PI}}\mid\mathcal{U}\right)\right] &\leq \tfrac{q-p}{p(n_c-1)}\cdot\widehat{\mathrm{Var}}\big(\bar{L}_\pi\big)+\left(\tfrac{d\beta Y_{\max}}{n_c}+\tfrac{d^2Y_{\max}}{n_c}\right)\left(\tfrac{q^2}{p^2n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\right)\\
&\leq \tfrac{q-p}{p(n_c-1)}\cdot\widehat{\mathrm{Var}}\big(\bar{L}_\pi\big)+\left(\tfrac{d\beta}{n_c}+\tfrac{d^2}{n_c}\right)Y_{\max}\left(\tfrac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\right)\\
&\leq \tfrac{q-p}{p(n_c-1)}\cdot\widehat{\mathrm{Var}}\big(\bar{L}_\pi\big)+\left(\tfrac{d\beta}{n_c}+\tfrac{d^2}{n_c}\right)\cdot Y_{\max}C(\delta(\Pi))\\
&\leq \tfrac{q-p}{p(n_c-1)}\cdot\widehat{\mathrm{Var}}\big(\bar{L}_\pi\big)+\tfrac{2d^2}{n_c}\cdot Y_{\max}C(\delta(\Pi)).
\end{aligned}$$

Here, the second inequality uses the fact that

$$\Pr(\mathcal{S}\subseteq\mathcal{U})\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big)\leq (p/q)^2\cdot\mathbb{I}\big(|\Pi(\mathcal{S})|\geq 2\big).$$

□

Theorem 4.7

Proof. When $\beta = 1$, the estimator simplifies to

$$\widehat{\mathrm{TTE}} = \tfrac{1}{np}\sum_{i\in[n]}\Big(Y_i(\mathbf{z}^1)-Y_i(\mathbf{0})\Big) = \tfrac{1}{np}\sum_{j\in[n]}L_jz_j^1 = \tfrac{1}{np}\sum_{\pi}\sum_{j\in\pi}L_jz_j^1.$$

Conditioning on $\mathcal{U}$, the estimator becomes

$$\tfrac{1}{np}\sum_{j\in\mathcal{U}}L_jz_j^1 = \tfrac{1}{q|\mathcal{U}|}\sum_{j\in\mathcal{U}}L_jz_j^1,$$

where here we use the fact that $|\mathcal{U}| = \frac{np}{q}$. Since $\mathbf{z}_{\mathcal{U}} \sim \text{CRD}(q|\mathcal{U}|, q)$, we may use Lemma 4.6 to obtain an expression for the conditional variance:

$$\underset{\mathbf{z}|\mathcal{U}}{\text{Var}}\left(\widehat{\text{TTE}}_{\text{PI}}\right) = \tfrac{1-q}{q(|\mathcal{U}|-1)} \cdot \left[\tfrac{1}{|\mathcal{U}|}\sum_{j\in\mathcal{U}} L_j^2 \;-\; \left(\tfrac{1}{|\mathcal{U}|}\sum_{j\in\mathcal{U}} L_j\right)^2\right].$$

Taking the expectation of this conditional variance with respect to $\mathcal{U}$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{U}}\left[\underset{\mathbf{z}|\mathcal{U}}{\text{Var}}\left(\widehat{\text{TTE}}_{\text{PI}}\right)\right] &= \tfrac{1-q}{q(|\mathcal{U}|-1)}\left[\tfrac{1}{|\mathcal{U}|}\sum_{j\in[n]} L_j^2 \cdot \Pr(j\in\mathcal{U}) \;-\; \tfrac{1}{|\mathcal{U}|^2}\sum_{j\in[n]}\sum_{j'\in[n]} L_j L_{j'} \cdot \Pr(j,j'\in\mathcal{U})\right] \\
&= \tfrac{1-q}{np-q}\left[\tfrac{q}{np}\sum_{j\in[n]} L_j^2 \cdot \Pr(j\in\mathcal{U}) \;-\; \tfrac{q^2}{n^2p^2}\sum_{j\in[n]}\sum_{j'\in[n]} L_j L_{j'} \cdot \Pr(j,j'\in\mathcal{U})\right] \\
&= \tfrac{1-q}{np-q}\left[\tfrac{1}{n}\sum_{j\in[n]} L_j^2 - \tfrac{q}{n^2p}\sum_{j\in[n]}\sum_{j'\in\pi(j)} L_j L_{j'} - \tfrac{pn_c-q}{n^2p(n_c-1)}\sum_{j\in[n]}\sum_{j'\notin\pi(j)} L_j L_{j'}\right] \\
&= \tfrac{1-q}{np-q}\left[\tfrac{1}{n}\sum_{j\in[n]} L_j^2 - \tfrac{q}{n^2p}\sum_{\pi\in\Pi}\Big(\sum_{j\in\pi} L_j\Big)^2 - \tfrac{pn_c-q}{n^2p(n_c-1)}\Big[\Big(\sum_{\pi\in\Pi}\sum_{j\in\pi} L_j\Big)^2 - \sum_{\pi\in\Pi}\Big(\sum_{j\in\pi} L_j\Big)^2\Big]\right] \\
&= \tfrac{1-q}{np-q}\left[\tfrac{1}{n}\sum_{j\in[n]} L_j^2 - \tfrac{q}{n^2p}\sum_{\pi\in\Pi}\Big(\sum_{j\in\pi} L_j\Big)^2 - \tfrac{pn_c-q}{n^2p(n_c-1)}\Big(\sum_{j\in[n]} L_j\Big)^2 + \tfrac{pn_c-q}{n^2p(n_c-1)}\sum_{\pi\in\Pi}\Big(\sum_{j\in\pi} L_j\Big)^2\right] \\
&= \tfrac{1-q}{np-q}\left[\tfrac{1}{n}\sum_{j\in[n]} L_j^2 + \tfrac{(p-q)n_c}{n^2p(n_c-1)}\sum_{\pi\in\Pi}\Big(\sum_{j\in\pi} L_j\Big)^2 - \tfrac{pn_c-q}{p(n_c-1)}\Big(\tfrac{1}{n}\sum_{j\in[n]} L_j\Big)^2\right] \\
&= \tfrac{1-q}{np-q}\left[\Big[\tfrac{1}{n}\sum_{j\in[n]} L_j^2 - \Big(\tfrac{1}{n}\sum_{j\in[n]} L_j\Big)^2\Big] + \tfrac{p-q}{p(n_c-1)}\cdot\tfrac{1}{n_c}\sum_{\pi\in\Pi}\Big(\tfrac{n_c}{n}\sum_{j\in\pi} L_j\Big)^2 - \tfrac{p-q}{p(n_c-1)}\Big(\tfrac{1}{n_c}\sum_{\pi\in\Pi}\tfrac{n_c}{n}\sum_{j\in\pi} L_j\Big)^2\right] \\
&= \tfrac{1-q}{np-q}\left[\widehat{\underset{j\in[n]}{\text{Var}}}\big(L_j\big) + \tfrac{p-q}{p(n_c-1)}\Big[\tfrac{1}{n_c}\sum_{\pi\in\Pi}\big(\bar{L}_\pi\big)^2 - \Big(\tfrac{1}{n_c}\sum_{\pi\in\Pi}\bar{L}_\pi\Big)^2\Big]\right] \\
&= \tfrac{1-q}{np-q}\left[\widehat{\underset{j\in[n]}{\text{Var}}}\big(L_j\big) + \tfrac{p-q}{p(n_c-1)}\cdot\widehat{\underset{\pi\in\Pi}{\text{Var}}}\big(\bar{L}_\pi\big)\right].
\end{aligned}$$

The conditional expectation is given by

$$\mathop{\mathbb{E}}_{\mathbf{z}}\left[\widehat{\text{TTE}} \,\middle|\, \mathcal{U}\right] = \tfrac{q}{pn_c}\sum_{\pi\in\Pi}\Big(\tfrac{n_c}{n}\sum_{j\in\pi} L_j\Big)\cdot\mathbb{I}(\pi\subseteq\mathcal{U}) = \tfrac{q}{pn_c}\sum_{\pi\in\Pi}\bar{L}_j\cdot\mathbb{I}(\pi\subseteq\mathcal{U}).$$

Since these indicator random variables are sampled in Stage 1 according to a $\text{CRD}(pn_c/q, n_c)$ distribution, we may apply Lemma 4.6 to conclude that

$$\text{Var}\left(\mathop{\mathbb{E}}_{\mathbf{z}|\mathcal{U}}\left[\widehat{\text{TTE}}_{\text{PI}}\right]\right) = \tfrac{1-(p/q)}{(p/q)(n_c-1)}\cdot\widehat{\underset{\pi\in\Pi}{\text{Var}}}\big(\bar{L}_\pi\big) = \tfrac{q-p}{p(n_c-1)}\cdot\widehat{\underset{\pi\in\Pi}{\text{Var}}}\big(\bar{L}_\pi\big).$$

Putting this together, we find that

$$\text{Var}\left(\widehat{\text{TTE}}\right) = \tfrac{1-q}{np-q}\cdot\widehat{\underset{j\in[n]}{\text{Var}}}\big(L_j\big) + \tfrac{(p-q)(1-np)}{p(np-q)(n_c-1)}\cdot\widehat{\underset{\pi\in\Pi}{\text{Var}}}\big(\bar{L}_\pi\big).$$

□

B Experiment Details

B.1 Potential Outcomes Model

We generate synthetic potential outcomes based on a generalization of the response model from Ugander and Yin (2023) to incorporate β-order interactions:

$$Y_i(\mathbf{z}) = Y_i(\mathbf{0})\cdot\left(1+\delta z_i + \sum_{k=1}^{\beta}\gamma_k\cdot\binom{d_i}{k}^{-1}\sum_{\substack{\mathcal{S}\in\mathcal{S}_i^\beta\\|\mathcal{S}|=k}}\prod_{j\in\mathcal{S}} z_j\right), \qquad Y_i(\mathbf{0}) = \left(a + b\cdot h_i + \varepsilon_i\right)\cdot\tfrac{d_i}{\bar{d}}.$$

In this model:

- a is a baseline effect. We select $a = 1$.
- $\mathbf{h} \in \mathbb{R}^n$ is a Fiedler vector of the graph Laplacian of the network which has undergone an affine transformation so that $\min(\mathbf{h}) = -1$ and $\max(\mathbf{h}) = 1$. This models network homophily effects.
- b controls the magnitude of the homophily effect. We select $b = 0$. We also ran the experiments with $b = 0.5$, to compare no homophily with some homophily, but the analysis and conclusions do not change. These are included later in the appendix.
- $\varepsilon_i \underset{\text{iid}}{\sim} N(0, \sigma)$ is a random perturbation of the baseline effect. We select $\sigma = 0.1$.
- d_i is the in-degree of vertex i. $\overline{d}$ is the average in-degree.
- δ is uniform direct effect on treated individuals. We select $\delta = 0.5$.
- γ_k is the effect of treated subsets of size k. We select $\gamma_k = 0.5^{k-1}$, which models marginal effects that decay with the size of the treated set.

B.2 Details of Real-World Networks

Here, we provide more details of the three real-world data sets we use in our analysis. We include all the raw data files, cleaned data, and processing scripts in our provided source code. A summary of the datasets is given in the following table.

Dataset	Vertices	Edges	Degree	Features
EMAIL Leskovec et al. (2007b); Yin et al. (2017); Leskovec and Krevl (2014)	employees $n = 1,005$	correspondence directed $\lvert E\rvert = 25,571$	min: 1 max: 334 average: 25	department $\lvert F\rvert = 42$
BLOGCATALOG Rossi and Ahmed (2015); Tang and Liu (2009b,a)	bloggers $n = 10,312$	connections undirected $\lvert E\rvert = 333,983$	min: 1 max: 3,992 average: 65	interests $\lvert F\rvert = 39$
AMAZON Leskovec et al. (2007a); Leskovec and Krevl (2014)	products $n = 14,436$	co-purchases directed $\lvert E\rvert = 70,832$	min: 1 max: 247 average: 5	category $\lvert F\rvert = 13,591$

Email

The EMAIL dataset is publicly available at `https://snap.stanford.edu/data/email-Eu-core.html` and is licensed under the BSD license[1]. This dataset models the email communications between members of a European research institution. The $n = 1,005$ vertices of the network are (anonymized) institution members, and there is a directed edge from individual i to individual j if i has sent at least one email to individual j.

It has a minimum degree of 1, a maximum degree of 212, and an average degree of 25.8, and its degree distribution is visualized in Figure 5; the support of the histogram has been cropped to remove some large outliers. The largest weakly connected component in the network contains 986 vertices, and the largest strongly connected component contains 803 vertices.

Each individual in the network has been assigned one of 42 department labels. The sizes of these departments vary greatly, with the smallest department including a single individual and the largest department including 109 individuals. The average department size is 23.9. In the EMAIL network, each vertex is assigned to exactly one department, and we use these assignments as our clustering. To pre-process this data for use in our experiments, we added self-loops to each node in the original dataset to represent the direct effect of the node's treatment on their outcome (See Section 2).

BlogCatalog

The BLOGCATALOG dataset is publicly available at `https://networkrepository.com/soc-BlogCatalog-ASU.php` and is licensed under a Creative Commons Attribution-ShareAlike License[2]. This dataset models the rela-

[1] For more information, see `https://snap.stanford.edu/snap/license.html` and `https://groups.google.com/g/snap-datasets/c/52MRzGbMkFg/m/FIFy_6qOCAAJ`

[2] For more information, see `https://networkrepository.com/policy.php`

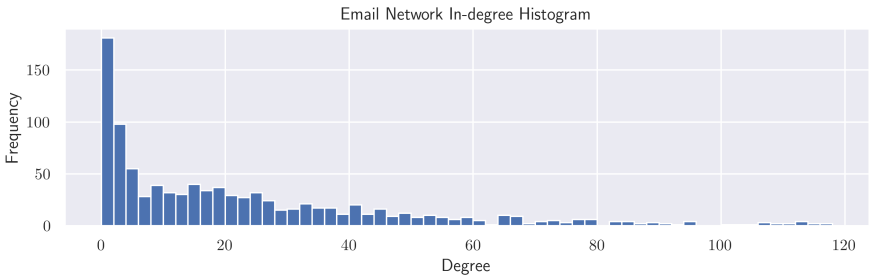


Figure 5: The degree distribution of the EMAIL graph

tionships between bloggers on the (now defunct) blogging website http://www.blogcatalog.com. The $n = 10,312$ nodes represent bloggers and the (undirected) edges represent the social network of the bloggers.

The network has a minimum degree of 1, a maximum degree of 3,992, and an average degree of 65, and its degree distribution is visualized in Figure 6; the support of the histogram has been cropped to remove some large outliers. The average clustering coefficient is approximately 0.46.

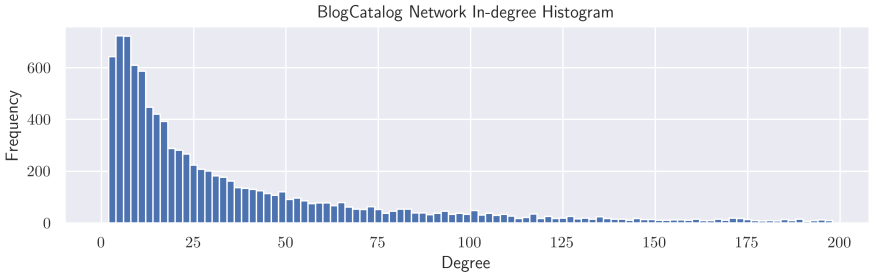


Figure 6: The degree distribution of the BLOGCATALOG graph

Each blogger in the network has an associated blog. Blogs (and thus, bloggers) are organized under interest categories specified by the website and can be listed under multiple categories. There are 39 such categories in this dataset and on average; on average, each blogger is listed under 1.6 categories. As part of the data pre-processing for our experiments, we added self-loops to each node in the original dataset, as we did with the EMAIL dataset.

Amazon

The AMAZON dataset is publicly available at `https://snap.stanford.edu/data/amazon-meta.html` and is licensed under the BSD license[3]. This dataset models an Amazon product co-purchasing network. The $n = 14,436$ nodes represent products and each node has outgoing edges to the top 5 products with which it is a frequent co-purchase. Thus, in addition to the self-loop at each node, each node has exactly 5 outgoing edges.

The network has a minimum in-degree of 1, a maximum in-degree of 247, and an average in-degree of 5; its in-degree distribution is visualized in Figure 7; the support of the histogram has been cropped to remove some large outliers.

Products are organized into categories (which correspond to attributes such as the genre, setting, and actors in the film, as well as marketplace data such as the inclusion of these titles in certain deals or promotions) but can belong to multiple categories. There are 13,591 possible product categories; on average each product belongs to 13.2 categories. As part of the data pre-processing for our experiments, we added all self-loops and restricted the original dataset to only the product nodes labeled as DVDs.

[3]For more information, see `https://snap.stanford.edu/snap/license.html` and `https://groups.google.com/g/snap-datasets/c/52MRzGbMkFg/m/FIFy_6qOCAAJ`

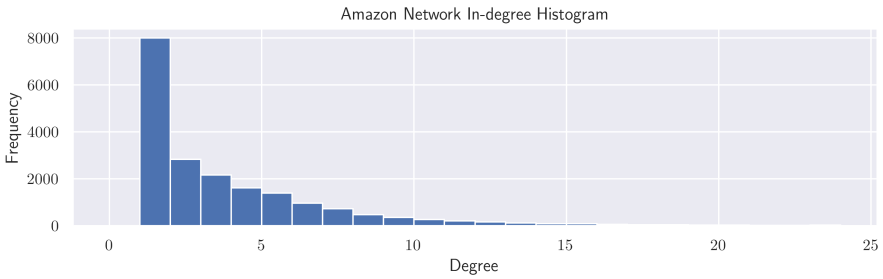


Figure 7: The degree distribution of the AMAZON graph

B.3 Other Estimators

Here, we provide additional details of the other estimators used in our experiments Section 5.

Difference-in-Means

The difference-in-means (DM) approach estimates the total treatment effect as the difference in the average outcome of a treated individual and the average outcome of an untreated individual.

$$\widehat{\mathrm{TTE}}_{\mathrm{DM}} = \frac{\sum_i z_i \cdot Y_i(\mathbf{z})}{\sum_i z_i} - \frac{\sum_i (1-z_i) \cdot Y_i(\mathbf{z})}{\sum_i (1-z_i)}.$$

This estimator does not utilize any knowledge of the underlying graph. It only requires knowledge of the treatment assignments and the observed outcomes, which are always known to the experimenter. Under SUTVA, this is an unbiased estimator (since an individual's outcome is a function only of their treatment). However, it is not unbiased under interference; untreated individuals may have treated neighbors which impacts their outcome, introducing bias into our signal of the baseline effects.

To counteract this bias, at the expense of requiring network knowledge, we can limit the set of individuals used in the estimator to those whose neighbors' treatment assignments largely align with theirs. We refer to this as a thresholded DM estimator.

Thresholded Difference-in-Means:

This family of estimators is parameterized by a value $\gamma \in [0,1]$, which can be viewed as a stringency requirement that we place on the treatment assignments within one's neighborhood. In particular, we only include an individual in the "treated" set in this DM estimator if they are treated and at least a γ fraction of their neighbors are also treated. Similarly, we will only include an individual in the "untreated" set in this DM estimator if they are untreated and at most a $(1-\gamma)$ fraction of their neighbors are treated.

$$\widehat{\mathrm{TTE}}_{\mathrm{DM}(\gamma)} = \frac{\sum_i z_i \cdot Y_i(\mathbf{z}) \cdot \mathbb{I}\big(\sum_{j\in\mathcal{N}_i} z_i \geq \gamma d_i\big)}{\sum_i z_i \cdot \mathbb{I}\big(\sum_{j\in\mathcal{N}_i} z_i \geq \gamma d_i\big)} - \frac{\sum_i (1-z_i) \cdot Y_i(\mathbf{z}) \cdot \mathbb{I}\big(\sum_{j\in\mathcal{N}_i} z_i \leq (1-\gamma) d_i\big)}{\sum_i (1-z_i) \cdot \mathbb{I}\big(\sum_{j\in\mathcal{N}_i} z_i \leq (1-\gamma) d_i\big)}$$

Note that these estimators for $\gamma > 0$ require network knowledge to calculate the neighborhood treatment proportions, and they are biased under interference for the same reasoning as the standard DM estimator. Note that DM(0) (i.e., the thresholded DM estimator with parameter $\gamma = 0$) coincides with the ordinary DM estimator. The DM(1) estimator will only consider individuals with fully treated or untreated neighborhoods. As such, the DM(1) estimator will, under simpler randomization schemes like Bernoulli design, include very few individuals in its "treated" and "untreated" sets. The Horvitz-Thompson and Hájek estimators also exhibit this phenomenon.

Horvitz-Thompson:

The Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) uses inverse probability weighting to construct an unbiased TTE estimator under *arbitrary* potential outcomes models. To do this, it must only incorporate the outcomes from an individual's neighborhoods that are either fully treated or fully untreated, as these are the only outcomes that appear in the TTE estimand. The estimator has the form,

$$\widehat{\mathrm{TTE}}_{\mathrm{HT}} = \tfrac{1}{n} \sum_{i\in[n]} \frac{Y_i(\mathbf{z}) \cdot \mathbb{I}(\mathcal{N}_i \text{ fully treated})}{\Pr(\mathcal{N}_i \text{ fully treated})} - \tfrac{1}{n} \sum_{i\in[n]} \frac{Y_i(\mathbf{z}) \cdot \mathbb{I}(\mathcal{N}_i \text{ fully untreated})}{\Pr(\mathcal{N}_i \text{ fully untreated})}$$

$$= \frac{1}{n} \sum_{i \in [n]} \left[\frac{Y_i(\mathbf{z}) \cdot \prod_{j \in \mathcal{N}_i} z_j}{\Pr\left(\prod_{j \in \mathcal{N}_i} z_j = 1 \right)} - \frac{Y_i(\mathbf{z}) \cdot \prod_{j \in \mathcal{N}_i} (1 - z_j)}{\Pr\left(\prod_{j \in \mathcal{N}_i} (1 - z_j) = 1 \right)} \right]$$

This estimator is unbiased, but it relies on network knowledge to compute the exposure probabilities in the denominators of each fraction. A related inverse probability weighted estimator is the Hájek estimator.

Hájek

Since the HT estimator only considers individuals with fully treated and fully untreated neighborhoods, most of the bracketed terms within its summation will be 0. To compensate for this, we can adjust the $1/n$ normalization on the summation to use the expected number of non-zero entries corresponding to both terms in the bracketed expression. This change gives the Hájek estimator (Basu, 2011).

$$\widehat{\mathrm{TTE}}_{\text{Hájek}} = \frac{\sum_{i \in [n]} \frac{Y_i(\mathbf{z}) \cdot \prod_{j \in \mathcal{N}_i} z_j}{\Pr\left(\prod_{j \in \mathcal{N}_i} z_j = 1 \right)}}{\sum_{i \in [n]} \frac{\prod_{j \in \mathcal{N}_i} z_j}{\Pr\left(\prod_{j \in \mathcal{N}_i} z_j = 1 \right)}} - \frac{\sum_{i \in [n]} \frac{Y_i(\mathbf{z}) \cdot \prod_{j \in \mathcal{N}_i} (1 - z_j)}{\Pr\left(\prod_{j \in \mathcal{N}_i} (1 - z_j) = 1 \right)}}{\sum_{i \in [n]} \frac{\prod_{j \in \mathcal{N}_i} (1 - z_j)}{\Pr\left(\prod_{j \in \mathcal{N}_i} (1 - z_j) = 1 \right)}}$$

The Hájek estimator trades off a reduction in the variance over the HT estimator for the introduction of some bias (a thorough discussion of this tradeoff is given by Khan and Ugander (2023)). As with the Horvitz-Thompson estimator, the calculation of exposure probabilities in this estimator requires knowledge of the interference network.

Two-Stage Estimator when $q = 1$

The one-stage estimator from Cortez et al. (2022) is

$$\widehat{\mathrm{TTE}}^{\beta}_{\text{1-Stage}} := \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{\beta} \Big(\ell_{t,p}(1) - \ell_{t,p}(0) \Big) \cdot Y_i(\mathbf{z}^t), \quad \ell_{t,p}(x) = \prod_{\substack{s=0 \\ s \neq t}}^{\beta} \tfrac{\beta x - ps}{pt - ps}. \tag{B.1}$$

When evaluating the estimator with $\beta = 1$, the estimator is simply

$$\widehat{\mathrm{TTE}}^{\beta=1}_{\text{1-Stage}} = \frac{1}{np} \sum_{i=1}^{n} \Big(Y_i(\mathbf{z}^1) - Y_i(\mathbf{z}^0) \Big) \tag{B.2}$$

In what follows, we show that the two-stage rollout estimator with $q = 1$ is equivalent to $\widehat{\mathrm{TTE}}^{\beta=1}_{\text{1-Stage}}$.

Theorem B.1. *Under a Two-Stage Rollout Design with budget p and effective treatment budget $q = 1$, the two-stage estimator defined in equation* (3.1) *under a model with degree β is equivalent to the estimator defined in* (B.2)*.*

Proof. Under a Two-Stage Rollout Design with $q = 1$, we have

$$h_{t,q} = h_{t,1} = \prod_{\substack{s=0 \\ s \neq t}}^{\beta} \tfrac{\beta - s}{t - s} \; - \; \prod_{\substack{s=0 \\ s \neq t}}^{\beta} \tfrac{-s}{t - s}.$$

Notice that when $t \in \{1, 2, \ldots, \beta - 1\}$, i.e. $t \neq 0$ and $t \neq \beta$, at some point we have a term corresponding to $s = 0$ and $s = \beta$ in the products above. Thus, both products are 0.

When $t = 0$, we have $h_{0,1} = \prod_{s=1}^{\beta} \frac{\beta - s}{0 - s} - \prod_{s=1}^{\beta} \frac{-s}{0 - s} = \prod_{s=1}^{\beta} \frac{\beta - s}{-s} - 1 = -1$ because the product equals 0 due to the $s = \beta$ term. Similarly, when $t = \beta$, we have $h_{\beta,1} = \prod_{s=0}^{\beta-1} \frac{\beta - s}{\beta - s} - \prod_{s=0}^{\beta-1} \frac{-s}{\beta - s} = 1$ since the second product will equal 0 due to the $s = 0$ term. To summarize, we have

$$h_{t,q} = \begin{cases} -1 & t = 0 \\ 0 & 1 \leq t \leq \beta - 1 \\ 1 & t = \beta \end{cases}.$$

□

B.4 Additional Experiments: Comparing Different Estimators

In this section, we have figures showing the MSE of different estimators as we vary the treatment budget p from 0.1 to 0.5 for different model degrees β and different real-world networks. As a reminder, we compare the following estimators:

- The two-stage polynomial interpolation estimator with $q = 0.5$ and no clustering, 2-Stage
- The two-stage polynomial interpolation estimator with $q = 1$ and no clustering, q=1
- The one-stage polynomial interpolation estimator, PI, from Cortez et al. (2022)
- The simple difference-in-means estimator, DM
- The thresholded difference-in-means estimator with parameter 0.75, DM(0.75)
- The Hájek estimator, Hájek

The first three estimators in this list are based on polynomial interpolation (PI), so we refer to them as the PI estimators. We refer to the others as the non-PI estimators. In all MSE plots, the lines indicate the empirical MSE over 1000 replications. In all bias and standard deviation plots, the bold line indicates the mean over 1000 replications, and the shading indicates the experimental standard deviation, calculated by taking the square root of the experimental variance over all replications.

In Figure 8, we show the MSE corresponding to Figure 2 from Section 5. The column faceting indicates model degree; note that the y-axis limits differ across these subplots. When $\beta = 1$, PI and q=1 are equivalent and have slightly lower MSE compared with 2-Stage. However, the difference is hard to see without zooming in further since the lines almost overlap. Note that the difference-in-means estimators have MSE outside the bounds of the plots. When $\beta = 2$, the results are similar but you can start to see the difference between the three PI estimators, which all have lower MSE when compared with the non-PI estimators. For smaller values of p, the estimator PI has slightly lower MSE, followed by 2-Stage, followed by q=1. Again, the difference is quite small. In this case, as noted in the main body of the paper, we are in a setting where using the one-stage rollout and estimator is preferable. When $\beta = 3$, the difference between the three PI estimators is more pronounced. The 2-Stage has the lowest MSE, especially for lower values of p. The q=1 estimator has MSE relatively close to it for all p-values, but does slightly worse, although better than PI for small p values. In this case, we have a setting where the two-stage approach is valuable as it outperforms the other methods.

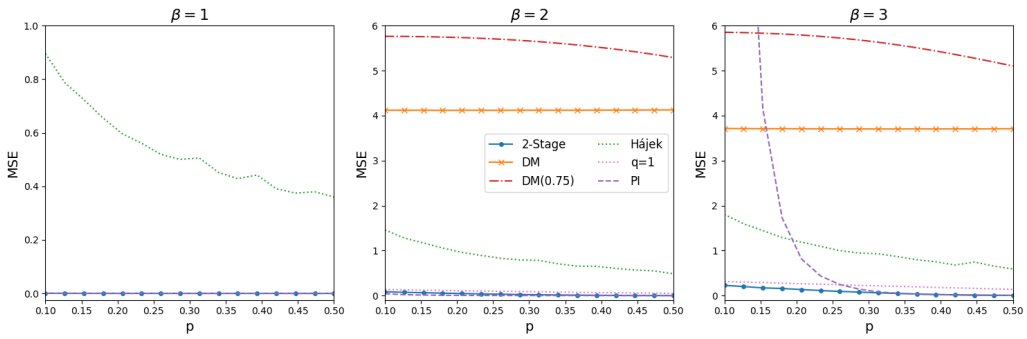
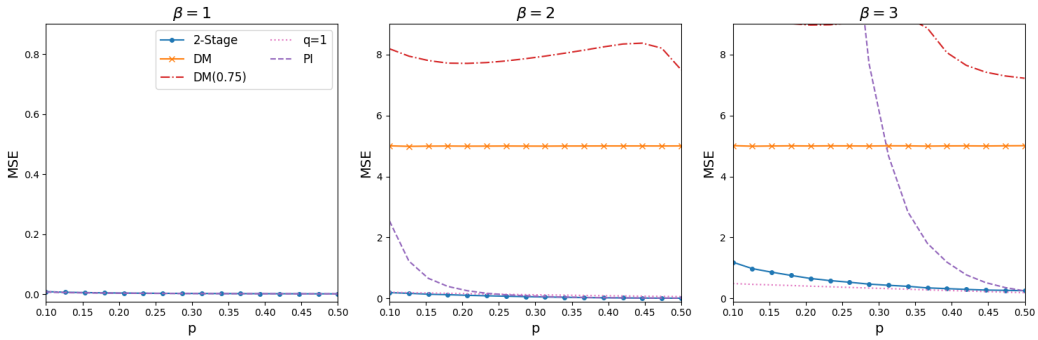
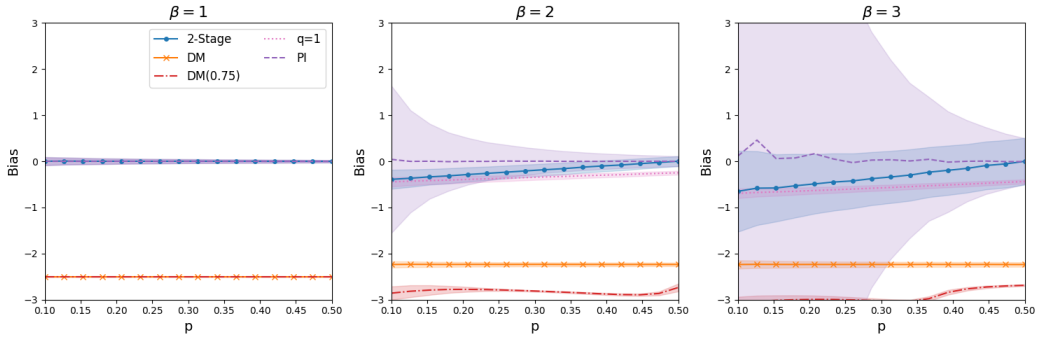


Figure 8: Amazon Network. MSE of different estimators as a function of treatment budget p.

In Figure 9, we show the MSE and the bias and standard deviation of the different estimators under the BlogCatalog network. We omit the Hájek estimator because the network degree is very high; under unit randomization, the estimator is often undefined. In all cases, the two difference-in-means estimators are very biased, so their MSE is much worse than the PI estimators. Similar to the Amazon network, when $\beta = 1$ the PI estimators are almost indistinguishable, with 2-Stage coming in with slightly higher MSE due to a small increase in variance. The difference-in-means estimators do not appear on the plot since their MSE exceeds the plotting range. When $\beta = 2$, we see that PI has higher MSE for smaller values of p due to an increased variance (the estimator is unbiased). 2-Stage has much lower variance than PI and its bias decreases as p approaches 0.5. q=1 has even lower variance than 2-Stage and similar bias, but its bias remains worse than 2-Stage. However, their MSE remains comparable. When $\beta = 3$, there is a much clearer difference between the PI estimators. For most



(a) MSE of different estimators as a function of treatment budget p.



(b) Bias and variance of different estimators as a function of treatment budget p.

Figure 9: BlogCatalog Network.

values of p, PI has the worst variance (due to extrapolation with a richer model), while q=1 has the worst bias (due to the heavier reliance on the subsampling). Meanwhile, 2-Stage is in between, with lower bias, but larger variance than q=1 and with higher bias, but lower variance than PI. In terms of MSE, q=1 outperforms the other estimators. Recall that the q=1 estimator is equivalent to the $\beta = 1$ version of the one-stage PI estimator. This setting shows that although the two-stage approach can greatly reduce error over the one-stage approach even without clustering, an even simpler design (one-stage rollout over just two time steps) and estimator (using observations from the two time steps) can still outperform.

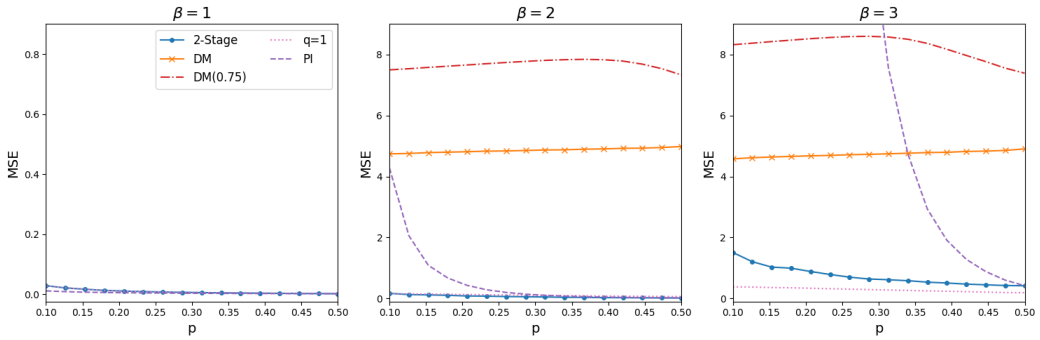
The results from Figure 10 are the same as Figure 9.

B.5 Additional Experiments: Comparing Different Clusterings

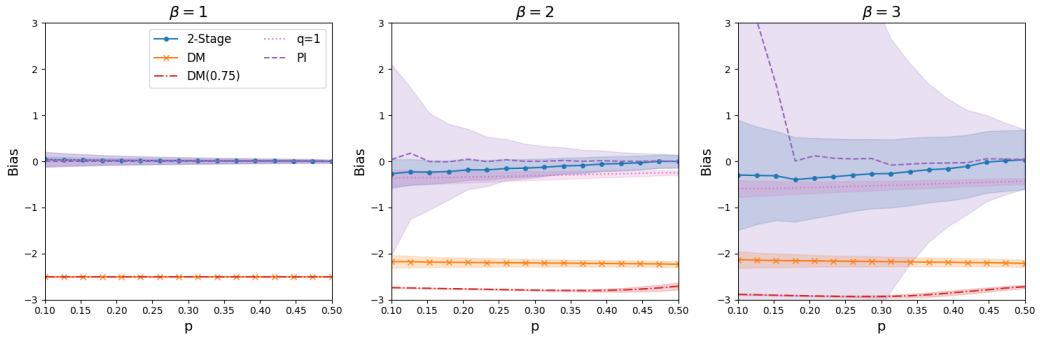
We compare the performance of the 2-Stage estimator under two clustering methods versus no clustering in the first stage of the experimental design. In the **clustering with full graph knowledge**, we cluster the true underlying graph using the METIS clustering library Karypis and Kumar (1998). In the **clustering with covariate knowledge**, clusters are based on features. When each vertex is assigned to one feature, we use these assignments as the clustering. When vertices may have multiple features we form a feature graph — a weighted graph, where the weight of edge (i, j) is the number of feature labels shared by i and j — and cluster this feature graph using METIS. In all plots, the column faceting indicates the type of clustering and the y-axis varies q on the interval $[p, 1]$, where $p = 0.1$.

We also include tables with various pieces of information pertaining to the performance of the two-stage design and estimator, including clustering metrics such as number of cut edges, the cut effect $C(\delta(\Pi))$, and the empirical variance across clusters of cluster average influences $\widehat{\mathrm{Var}}(\bar{L}_\pi)$. The latter two metrics are defined in Section 3. In each row, $q_{\min}$ is the value of q that minimizes the MSE and the column MSE($q_{\min}$) contains that value.

In Figure 11, we show the results for the BlogCatalog network under a model with degree $\beta = 2$. In this case, the clusterings each have $n_c = 50$ clusters. For this network, clustering does not appear to be of any help in reducing the bias and at worst, under a clustering that uses full graph knowledge, increases variance. Taking a look at the first two rows of Table 4 sheds some light on this. We see that the $\widehat{\mathrm{Var}}(\bar{L}_\pi)$ under a clustering that



(a) MSE of different estimators as a function of treatment budget p.



(b) Bias and variance of different estimators as a function of treatment budget p.

Figure 10: Email Network.

Table 4: Clustering Metrics for BlogCatalog Network

β	Cluster	$\widehat{\mathrm{Var}}(\bar{L}_\pi)$	$C(\delta(\Pi))$	Cut Edges	$q_{\min}$	MSE($q_{\min}$)
2	Full	0.697	0.471	604504	0.5	0.260
2	Covariate	0.059	0.486	643080	0.5	0.190
3	Full	0.703	0.717	604504	1	0.610
3	Covariate	0.060	0.734	643080	1	0.486

uses full graph knowledge is more than ten times higher than a clustering that only uses covariate information. The number of cut edges is similar under both clusterings and thus so is the cut effect. In this example, it would appear that one is better off not clustering at all.

In Figure 12, we show the results for the Email network under a model with degree $\beta = 2$. In this case, the clusterings each have $n_c = 42$ clusters. We see an advantage with clustering on covariates in particular. The highest bias, but lowest variance, is under no clustering. The clustering with full knowledge certainly decreases variance versus no clustering, but at the expense of incurring a lot of variance. The lowest MSE is achieved by the covariate knowledge clustering at $q = 1$, which strikes a balance between bias and variance. Taking a look at Table 5, we see that the $\widehat{\mathrm{Var}}(\bar{L}_\pi)$ term is similar under both clusterings. However, the covariate clustering cuts about a quarter less edges than the full knowledge clustering and thus has a smaller cut effect.

Table 5: Clustering Metrics for Email Network

β	Cluster	$\widehat{\mathrm{Var}}(\bar{L}_\pi)$	$C(\delta(\Pi))$	Cut Edges	$q_{\min}$	MSE($q_{\min}$)
2	Full	0.399	0.442	21756	0.5	0.232
2	Covariate	0.398	0.372	16284	1	0.133
3	Full	0.417	0.686	21756	1	0.483
3	Covariate	0.412	0.591	16284	1	0.288

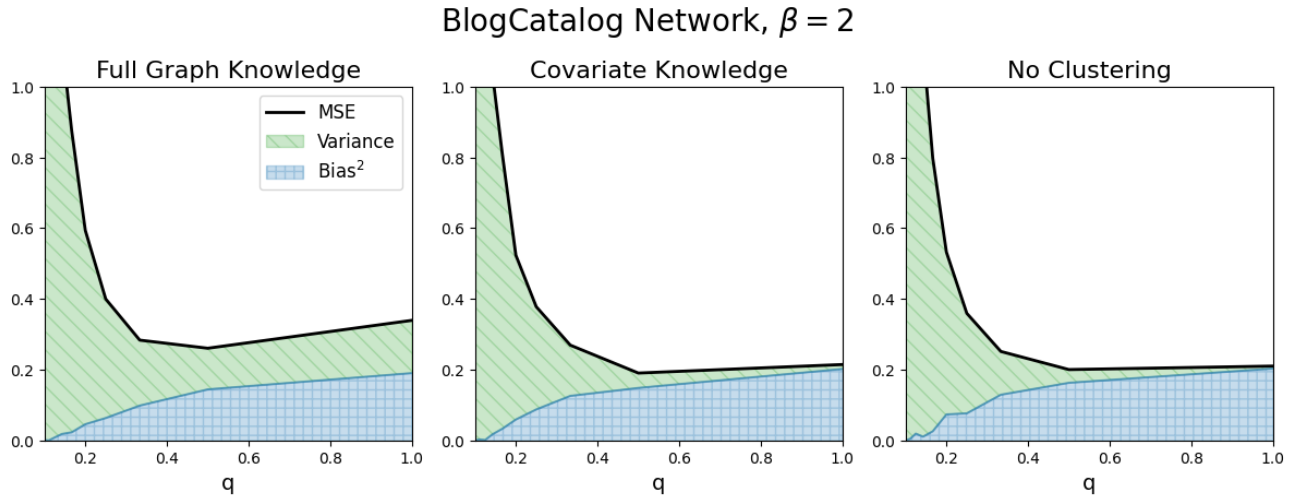


Figure 11: BlogCatalog Network. MSE of 2-Stage estimator under two different clusterings versus no clustering, as a function of q.

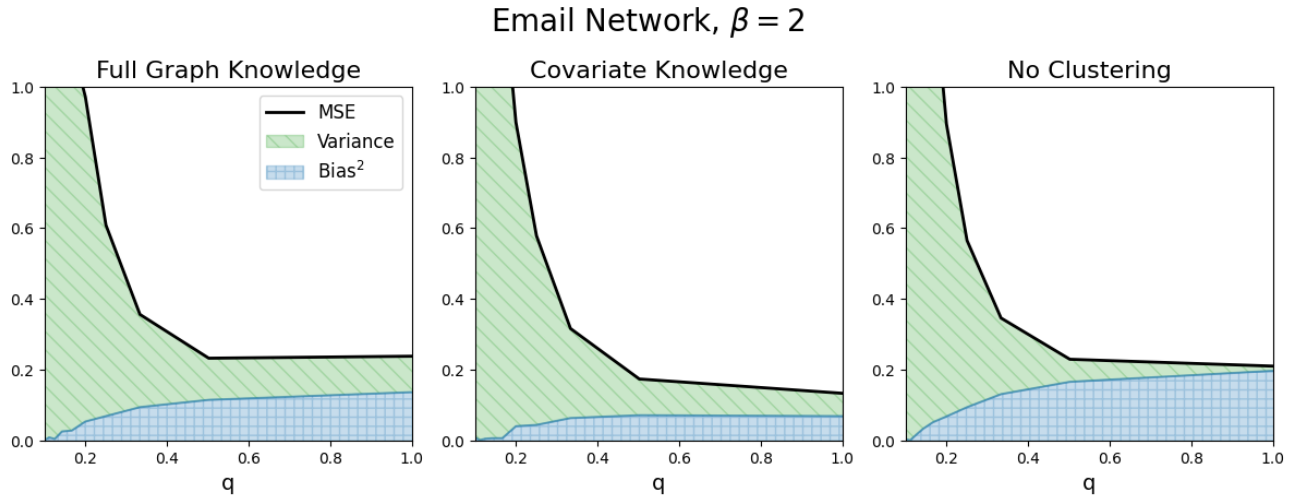
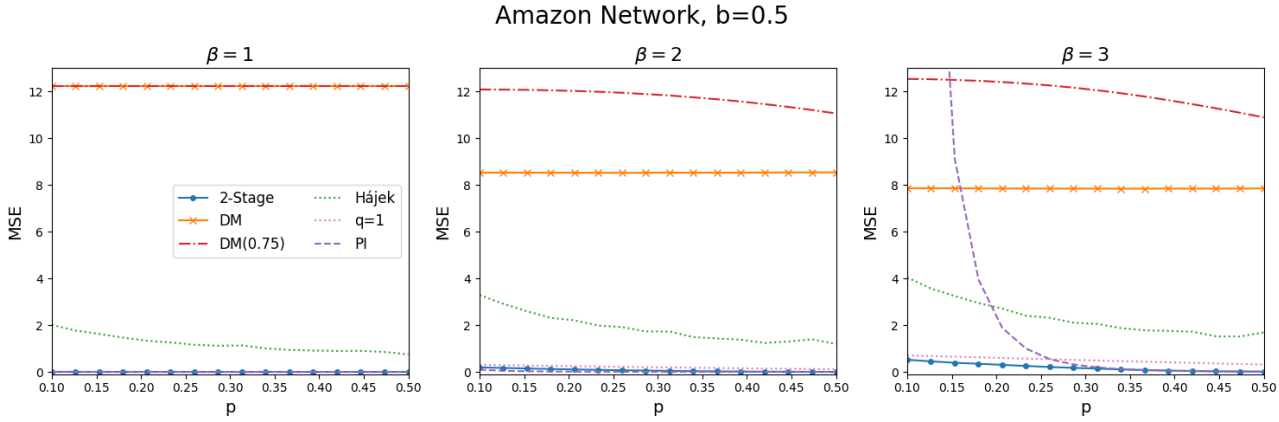


Figure 12: Email Network. MSE of 2-Stage estimator under two different clusterings versus no clustering, as a function of q.

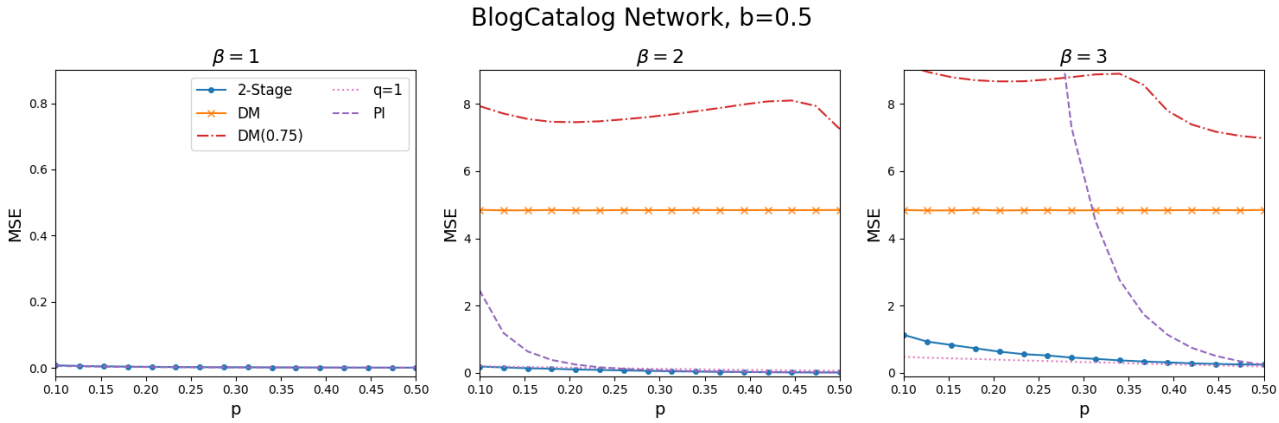
B.6 Additional Experiments: Homophily Parameter $b = 0.5$.

In this section, we show some results when the model exhibits homophily by setting the parameter $b = 0.5$. All other parameters are set to the same values as previous plots. Although there are some small visual differences between the plots in this section and the plots throughout the rest of this work, the analyses and conclusions remain the same. For example, we can compare Figure 8 (where $b = 0$) with 13a (where $b = 0.5$). Both of these show the MSE of different estimators for different values of treatment budgets p and different model degrees β. Notice the difference in the scaling on the y-axis, particularly for $\beta = 2$ and $\beta = 3$. However, the patterns are the same: for most values of p, the two difference in means estimators have the highest MSE, followed by the Hájek estimator, and then followed by the three PI estimators. When $\beta = 3$, the MSE of the vanilla PI estimator is extremely high for small values of p, but gets smaller than the non-PI estimators around $p = 0.2$. In general, for $\beta = 3$, 2-Stage tends to outperform PI for many paramter values and for some networks, q=1 has the smallest MSE in some cases. When $\beta = 2$, we see that the two stage approach improves over the one stage approach under the BlogCatalog and Email networks for small values of p. When $\beta = 1$ the performances of the PI estimators are similar, with 2-Stage performing ever so slightly worse.

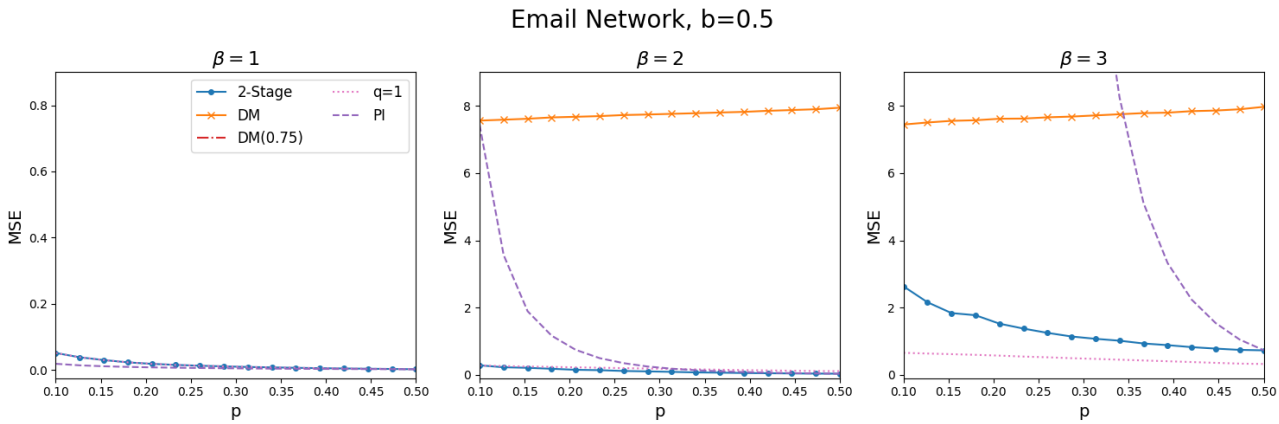
In Figure 14, we show the performance of the two stage approach under two different clustering versus no clustering for three different networks. The BlogCatalog and Email network results are very similar to those in Figures 11 and 12. The Amazon network result sheds some light onto why the scaling is different in the Amazon MSE plots: the bias has a larger magnitude. Part of this is likely attributable to the fact that switching from $b = 0$ to $b = 0.5$ changed the magnitude of the baseline outcomes, and therefore all outcomes.



(a)

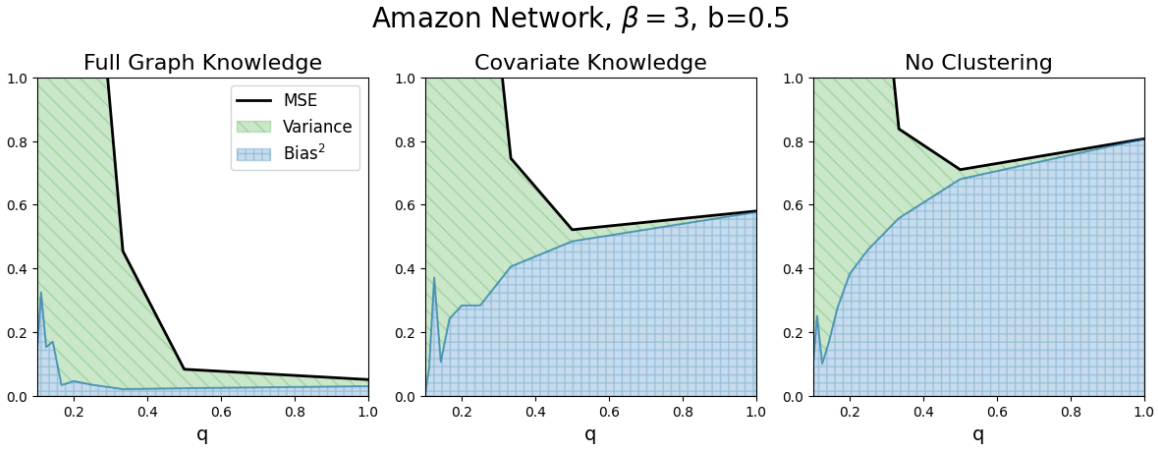


(b)

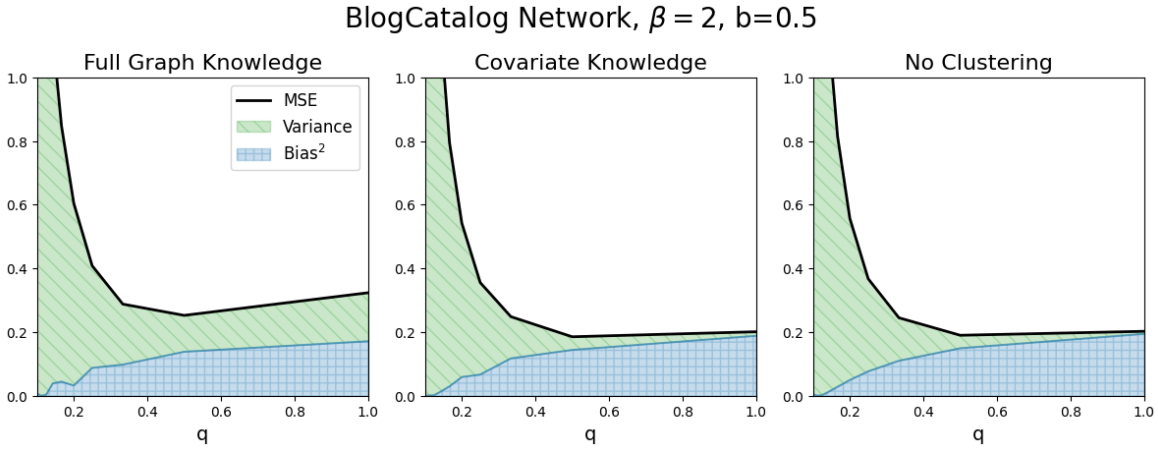


(c)

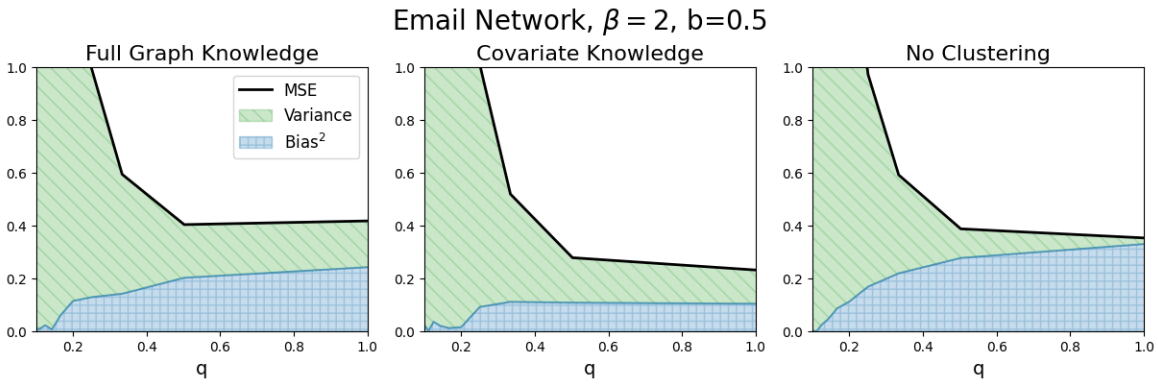
Figure 13



(a)



(b)



(c)

Figure 14