

Graph Agnostic Randomized Experimental Design under Heterogeneous Linear Network Interference

Christina Lee Yu ^{*} Edoardo M Airolidi [†] Christian Borgs [‡] Jennifer T Chayes [§]

Abstract

Randomized experiments are widely used to estimate causal effects of proposed “treatments” in domains spanning across physical sciences, social sciences, medicine, and technology industries. However, classical approaches to experimental design rely on critical independence assumptions that are violated when the outcome of an individual a may be affected by the treatment of another individual b , referred to as network interference. Under such network interference, naïvely using popular estimators and randomized experimental designs can result in significant bias and loss of efficiency. We consider a heterogeneous linear outcomes model that can capture network interference that arises from spillover, peer effects, and contagion. Under this model, we characterize the limitations and possibilities for estimating the total treatment effect, average direct treatment effect, and average interference effect. Given access to average historical baseline measurements prior to the experiment, we propose simple estimators and randomized designs that output unbiased estimates with low variance for these three estimands. Furthermore, our solution and statistical guarantees do not require knowledge of the underlying network structure, and thus can be used for scenarios where the network is unknown and complex. We believe our results are poised to impact current randomized experimentation strategies due to its ease of interpretation and implementation, alongside its provable statistical guarantees under heterogeneous network effects.

1 Introduction

The measurement of treatment effects via randomized experiments is a fundamental tool used across all fields of scientific disciplines and beyond. For example, consider a public health campaign to increase public awareness of the importance of wearing masks during a global pandemic. The administrator in charge of running the public health campaign would like to determine which proposed banner ad would be most effective for displaying on a public billboard. In particular, the administrator would like to estimate the “total treatment effect”, i.e. the change in behavior of the population at large that results from viewing the proposed banner ad. The “total treatment effect” is a causal effect, as it described the change in behavior that is caused by the treatment. The experimental units in this example are the individuals in the population, and outcomes refer to some measurable behavior of individuals, such as whether an individual is wearing a mask or not at the grocery store.

The classical approach to estimating causal effects involves running a randomized experiment, where one randomly partitions the population into a treatment group and a control group. The

^{*}School of Operations Research and Information Engineering, Cornell University, Ithaca NY

[†]Fox School of Business, Temple University, Philadelphia, PA

[‡]Electrical Engineering and Computer Science Department, UC Berkeley, Berkeley, CA

[§]Electrical Engineering and Computer Science Department, UC Berkeley, Berkeley, CA

treatment of interest is administered to each individual in the treatment group, and a placebo is administered to each individual in the control group. The causal effect is then approximated by the difference in measured outcomes or behaviors between the treatment and control groups after the treatment has been administered. This approach results in an efficient unbiased estimate for the desired causal effect under a critical assumption that the outcome of an individual is not affected by the treatment assignment of any other individual; this assumption is referred to in the literature as the Stable Unit Treatment Value Assumption (SUTVA) [12, 27, 23].

Unfortunately, SUTVA is violated in many applications, as individuals are connected in a complex social network that mediates communication, influence, or spread of disease, resulting in network interference that couple the outcomes of individuals. The treatment of individual A may impact the outcome of individual B, violating SUTVA and introducing significant bias to the estimates resulting from the classical experimental approach. As a result, we need new theory for experimental design which can account for these network interference effects, and yet are simple and practical to implement. Next we illustrate a few motivating scenarios.

Example 1 (Social Media Platforms). As humans are social creatures, we influence each other over a variety of different types of communication networks: mobile networks, social media platforms, email exchange networks, collaboration communities, etc. As a result, an experiment to measure the impact of an advertising campaign or a new platform feature may need to account for the social network interference effects that result from the treatment. For example, consider a social media platform such as LinkedIn, which would like to estimate the total treatment effect of a proposed change in the recommendation engine for a user’s news feed, with a goal of increasing user engagement. The change in engagement level of individual A as a result of being exposed to the proposed change could subsequently impact the engagement level of others in individual A’s social network, resulting in a positive or negative network effect.

Example 2 (Targeted Public Health Campaigns). Consider a public health campaign which seeks to educate individuals in the society about some health concern. As information is shared through relationships and common communities, the efficacy of a public health campaign could heavily depend on the interaction between the targetting mechanism and the underlying social network. For example, consider estimating the total treatment effect of a public health campaign to increase the use of masks in public during a pandemic. Suppose that individual A, a senior citizen, is shown the proposed banner ad, and thus decides to wear a mask in public. Individual A’s behavior could cause a positive network effect on their friends, who then decide to also wear a mask even though they did not see the original banner ad. In contrast, individual A’s behavior may have a negative effect on a teenager, who may think that wearing masks must be for the elderly and thus not “cool”.

Example 3 (Vaccine Efficacy for Infectious Diseases). Consider running a clinical trial to estimate the total treatment effect of a proposed vaccine for COVID-19, i.e. how much would the overall rate of cases contracted in the public at large decrease as a result of everyone receiving the vaccine. Since COVID-19 is transmitted via an underlying social contact network, the impact of individual A receiving the vaccine may not only reduce individual A’s chance of contracting the disease, but also may reduce the risk of exposure of other individuals in the connected to A in the network. This network effect is heterogeneous as the frequency of time individual A spends with others in its contact network may vary. An added complexity in this example is that the contact network is dynamic and often unobserved.

2 Preliminaries

Consider a population of n individuals. We denote the treatment vector by $\mathbf{z} = (z_1, z_2, \dots, z_n) \in \{0, 1\}^{[n]}$, where $z_i = 1$ if individual i receives the treatment and $z_i = 0$ if individual i is in the control group. Let \mathbf{e}_i denote the standard basis vector which takes value one at coordinate i and is zero everywhere else. The randomized design refers to the probability distribution governing how the treatment vector \mathbf{z} is generated. $Y_i(\mathbf{z})$ denotes the potential outcome of individual i in the event that treatment vector \mathbf{z} is implemented. Only the outcomes for the implemented treatment vector \mathbf{z} are observed, and thus all other “potential outcomes” that would result from other realizations of the treatment vector are unobserved. Under the Stable Unit Treatment Value Assumption, the potential outcome of individual i only depends on z_i and not on the treatment of any other individual [12, 27, 23]. Under this assumption, $Y_i(\mathbf{z}) = Y_i(z_i \mathbf{e}_i)$ for all \mathbf{z} . In the presence of general arbitrary network interference, the outcome of individual i may depend on the full treatment vector.

There are a few different estimands that we may be interested in. The total treatment effect (TTE) is the difference between the average outcome if all individuals were treated and the average outcome if nobody were treated,

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

where $\mathbf{1}$ denotes the vector of all ones and $\mathbf{0}$ denotes the vector of all zeros. The total treatment effect is particularly relevant in scenarios in which a decision maker desires to use the outcome of the randomized experiment to determine whether to apply the treatment to the entire population at large, or to stay with the status quo. The average treatment effect (ATE), also referred to as the direct treatment effect, captures the change in outcomes of an average individual due to only itself being treated and not its neighbors,

$$\text{ATE} := \frac{1}{n} \sum_{i \in [n]} (Y_i(\mathbf{e}_i) - Y_i(\mathbf{0})).$$

The average interference effect (AIE), also referred to as the network interference or spillover effect, captures the change in outcomes of an average individual due to the network (excluding itself) being treated,

$$\text{AIE} := \frac{1}{n} \sum_{i \in [n]} (Y_i(e_{[n] \setminus \{i\}}) - Y_i(\mathbf{0})).$$

The total treatment effect can be decomposed into a sum of the average (direct) treatment effect, the average (network) interference effect, and an interaction effect [31]. The interaction effect is nonzero in scenarios when the effect of interference on an individual may depend on whether the individual is treated or not. We will present results for all three estimators, with an emphasis on total treatment effect, as it is particularly relevant to real-world scenarios where the goal is to make an optimal policy decision. Furthermore, the majority of positive results in the literature have been for estimating the ATE which is easier to estimate. In contrast, existing results for estimating the TTE and AIE either assume strong properties on the network, are computationally costly in general networks, or have prohibitively high variance.

3 Related Literature

Some previous work has focused on estimating the average (direct) treatment effect [7, 17, 29, 31, 19, 22], or testing for the presence of network interference [1, 9, 3, 25, 28]; these methods do not produce an estimate of the total treatment effect, which is significantly harder. A critical

challenge for estimating the total treatment effect is that we only observe $\{Y_i(\mathbf{z})\}_{i \in [n]}$ for a single fixed treatment vector \mathbf{z} , which is not $\mathbf{1}$ or $\mathbf{0}$. As a result, we may not observe any of the terms in the expression of interest.

Under a fully general arbitrary interference model, it has been repeatedly shown that it is impossible to estimate any desired causal estimands as the model is not fully identifiable [23, 2, 6, 18]. As a result, there have been many proposed models that impose assumptions on exposure functions [23, 2, 36, 4, 20], interference neighborhoods [34, 5, 31, 8], parametric structure [33, 7, 10, 14, 13], or a combination of these. Each of these assumptions lead to different solution concepts. The art in choosing a good model is balancing the tension between strong assumptions which facilitate simple solutions, and weak assumptions which can more flexibly encompass real world applications. Additionally there have been studies showing that one must take caution in choosing model assumptions, as the results may be sensitive to model misspecification [2, 18]. In particular, a majority of these approaches rely on knowledge of a network mediating the interference effects, which is often not available in practice. While we do not have space to discuss all previously proposed models, we highlight a few of the most common models to highlight the strengths and weaknesses of each model and the existing results.

3.1 Partial Interference

Partial interference assumes that the population can be partitioned into disjoint groups, such that all network interference effects can only occur within but not across the pre-specified groups [30, 26, 16, 32, 21, 35, 8, 4]. Specifically, the outcome of individual i can only depend on the treatment of others in the same group as individual i , and is independent from the treatment assignments of individuals in other groups. Under this assumption, we can randomize treatments over the groups jointly to assign groups to different treatment levels. This would allow us to observe the average outcomes on individuals that have differing fraction of the network treated, which can then be used to infer the impact of network effect on the outcomes.

In particular, if we are estimating the total treatment effect, we could assign each group to be either fully treated or fully control, where all individuals in the group are given the same treatment assignment. As a result, $Y_i(\mathbf{z}) = Y_i(\mathbf{1})$ for all i such that $z_i = 1$, and $Y_i(\mathbf{z}) = Y_i(\mathbf{0})$ for all i such that $z_i = 0$. Unfortunately, this approach does not extend when the network could be highly connected, limiting its use in practice. The bias of standard estimators will scale with the number of edges across clusters, leading to proposed cluster randomized designs that randomize over clusters that are constructed to minimize edges between clusters [14, 13].

3.2 Neighborhood Interference

Neighborhood interference assumes that an individual's outcome can only depend on the treatment assignments of its direct neighbors in a known specified graph [34, 31]. This assumption guarantees that $Y_i(\mathbf{z}) = Y_i(\mathbf{1})$ for any individual in the treatment group whose neighbors are also all in the treatment group; we denote this set of individuals $\mathcal{S}_1(\mathbf{z})$. Similarly, $Y_i(\mathbf{z}) = Y_i(\mathbf{0})$ for any individual in the control group whose neighbors are also all in the control group; we denote this set $\mathcal{S}_0(\mathbf{z})$. A natural estimate for the total treatment effect is the difference in average outcomes between groups $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_0(\mathbf{z})$, or an inverse propensity weighted estimator when the probability of being in group $\mathcal{S}_1(\mathbf{z})$ or $\mathcal{S}_0(\mathbf{z})$ may vary across individuals [2]. Without further structure on the interference, one cannot use measurements from individuals not in either sets $\mathcal{S}_1(\mathbf{z})$ or $\mathcal{S}_0(\mathbf{z})$, as the relationship between $Y_i(\mathbf{z})$ and $Y_i(\mathbf{1})$ or $Y_i(\mathbf{0})$ is unknown.

Under naïve randomized designs such as a Bernoulli design that assigns each individual independently to treatment with probability p or control with probability $1 - p$, the variance of the inverse propensity weighted estimator will go to infinity with n for well-connected networks such as the Erdos-Renyi graph with average degree larger than \sqrt{n} [6]; this results from the fact that with high probability the sizes of sets $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_0(\mathbf{z})$ will be small for highly connected networks. As a result, [34] proposes a graph cluster randomized design that aims to jointly assign individuals and their neighbors to treatment or control in order to minimize variance. Unfortunately this requires detailed knowledge of the network, and constructing optimal clusters can be computationally expensive for non-trivial well-connected networks. As a result, this approach has not translated into practical solutions.

3.3 Linear Outcomes with few parameters

While the above models impose network-based conditions on the interference, an alternate approach is to impose parametric structure on the form of the potential outcomes. The most common assumption is that the potential outcomes are linear with respect to a specified statistic of the local neighborhood [33, 14, 7, 10, 24, 11]. For example, [10] assumes the outcome is linear in the fraction of treated neighbors, such that

$$Y_i(\mathbf{z}) = \alpha + \beta z_i + \gamma \left(\frac{\sum_{k \in [n]} A_{ki} z_k}{\sum_{k \in [n]} A_{ki}} \right),$$

where A is a known adjacency matrix representing edges in the network. Similarly, [24] assumes linearity with respect to the absolute number of treated neighbors. Threshold models also can be expressed with a linear model using indicator statistics. For example [14] assumes that network effects arise when at least θ neighbors are treated, where θ is assumed to be known,

$$Y_i(\mathbf{z}) = \alpha + \beta z_i + \gamma \mathbb{I} \left(\sum_{k \in [n]} A_{ki} z_k \geq \theta \right).$$

One can extend these models to incorporate covariate types, such that the total number of unknown parameters is three times the number of different covariate types, assuming that each covariate type is associated to a set of parameters α, β, γ .

What is characteristic of this approach is that the assumptions reduce the number of unknown parameters in the potential outcomes models to a fixed dimension that does not grow with the population size, reducing the inference task to linear regression. As a result, the natural solution is to use a least squares estimate, shifting the focus to constructing randomized designs that minimize the variance of the estimate. A limitation of this approach is that it requires the correct choice of the the statistic governing the linearity, and it requires precise knowledge of the network structure to compute these neighborhood statistics. Furthermore, it assumes knowledge of the relevant covariate types that differentiate individual responses, or otherwise assumes homogeneity in the effects.

4 Heterogeneous Linear Potential Outcomes Model

For the remainder of the paper, we will assume that the outcome of each individual is affected by an additive term for each treated individual in the population; most notably the network interference effects are allowed to be fully heterogeneous for each pair of individuals. This is also referred to as the linear model in [13], the saturated structural linear model in [15], or the joint assumptions of additivity of main effects and interference effects as defined in [31]. The potential outcomes can be expressed by

$$Y_i(\mathbf{z}) = \alpha_i + \beta_i z_i + \sum_{k \in [n]} \gamma_{ki} z_k,$$

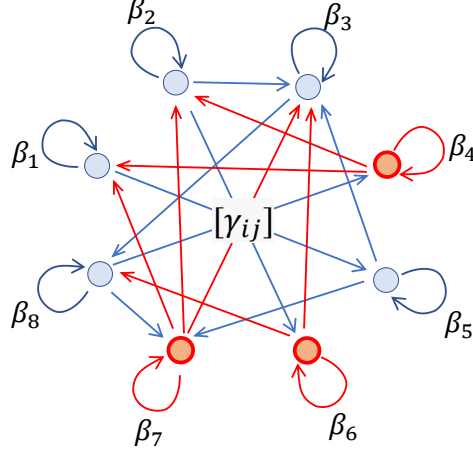


Figure 1: *Depiction of heterogeneous linear outcomes model: Vertex weights correspond to baseline outcomes. When individual i is treated (red), its outgoing edges are activated. Individual j 's outcome $Y_j(\mathbf{z})$ is the sum of its own baseline plus any incoming activated edges (red).*

where α_i represents the individual baseline, β_i represents the individual direct treatment effect, and γ_{ki} represents the additive network interference effect over the directed pair (k, i) . We will use \mathcal{E} to denote the set of pairs (k, i) for which $\gamma_{ki} \neq 0$. Without imposing constraints on the sparsity of \mathcal{E} , $|\mathcal{E}|$ could be as large as $n(n-1)$, such that the total number of parameters in this model can be as large as $n + n^2$. This heterogeneous model is much more expressive than fixed parameter models as it allows for individualized effects for each individual and potential edge in the network.

This model can capture network interference that arises from spillover, peer effects, and contagion. Spillover refers to the interference that arises from individual j 's treatment affecting individual i 's outcome. Typically the spillover effect is assumed to be mediated by the network, such that γ_{ji} is non-zero only if (j, i) is an edge in the network. By relaxing constraints on the sparsity of the γ_{ki} parameters, the heterogeneous linear outcomes model can also capture long-range spillover effects mediated by multi-hop paths in the network.

Contagion or peer effects refer to interference that arises from individual j 's outcome affecting individual i 's outcome. When the contagion effect is linear, then this translates into long-range interference effects over multi-hop paths in the network. For example, consider a path $\ell \rightarrow k \rightarrow j \rightarrow i$ in the network. Under contagion, individual ℓ 's treatment affects individual ℓ 's outcome, which subsequently affects individual k 's outcome, which subsequently affects individual j 's outcome, which then affects individual i 's outcome, as described by

$$Y_i(\mathbf{z}) = a_i + b_i z_i + \sum_{k \in [n]} c_{ki} Y_k(\mathbf{z}).$$

The potential outcomes model can be derived by solving a system of linear equations for the outcomes vector given an assigned treatment vector, which results in the following potential outcomes model

$$\mathbf{Y}(\mathbf{z}) = (I - C)^{-1} \mathbf{a} + \sum_{t=0}^{\infty} C^t \cdot \text{diag}(\mathbf{b}) \cdot \mathbf{z},$$

where C is a matrix with the (k, i) -th entry equal to c_{ki} , $\text{diag}(\mathbf{b})$ is a diagonal matrix with diagonal entries taking values from \mathbf{b} , and \mathbf{a} and \mathbf{b} are vectors corresponding to the parameters $\{a_i\}_{i \in [n]}$

and $\{b_i\}_{i \in [n]}$. Written in this form, we can verify that the potential outcomes could be described via a heterogeneous linear outcomes model with dense network effects [13, 15].

As there are more unknown model parameters than measurements, we cannot hope to identify the model via regression, and thus a randomized experimental design will be critical to any solution. Previous attempts at causal inference under this model either involve complicated network dependent randomized designs [13], incur potentially high network-dependent biases [13], or impose Bayesian priors on the unknown parameters that reduce the statistical estimation task to again estimating a model with a fixed number of parameters [33, 7].

Scenarios that violate linearity include when network effects saturate after a certain number of neighbors are treated, are sublinear in the number of treated neighbors, or are only present after a minimum number of neighbors are treated.

5 Summary of Results

Our results focus on estimating the TTE, ATE, and AIE under the heterogeneous linear outcomes model, which is significantly more expressive than fixed parameter models or sparse neighborhood interference models. Under our model, each of these estimands scale linearly in the fraction of treated individuals; our approach exploits this linearity for a simple and efficient solution. We believe this combination of a practical solution with a flexible model positions our results to have impact in the broader scientific community.

The primary research question is: *Is there a simple and efficient approach to estimate causal effects in the presence of network effects without critically relying on knowledge of the network structure or restrictive network properties?* While this has previously remained elusive, our results provide a clear and simple answer, including both a negative scenario in which there can be no simple solution, and a positive scenario in which we outline a simple efficient solution which can easily translate into practice.

First, we show that in the presence of additive network effects, any individually weighted linear estimator for the total treatment effect (TTE) is necessarily biased unless the network can be perfectly partitioned into small disjoint subsets with no interfering edges. Furthermore, this bias can be large depending on the relative magnitude of the network effects. Similarly, we show that any linear estimator for the average interference effect (AIE) is necessarily biased as well. On the other hand the difference in means estimator is an unbiased estimate for the average treatment effect (ATE) for any network under Bernoulli randomization. This negative result suggests that the linearity arising from the additive model is not sufficient in itself to admit simple solutions. The primary reason is that it is difficult for simple estimators to distinguish between the response due to baseline values $\{\alpha_i\}_{i \in [n]}$ as opposed to network effects $\{\gamma_{ji}\}_{(j,i) \in \mathcal{E}}$.

Second, we consider the scenario when we have access to an estimate of the average individual baselines; in practice this could be constructed from historical data or pilot studies. Given baseline estimates, we present simple estimators for TTE, ATE, and AIE, which are unbiased for any randomized design satisfying symmetry conditions in the treatment assignments of individuals and their neighbors. In particular these symmetry conditions are satisfied by completely randomized design and Bernoulli randomization. These estimators are extremely easy to compute, and neither the randomized designs nor the estimate themselves require knowledge of the underlying network, which is often not available in practice.

Third, we show that our proposed approach has low variance under a simple completely randomized design. In particular, the estimator for total treatment effect is consistent as long as the fraction of treated individuals is asymptotically larger than d_{\max}^2/n , where d_{\max} is the maxi-

mum out-degree of any individual in the network, i.e. the maximum influence of any individual in the network. Furthermore, we provide analytical expression for the variance of our estimator under commonly used randomized designs, including completely randomized and cluster randomized design, as well as uniform and varying saturation designs. These variance expressions provide insight for designing randomized designs that minimize variance by matching individuals based on estimated network influence.

6 Linear Estimators not Sufficient for Unbiased Estimators

Assuming heterogeneous linear network effects implies that the total treatment effect scales linearly in the number of treated individuals. A natural question is whether this linearity is sufficient to admit simple unbiased estimators for the causal estimands of interest. In this section we provide some results in the negative by considering the restricted class of individually weighted linear estimators. These estimators have the form

$$\hat{\text{est}}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} (w_i z_i + v_i (1 - z_i)) Y_i(\mathbf{z}),$$

where the weights $\mathbf{w} = (w_1, w_2, \dots, w_n)$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)$ are deterministic and not a function of the treatment \mathbf{z} . Most notably, the weight w_i or v_i are only selected based on whether individual i is treated or not, and does not depend on the treatment configuration of its neighbors. This limitation keeps the estimator simple to compute, and furthermore helps us focus on the question of whether there exists a solution that does not require knowledge of the network.

Linear estimators are often used in practice due to their simplicity, and the restricted class of individually weighted linear estimators has an added benefit of not requiring network knowledge and furthermore encompassing many commonly used estimators. For example the simple Horvitz-Thompson inverse propensity score weighted estimator sets $w_i = 1/(n\mathbb{E}[z_i])$ and $v_i = 1/(n\mathbb{E}[1 - z_i])$. The difference in means estimator sets $w_i = 1/\sum_{j \in n} z_j$ and $v_i = 1/\sum_{j \in n} (1 - z_j)$, which can be approximated by deterministic quantities $n/\mathbb{E}[\sum_{j \in n} z_j]$ and $n/\mathbb{E}[\sum_{j \in n} (1 - z_j)]$ when n is large.

6.1 Total Treatment Effect

Recall that the total treatment effect (TTE) measures the difference in average outcomes if all individuals were treated versus if nobody were treated. Under the heterogeneous linear outcomes model, it is equal to

$$\text{TTE} = \frac{1}{n} \sum_{i \in [n]} (\beta_i + \sum_{k \in [n]} \gamma_{ki}).$$

Theorem 4. *Under heterogeneous linear network effects, any unbiased individually weighted linear estimator for total treatment effect must have the form*

$$\widehat{\text{TTE}} = \frac{1}{n} \sum_{i \in [n]} \left(\frac{z_i}{\mathbb{E}[z_i]} - \frac{1 - z_i}{\mathbb{E}[1 - z_i]} \right) Y_i(\mathbf{z}),$$

and the randomized design must satisfy $\mathbb{P}(z_k = z_i) = 1$ for all $(k, i) \in \mathcal{E}$. As a result, there does not exist an unbiased linear estimator for the total treatment effect if the network is fully connected.

Proof. Under the heterogeneous linear network effects model, an individually weighted linear estimator takes the value

$$\hat{\text{est}}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} (w_i z_i + v_i (1 - z_i)) \alpha_i + \sum_{i \in [n]} w_i z_i \beta_i$$

$$+ \sum_{(k,i) \in \mathcal{E}} (w_i z_i + v_i (1 - z_i)) z_k \gamma_{ki}.$$

This is an unbiased estimator for the total treatment effect only if $\mathbb{E}[\hat{\text{est}}(\mathbf{w}, \mathbf{v})] = \frac{1}{n} \sum_{i \in [n]} \beta_i + \frac{1}{n} \sum_{(k,i) \in \mathcal{E}} \gamma_{ki}$ is satisfied for any configuration of $\{\alpha_i\}_{i \in [n]}$, $\{\beta_i\}_{i \in [n]}$, and $\{\gamma_{ki}\}_{(k,i) \in \mathcal{E}}$. This requirement results in the following $2n + |\mathcal{E}|$ constraints, which arise from matching coefficients for each of the parameters,

- α : for all $i \in [n]$, $w_i \mathbb{E}[z_i] + v_i \mathbb{E}[1 - z_i] = 0$,
- β : for all $i \in [n]$, $w_i \mathbb{E}[z_i] = \frac{1}{n}$,
- γ : for all $(k, i) \in \mathcal{E}$, $w_i \mathbb{E}[z_i z_k] + v_i \mathbb{E}[(1 - z_i) z_k] = \frac{1}{n}$.

The second set of constraints for the direct treatment effects imply that the weights $w_i = 1/(n \mathbb{E}[z_i])$. As a result, combining this with the first set of constraints for the baselines imply that $v_i = -1/(n \mathbb{E}[1 - z_i])$. After fixing the values of all the weights \mathbf{w}, \mathbf{v} , the third set of constraints become difficult to satisfy. We can rewrite the third set of constraints as

$$w_i \mathbb{E}[z_i] \mathbb{P}(z_k = 1 | z_i = 1) + v_i \mathbb{E}[1 - z_i] \mathbb{P}(z_k = 1 | z_i = 0) = \frac{1}{n},$$

for all $(k, i) \in \mathcal{E}$. Most notably, it is a linear combination of the two terms that show up in the first and second set of constraints, multiplied by probabilities that must be in $[0, 1]$. By plugging in the values for w_i and v_i that arise from the first two constraints, it follows that the third constraint is only satisfied when $\mathbb{P}(z_k = 1 | z_i = 1) = 1$ and $\mathbb{P}(z_k = 1 | z_i = 0) = 0$ for all $(k, i) \in \mathcal{E}$. This is equivalent to requiring that the randomized design always assigns connected individuals to the same treatment, i.e. $\mathbb{P}(z_k = z_i) = 1$ for all $(k, i) \in \mathcal{E}$. \square

The constraint on the randomized design implies that every pair of connected individuals in the population must be either both treated or both in control. This restricts the valid randomized designs to a cluster randomized design where the clusters are defined by the connected components of the graph. Theorem 4 highlights that the imposed structure from heterogeneous linear network effects is insufficient to remove the complex dependence on the network. Even under linearity, we still need to deal with either imposing strong assumptions on the connectivity structure of the network, or we will need to use more complex estimators that utilize knowledge of the network, bringing us back to the same challenges present in the fully general model.

When the conditions for unbiasedness are not satisfied, the bias of the above simple estimator will scale with the average network effect across the edges between the treated and control groups, given by the expression

$$\mathbb{E}[\widehat{\text{TTE}}] - \text{TTE} = \frac{1}{n} \sum_{(k,i) \in \mathcal{E}} \left(\frac{\text{Cov}[z_i, z_k]}{\text{Var}[z_i]} - 1 \right) \gamma_{ki}.$$

If the randomized design enforces high correlation across pairs of connected individuals in the network, then $\mathbb{E}[\widehat{\text{TTE}}]$ is close in expectation to the total treatment effect. If the design enforces independence of treatments across edges in the network, then $\mathbb{E}[\widehat{\text{TTE}}]$ only captures the direct treatment effects and not the network effects.

The restrictive unbiasedness conditions result from the fact that it is difficult to set the coefficient for the baseline parameters to 0 while maintaining that the coefficients on the network effects are $1/n$, as the expressions for both are very similar. Essentially, it is difficult for the model to distinguish between the effects arising from individual baselines as opposed to the ambient network effects from treated neighbors. Given this insight, in the next section we consider the scenario where we have access to estimates of the average individual baselines.

6.2 Average Treatment Effect

Recall that the average treatment effect (ATE), also referred to as the direct treatment effect, measures the average difference in outcomes for individuals that is caused only by their own treatments, not including any network effects. Under the heterogeneous linear outcomes model, it is equal to

$$\text{ATE} = \frac{1}{n} \sum_{i \in [n]} \beta_i.$$

Theorem 5. *under heterogeneous linear network effects, any unbiased individually weighted linear estimator for average treatment effect must have the form*

$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i \in [n]} \left(\frac{z_i}{\mathbb{E}[z_i]} - \frac{1 - z_i}{\mathbb{E}[1 - z_i]} \right) Y_i(\mathbf{z}),$$

and the randomized design must satisfy $z_k \perp\!\!\!\perp z_i$ for all $(k, i) \in \mathcal{E}$.

Proof. Under the heterogeneous linear network effects model, an individually weighted linear estimator takes the value

$$\begin{aligned} \widehat{\text{est}}(\mathbf{w}, \mathbf{v}) &= \sum_{i \in [n]} (w_i z_i + v_i (1 - z_i)) \alpha_i + \sum_{i \in [n]} w_i z_i \beta_i \\ &\quad + \sum_{(k, i) \in \mathcal{E}} (w_i z_i + v_i (1 - z_i)) z_k \gamma_{ki}. \end{aligned}$$

This is unbiased for the average treatment effect only if $\mathbb{E}[\widehat{\text{est}}(\mathbf{w}, \mathbf{v})] = \frac{1}{n} \sum_{i \in [n]} \beta_i$ is satisfied for any configuration of $\{\alpha_i\}_{i \in [n]}$, $\{\beta_i\}_{i \in [n]}$, and $\{\gamma_{ki}\}_{(k, i) \in \mathcal{E}}$. This requirement results in the following $2n + |\mathcal{E}|$ constraints, which arise from matching coefficients for each of the parameters,

- α : for all $i \in [n]$, $w_i \mathbb{E}[z_i] + v_i \mathbb{E}[1 - z_i] = 0$,
- β : for all $i \in [n]$, $w_i \mathbb{E}[z_i] = \frac{1}{n}$,
- γ : for all $(k, i) \in \mathcal{E}$, $w_i \mathbb{E}[z_i z_k] + v_i \mathbb{E}[(1 - z_i) z_k] = 0$.

Solving for the weights given the first two constraints results in

$$w_i = \frac{1}{n \mathbb{E}[z_i]} \text{ and } v_i = -\frac{1}{n \mathbb{E}[1 - z_i]}.$$

By plugging in these values of w_i and v_i into the third set of constraints arising from γ , it follows that we must satisfy

$$\frac{\mathbb{E}[z_i z_k]}{n \mathbb{E}[z_i]} - \frac{\mathbb{E}[(1 - z_i) z_k]}{n \mathbb{E}[1 - z_i]} = \frac{\mathbb{P}(z_k = 1 | z_i = 1)}{n} - \frac{\mathbb{P}(z_k = 1 | z_i = 0)}{n} = 0,$$

which implies that $\mathbb{P}(z_k = 1 | z_i = 1) = \mathbb{P}(z_k = 1 | z_i = 0)$ such that $z_k \perp\!\!\!\perp z_i$ for all $(k, i) \in \mathcal{E}$. \square

This independence constraint on the randomized design is quite restrictive, yet is easily satisfied by simple Bernoulli randomization. Although many other randomizations may not satisfy independence, if they are “almost” independent, for example in completely randomized design with a large population, then the bias will still be small.

Corollary 6. *Under heterogeneous linear network effects, the Horvitz-Thompson estimator with Bernoulli randomization is an unbiased estimator for the average treatment effect, even when the interference effects are fully dense.*

Proof. This follows from the fact that by definition of Bernoulli randomization, $z_k \perp\!\!\!\perp z_i$ for all $i \neq k$. \square

6.3 Average Interference Effect

Recall that the average interference effect (AIE), also referred to as the network interference effect, measures the average difference in outcomes of individuals that is caused only due to network effects but not their own direct treatment effects. Under the heterogeneous linear outcomes model, it is equal to

$$\text{AIE} = \frac{1}{n} \sum_{(k,i) \in \mathcal{E}} \gamma_{ki}.$$

Theorem 7. *Under heterogeneous linear network effects, there does not exist an unbiased individually weighted linear estimator for the average interference effect.*

Proof. Under the heterogeneous linear network effects model, an individually weighted linear estimator is unbiased for the average interference effect only if $\mathbb{E}[\hat{\text{est}}(\mathbf{w}, \mathbf{v})] = \frac{1}{n} \sum_{(k,i) \in \mathcal{E}} \gamma_{ki}$ is satisfied for any configuration of $\{\alpha_i\}_{i \in [n]}$, $\{\beta_i\}_{i \in [n]}$, and $\{\gamma_{ki}\}_{(k,i) \in \mathcal{E}}$. This requirement results in the following $2n + |\mathcal{E}|$ constraints, which arise from matching coefficients for each of the parameters,

- α : for all $i \in [n]$, $w_i \mathbb{E}[z_i] + v_i \mathbb{E}[1 - z_i] = 0$,
- β : for all $i \in [n]$, $w_i \mathbb{E}[z_i] = 0$,
- γ : for all $(k, i) \in \mathcal{E}$, $w_i \mathbb{E}[z_i z_k] + v_i \mathbb{E}[(1 - z_i) z_k] = \frac{1}{n}$.

The first two constraints together require that the weights are all zero, i.e. $w_i = 0$ and $v_i = 0$. However, with zero weights it is impossible to satisfy the constraint arising from γ . \square

6.4 Discussion

By comparing Corollary 6 with Theorem 4 and Theorem 7, we can see that the average treatment effect is easier to estimate, as there exists a simple estimator that leads to an unbiased estimator for ATE under Bernoulli randomization. In contrast, the existence of an unbiased estimator for the TTE depends on restrictive properties of the network, and there does not exist an unbiased individually weighted linear estimator for the AIE in the heterogeneous linear outcomes model.

Theorems 4 and 5 show that the only unbiased individually weighted linear estimator for the total treatment effect and average treatment effect is the Horvitz-Thompson inverse propensity estimator, which takes the form

$$\hat{\text{est}}_{HT} = \sum_i \frac{z_i Y_i(z)}{n \mathbb{E}[z_i]} - \sum_i \frac{(1 - z_i) Y_i(z)}{n \mathbb{E}[1 - z_i]}.$$

This uniqueness results from satisfying unbiasedness constraints arising from the α and β parameters. The additional constraints on the randomized design are also very limiting. In a setting where one does not have much control over the randomization due to regulatory policies, then the bias of the Horvitz-Thompson estimator could be estimated as a function of the randomized design. In particular, the expected value of the Horvitz-Thompson estimator is given by

$$\mathbb{E}[\hat{\text{est}}_{HT}] = \frac{1}{n} \sum_{i \in [n]} \beta_i + \frac{1}{n} \sum_{(k,i) \in \mathcal{E}} \left(\frac{\text{Cov}[z_i, z_k]}{\text{Var}[z_i]} \right) \gamma_{ki}.$$

If the randomized design enforces high correlation across pairs of connected individuals in the network, then the estimator is closer to the TTE. If the design enforces independence of treatments across edges in the network, then the estimator is closer to the ATE.

For practitioners, one can design a randomized design as a function of the graph to adjust the Horvitz-Thompson estimator to approximate either the TTE or ATE as desired. If the treatment

across edges is highly correlated such that endpoints of most edges are either both treated or both control, then the estimator will approximate the TTE. Unfortunately, given the connectivity of the graph and fraction of treated units, it may not even be possible to implement such a randomized design, in which case one should keep in mind that the bias may be proportional to the edges for which the endpoints have different treatments. If the treatment of pairs of units sharing an edge is nearly independent, then the estimator will approximate the ATE. Since there are simple randomizations (CRD and Bernoulli) which can implement near independence across all pairs, it is not difficult to extend previous estimators for the ATE in presence of network interference effects.

6.5 Partially Heterogeneous Model Parameters

The above results arise from assuming a worst case full heterogeneity of the model parameters which imposes $2n + |\mathcal{E}|$ constraints for an unbiased estimator. If in reality we knew that there were clusters of parameters which took the same value, e.g. assuming that the parameters were related to known covariate structure amongst the units, then the number of constraints would reduce significantly. For example, consider a model in which each unit i is associated to a covariate type $r_i \in [m]$, such that $\alpha_i = a_{r_i}$, $\beta_i = b_{r_i}$, and $\gamma_{ji} = g_{r_i, r_j}$. The space of model parameters reduces to a m -dimensional vector a , a m -dimensional vector b , and a $m \times m$ -dimensional matrix g . In this reduced model there are only $2m + m^2$ constraints on an unbiased estimator. If $2m + m^2$ is smaller than $2n$, there may be multiple choices of weight vectors that would satisfy the unbiasedness constraints, potentially resolving the negative result presented in the fully heterogeneous setting.

If we however restrict ourselves to a setting in which the weights given to units in the estimator is equal for all individuals of the same covariate type, and if we restrict to “fair” randomized designs in the sense that sets of units with same covariate types should have the same treatment probability distributions, then we reduce the flexibility of the class of estimators such that we recover the previous impossibility results.

7 Simple Unbiased Estimators given Prior Baseline Data

In practice there are many applications in which we do have access to additional information from historical data or pilot studies that could be used to construct estimates of the average baseline $\frac{1}{n} \sum_{i \in [n]} Y_i(\mathbf{0}) = \frac{1}{n} \sum_{i \in [n]} \alpha_i$. For example, a social media platform such as LinkedIn is constantly monitoring the engagement level of its users, such that it always has access to the current status quo baselines at an individual level before deploying randomized trials for a newly proposed feature. As another example, a pharmaceutical company could use reported data from state and national level public health departments to estimate the average baseline rates of contracting COVID-19 before beginning its clinical trials. Even when historical data may not be available, it is typically easy to conduct small scale surveys to estimate the baseline outcome levels before beginning the randomized experiment. The only required condition is that the data is collected before the experiment begins such that no one has yet received the treatment; as such the measurements will accurately reflect the baseline as there will be no network effects. We will show that this commonly available data can be used to significantly simplify the estimation of causal effects in the presence of heterogeneous linear network effects.

Let’s first assume that we have access to the full individual baselines; it follows naturally to then subtract the baseline α_i from the measurement $Y_i(\mathbf{z})$ to remove all contributions of the baseline effects from the linear estimator, resulting in

$$\hat{\text{est}}_{-\alpha}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} (w_i z_i + v_i (1 - z_i)) (Y_i(\mathbf{z}) - \alpha_i).$$

To characterize conditions for unbiased linear estimators, we use the same approach of equating the coefficients of the direct effects and the network effects between the expected value of this estimator and the total treatment effect. Subtracting out the baselines removes the set of constraints for unbiasedness associated to the baseline parameters, leaving us with $n + |\mathcal{E}|$ constraints. While this is still significantly more than the number of measurements, it turns out that there are still many reasonable randomized designs under which we are able to satisfy these constraints.

7.1 Total Treatment Effect

Theorem 8 presents the necessary conditions for unbiased linear estimators for total treatment effect given baseline estimates.

Theorem 8. *For any randomized design such that $\frac{\mathbb{P}(z_k=0 \mid z_i=1)}{\mathbb{P}(z_k=1 \mid z_i=0)} = \rho_i$ for all $(k, i) \in \mathcal{E}$ for some values of $\{\rho_i\}_{i \in [n]}$, the following simple estimator*

$$\widehat{TTE}_{-\alpha} = \frac{1}{n} \sum_{i \in [n]} \left(\frac{z_i}{\mathbb{E}[z_i]} - \frac{(1 - z_i)\rho_i}{\mathbb{E}[1 - z_i]} \right) (Y_i(z) - \alpha_i),$$

produces an unbiased estimate for the total treatment effect under heterogeneous linear network effects.

Proof. After subtracting out the baseline parameters, the constraints for unbiasedness reduce to

- β : for all $i \in [n]$, $w_i \mathbb{E}[z_i] = \frac{1}{n}$,
- γ : for all $(k, i) \in \mathcal{E}$, $w_i \mathbb{E}[z_i z_k] + v_i \mathbb{E}[(1 - z_i)z_k] = \frac{1}{n}$.

Satisfying the constraints arising from β and γ results in

$$w_i = \frac{1}{n \mathbb{E}[z_i]} \text{ and } v_i = \frac{\mathbb{E}[z_i(1 - z_k)]}{n \mathbb{E}[z_i] \mathbb{E}[(1 - z_i)z_k]} \text{ for all } (k, i) \in \mathcal{E}.$$

In order to ensure that such a valid v_i exists, we would need that $\frac{\mathbb{P}(z_k=0 \mid z_i=1)}{\mathbb{P}(z_k=1 \mid z_i=0)} = \rho_i$ for all $(k, i) \in \mathcal{E}$. Under this condition,

$$\frac{\mathbb{E}[z_i(1 - z_k)]}{\mathbb{E}[z_i] \mathbb{E}[(1 - z_i)z_k]} = \frac{\mathbb{E}[z_i] \mathbb{P}(z_k = 0 \mid z_i = 1)}{\mathbb{E}[z_i] \mathbb{E}[(1 - z_i)] \mathbb{P}(z_k = 1 \mid z_i = 0)} = \frac{\rho_i}{\mathbb{E}[(1 - z_i)]}.$$

□

The condition on the randomization is equivalent to imposing that $\frac{\mathbb{E}[z_i(1 - z_k)]}{\mathbb{E}[(1 - z_i)z_k]} = \frac{\mathbb{E}[z_i(1 - z_j)]}{\mathbb{E}[(1 - z_i)z_j]}$ for all triplets (i, j, k) such that $(k, i) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$. Essentially this boils down to symmetry conditions on the second moments of the treatment vector across edges in the network. Such a symmetry condition would be satisfied by ensuring that for all i , the neighbors that influence i are treated equally in the distribution of the assigned treatments. A sufficient condition to satisfy this symmetry is to impose that the marginals are equal across edges, i.e. $\mathbb{E}[z_i] = \mathbb{E}[z_k]$ for all $(k, i) \in \mathcal{E}$. This is an easy condition to satisfy, and leads to a simplified result as stated below.

Corollary 9. *For any randomized design such that $\mathbb{E}[z_i] = \mathbb{E}[z_k]$ for all $(k, i) \in \mathcal{E}$, the following simple estimator*

$$\widehat{TTE}_{-\alpha} = \frac{1}{n} \sum_{i \in [n]} \frac{Y_i(\mathbf{z}) - \alpha_i}{\mathbb{E}[z_i]}$$

produces an unbiased estimate for the total treatment effect under heterogeneous linear network effects. When $\mathbb{E}[z_i] = p$ for all $i \in [n]$, the estimator further simplifies to

$$\widehat{TTE}_{-\alpha} = \frac{1}{p} \left(\frac{1}{n} \sum_{i \in [n]} Y_i(\mathbf{z}) - \frac{1}{n} \sum_{i \in [n]} \alpha_i \right).$$

Proof. First we can verify that if $\mathbb{E}[z_i] = \mathbb{E}[z_k]$, then

$$\frac{\mathbb{P}(z_k = 0 \mid z_i = 1)}{\mathbb{P}(z_k = 1 \mid z_i = 0)} = \frac{\mathbb{E}[z_i(1 - z_k)]\mathbb{E}[1 - z_i]}{\mathbb{E}[z_i]\mathbb{E}[z_k(1 - z_i)]} = \frac{(\mathbb{E}[z_i] - \mathbb{E}[z_i z_k])\mathbb{E}[1 - z_i]}{\mathbb{E}[z_i](\mathbb{E}[z_k] - \mathbb{E}[z_k z_i])} = \frac{\mathbb{E}[1 - z_i]}{\mathbb{E}[z_i]} =: \rho_i.$$

The estimator then results from plugging in the expression for ρ_i into the estimator defined in Theorem 8. When $\mathbb{E}[z_i] = p$, rearranging the expression then results in the simplified form

$$\widehat{TTE}_{-\alpha} = \frac{1}{p} \left(\frac{1}{n} \sum_{i \in [n]} Y_i(\mathbf{z}) - \frac{1}{n} \sum_{i \in [n]} \alpha_i \right),$$

highlighting that the only knowledge of the baseline parameters needed is the population average baseline. \square

When the marginal treatment probability is equal for all individuals, i.e. $\mathbb{E}[z_i] = p$ for all $i \in [n]$, the resulting estimator in fact only needs knowledge of the average population baselines rather than individual baseline parameters. While there are a few settings for which individual baseline parameters are observed from historical data, such as experimentation on social media platforms, data of such granularity is not realistic in general. On the other hand, having access an accurate estimate of the average population wide baselines is realistic for a broad variety of applications across public health and social sciences, as the population wide statistic could be estimated from small scale pilot studies.

Many simple classical randomized designs satisfy the property that all individuals have an equal marginal probability of treatment, in particular, this includes completely randomized design, which assigns a p fraction of individuals uniformly at random from the population to the treatment group. An important property is that *neither our estimator nor the appropriate randomized designs need to have knowledge of the underlying network*. In fact, all previously proposed solutions required knowledge of the network either for the randomized design or to compute the estimator. In applications where the network is not fully observed, our proposed estimators will still output in an unbiased estimate for the total treatment effect with simple randomizations that can be implemented without knowledge of the underlying network. This provides positive guarantees for settings in which the randomization may be limited due to regulatory policies or lack of precise network information.

7.2 Average Treatment Effect

Theorem 10 presents the necessary conditions for unbiased linear estimators for average treatment effect given baseline estimates.

Theorem 10. *For any randomized design such that $\mathbb{P}(z_i = 1 \mid z_k = 1) = \rho_i$ for all $(k, i) \in \mathcal{E}$ for some values of $\{\rho_i\}_{i \in [n]}$, the following simple estimator*

$$\widehat{ATE}_{-\alpha} = \frac{1}{n} \sum_{i \in [n]} \left(\frac{z_i}{\mathbb{E}[z_i]} - \frac{(1 - z_i)\rho_i}{\mathbb{E}[z_i](1 - \rho_i)} \right) (Y_i(z) - \alpha_i),$$

produces an unbiased estimate for the average treatment effect under heterogeneous linear network effects.

The condition that $\mathbb{P}(z_i = 1|z_k = 1) = \rho_i$ for all $(k, i) \in \mathcal{E}$ imposes symmetry in the randomized assignment amongst the neighbors of an individual. Completely randomized design or bernoulli randomization would satisfy this constraint. Alternatively a cluster based randomization would satisfy this constraint if the neighbors of a unit had equal chance of being assigned to the same versus different cluster.

Proof. After subtracting out the baseline parameters, the constraints for unbiasedness reduce to

- β : for all $i \in [n]$, $w_i \mathbb{E}[z_i] = \frac{1}{n}$,
- γ : for all $(k, i) \in \mathcal{E}$, $w_i \mathbb{E}[z_i z_k] + v_i \mathbb{E}[(1 - z_i) z_k] = 0$.

Satisfying the constraints arising from β and γ results in

$$w_i = \frac{1}{n \mathbb{E}[z_i]} \text{ and } v_i = -\frac{\mathbb{E}[z_i z_k]}{n \mathbb{E}[z_i] \mathbb{E}[(1 - z_i) z_k]} \text{ for all } (k, i) \in \mathcal{E}.$$

In order to ensure that such a valid v_i exists, we would need that there exists values ρ_i such that $\mathbb{P}(z_i = 1|z_k = 1) = \rho_i$ for all $(k, i) \in \mathcal{E}$. Under this condition,

$$\frac{\mathbb{E}[z_i z_k]}{\mathbb{E}[z_i] \mathbb{E}[(1 - z_i) z_k]} = \frac{\rho_i}{\mathbb{E}[z_i] (1 - \rho_i)}.$$

□

Under completely randomized design with a budget of treating p fraction of the population, $\mathbb{E}[z_i] = p$ and $\rho_i = \frac{pn-1}{n-1}$ for all $i \in [n]$. As a result,

$$\widehat{\text{ATE}}_{-\alpha} = \sum_{i \in [n]} \left(\frac{z_i(n-1) - pn + 1}{(1-p)pn^2} \right) Y_i(z) - \frac{(n-1)}{n(1-p)} \left(\frac{1}{pn} \sum_{i \in [n]} z_i \alpha_i \right) + \frac{pn-1}{(1-p)pn} \left(\frac{1}{n} \sum_{i \in [n]} \alpha_i \right).$$

Furthermore, for large enough n the empirical average baseline of the treated individuals would be similar to the population average baseline, such that this estimator could be computed using the population baseline estimates by

$$\widehat{\text{ATE}}_{-\alpha} \approx \sum_{i \in [n]} \left(\frac{z_i(n-1) - pn + 1}{(1-p)pn^2} \right) Y_i(z) - \frac{1}{pn^2} \sum_{i \in [n]} \alpha_i.$$

7.3 Average Interference Effect

Theorem 11 presents the necessary conditions for unbiased linear estimators for average interference effect given baseline estimates.

Theorem 11. *For any randomized design such that $\mathbb{P}(z_k = 1|z_i = 0) = \rho_i$ for all $(k, i) \in \mathcal{E}$ for some values of $\{\rho_i\}_{i \in [n]}$, the following simple estimator*

$$\widehat{\text{AIE}}_{-\alpha} = \frac{1}{n} \sum_{i \in [n]} \frac{(1 - z_i)}{\rho_i \mathbb{E}[1 - z_i]} (Y_i(z) - \alpha_i),$$

produces an unbiased estimate for the average interference effect under heterogeneous linear network effects.

Proof. After subtracting out the baseline parameters, the constraints for unbiasedness reduce to

- β : for all $i \in [n]$, $w_i \mathbb{E}[z_i] = 0$,
- γ : for all $(k, i) \in \mathcal{E}$, $w_i \mathbb{E}[z_i z_k] + v_i \mathbb{E}[(1 - z_i) z_k] = \frac{1}{n}$.

Satisfying the constraints arising from β and γ results in

$$w_i = 0 \text{ and } v_i = \frac{1}{n \mathbb{E}[(1 - z_i) z_k]} \text{ for all } (k, i) \in \mathcal{E}.$$

In order to ensure that such a valid v_i exists, we would need that there exists some value ρ_i for which $\mathbb{P}(z_k = 1 | z_i = 0) = \rho_i$ for all $(k, i) \in \mathcal{E}$. Under this condition,

$$\mathbb{E}[(1 - z_i) z_k] = \rho_i \mathbb{E}[1 - z_i].$$

□

The condition that $\mathbb{P}(z_k = 1 | z_i = 0) = \rho_i$ for all $(k, i) \in \mathcal{E}$ imposes symmetry in the randomized assignment amongst the neighbors of an individual. Completely randomized design or bernoulli randomization would satisfy this constraint. Alternatively a cluster based randomization would satisfy this constraint if the cluster assignments were also randomized, and neighbors of a unit had equal chance of being assigned to the same versus different cluster.

Under completely randomized design with a budget of treating p fraction of the population, $\mathbb{E}[z_i] = p$ and $\rho_i = \frac{pn}{n-1}$ for all $i \in [n]$. Furthermore, for large enough n the empirical average baseline of the treated individuals would be similar to the population average baseline, such that this estimator could be computed using the population baseline estimates by

$$\begin{aligned} \widehat{\text{AIE}}_{-\alpha} &= \frac{n-1}{(1-p)pn^2} \sum_{i \in [n]} (1 - z_i) Y_i(z) - \frac{n-1}{(1-p)pn^2} \sum_{i \in [n]} (1 - z_i) \alpha_i \\ &\approx \frac{n-1}{(1-p)pn^2} \sum_{i \in [n]} (1 - z_i) Y_i(z) - \frac{n-1}{pn^2} \sum_{i \in [n]} \alpha_i. \end{aligned}$$

7.4 Discussion

The above examples show that once we subtract the baselines, this opens up a significantly larger set of feasible unbiased individually weighted linear estimators for TTE, ATE, and AIE. Furthermore, the constraints on the randomized designs can be satisfied by standard simple randomizations such as completely randomized design or a matched pair design (i.e. cluster stratified randomized design with small clusters such that most edges are across clusters). In particular, we can implement many of these estimators without requiring perfect observation of the underlying network, which was a key limitation of the Horvitz-Thompson estimator for total treatment effect (requiring edges to be treated or placed in control together). This is a very useful property, as in practice the network could be only partially or noisily observed, or in some contexts it may not be observed at all. The above proposed estimators will still result in an unbiased estimate with simple randomizations that can be implemented without knowledge of the underlying network. This provides positive guarantees for settings in which the randomization may be limited due to regulatory policies or lack of precise network information.

Suppose the network is generated from a stochastic block model for which there are k communities. The community type of unit i is denoted by $q_i \in [k]$, and the probability of edge (k, i) is

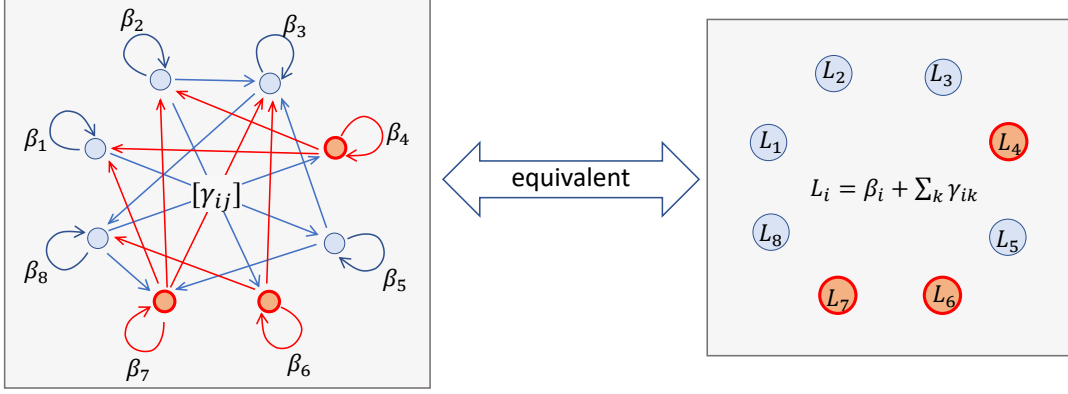


Figure 2: *By incorporating baseline estimates, network causal inference reduces to estimating population mean influence. (Left) The total treatment effect (TTE) under heterogeneous linear outcomes is equivalent to the weighted sum of all the edges divided by the total number of individuals. The vertices corresponding to treated individuals are colored red, and all the outgoing edges from treated individuals are colored red. The proposed estimator $\widehat{TTE}_{-\alpha}$ is equal to the weighted sum of all red edges divided by the number of treated individuals. (Right) For each vertex i , we can sum the weights of outgoing edges into an influence term L_i . As a result this is equivalent to a parallel universe in which the total treatment effect is the average of individual influences L_i , and we observe the sum of the influences of all treated individuals.*

$B_{q_k q_i}$, where $B \in [0, 1]^{k \times k}$ is the connectivity parameter matrix. We can rearrange the estimator to show how it depends on the set of edges between any pair of communities,

$$\widehat{\text{est}}_{-\alpha}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} w_i z_i \beta_i + \sum_{\tau, \tau' \in [k]} \sum_{k: q_k = \tau} \sum_{i: q_i \in \tau'} (w_i z_i + v_i (1 - z_i)) z_k \gamma_{ki} \mathbb{I}((k, i) \in \mathcal{E}).$$

Suppose that we had two networks \mathcal{E} and \mathcal{E}' drawn from the same underlying stochastic block model community structure. One can consider that these are noisy or partially observed samples of the true network. Alternatively, one can consider this to be a model of stochasticity over the active links in the network, capturing the dynamics of a network that may appear different at different time points. Although the underlying community structure describes the overall relationships, at any given time the active network may be a sample from the corresponding stochastic block model. The stability of the proposed estimator could be captured by the closeness of the estimate between the two sampled instances of the network, \mathcal{E} and \mathcal{E}' . Since each edge is sampled independently with a fixed probability as specified by the community structure, the concentration of this estimator with respect to a randomly sampled graph will follow from the concentration of a sum of weighted independent random variables.

When we may have the access to the network structure, and when we have the flexibility to design network dependent randomizations, we may be able to exploit the structure to reduce the variance of the estimator. In subsequent sections we provide variance calculations that give intuition for how the network connectivity structure may influence the variance of the estimator.

8 Reduction to Estimating Population Mean Influence

For the remainder of the paper we will focus on the following estimator for the Total Treatment Effect introduced in Corollary 9,

$$\widehat{\text{TTE}}_{-\alpha} = \frac{1}{n} \sum_{i \in [n]} \frac{Y_i(\mathbf{z}) - \alpha_i}{\mathbb{E}[z_i]}.$$

It is easy to verify that the bias of the estimator is given by

$$\mathbb{E}[\widehat{\text{TTE}}_{-\alpha}] - \text{TTE} = \frac{1}{n} \sum_{(i,k) \in \mathcal{E}} \left(\frac{\mathbb{E}[z_i]}{\mathbb{E}[z_k]} - 1 \right) \gamma_{ik},$$

which is zero when the marginal treatment probabilities are equal across individuals. This estimator is particularly simple because the weights are chosen such that $w_i = v_i$, i.e. each outcome is incorporated to the estimator with the same weight regardless of whether it is treated or not.

We define the “influence” of individual i on the estimate as

$$L_i = \beta_i + \sum_{k \in [n]} \frac{\mathbb{E}[z_i] \gamma_{ik}}{\mathbb{E}[z_k]}.$$

We refer to this as “influence” because L_i captures the contribution that individual i has towards the estimate $\widehat{\text{TTE}}_{-\alpha}$ when i is treated, including the interference effect it has on other individuals as well as the direct effect it has on itself. L_i does not depend on the realization of the treatment vector.

Under the heterogeneous linear outcomes model,

$$\widehat{\text{TTE}}_{-\alpha} = \frac{1}{n} \sum_{i \in [n]} \left(\frac{\beta_i}{\mathbb{E}[z_i]} + \sum_{k \in [n]} \frac{\gamma_{ik}}{\mathbb{E}[z_k]} \right) z_i =: \frac{1}{n} \sum_{i \in [n]} \frac{L_i z_i}{\mathbb{E}[z_i]}.$$

Written in this form, it is clear that the $\widehat{\text{TTE}}_{-\alpha}$ is simply an inverse propensity weighted estimator, although the terms L_i are not actually observed. As a result, analytical expressions of the variance of our estimate will follow from direct calculations over the inverse propensity weighted estimates.

Furthermore, under the sufficient conditions for unbiasedness when $\mathbb{E}[z_i] = \mathbb{E}[z_k]$ for all $(i, k) \in \mathcal{E}$, the total treatment effect is equal to the population mean of the influence terms,

$$\text{TTE} = \frac{1}{n} \sum_{i \in [n]} L_i.$$

As a result, under the mild assumption of having access to baseline estimates, we have reduced the complex task of network causal inference to a simple task of estimating a population mean of the influence via the sample average of the influence of treated individuals. This removes all complexity of network interference as the influence terms do not interact. Although our randomized designs and estimators do not require any knowledge of the network, the distribution of the influence will depend on the network, and will subsequently affect the variance of our estimator.

9 Variance of Estimate for Total Treatment Effect

As there are still many different randomized treatment designs one could use, we compare the variance of our proposed estimator for the total treatment effect under some commonly used randomized designs. We present similar variance calculations in Appendix C for general estimators, which apply to our estimators for the ATE and AIE as well.

Let p denote the treatment budget, such that the size of the treatment group can be at most p fraction of the population. As our estimator $\widehat{\text{TTE}}_{-\alpha}$ corresponds to approximating the total treatment effect with the average of the influence of the treated individuals, this estimator directly inherits properties from the analysis of a sample average estimator for the population mean of the influence. The variance of the estimator is

$$\text{Var}[\widehat{\text{TTE}}_{-\alpha}] = \sum_{i,j} \frac{L_i L_j \text{Cov}(z_i, z_j)}{n^2 \mathbb{E}[z_i] \mathbb{E}[z_j]}.$$

The randomized design affects the variance through the covariance matrix of the treatment vector.

9.1 Completely Randomized Design (CRD)

The completely randomized design generates the treatment assignment vector \mathbf{z} by selecting a subset of pn units to treat uniformly at random out of the size n population. This randomized design is commonly used in the classical setting without network interference. Due to the uniform sampling, $\mathbb{E}[z_i] = p$ for all $i \in [n]$, and

$$\text{Var}[\widehat{\text{TTE}}_{-\alpha}] = \frac{1-p}{p(n-1)} \left(\frac{1}{n} \sum_{i \in [n]} L_i^2 - \left(\frac{1}{n} \sum_{i \in [n]} L_i \right)^2 \right).$$

The inner expression is equal to the population variance of the influence terms $\{L_i\}_{i \in [n]}$, which is bounded by $B^2 d_{\max}^2$, where d_{\max} denotes the maximum out-degree of the network, and B is a bound on the direct effect and network effect parameters. A simple bound on the variance thus scales as $B^2 d_{\max}^2 / pn$. As a result, the variance will converge to zero for large n as long as the number of treated individuals pn divided by a constant $B^2 d_{\max}^2$ goes to infinity with large n . This is an optimally efficient rate when B and d_{\max} are constants, as this only requires the number of treated individuals to be growing larger than a constant as n grows. In fact, even optimal estimators for causal effects under SUTVA, without network interference, results in a variance scaling as $1/pn$. This means that given the mild assumption of having access to baseline estimates, our approach under fully heterogeneous network effects attains a simple, unbiased, and optimally efficient estimate for the total treatment effect, under the simplest randomized design.

9.2 Cluster Randomized Design (cluster RD)

The cluster randomized design partitions the population into clusters, and all individuals in each cluster are either jointly placed in the treatment group or the control group. This is also referred to in the literature as block randomized design, where blocks refer to clusters. The treatment assignment vector is generated by selecting a subset of pT clusters to treat uniformly at random amongst the T clusters. In contrast to completely randomized design, the treatment of individuals within a cluster are perfectly correlated. This randomized design is commonly used in the network interference setting, where the clusters are additionally constructed to minimize edges across clusters, so that an individual and its local neighbors are jointly assigned to treatment or control as much as possible. In our setting, we do not require such conditions on the construction of the clusters, and thus our clusters may not correspond to tightly connected communities in the network.

The probability an individual is treated is equal to the probability that the cluster it belongs to is treated. Due to the uniform sampling across the clusters, it follows that $\mathbb{E}[z_i] = p$ for all $i \in [n]$. Let T denote the number of clusters, assuming clusters of uniform size n/T for simplicity. Let $\pi : [n] \rightarrow [T]$ denote the mapping that assigns individuals to clusters. Let L'_τ denote the average

value of the influence terms within cluster τ , $L'_\tau = \frac{T}{n} \sum_{i:\pi(i)=\tau} L_i$. The variance of our estimator under this randomized design is given by

$$\text{Var}[\widehat{\text{TTE}}_{-\alpha}] = \frac{1-p}{p(T-1)} \left(\frac{1}{T} \sum_{\tau \in [T]} L'_\tau{}^2 - \left(\frac{1}{T} \sum_{\tau \in [T]} L'_\tau \right)^2 \right).$$

Observe that the expression is very similar to the variance under the completely randomized design, except we are randomizing over clusters rather than individuals. The inner expression is equal to the variance across clusters of L'_τ , which is the average influence of individuals in cluster τ , bounded by $B^2(\max_{\tau \in [T]} \bar{d}_\tau)^2$, where \bar{d}_τ denotes the average out-degree of individuals in cluster τ . A bound on the variance thus scales as $B^2(\max_{\tau \in [T]} \bar{d}_\tau)^2/pT$.

If d_{\max} is constant, and if T is asymptotically smaller than n , i.e. the size of each cluster is not constant with respect to the population size n , then the variance under cluster randomized design is larger than the variance under completely randomized design. When d_{\max} may be large or even growing with n then the variance might be improved by using cluster randomized design with an optimal choice of clusters. In particular, there would be a tradeoff between the choice of cluster size, and the gain from smoothing out the degree distribution, i.e. the influence, across clusters. In particular, if there is high variation in the influence amongst individuals, then the variance would be minimized by splitting large influence individuals across different clusters, and grouping them with low influence individuals, in order to try to even out the average influence of each cluster. This requires detailed knowledge of the network however, which is often not available.

9.3 Saturation Randomized Design (SRD)

The saturation randomized design also assumes that the population is partitioned into clusters, but instead each cluster is treated at a specified saturation level, specifying the percentage of individuals in that cluster that are treated. This is also referred to as a cluster stratified randomized design. For a cluster τ , let p_τ denote the fraction of individuals treated in cluster τ , satisfying $\sum_{\tau \in [T]} n_\tau p_\tau = np$. Let T denote the number of clusters, and let $\pi : [n] \rightarrow [T]$ map individuals to clusters. Let $n_\tau = |\{i : \pi(i) = \tau\}|$ denote the size of cluster τ . Treatment across different clusters are assigned independently. For each cluster τ , a set of $p_\tau n_\tau$ individuals within the cluster are selected uniformly at random to be treated. Let V_τ denote the variance of the influence terms within cluster τ ,

$$V_\tau = \frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i^2 - \left(\frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i \right)^2.$$

Under uniform saturation, i.e. $p_\tau = p$ for all τ , the estimator is unbiased as $\mathbb{E}[z_i] = p$ for all i , and the variance is

$$\text{Var}[\widehat{\text{TTE}}_{-\alpha}] = \frac{1-p}{pn} \sum_{\tau \in [T]} \frac{n_\tau^2}{n(n_\tau - 1)} V_\tau.$$

This expression essentially scales as V_{avg}/pn , where V_{avg} denotes the weighted average across cluster variances V_τ . In particular, to minimize the variance, each cluster should be chosen to be as homogeneous as possible, so that there is little variation in the influence parameters within each cluster. If V_{avg} is significantly smaller than the overall population variance over the influence terms $\{L_i\}_{i \in [n]}$, then the uniform saturation randomized design improves in efficiency upon CRD. An extreme special case of this randomized design would be the matched pair randomized design, where the pairs correspond to the clusters, and $p = 1/2$. The pairs are selected to be as similar as possible on known features.

Constructing such clusters requires additional knowledge of covariates or network structure which is not always available; however, this analysis provides motivation that whenever we do have such auxiliary information at hand, we can only benefit by controlling for the variance which may be related to the auxiliary information. In particular, by grouping similar individuals together, we ensure that the distribution over the auxiliary information in the treated group is as similar as possible to the control group.

For a general choice of saturation levels, the estimator may be biased, and the bias will scale proportionally with the sum of the network effect of edges across clusters with different saturation levels. The variance of the estimator is given by

$$\text{Var}[\widehat{\text{TTE}}_{-\alpha}] = \sum_{\tau \in [T]} \frac{(1 - p_{\tau})n_{\tau}^2}{p_{\tau}n^2(n_{\tau} - 1)} V_{\tau}.$$

While varying the saturation levels would introduce bias into the estimator, it may be able to reduce the variance by allocating a larger treated fraction to clusters that are larger or that have larger within-cluster variance V_{τ} . the reduction in variance would need to be carefully balanced with the introduced bias however, which may be difficult to do since it would require auxiliary information about the cluster variances.

9.4 Discussion

One way to explain the relationship amongst the variances of different randomizations is that the variance is reduced when the randomization enforces that the distribution of types of units with respect to the expression L_i within treated and control are as similar as possible. This results from the observation that all of the above variance calculations boil down to variations of estimating the empirical variance of L_i amongst different subsets of the populations. For example, in cluster randomized design, each cluster is fully treated or control. In order to guarantee that the resulting treated population has a similar distribution of L_i to the control population, we should partition clusters so that the cluster average L_{τ} is similar across different clusters. In contrast, for cluster stratified randomized design, since a uniform fraction is treated within each cluster, in order to guarantee that the resulting treated population is most similar to the control population, we would want to partition clusters such that the units within a cluster are most similar or homogeneous as possible in terms of their L_i values.

Note that L_i is both a function of the parameters β_i, γ_{ik} , but also depends on the graph and degree of connectivity. If these parameters were linked to associated covariates, we could expect nodes with similar covariates as well as connectivity structure to have similar values for L_i . In this setting the relevant connectivity structure would be the distribution of edges that a node has to neighbors of different covariate types.

In the supplementary material, we provide variance expressions in the case when the model parameters and network are drawn stochastically from priors, which can help elucidate when the variation primarily results from the variability in the edge connectivity structure as opposed to heterogeneity in the model parameters. We also provide variance calculations for the general weighted linear estimator under standard randomizations, which would include the corresponding estimators for AIE and ATE.

Recall that $\widehat{\text{TTE}}_{-\alpha}$ can be rearranged into a sum of individual terms $L_i z_i$ across all i . For a sufficiently large population, the estimator and estimand are approximately Gaussian, and the variance calculations can be used to design a test and compute p-values. In practical applications, we cannot directly calculate the variance as given in the above expressions because the parameters

$\{\beta_i\}_{i \in [n]}$ and $\{\gamma_{ij}\}_{i,j \in [n]}$ are unknown. However, if one had prior domain knowledge which indicated that the parameters were correlated with certain covariates, these covariates could be used to design a clustering to guarantee that the distribution of these covariates between treated and control are as similar as possible, which would reduce the variance of the estimator.

10 Simulated Experiments

We test the proposed estimator on simulated data to verify the insights gained from the analysis. We generate networks from a stochastic block model (SBM) which exhibits denser connectivity within community than across communities. Each individual i is a member of one of T communities, denoted by $q_i \in [T]$. $B \in [0, 1]^{T \times T}$ is the connection probability matrix, such that $B_{q_i q_j}$ is the probability of an edge between individuals i and j , i.e. $(i, j) \in \mathcal{E}$. The parameters $\{\alpha_i, \beta_i\}_{i \in [n]}, \{\gamma_{ij}\}_{(i,j) \in \mathcal{E}}$ of the outcome model are drawn iid from a prior distribution. We fix the variance of the observation noise ϵ_i to 1. We explore the interplay between different parameter settings and choices of randomized designs and estimators.

10.1 Bias of Horvitz-Thompson Estimator

Recall that the Horvitz-Thompson estimator arose naturally from the unbiasedness conditions. In the presence of network interference effects, the behavior of the Horvitz-Thompson estimator depends crucially on the randomized design, in particular whether the units on endpoints of a network edge were assigned to treatment or control independently or in a correlated fashion.

In this experiment we illustrate this phenomenon by varying the fraction of edges that are either assigned independently or assigned together to treatment or control. We illustrate this dependence this by varying the connectivity structure of the communities in the graph and comparing completely randomized design, cluster randomized design, and stratified block randomized design. The clusters used in the randomized designs are chosen to align exactly with the stochastic block model (SBM) communities, such that they represent clustering by similar connectivity structures. We vary the strength of community clustering along the x-axis of the plot, such that the x-axis denotes the fraction of a unit's edges that are within the same community as opposed to across different communities. The within community edge probability is set to $B_{qq} = 0.7x$, and the across community edge probability is set to $B_{qq'} = 0.7(1-x)(n-k)/(nk-n)$ for $q \neq q'$. For any given unit i , the expected number of edges to units within the same community will be $0.7x(n-k)/k$, and the expected number of edges to units across a different community will be $0.7(1-x)(n-k)/k$. For $x \in [0, 1]$, the within and across community edge probabilities are varied to maintain the same expected overall degree. The treatment level is set to 50% (half treated, half control) for all randomized designs.

Each data point on the plot is generated by first sampling a graph of population size 400 units with 20 communities according to a SBM which exhibits the desired fraction of within-community vs. across-community edges. Then the outcome model parameters are drawn from a normal distribution, where the mean of α_i is 1, the mean of β_i is 5, the mean of γ_{ij} is 5, and the variances are all set to 1. Given the network and outcome model, for each randomized design that we are testing, we simulate 1000 instances of the treatment vector, which are used then to compute the empirical mean and variance of the corresponding estimator.

Figure 3 shows the mean and standard deviation of the estimates (along with the true TTE and ATE values). "CRD" refers to completely randomized design with Horvitz-Thompson estimator $\hat{\text{est}}_{HT}$, "block RD" refers to a cluster randomized design where each cluster is either treated or not along with the Horvitz-Thompson estimator $\hat{\text{est}}_{HT}$, "stratified RD" refers to the design where

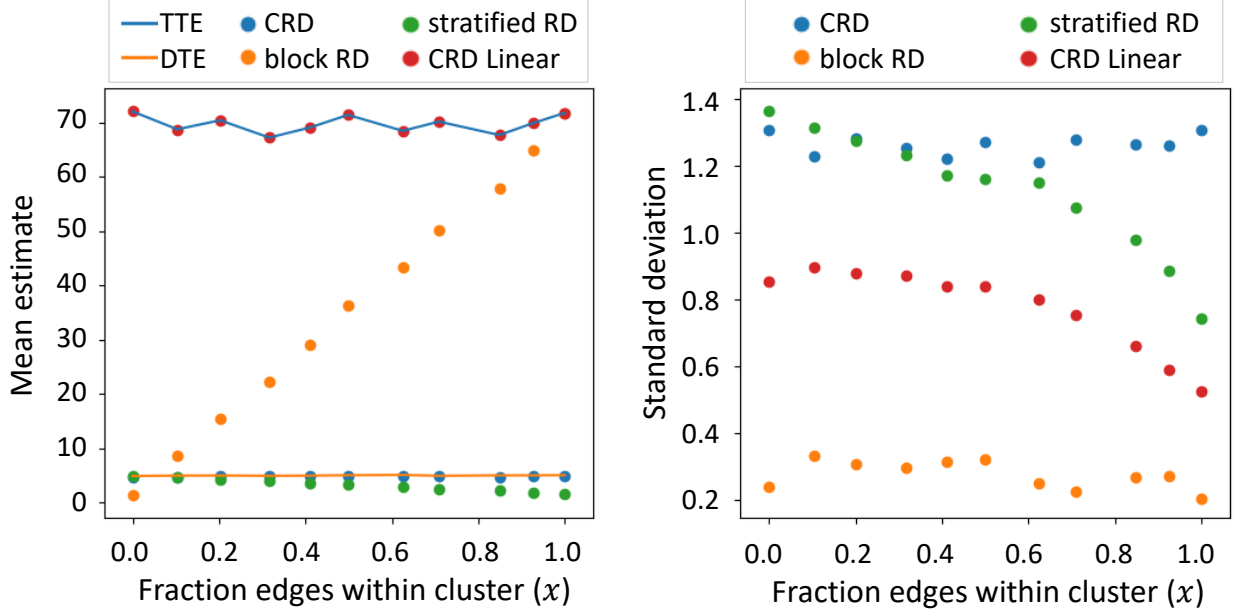


Figure 3: Illustrating the bias of the Horvitz-Thompson estimator with different randomized designs.

50% of the units in each cluster are treated along with the Horvitz-Thompson estimator $\hat{\text{est}}_{HT}$, and “CRD Linear” is the proposed estimator $\widehat{\text{TTE}}_{-\alpha}$ (which scales the total measured effect) with the completely randomized design.

This figure confirms that the proposed estimator $\widehat{\text{TTE}}_{-\alpha}$ as depicted by “CRD Linear” is an unbiased estimator of the TTE. The Horvitz-Thompson estimator $\hat{\text{est}}_{HT}$ with block randomized design, i.e. “block RD”, is a good estimate for the TTE if no edges are cut (i.e. clusters are very tight). In contrast, it is a good estimate for the ATE when almost all edges are cut across clusters, linearly extrapolating in between these two points depending on the fraction of edges within cluster verses across cluster. The Horvitz-Thompson estimator $\hat{\text{est}}_{HT}$ with completely randomized design, i.e. “CRD”, is a good estimate for the ATE, as the assignments across edges are nearly independent. The Horvitz-Thompson estimator $\hat{\text{est}}_{HT}$ with stratified cluster randomized design, i.e. “stratified RD”, is a good estimate for the ATE, although there is a small bias when there are more edges within cluster, as consistent with our calculations.

10.2 Variance as a function of Clustering

In the following set of experiments, we illustrate properties of the variance for our proposed estimator $\widehat{\text{TTE}}_{-\alpha}$ under different randomized designs. Recall that the variance calculations depend on balancing the parameter L_i between the treated and control populations, where recall that L_i represents the incremental change in the total measured outcome amongst all units due to unit i being treated. The variance can be reduced by a randomized design that ensures the treated and control population have similar distributions over the parameter L_i . This can be implemented by a cluster randomized design which balances the clusters to minimize variance of average L_i across-clusters, or a cluster stratified randomized design which minimizes variance within-cluster. The choice of clusters needs to appropriately match the randomized design. The term L_i depends both on the model parameters $\beta_i, \{\gamma_{ik}\}_{k:(i,k) \in \mathcal{E}}$ as well as the edge connectivity structure $\{k : (i, k) \in \mathcal{E}\}$.

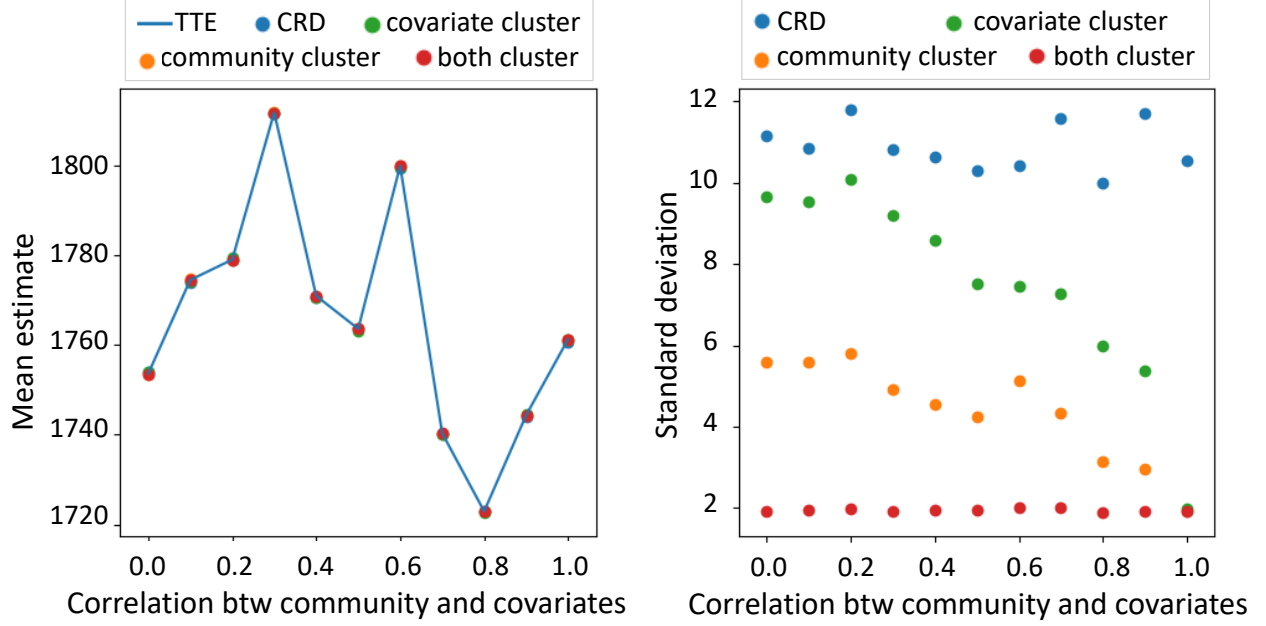


Figure 4: Illustrating the variance of the proposed estimator with different clustering schemes as a function of the correlation between community and covariate types.

In this simulated experiment we explore the impact of the clustering design on the variance of the proposed estimator as a function of the variation between the parameters and edge connectivity structure across the population.

To illustrate the variation of parameters within different subgroups of the population, we let each unit i be associated both to a community type $q_i \in [T]$ which governs its connectivity structure, as well as a covariate type $r_i \in [m]$ which governs its model parameters. The variation of parameters within a covariate type is smaller than the variation of parameters across different covariate types. Let σ_B^2 denote the variance amongst the connectivity parameter matrix B , and let σ_R^2 denote the variance amongst the parameters across different covariate types.

We sample B according to $B_{aa} \sim \text{Proj}(N(0.9, 2\sigma_B^2), [0, 1])$ and $B_{ab} \sim \text{Proj}(N(0.2, \sigma_B^2), [0, 1])$ for $a \neq b$, where $\text{Proj}(X, [0, 1])$ simply projects X onto the $[0, 1]$ interval. This creates a community structure where within-community edges are typically denser than across-community, yet there is variation in that some communities are tightly connected and others are more loosely connected.

For covariate types $r, r' \in [m]$, we draw iid parameters $\mu_\alpha(r) \sim N(1, \sigma_R^2)$, $\mu_\beta(r) \sim N(5, \sigma_R^2)$, and $\mu_\gamma(r, r') \sim N(5, \sigma_R^2)$. Then the individual parameters are drawn such that $\alpha_i \sim N(\mu_\alpha(r_i), 0.1)$, $\beta_i \sim N(\mu_\beta(r_i), 0.1)$, $\gamma_{ij} \sim N(\mu_\gamma(r_i, r_j), 0.1)$. Typically σ_R^2 is larger than 0.1 such that there is lower variance within any given covariate type than across different covariate types.

10.2.1 Correlation between Community and Covariate types

In the simulations depicted in figure 4, we vary the correlation between community types and covariate types. We compare the completely randomized design (CRD) with the cluster stratified randomized design for 3 different clusterings, clustering according to community type, covariate type, and both community and covariate type. We choose population size $n = 500$, communities $T = 5$, covariate types $m = 5$, and model variances $\sigma_B^2 = 0.25$ and $\sigma_R^2 = 0.8$.

To generate each data point, we first assign units to covariate types and community types to

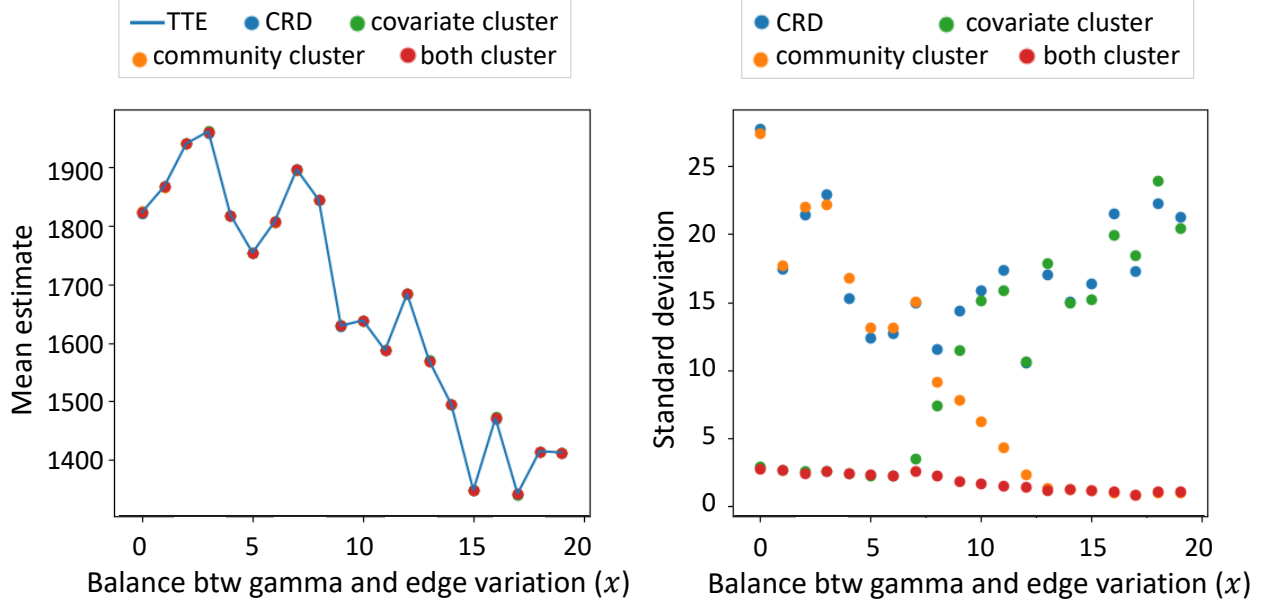


Figure 5: Illustrating the variance of the proposed estimator with different clustering schemes as the variation of the network effect and block model edge parameters vary.

achieve the desired correlation between covariates and communities. On the right side of the x -axis, the covariate types and community types are fully aligned, i.e. $q_i = r_i$, and on the left side they are fully independent in that each community has an even distribution of all the covariates, i.e. $q_i \perp r_i$. We sample 15 instances of the model specified by $\mathcal{E}, \{\alpha_i, \beta_i\}_{i \in [n]}, \{\gamma_{ij}\}_{i,j \in [n]}$ from the above defined prior distributions which also depend on first sampling $B, \{\mu_\alpha(r), \mu_\beta(r)\}_{r \in [m]}, \{\mu_\gamma(r, r')\}_{r,r' \in [m]}$. Given the network and outcome model, for each randomized design that we are testing, we simulate 100 instances of the treatment vector, which are used then to compute the empirical mean and variance of the corresponding estimator, averaging over the 15 model instances.

Consistent to our expectations from the variance calculations, figure 4 shows that the completely randomized design has the largest variance, as it does not exploit the underlying structure. Cluster stratified randomized design which clusters according to both community and covariate type has the smallest variance as expected. For the parameter settings we chose, clustering by community type has lower variance than clustering by covariate type, and the difference would depend on if the variation due to the parameters varying across covariate type is larger or smaller than the variation due to the edge density for different communities. As the covariate and community types align, this difference becomes smaller and eventually the variances converge for the different clustering schemes.

For a given application, we may not know in advance whether the variation across covariate types or community types is more significant, so clustering as finely as possible seems to be the best strategy, which taken to its extreme becomes similar to implementing a “matched pairs” randomization. Although it reduces variance, it may be more computationally expensive to form the fine clusters as opposed to simply clustering on meta-characteristics.

10.2.2 Variation due to Covariate Parameters or Edge Connectivity

In figure 5, we compare models with different relative ratios between the variation in model parameters σ_R^2 or the variation in edge connectivity patterns σ_B^2 , such that clustering either by community type, covariate type, or both illustrate the gains in variance due to reducing variation in model parameters or edge connectivity. We assume the covariate types $\{r_i\}_{i \in [n]}$ are uncorrelated with the community types $\{q_i\}_{i \in [n]}$. We compare the completely randomized design (CRD) with the cluster stratified randomized design for 3 different clusterings, clustering according to community type, covariate type, and both community and covariate type. We choose population size $n = 500$, communities $T = 5$ and covariate types $m = 5$.

To generate each data point, we first assign units to covariate types and community types such that each community has an equal distribution of covariate types. The x -axis is a parameter $x \in [0, 19]$ which trades off between σ_R^2 and σ_B^2 , defined according to

$$\begin{aligned}\sigma_R^2 &= \max(0, -2 + 0.3(20 - x)) \\ \sigma_B^2 &= \max(0, -1 + 0.15x),\end{aligned}$$

such that σ_B^2 is increasing in x and σ_R^2 is decreasing in x . On the left of the x -axis, there is large variation in parameters across different covariate types, but the community connectivity is fixed for within and across community edge probabilities. On the right of the x -axis, the expected parameters across covariate types are equal, but there is variation in the strength of connectivity for different communities. For each x , we sample 15 instances of the model specified by $\mathcal{E}, \{\alpha_i, \beta_i\}_{i \in [n]}, \{\gamma_{ij}\}_{i,j \in [n]}$ from the above defined prior distributions which also depend on first sampling $B, \{\mu_\alpha(r), \mu_\beta(r)\}_{r \in [m]}, \{\mu_\gamma(r, r')\}_{r,r' \in [m]}$. Given the network and outcome model, for each randomized design that we are testing, we simulate 100 instances of the treatment vector, which are used then to compute the empirical mean and variance of the corresponding estimator, averaging over the 15 model instances.

Consistent to our expectations from the variance calculations, figure 5 shows that if the variation is largely due to the model parameters across different covariate types, then a block stratified randomized design performs better if it clusters by covariates as opposed to community types. If the variation is largely due to differences in connectivity strength across different communities, then a block stratified randomized design performs better if it clusters by community types as opposed to covariates. Again the completely randomized design has the largest variance, as it does not exploit the underlying structure, and the block stratified randomized design which clusters according to both community and covariate type has the smallest variance as it controls for the variation in both community connectivity and covariate effect parameters.

11 Conclusion

Estimating causal effects under network interference is an important yet challenging problem. Previous solutions under general models are often too computationally or statistically costly, or are limited to very simplistic network structures, inhibiting adoption in practice. As a result, many practical solutions consider strong assumptions on the network effects, which end up reducing the estimation task to a simple regression problem. In contrast, we consider the heterogeneous linear outcomes model, which imposes additive network effects, but allows for full heterogeneity in the edge level network effect parameters. This model is significantly more flexible than the simple linear models used in practice. We analyze the properties of weighted linear estimators under our model,

and our results directly translate into the following insights that are simple to apply in practice. Given baseline estimates, we show that *network causal inference is as easy as estimating a population mean!* Most notably, our solution *does not require knowledge of the underlying network*, nor does it critically require strong structural conditions on the network.

Insight 1: Prior information from historical data or pilot studies is incredibly valuable; without such information, any unbiased estimate must use knowledge of the network.

We showed that without prior information, there does not exist any unbiased estimate for the total treatment effect which does not use network structure. In particular, we restricted to linear estimators with weights that depend only on whether an individual is treated or not, and not on the treatment of its neighbors. We showed that an unbiased linear estimator only exists if the network can be fully partitioned into disconnected components, such that the randomized design must jointly treat or not treat all individuals in each connected component. Furthermore, there does not exist any unbiased linear estimator for the average interference effect under any randomized design.

Insight 2: Use historical data or pilot studies to estimate the population baseline; use the following simple unbiased estimators to approximate the desired causal effects.

We proposed the following unbiased estimator for any randomized design for which the marginal treatment probability of each individual is p ,

$$\widehat{\text{TTE}}_{-\alpha} = \frac{1}{p} \left(\frac{1}{n} \sum_{i \in [n]} Y_i(\mathbf{z}) - \frac{1}{n} \sum_{i \in [n]} \alpha_i \right).$$

Under completely randomized design, the following estimators for the ATE and AIE are unbiased,

$$\begin{aligned} \widehat{\text{ATE}}_{-\alpha} &= \frac{n-1}{pn} \left(\frac{1}{(1-p)n} \sum_{i \in [n]} \left(\frac{(1-p)n}{n-1} z_i - \frac{pn-1}{n-1} (1-z_i) \right) Y_i(\mathbf{z}) - \frac{1}{n} \sum_{i \in [n]} \alpha_i \right), \\ \widehat{\text{AIE}}_{-\alpha} &= \frac{n-1}{pn} \left(\frac{1}{(1-p)n} \sum_{i \in [n]} (1-z_i) Y_i(\mathbf{z}) - \frac{1}{n} \sum_{i \in [n]} \alpha_i \right). \end{aligned}$$

Insight 3: The statistical properties of the total treatment effect estimator depend on the population distribution of individual influence. Under our additive network effects model, our proposed estimator for the total treatment effect takes the form of

$$\widehat{\text{TTE}}_{-\alpha} = \frac{1}{n} \sum_{i \in [n]} \frac{L_i z_i}{\mathbb{E}[z_i]} \quad \text{for } L_i = \beta_i + \sum_{k \in [n]} \frac{\mathbb{E}[z_i] \gamma_{ik}}{\mathbb{E}[z_k]},$$

where L_i quantifies individual i 's influence on the total treatment effect. This characterization shows that for a sufficiently large population, under simple randomized designs, the distribution of the estimator will also be approximately Gaussian. As a result, variance estimates could be used to design hypothesis tests and compute p-values.

Insight 4: Using completely randomized design (CRD) results in optimally efficient estimation of the total treatment effect when the effect sizes and out-degrees are

bounded. Under CRD, the variance of our estimator is roughly equal to $1/pn$ times the population variance of the influence terms $\{L_i\}_{i \in [n]}$, which is bounded by $B^2 d_{\max}^2$ when the causal effect parameters are bounded by B and the out-degree of each individual is bounded by d_{\max} . As a result, the estimator is consistent as long as the number of treated individuals pn is larger than a constant with respect to the population size n .

Insight 5: Utilize any auxiliary information about network structure or covariates of individuals in the population to control for variance that may arise from heterogeneity amongst individuals. The variance of the estimator is minimized by constructing a randomized design under which the distribution of the influence terms within treated and control are as similar as possible. The influence of individual i , denoted by L_i , is a function of the causal effect parameters β_i, γ_{ik} , but also depends on the network via its local neighborhood structure. If the causal effect parameters or the network structure were related to observed covariates, we could reduce the variance of the estimator by using a uniform saturation randomized design, where we group together individuals that are similar with respect to observed covariates and local neighborhood structure in the network.

Acknowledgements

This research was supported in part by Microsoft Research New England. We also gratefully acknowledge funding from the NSF under grants CCF-1948256 and CNS-1955997. Christina Lee Yu is also supported by an Intel Rising Stars Award.

References

- [1] Peter M. Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.
- [2] Peter M Aronow, Cyrus Samii, et al. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.
- [3] Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.
- [4] Eric Auerbach and Max Tabord-Meehan. The local approach to causal inference under network interference. Technical report, 2021.
- [5] Falco J Bargagli-Stoffi, Costanza Tortù, and Laura Forastiere. Heterogeneous treatment and spillover effects under clustered network interference. Technical report, 2020.
- [6] Guillaume W. Basse and Edoardo M. Airoidi. Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology*, 48(1):136–151, 2018.
- [7] Guillaume W Basse and Edoardo M Airoidi. Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858, 2018.

- [8] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 1028–1038. PMLR, 22–25 Jul 2020.
- [9] Jake Bowers, Mark M. Fredrickson, and Costas Panagopoulos. Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97–124, 2013.
- [10] Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.
- [11] Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2), 2019.
- [12] David Roxbee Cox. *Planning of experiments*. Wiley, 1958.
- [13] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.
- [14] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. Network a/b testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pages 399–409. International World Wide Web Conferences Steering Committee, 2015.
- [15] Yuchen Hu, Shuangning Li, and Stefan Wager. Average treatment effects in the presence of interference. Technical report, 2021.
- [16] Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103:832–842, 2008.
- [17] Ravi Jagadeesan, Natesh S Pillai, and Alexander Volfovsky. Designs for estimating the treatment effect in networks with interference. *The Annals of Statistics*, 48(2):679–712, 2020.
- [18] Vishesh Karwa and Edoardo M Airolidi. A systematic investigation of classical causal inference strategies under mis-specification due to network interference. Technical report, 2018.
- [19] Michael P Leung. Causal inference under approximate neighborhood interference. Technical report, 2019.
- [20] Wenrui Li, Daniel L Sussman, and Eric D Kolaczyk. Causal inference under network interference with noise. Technical report, 2021.
- [21] Lan Liu and Michael G. Hudgens. Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109(505):288–301, 2014. PMID: 24659836.
- [22] Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*, pages 3700–3708. PMLR, 2021.
- [23] Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1), 2013.

- [24] Ben M. Parker, Steven G. Gilmour, and John Schormans. Optimal design of experiments on connected units with application to social networks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):455–480, 2017.
- [25] Jean Pouget-Abadie, Martin Saveski, Guillaume Saint-Jacques, Weitao Duan, Ya Xu, Souvik Ghosh, and Edoardo Maria Airolidi. Testing for arbitrary interference on experimentation platforms. Technical report, 2017.
- [26] Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007.
- [27] Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34 – 58, 1978.
- [28] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airolidi. Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1027–1035. ACM, 2017.
- [29] Fredrik Sävje, Peter M Aronow, and Michael G Hudgens. Average treatment effects in the presence of unknown interference. *The Annals of Statistics*, 49(2):673–701, 2021.
- [30] Michael E Sobel. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.
- [31] Daniel L Sussman and Edoardo M Airolidi. Elements of estimation theory for causal effects in the presence of network interference. Technical report, 2017.
- [32] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012. PMID: 21068053.
- [33] Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International Conference on Machine Learning*, pages 1489–1497, 2013.
- [34] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337. ACM, 2013.
- [35] Tyler J. VanderWeele, Eric J. Tchetgen Tchetgen, and M. Elizabeth Halloran. Interference and sensitivity analysis. *Statist. Sci.*, 29(4):687–706, 11 2014.
- [36] Davide Viviano. Experimental design under network interference. Technical report, 2020.

A Optimal Individually Weighted Linear Estimators for Fixed Randomized Designs

There are some application domains in which the randomization is fixed due to regulatory policies, and thus we may want to find the individually weighted linear estimator that minimizes the mean squared error for a fixed randomized design. In fact we will show in this section that this corresponds to solving a quadratic program for the best set of weights to use. This follows from showing that

for linear estimators, the mean squared error is a quadratic function of the weights. Assume that the treatment vector z is sampled from a distribution P_z , and let τ denote the estimand. Let w and v denote the n -dimensional vectors consisting of the weights used in the estimator,

$$\hat{\text{est}}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} (w_i z_i + v_i (1 - z_i)) Y_i(\mathbf{z}). \quad (1)$$

We define the vectors $x, b \in \mathbb{R}^{2n}$ and the matrix $Q \in \mathbb{R}^{2n \times 2n}$ according to

$$x = \begin{bmatrix} w - v \\ v \end{bmatrix}, \quad (2)$$

$$b_i = \begin{cases} \mathbb{E}[z_i Y_i(\mathbf{z})] & \text{if } i \in [n] \\ \mathbb{E}[Y_{i-n}(\mathbf{z})] & \text{if } i \in [n+1, 2n] \end{cases}, \quad (3)$$

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad (4)$$

where $Q_{11}(i, j) = \mathbb{E}[z_i Y_i(\mathbf{z}) z_j Y_j(\mathbf{z})]$, $Q_{21}(i, j) = \mathbb{E}[Y_i(\mathbf{z}) z_j Y_j(\mathbf{z})]$, $Q_{12}(i, j) = \mathbb{E}[z_i Y_i(\mathbf{z}) Y_j(\mathbf{z})]$, and $Q_{22}(i, j) = \mathbb{E}[Y_i(\mathbf{z}) Y_j(\mathbf{z})]$. Using the above notation, we can show that the mean squared error is equal to

$$MSE := \mathbb{E}_{z \sim P_z}[(\hat{\text{est}}(\mathbf{w}, \mathbf{v}) - \tau)^2] = x^T Q x - 2\tau b^T x + \tau^2, \quad (5)$$

and the bias is linear in the weight vector,

$$\text{bias} := \mathbb{E}[\hat{\text{est}}(\mathbf{w}, \mathbf{v})] - \tau = b^T x - \tau. \quad (6)$$

Therefore the set of minimum variance unbiased estimators within this class of deterministic linear weighted estimators is given by solutions to the constrained quadratic program

$$\begin{aligned} & \underset{x}{\text{minimize}} && x^T Q x - 2\tau b^T x \\ & \text{subject to} && b^T x = \tau. \end{aligned} \quad (7)$$

Observe that Q is a symmetric matrix, and its form depends on the randomization P_z and the outcomes model for $Y_i(\mathbf{z})$. In general, the objective function would depend on the unknown parameters α, β, γ , thus this may be an interesting tool for studying theoretical properties of different models, but it may be difficult to optimize over in practice.

B Variance Calculations

We focus on the following estimator for the TTE:

$$\begin{aligned} \widehat{\text{TTE}}_{-\alpha} &= \sum_i \frac{(Y_i(z) - \alpha_i)}{n\mathbb{E}[z_i]} \\ &= \sum_i \frac{\beta_i z_i + \sum_j \gamma_{ji} z_j \mathbb{I}((j, i) \in \mathcal{E})}{n\mathbb{E}[z_i]} \\ &= \sum_i \left(\frac{\beta_i}{n\mathbb{E}[z_i]} + \sum_k \frac{\gamma_{ik}}{n\mathbb{E}[z_k]} \mathbb{I}((i, k) \in \mathcal{E}) \right) z_i. \end{aligned}$$

In expectation, this takes value

$$\mathbb{E}[\widehat{\text{TTE}}_{-\alpha}] = \frac{1}{n} \left(\sum_{i \in [n]} \beta_i + \sum_{(i,k) \in \mathcal{E}} \frac{\mathbb{E}[z_i]}{\mathbb{E}[z_k]} \gamma_{ik} \right).$$

For randomizations that satisfy $\mathbb{E}[z_i] = \mathbb{E}[z_k]$ for all $(k, i) \in \mathcal{E}$, this estimator is unbiased. We compare the variance calculations of a few standard randomizations below. The variance will depend on the expression

$$L_i = \beta_i + \sum_{k \in [n]} \frac{\mathbb{E}[z_i]}{\mathbb{E}[z_k]} \gamma_{ik} \mathbb{I}((i, k) \in \mathcal{E}).$$

When $\mathbb{E}[z_i] = \mathbb{E}[z_k]$, then the second term of L_i simply measures the influence that unit i has on its neighbors if unit i were treated, such that L_i can be interpreted as the incremental change in the total measured outcome amongst all units due to the unit i being treated. The estimator can be rewritten as

$$\widehat{\text{TTE}}_{-\alpha} = \sum_i \frac{L_i z_i}{n \mathbb{E}[z_i]}.$$

Essentially this boils down to estimating the average of L_i across the population when we can only observe the sum of a randomly selected subset of units. The variance of the estimator is given by

$$\text{Var}[\widehat{\text{TTE}}_{-\alpha}] = \sum_{i,j} \frac{L_i L_j \text{Cov}(z_i, z_j)}{n^2 \mathbb{E}[z_i] \mathbb{E}[z_j]}.$$

Therefore, for specific randomized designs, we simply plug in the expressions for $\text{Cov}(z_i, z_j)$. For $i = j$, $\text{Cov}(z_i, z_j) = \mathbb{E}[z_i](1 - \mathbb{E}[z_i])$.

B.1 Completely Randomized Design

The completely randomized design generates the treatment assignment vector \mathbf{z} by selecting a subset of pn units to treat uniformly at random out of the size n population. Therefore $\mathbb{E}[z_i] = p$ for all $i \in [n]$, and $\text{Cov}(z_i, z_j) = -\frac{p(1-p)}{n-1}$ for $i \neq j$.

$$\begin{aligned} & \sum_i \frac{L_i^2 (1-p)}{n^2 p} - \sum_{i \neq j} \frac{L_i L_j (1-p)}{n^2 (n-1) p} \\ &= \frac{1-p}{p} \frac{1}{n-1} \left(\frac{1}{n} \sum_i L_i^2 - \left(\frac{1}{n} \sum_i L_i \right)^2 \right). \end{aligned}$$

B.2 Cluster Randomized Design

The cluster randomized design partitions the population into clusters, and each cluster is either fully treated or fully control. the treatment assignment vector is generated by selecting a subset of pT clusters to treat uniformly at random amongst the T clusters. We can verify that $\mathbb{E}[z_i] = p$ for all $i \in [n]$. Let T denote the number of clusters assuming clusters of uniform size n/T , and let $\pi : [n] \rightarrow [T]$ map units to clusters. Let L_τ denote the average value of L_i within cluster τ ,

$$L_\tau = \frac{T}{n} \sum_{i:\pi(i)=\tau} L_i.$$

For $i \neq j$, if $\pi(i) = \pi(j)$ then $\text{Cov}(z_i, z_j) = p(1-p)$, and if $\pi(i) \neq \pi(j)$ then $\text{Cov}(z_i, z_j) = -\frac{p(1-p)}{T-1}$. The variance under this randomization is given by

$$\begin{aligned} & \frac{(1-p)}{T^2 p} \sum_{\tau} L_\tau^2 - \frac{(1-p)}{T^2 p(T-1)} \sum_{\tau \neq \tau'} L_\tau L_{\tau'} \\ &= \frac{1-p}{p} \frac{1}{(T-1)} \left(\frac{1}{T} \sum_{\tau \in [T]} L_\tau^2 - \left(\frac{1}{T} \sum_{\tau \in [T]} L_\tau \right)^2 \right). \end{aligned}$$

B.3 Cluster Stratified Randomized Design with Uniform Treatment Fraction

The cluster stratified randomized design partitions the population into clusters, and within each cluster a specified fraction of the units are treated uniformly at random amongst all units in that cluster. The assignments in different clusters are fully independent. Assume uniform treatment fraction such that p fraction of units are treated within each cluster, which guarantees that $\mathbb{E}[z_i] = p$ for all $i \in [n]$. Let T be the number of clusters, let $\pi : [n] \rightarrow [T]$ map units to clusters, and let $n_\tau = |\{i : \pi(i) = \tau\}|$ be the size of cluster τ . Recall that

$$L_\tau = \frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i.$$

For $i \neq j$, if $\pi(i) = \pi(j) = \tau$ then $\text{Cov}(z_i, z_j) = -\frac{p(1-p)}{n_\tau}$, and if $\pi(i) \neq \pi(j)$ then $\text{Cov}(z_i, z_j) = 0$. The variance under this randomized design is

$$\begin{aligned} & \sum_i \frac{L_i^2(1-p)}{n^2 p} - \sum_{\tau} \sum_{i \neq j: \pi(i)=\pi(j)=\tau} \frac{L_i L_j (1-p)}{n^2 p n_\tau} \\ &= \frac{1-p}{p} \sum_{\tau \in [T]} \frac{n_\tau^2}{n^2(n_\tau-1)} \left(\frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i^2 - \left(\frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i \right)^2 \right). \end{aligned}$$

B.4 Cluster Stratified Randomized Design with Varying Treatment Fraction

We can also vary the treatment fraction within clusters to satisfy the global constraint that np units are treated. Let p_τ denote the fraction of units treated in cluster τ , satisfying $\sum_{\tau \in [T]} n_\tau p_\tau = np$. Let T denote the number of clusters, let $\pi : [n] \rightarrow [T]$ map units to clusters, and let $n_\tau = |\{i : \pi(i) = \tau\}|$ be the size of cluster τ . Since $\mathbb{E}[z_i] = p_{\pi(i)}$ is no longer the same for all i , the estimator may no longer be unbiased, and its bias is given by

$$\begin{aligned} \mathbb{E}[\widehat{\text{TTE}}_{-\alpha} - \text{TTE}] &= \frac{1}{n} \sum_{(j,i) \in \mathcal{E}} \gamma_{ji} \left(\frac{p_{\pi(j)}}{p_{\pi(i)}} - 1 \right) \\ &= \frac{1}{n} \sum_{\tau, \tau'} \left(\frac{p_\tau}{p_{\tau'}} - 1 \right) \sum_{(k,i) \in \mathcal{E}_{\tau\tau'}} \gamma_{ki}. \end{aligned}$$

For $i \neq j$, if $\pi(i) = \pi(j) = \tau$ then $\text{Cov}(z_i, z_j) = -\frac{p_\tau(1-p_\tau)}{n_\tau}$, and if $\pi(i) \neq \pi(j)$ then $\text{Cov}(z_i, z_j) = 0$. The variance of the estimator is given by

$$\begin{aligned} & \sum_i \frac{L_i^2(1-p_{\pi(i)})}{n^2 p_{\pi(i)}} - \sum_{\tau} \sum_{i \neq j: \pi(i)=\pi(j)=\tau} \frac{L_i L_j (1-p_\tau)}{n^2 p_\tau n_\tau} \\ &= \sum_{\tau \in [T]} \frac{1-p_\tau}{p_\tau} \frac{n_\tau^2}{n^2(n_\tau-1)} \left(\frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i^2 - \left(\frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i \right)^2 \right). \end{aligned}$$

B.4.1 Expected Variance when Parameters Sampled from a Prior

We can simplify the above expressions to get further intuition by assuming that the parameters $\beta_i, \alpha_i, \gamma_{ji}$ and the graph \mathcal{E} are drawn independently from a prior. In particular, we assume that each vertex i is associated with a community type q_i and a response type r_i . Assume that the parameters are drawn from a distribution with means $\mu_\beta(r_i), \mu_\alpha(r_i), \mu_\gamma(r_j, r_i)$, and variances $\sigma_\beta^2(r_i), \sigma_\alpha^2(r_i), \sigma_\gamma^2(r_j, r_i)$. We assume that the graph is independently sampled from a directed stochastic block model with edge connectivity parameter matrix B such that the probability of an edge from i to j is $B_{q_i q_j}$. We assume that each direction of an edge (i, j) and (j, i) are sampled independently to simplify the calculations. Assume that the observation noise has uniform variance of σ_ϵ^2 . We can compute the expectation of the above variances under a randomly sampled graph and randomly sampled parameters. Recall that

$$L_i = \beta_i + \sum_{k \in [n]} \frac{\mathbb{E}[z_i]}{\mathbb{E}[z_k]} \gamma_{ik} \mathbb{I}((i, k) \in \mathcal{E})$$

$$L_\tau = \frac{1}{n_\tau} \sum_{i: \pi(i) = \tau} L_i.$$

We can show that

$$\begin{aligned} \mathbb{E}[L_i] &= \mu_\beta(r_i) + \sum_{k: k \neq i} \mu_\gamma(r_i, r_k) B_{q_i q_k} \frac{\mathbb{E}[z_i]}{\mathbb{E}[z_k]} \\ \text{Var}[L_i] &= \sigma_\beta^2(r_i) \\ &\quad + \sum_{k: k \neq i} \left(\frac{\mathbb{E}[z_i]}{\mathbb{E}[z_k]} \right)^2 (\sigma_\gamma^2(r_i, r_k) B_{q_i q_k} + \mu_\gamma^2(r_i, r_k) B_{q_i q_k} (1 - B_{q_i q_k})) \\ \mathbb{E}[L_\tau] &= \frac{1}{n_\tau} \sum_{i \in \tau} \mathbb{E}[L_i] \\ \text{Var}[L_\tau] &= \frac{1}{n_\tau^2} \sum_{i \in \tau} \text{Var}[L_i] \end{aligned}$$

By the law of total variance,

$$\text{Var}[\widehat{\text{TTE}}_{-\alpha}] = \mathbb{E}[\text{Var}[\widehat{\text{TTE}}_{-\alpha} | \alpha, \beta, \gamma]] + \text{Var}[\mathbb{E}[\widehat{\text{TTE}}_{-\alpha} | \alpha, \beta, \gamma]].$$

The second term is the same for all designs,

$$\text{Var}[\mathbb{E}[\widehat{\text{TTE}}_{-\alpha} | \alpha, \beta, \gamma]] = \text{Var}\left[\frac{1}{n} \sum_i L_i\right] = \frac{1}{n^2} \sum_i \text{Var}[L_i]$$

The first term depends on the randomized design, and we show the resulting calculations below.

B.4.2 Completely Randomized Design

$$\begin{aligned} &\mathbb{E}[\text{Var}[\widehat{\text{TTE}}_{-\alpha}]] \\ &= \frac{1-p}{p} \frac{1}{n-1} \left(\frac{1}{n} \sum_i \mathbb{E}[L_i^2] - \mathbb{E}\left[\left(\frac{1}{n} \sum_i L_i\right)^2\right] \right) + \frac{\sigma_\epsilon^2}{np^2} \\ &= \frac{1-p}{pn^2} \sum_i \text{Var}[L_i] + \frac{1-p}{pn(n-1)} \sum_i \left(\mathbb{E}[L_i] - \frac{1}{n} \sum_j \mathbb{E}[L_j] \right)^2 + \frac{\sigma_\epsilon^2}{np^2} \end{aligned}$$

B.4.3 Cluster Randomized Design

The variance under this randomization (assume size of all clusters is equal) is given by

$$\begin{aligned}
& \mathbb{E}[\text{Var}[\widehat{\text{TTE}}_{-\alpha}]] \\
&= \frac{1-p}{p} \frac{1}{(T-1)} \left(\frac{1}{T} \sum_{\tau \in [T]} \mathbb{E}[L_\tau^2] - \mathbb{E}\left[\left(\frac{1}{T} \sum_{\tau \in [T]} L_\tau\right)^2\right] \right) + \frac{\sigma_\epsilon^2}{np^2} \\
&= \frac{1-p}{pT^2} \sum_{\tau \in [T]} (\mathbb{E}[L_\tau^2] - \mathbb{E}[L_\tau]^2) + \frac{\sigma_\epsilon^2}{np^2} \\
&\quad + \frac{1-p}{pT(T-1)} \left(\sum_{\tau} \mathbb{E}[L_\tau]^2 - \frac{1}{T} \left(\sum_{\tau \in [T]} \mathbb{E}[L_\tau] \right)^2 \right) \\
&= \frac{1-p}{pn^2} \sum_i \text{Var}[L_i] + \frac{\sigma_\epsilon^2}{np^2} \\
&\quad + \frac{1-p}{pT(T-1)} \sum_{\tau} \left(\mathbb{E}[L_\tau] - \frac{1}{T} \sum_{\tau' \in [T]} \mathbb{E}[L_{\tau'}] \right)^2
\end{aligned}$$

B.4.4 Cluster Stratified Randomized Design with Uniform Treatment Fraction

$$\begin{aligned}
& \mathbb{E}[\text{Var}[\widehat{\text{TTE}}_{-\alpha}]] \\
&= \frac{1-p}{p} \sum_{\tau \in [T]} \frac{n_\tau^2}{n^2(n_\tau-1)} \left(\frac{1}{n_\tau} \sum_{i \in \tau} \mathbb{E}[L_i^2] - \mathbb{E}\left[\left(\frac{1}{n_\tau} \sum_{i \in \tau} L_i\right)^2\right] \right) \\
&\quad + \frac{\sigma_\epsilon^2}{np^2} \\
&= \frac{1-p}{p} \sum_{\tau \in [T]} \frac{n_\tau^2}{n^2(n_\tau-1)} \frac{n_\tau-1}{n_\tau^2} \sum_{i \in \tau} (\mathbb{E}[L_i^2] - \mathbb{E}[L_i]^2) + \frac{\sigma_\epsilon^2}{np^2} \\
&\quad + \frac{1-p}{p} \sum_{\tau \in [T]} \frac{n_\tau^2}{n^2(n_\tau-1)} \left(\frac{1}{n_\tau} \sum_{i \in \tau} \mathbb{E}[L_i]^2 - \left(\frac{1}{n_\tau} \sum_{i \in \tau} \mathbb{E}[L_i]\right)^2 \right) \\
&= \frac{1-p}{pn^2} \sum_i \text{Var}[L_i] + \frac{\sigma_\epsilon^2}{np^2} \\
&\quad + \sum_{\tau \in [T]} \frac{(1-p)n_\tau}{pn^2(n_\tau-1)} \sum_{i \in \tau} \left(\mathbb{E}[L_i] - \frac{1}{n_\tau} \sum_{j \in \tau} \mathbb{E}[L_j] \right)^2
\end{aligned}$$

B.4.5 Cluster Stratified Randomized Design with Varying Treatment Fraction

$$\begin{aligned}
& \mathbb{E}[\text{Var}[\widehat{\text{TTE}}_{-\alpha}]] \\
&= \sum_{\tau \in [T]} \frac{1-p_\tau}{p_\tau} \frac{n_\tau^2}{n^2(n_\tau-1)} \left(\frac{1}{n_\tau} \sum_{i \in \tau} \mathbb{E}[L_i^2] - \mathbb{E}\left[\left(\frac{1}{n_\tau} \sum_{i \in \tau} L_i\right)^2\right] \right) \\
&\quad + \sum_{\tau} \frac{n_\tau \sigma_\epsilon^2}{n^2 p_\tau^2} \\
&= \sum_{\tau \in [T]} \frac{1-p_\tau}{p_\tau n^2} \sum_{i \in \tau} \text{Var}[L_i] + \sum_{\tau} \frac{n_\tau \sigma_\epsilon^2}{n^2 p_\tau^2} \\
&\quad + \sum_{\tau \in [T]} \frac{(1-p_\tau)n_\tau}{p_\tau n^2(n_\tau-1)} \sum_{i \in \tau} \left(\mathbb{E}[L_i] - \frac{1}{n_\tau} \sum_{i \in \tau} \mathbb{E}[L_i] \right)^2
\end{aligned}$$

C Variance for General Linear Estimators

We can compute the variance of a linear weighted estimator for commonly used randomizations as a function of the weights w_i and v_i . This could help in choosing the randomization for a general linear weighted estimator, or choosing the weights to balance between bias and variance for a fixed randomization and estimand. As the expressions given are for a general estimator, this can be used to compute variance for the above presented estimators for the ATE and AIE.

Consider the fully general linear weighted estimator of the form

$$\hat{\text{est}}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} (w_i z_i Y_i(\mathbf{z}) + v_i (1 - z_i) Y_i(\mathbf{z})),$$

where w_i and v_i are not functions of the treatment vector \mathbf{z} . When we have baseline estimates available, we would instead subtract them from the outcomes when designing the estimator resulting in

$$\hat{\text{est}}_{-\alpha}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} (w_i z_i + v_i (1 - z_i)) (Y_i(\mathbf{z}) - \alpha_i).$$

The variance of $\hat{\text{est}}_{-\alpha}(\mathbf{w}, \mathbf{v})$ will in fact follow from our calculations of the variance of $\hat{\text{est}}(\mathbf{w}, \mathbf{v})$ with the simplifying condition that all baseline parameters α_i will be set to zero in the variance calculations since we have already subtracted them from the outcomes in the estimator $\hat{\text{est}}_{-\alpha}(\mathbf{w}, \mathbf{v})$. As such we provide the calculations for the variance of $\hat{\text{est}}(\mathbf{w}, \mathbf{v})$ as this is strictly more general. We define the following expressions

$$L_i = (w_i - v_i)\alpha_i + w_i\beta_i + \sum_{k \in [n]} v_k \gamma_{ik} \mathbb{I}((i, k) \in \mathcal{E})$$

$$H_{ij} = (w_i - v_i)\gamma_{ji} \mathbb{I}((j, i) \in \mathcal{E}) + (w_j - v_j)\gamma_{ij} \mathbb{I}((i, j) \in \mathcal{E}).$$

By expanding the expressions for $Y_i(\mathbf{z})$ from the heterogeneous linear outcomes model and rearranging terms, we can rewrite the estimator in terms of L_i and H_{ij} according to

$$\begin{aligned} \hat{\text{est}}(\mathbf{w}, \mathbf{v}) &= \sum_{i \in [n]} w_i z_i (\alpha_i + \beta_i + \sum_{k \in [n]} \gamma_{ki} \mathbb{I}((k, i) \in \mathcal{E}) z_k) + \sum_{i \in [n]} v_i (1 - z_i) (\alpha_i + \sum_{k \in [n]} \gamma_{ki} \mathbb{I}((k, i) \in \mathcal{E}) z_k) \\ &= \sum_{i \in [n]} v_i \alpha_i + \sum_{i \in [n]} ((w_i - v_i)\alpha_i + w_i \beta_i) z_i + \sum_{i \in [n]} \sum_{k \in [n]} (w_i - v_i) \gamma_{ki} \mathbb{I}((k, i) \in \mathcal{E}) z_i z_k \\ &\quad + \sum_{i \in [n]} v_i (\sum_{k \in [n]} \gamma_{ki} \mathbb{I}((k, i) \in \mathcal{E}) z_k) \\ &= \sum_{i \in [n]} v_i \alpha_i + \sum_{i \in [n]} ((w_i - v_i)\alpha_i + w_i \beta_i + \sum_{k \in [n]} v_k \gamma_{ik} \mathbb{I}((i, k) \in \mathcal{E})) z_i \\ &\quad + \sum_{i < j \in [n]^2} ((w_i - v_i)\gamma_{ji} \mathbb{I}((j, i) \in \mathcal{E}) + (w_j - v_j)\gamma_{ij} \mathbb{I}((i, j) \in \mathcal{E})) z_i z_j \\ &= \sum_{i \in [n]} v_i \alpha_i + \sum_{i \in [n]} L_i z_i + \sum_{i < j \in [n]^2} H_{ij} z_i z_j. \end{aligned}$$

As a result the variance of $\hat{\text{est}}(\mathbf{w}, \mathbf{v})$ is given by

$$\begin{aligned} &\text{Var}[\hat{\text{est}}(\mathbf{w}, \mathbf{v})] \\ &= \text{Var}[\sum_{i \in [n]} L_i z_i] + 2\text{Cov}[\sum_{i \in [n]} L_i z_i, \sum_{i < j \in [n]^2} H_{ij} z_i z_j] + \text{Var}[\sum_{i < j \in [n]^2} H_{ij} z_i z_j] \\ &= \sum_{i, j \in [n]^2} L_i L_j \text{Cov}[z_i, z_j] + 2 \sum_{i \in [n]} \sum_{j < k \in [n]^2} L_i H_{jk} \text{Cov}[z_i, z_j z_k] + \sum_{i < j \in [n]^2} \sum_{k < \ell \in [n]^2} H_{ij} H_{k\ell} \text{Cov}[z_i z_j, z_k z_\ell] \\ &= \sum_{i, j \in [n]^2} L_i L_j \text{Cov}[z_i, z_j] + 2 \sum_{i \in [n]} \sum_{j < k \in [n]^2} L_i H_{jk} \text{Cov}[z_i, z_j z_k] + \sum_{i < j \in [n]^2} \sum_{k < \ell \in [n]^2} H_{ij} H_{k\ell} \text{Cov}[z_i z_j, z_k z_\ell] \\ &= \sum_{i, j \in [n]^2} L_i L_j \text{Cov}[z_i, z_j] + 2 \sum_{i \in [n]} \sum_{j < k \in [n]^2} L_i H_{jk} \text{Cov}[z_i, z_j z_k] + \sum_{i < j \in [n]^2} \sum_{k < \ell \in [n]^2} H_{ij} H_{k\ell} \text{Cov}[z_i z_j, z_k z_\ell]. \end{aligned}$$

For specific randomized designs, we simply plug in the expressions for the moments of the assignment vector \mathbf{z} .

C.1 Completely Randomized Design

Consider the completely randomized design, which generates the treatment assignment vector \mathbf{z} by selecting a subset of pn units to treat uniformly at random out of the size n population. The second, third, and fourth moments take value:

$$\begin{aligned} \text{Cov}[z_i, z_j] &= \begin{cases} p(1-p) & \text{if } i = j \\ -\frac{p(1-p)}{n-1} & \text{if } i \neq j \end{cases} \\ \text{Cov}[z_i, z_j z_k] &= \begin{cases} -2p(1-p)(\frac{np-1}{(n-1)(n-2)}) & \text{if } i \notin \{j, k\}, j < k \\ p(1-p)(\frac{np-1}{n-1}) & \text{if } i \in \{j, k\}, j < k \end{cases} \\ \text{Cov}[z_i z_j, z_k z_\ell] &= \begin{cases} p(\frac{np-1}{n-1})(1-p(\frac{np-1}{n-1})) & \text{if } (i, j) = (k, \ell), i < j \\ p(\frac{np-1}{n-1})(\frac{np-2}{n-2} - p(\frac{np-1}{n-1})) & \text{if } |\{i, j, k, \ell\}| = 3 \\ p(\frac{np-1}{n-1})(\frac{(np-2)(np-3)}{(n-2)(n-3)} - p(\frac{np-1}{n-1})) & \text{if } |\{i, j, k, \ell\}| = 4 \end{cases} \end{aligned}$$

The variance calculations follow by plugging the expressions for the moments and rearranging the expression algebraically,

$$\begin{aligned} \text{Var}[\hat{\text{est}}(\mathbf{w}, \mathbf{v})] &= (1-p)p \frac{n^2}{n-1} \left(\frac{1}{n} \sum_i L_i^2 - \left(\frac{1}{n} \sum_i L_i \right)^2 \right) \\ &\quad + \frac{(1-p)p(np-1)}{(n-1)(n-2)} \left(n \sum_{j < k \in [n]^2} (L_j + L_k) H_{jk} - 2 \sum_i L_i \sum_{j < k \in [n]^2} H_{jk} \right) \\ &\quad + p \left(\frac{np-1}{n-1} \right) \left(\left(\frac{np-2}{n-2} \right) \left(\frac{np-3}{n-3} \right) - p \left(\frac{np-1}{n-1} \right) \right) \left(\sum_{i < j \in [n]^2} H_{ij} \right)^2 \\ &\quad + \frac{np(1-p)(np-1)(np-2)}{(n-1)(n-2)(n-3)} \sum_{i \in [n]} \left(\sum_{j \neq i \in [n]} H_{ij} \right)^2 \\ &\quad + \frac{np(1-p)(np-1)}{(n-1)(n-2)} \left(1 - \frac{np-2}{n-3} \right) \sum_{i < j \in [n]^2} H_{ij}^2. \end{aligned}$$

C.2 Cluster RD

Consider the cluster randomized design, which partitions the population into clusters. Each cluster is either fully treated or fully control. The treatment assignment vector is generated by selecting a subset of pT clusters to treat uniformly at random amongst the T clusters.

We can analyze cluster RD using the equations from CRD. Let $\mathbf{z}' \in \{0, 1\}^T$ denote the cluster treatment vector such that $z_i = z'_{\pi(i)}$. Then the estimator can be written as a sum of z'_τ for $\tau \in T$, in particular

$$\sum_{i \in [n]} L_i z_i = \sum_{i \in [n]} L_i z_i \sum_{\tau \in [T]} \mathbb{I}(\pi(i) = \tau) = \sum_{\tau \in [T]} \left(\sum_{i: \pi(i) = \tau} L_i \right) z'_\tau$$

and

$$\begin{aligned} \sum_{i < j \in [n]^2} H_{ij} z_i z_j &= \sum_{i < j \in [n]^2} H_{ij} z_i z_j \sum_{\tau \leq \tau' \in [T]^2} \mathbb{I}(\pi(i) = \tau, \pi(j) = \tau' \text{ or } \pi(i) = \tau', \pi(j) = \tau) \\ &= \sum_{\tau \in [T]} \left(\sum_{i < j: \pi(i) = \pi(j) = \tau} H_{ij} \right) z'_\tau + \sum_{\tau < \tau' \in [T]^2} \left(\sum_{i < j: \{\pi(i), \pi(j)\} = \{\tau, \tau'\}} H_{ij} \right) z'_\tau z'_{\tau'}. \end{aligned}$$

If we define the notation

$$L'_\tau = \sum_{i: \pi(i) = \tau} L_i \quad \text{and} \quad H'_{\tau\tau'} = \sum_{i < j: \{\pi(i), \pi(j)\} = \{\tau, \tau'\}} H_{ij},$$

it follows that

$$\hat{\text{est}}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} v_i \alpha_i + \sum_{\tau \in [T]} (L'_\tau + H'_{\tau\tau}) z'_\tau + \sum_{\tau < \tau' \in [T]^2} H'_{\tau\tau'} z'_\tau z'_{\tau'}.$$

We can see that the estimator looks essentially the same as an estimator which operates at the level of the clusters with the modified parameters L' and H' . Therefore, the variance can be computed by substituting the relevant quantities into the variance calculations for completely randomized design.

$$\begin{aligned} \text{Var}[\hat{\text{est}}(\mathbf{w}, \mathbf{v})] &= (1-p)p \frac{T^2}{T-1} \left(\frac{1}{T} \sum_{\tau} (L'_\tau + H'_{\tau\tau})^2 - \left(\frac{1}{T} \sum_{\tau} L'_\tau \right)^2 \right) \\ &\quad + \frac{T(1-p)p(Tp-1)}{(T-1)(T-2)} \sum_{\tau < \tau' \in [T]^2} (L'_\tau + H'_{\tau\tau} + L'_{\tau'} + H'_{\tau'\tau'}) (H'_{\tau,\tau'} + H'_{\tau',\tau}) \\ &\quad - \frac{2(1-p)p(Tp-1)}{(T-1)(T-2)} \sum_{\tau''} (L'_{\tau''} + H'_{\tau''\tau''}) \sum_{\tau < \tau' \in [T]^2} (H'_{\tau,\tau'} + H'_{\tau',\tau}) \\ &\quad + p \left(\frac{Tp-1}{T-1} \right) \left(\left(\frac{Tp-2}{T-2} \right) \left(\frac{Tp-3}{T-3} \right) - p \left(\frac{Tp-1}{T-1} \right) \right) \left(\sum_{\tau < \tau' \in [T]^2} (H'_{\tau\tau'} + H'_{\tau'\tau}) \right)^2 \\ &\quad + \frac{Tp(1-p)(Tp-1)(Tp-2)}{(T-1)(T-2)(T-3)} \sum_{\tau \in [T]} \left(\sum_{\tau' \neq \tau \in [T]} (H'_{\tau\tau'} + H'_{\tau'\tau}) \right)^2 \\ &\quad + \frac{Tp(1-p)(Tp-1)}{(T-1)(T-2)} \left(1 - \frac{Tp-2}{T-3} \right) \sum_{\tau < \tau' \in [T]^2} (H'_{\tau\tau'} + H'_{\tau'\tau})^2. \end{aligned}$$

C.3 Cluster stratified RD

Consider the cluster stratified randomized design, which partitions the population into clusters, and within each cluster a specified fraction of the units are treated uniformly at random amongst all units in that cluster. The assignments in different clusters are fully independent. We can rearrange the estimator and use the independence of assignments across different clusters to simplify the variance calculations. Using the definitions of L_i and H_{ij} ,

$$\hat{\text{est}}(\mathbf{w}, \mathbf{v}) = \sum_{i \in [n]} v_i \alpha_i + \sum_{\tau \in [T]} \left(\sum_{i: \pi(i) = \tau} L_i z_i + \sum_{i < j: \pi(i) = \pi(j) = \tau} H_{ij} z_i z_j \right) + \sum_{\tau < \tau' \in [T]^2} \left(\sum_{i < j: \{\pi(i), \pi(j)\} = \{\tau, \tau'\}} H_{ij} z_i z_j \right).$$

Let us define the notation

$$A_\tau = \sum_{i:\pi(i)=\tau} L_i z_i + \sum_{i < j: \pi(i)=\pi(j)=\tau} H_{ij} z_i z_j \quad \text{and} \quad B_{\tau\tau'} = \sum_{i < j: \{\pi(i), \pi(j)\} = \{\tau, \tau'\}} H_{ij} z_i z_j.$$

Because the assignments are independent across different clusters, the variance calculation reduces to

$$\begin{aligned} \text{Var}[\hat{\text{est}}(\mathbf{w}, \mathbf{v})] &= \sum_{\tau \in [T]} \text{Var}[A_\tau] + \sum_{\tau \in [T]} \sum_{\tau' \neq \tau \in [T]} \text{Cov}[A_\tau, B_{\tau\tau'}] + \sum_{\tau < \tau' \in [T]^2} \text{Var}[B_{\tau\tau'}] \\ &\quad + \sum_{\tau} \sum_{\tau' \neq \tau'', \tau \notin \{\tau' \tau''\}} \text{Cov}[B_{\tau\tau'}, B_{\tau\tau''}]. \end{aligned}$$

Observe that $\text{Var}[A_\tau]$ is the same as the variance calculations for completely randomized design restricted to the cluster τ , such that

$$\begin{aligned} \text{Var}[A_\tau] &= (1 - p_\tau) p_\tau \frac{n_\tau^2}{n_\tau - 1} \left(\frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i^2 - \left(\frac{1}{n_\tau} \sum_{i:\pi(i)=\tau} L_i \right)^2 \right) \\ &\quad + \frac{2(1 - p_\tau) p_\tau (n_\tau p_\tau - 1)}{(n_\tau - 1)(n_\tau - 2)} \left(n_\tau \sum_{i < j: \pi(i)=\pi(j)=\tau} L_i H_{ij} - \sum_{k:\pi(k)=\tau} L_k \sum_{i < j: \pi(i)=\pi(j)=\tau} H_{ij} \right) \\ &\quad + p_\tau \left(\frac{n_\tau p_\tau - 1}{n_\tau - 1} \right) \left(\left(\frac{n_\tau p_\tau - 2}{n_\tau - 2} \right) \left(\frac{n_\tau p_\tau - 3}{n_\tau - 3} \right) - p_\tau \left(\frac{n_\tau p_\tau - 1}{n_\tau - 1} \right) \right) \left(\sum_{i < j: \pi(i)=\pi(j)=\tau} H_{ij} \right)^2 \\ &\quad + \frac{n_\tau p_\tau (1 - p_\tau) (n_\tau p_\tau - 1) (n_\tau p_\tau - 2)}{(n_\tau - 1)(n_\tau - 2)(n_\tau - 3)} \sum_{i:\pi(i)=\tau} \left(\sum_{j \neq i: \pi(j)=\tau} H_{ij} \right)^2 \\ &\quad + \frac{n_\tau p_\tau (1 - p_\tau) (n_\tau p_\tau - 1)}{(n_\tau - 1)(n_\tau - 2)} \left(1 - \frac{n_\tau p_\tau - 2}{n_\tau - 3} \right) \sum_{i < j: \pi(i)=\pi(j)=\tau} H_{ij}^2. \end{aligned}$$

For $\tau \neq \tau'$,

$$\begin{aligned} \text{Cov}[A_\tau, B_{\tau\tau'}] &= \frac{p_\tau (1 - p_\tau) p_{\tau'}}{n_\tau - 1} \left(n_\tau \sum_{i, k: \pi(i)=\tau, \pi(k)=\tau'} L_i H_{ik} - \sum_{i:\pi(i)=\tau} L_i \sum_{h, k: \pi(h)=\tau, \pi(k)=\tau'} H_{hk} \right. \\ &\quad + \frac{n_\tau (n_\tau p_\tau - 1)}{n_\tau - 2} \sum_{i < j: \pi(i)=\pi(j)=\tau} \sum_{k: \pi(k)=\tau'} H_{ij} (H_{ik} + H_{jk}) \\ &\quad - \frac{2(n_\tau p_\tau - 1)}{n_\tau - 2} \sum_{i < j: \pi(i)=\pi(j)=\tau} H_{ij} \sum_{h, k: \pi(h)=\tau, \pi(k)=\tau'} H_{hk} \\ &\quad \left. - \frac{2(n_\tau p_\tau - 1)}{n_\tau - 2} \sum_{i < j: \pi(i)=\pi(j)=\tau} H_{ij} \sum_{h, k: \pi(h)=\tau, \pi(k)=\tau'} H_{hk} \right). \end{aligned}$$

For $\tau \neq \tau'$,

$$\begin{aligned}
\text{Var}[B_{\tau\tau'}] &= \frac{n_\tau p_\tau (1-p_\tau) n_{\tau'} p_{\tau'} (1-p_{\tau'})}{(n_\tau - 1)(n_{\tau'} - 1)} \sum_{h,k:\pi(h)=\tau, \pi(k)=\tau'} H_{hk}^2 \\
&\quad + \frac{p_\tau p_{\tau'} n_{\tau'} (1-p_{\tau'}) (n_\tau p_\tau - 1)}{(n_\tau - 1)(n_{\tau'} - 1)} \sum_{k:\pi(k)=\tau'} \left(\sum_{i:\pi(i)=\tau} H_{ik} \right)^2 \\
&\quad + \frac{p_\tau p_{\tau'} n_\tau (1-p_\tau) (n_{\tau'} p_{\tau'} - 1)}{(n_\tau - 1)(n_{\tau'} - 1)} \sum_{i:\pi(i)=\tau} \left(\sum_{k:\pi(k)=\tau'} H_{ik} \right)^2 \\
&\quad + p_\tau p_{\tau'} \left(\left(\frac{n_{\tau'} p_{\tau'} - 1}{n_{\tau'} - 1} \right) \left(\frac{n_\tau p_\tau - 1}{n_\tau - 1} \right) - p_\tau p_{\tau'} \right) \left(\sum_{i,j:\pi(i)=\tau, \pi(j)=\tau'} H_{ij} \right)^2.
\end{aligned}$$

For a distinct triple $\tau \neq \tau' \neq \tau''$,

$$\begin{aligned}
\text{Cov}[B_{\tau\tau'}, B_{\tau\tau''}] &= \frac{p_\tau (1-p_\tau) p_{\tau'} p_{\tau''}}{n_\tau - 1} \left(n_\tau \sum_{i,k,h:\pi(i)=\tau, \pi(k)=\tau', \pi(h)=\tau''} H_{ik} H_{ih} \right. \\
&\quad \left. - \sum_{i,k:\pi(i)=\tau, \pi(k)=\tau'} H_{ik} \sum_{j,h:\pi(j)=\tau, \pi(h)=\tau''} H_{ih} \right).
\end{aligned}$$

The final variance of the estimator results from combining these expressions together.