Graph Agnostic Estimators with Staggered Rollout Designs under Network Interference

Anonymous Author(s)
Affiliation
Address
email

Abstract

Randomized experiments are widely used to estimate causal effects across a variety of domains. However, classical causal inference approaches rely on critical independence assumptions that are violated by *network interference*, when the treatment of one individual influences the outcomes of others. All existing approaches require at least approximate knowledge of the network, which may be unavailable and costly to collect. We consider the task of estimating the total treatment effect (TTE), or the average difference between the outcomes when the whole population is treated versus when the whole population is untreated. By leveraging a *staggered rollout* design, in which treatment is incrementally given to random subsets of individuals, we derive unbiased estimators for TTE that *do not rely on any prior structural knowledge of the network*, as long as the network interference effects are constrained to low-degree interactions among neighbors of an individual. We derive bounds on the variance of the estimators, and we show in experiments that our estimator performs well against baselines on simulated data. Central to our theoretical contribution is a connection between staggered rollout observations and polynomial extrapolation.

1 Introduction

A cornerstone of much of the classic causal inference literature is the stable unit treatment value assumption (SUTVA), which posits that an individual's potential outcome is a function only of their assigned treatment; there are no spillover effects due to the treatment of others. Such an assumption fails to account for the ways in which individuals interact in many real-world experimental settings. For instance, new features rolled out on social networking sites such as LinkedIn may alter these users' behaviors, which in turn affect how their connections (who do not have access to the feature) interact with the platform. Individuals receiving a vaccine against an infectious disease may reduce the transmission probability of the disease to others they interact with. Implementing a different pricing policy for a subset of individuals in an online marketplace such as Airbnb or a platform such as Uber could impact the experience of other users, as they compete for the same resources or same customers. Public health measures instituted in one city can limit travel to nearby communities, indirectly affecting their health outcomes or transit related outcomes. These examples illustrate how network interference may arise naturally from the connectedness of our society. Unfortunately, the standard causal inference techniques which do not account for network interference may result in aribtrarily biased estimates.

As these issues come into greater focus, there is a growing research area in developing new tools for causal inference under *network interference*, when the outcome of an individual can be affected by the treatment of another. Many approaches either propose complex graph-based cluster randomized designs, or require strong parametric assumptions on the network interference effects. A limitation

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

is that all these approaches require at least partial knowledge of the underlying network in order to implement the randomization or to compute the estimator. While structural knowledge is available to online social networks, other applications such as public health must reason about an unknown or potentially transient network. The additional effort required to collect or model network structure is both difficult and costly.

In this work, we explore the value of additional measurements which arise from a staggered rollout randomized design, in which the treatments are administered over a span of a few timepoints. For example, the experimentation team at LinkedIn may roll out an experiment over 5 days, increasing the fraction of treated individuals according to a schedule of 1%, 2%, 5%, 10%, 20%, where it continuously collects data and measurements before and during each day of the experiment. Not only is this type of experiment easy to implement in such applications, it is often desirable to implement treatments according to such a staggered rollout design as it allows the system to first ensure safety of the proposed treatment on a smaller test group before implementing it on larger groups. This type of experimental design is also common for trials involving healthcare and medicine due to the requirement of ensure safety considerations before testing for efficacy. A key contribution is that we show the additional measurements from a *staggered rollout* design enable graph agnostic causal inference, lifting all requirements on knowledge of the network.

We focus on estimating the *total treatment effect* (TTE), informally defined as the difference in average outcomes across the population between two scenarios: when all individuals are treated and when no individuals are treated. It has also been referred to as the global average treatment effect (GATE). The TTE is particularly pertinent to applications where the decision maker must choose between entirely adopting the new treatment or remaining with the status quo. For example, LinkedIn would like to choose a single news feed recommendation algorithm, and Airbnb and Uber would like to choose a single dynamic pricing algorithm. We assume *neighborhood interference*, where each individual is only affected by the treatments of its direct neighbors; this is only mildly restrictive as the neighborhood can be defined with respect to an unknown network, which is neither used for the estimator nor the randomized design.

Related Work. In addressing the challenges that arise from network interference, a key tension arises between the model assumptions and the simplicity and efficiency of the proposed estimator. Previously proposed model assumptions can be generally classified into assumptions on exposure functions [14, 1, 24, 2, 12], interference neighborhoods [22, 3, 18, 5], parametric structure [20, 4, 6, 9, 8], or a combination of these. Each of these assumptions lead to different solution concepts. All of these approaches rely on knowledge of the network mediating the interference effects.

One class of approaches relies on assumptions about the network structure. They assume *partial interference*, meaning 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 [17, 16, 10, 19, 13, 23, 5, 2]. This assumption is motivated by scenarios where the network is naturally strongly clustered. A natural solution is to randomize treatments over the groups jointly, such that each group is assigned to either fully treated or fully control. A drawback of this approach is that many networks are well-connected such that there is no clear clustering of the network which does not cut a significant fraction of the edges. The bias of standard estimators will scale with the number of edges that cross between groups, leading to proposed cluster randomized designs that randomize over clusters that are constructed to minimize edges between clusters [9, 8]. Constructing good clusters itself can be computationally intensive. Additionally some applications may prohibit such nonuniform treatment assignment probabilities due to fairness considerations. Under *neighborhood interference* assumptions, [22, 21] analyze the Horvitz-Thompson estimator alongside a cluster randomized design, which involves both clustering the graph and computing probabilities of entire neighborhoods being assigned to treatment or control over the distribution of clusterings, which is computationally intensive. When one is willing to impose a distributional model on the network itself, [11] provide central limit theorem convergence results for a related but weaker estimand measuring the change in outcomes under small perturbations of the fraction of treated individuals.

An alternate approach is to impose structure on the form of the network interference effects. The most common assumption is that the network effects are linear with respect to a specified statistic of the local neighborhood [20, 9, 4, 6, 15, 7]. The assumptions reduce the number of unknown parameters in the model 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 network effects.

The most similar work to our paper is the solution proposed in [25], which provides an estimator for the TTE under a heterogeneous linear interference model [8], also referred to as the joint assumptions of additivity of main effects and interference effects in [18]. Their estimator does not require knowledge of the network, but requires measurements over two time steps. Our work generalizes their results beyond linear to polynomial models, and we show that the staggered rollout experimental design enables graph agnostic causal inference.

Contributions. We show that under a staggered rollout experimental design, the task of estimating the total treatment effect reduces to polynomial extrapolation, where the degree of the polynomial is governed by the cardinality of interactions in the neighborhood interference model, bounded above by the degree of the graph. Our approach is the first in the literature to propose an estimator and randomized design that does not require any knowledge of the network structure, and yet is unbiased and consistent. We provide variance bounds on the estimator, showing that the variance only grows polynomially in the degree as opposed to the exponential growth that is exhibited in the Horvitz-Thompson estimator under simple Bernoulli randomized designs. We provide experiments that also illustrate that naively using regression models without allowing for heterogeneity could lead to significant bias, whereas our estimator is unbiased with significantly lower variance than the bias incurred due to a misspecified model. We are also the first to study the value of a staggered rollout experimental design in the presence of network interference, and we believe the overall framework could extend beyond polynomial models to other function classes, opening a new approach for handling network interference while allowing for flexible heterogeneity in the network effects.

2 Setup

Consider a population of n individuals, and assume that the network interference can be represented via an unknown directed graph with edge set $E \subset [n] \times [n]$. An edge $(j, i) \in E$ represents that individual i is affected by the treatment assignment of individual j; as such, self-loops are expected. The in-neighborhood of individual i is denoted by $\mathcal{N}_i = \{j \in [n] : (j, i) \in E\}$, and we let d_{in} denote the maximum in-degree, d_{out}, and $d = \max\{d_{\text{in}}, d_{\text{out}}\}$. We posit that the outcome of individual i as a function of the entire population's exposure to treatment can be expressed by the potential outcomes function $Y_i : \{0, 1\}^n \to \mathbb{R}$.

Our task is to estimate the total treatment effect (TTE), which represents the difference in average outcomes when the entire population is fully under treatment as opposed to fully under control, denoted as

$$\text{TTE} := \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \big(Y_i(\mathbf{1}) - Y_i(\mathbf{0})\big). \tag{2.1}$$

We use $\mathbf{z} \in \{0, 1\}^n$ to denote the treatment assignment vector, where $z_i = 1$ if individual i is assigned to treatment, and $z_i = 0$ if i is assigned to control. By definition of E, it follows that the potential outcomes functions satisfy neighborhood interference with respect to the graph defined by E.

Assumption 1 (Neighborhood Interference)**.** *$Y_i(\mathbf{z})$ only depends on the treatment of individuals in $\mathcal{N}_i$ (including i). Equivalently, $Y_i(\mathbf{z}) = Y_i(\mathbf{z}')$ for any $\mathbf{z}$ and $\mathbf{z}'$ such that $z_j = z'_j$ for all $j \in \mathcal{N}_i$.*

Additionally, as the treatment variables z_i are binary, any potential outcomes function satisfying neighborhood interference can be written as a polynomial in the neighborhood treatment variables:

$$Y_i(\mathbf{z}) = \sum_{\mathcal{S} \subseteq \mathcal{N}_i} a_{\mathcal{S}} \prod_{j \in \mathcal{S}} x_i \prod_{j' \in \mathcal{N}_i \setminus \mathcal{S}} (1 - x_i),$$

for some coefficients $\{a_{\mathcal{S}}\}$. We use the degree of the polynomial to quantify the complexity of the model. In full generality, any model satisfying the neighborhood interference assumption will have polynomial degree bounded by $\max_i |\mathcal{N}_i|$, the maximum in-degree of the graph. In this work we consider the scenario where the polynomial degree may be significantly smaller $\max_i |\mathcal{N}_i|$.

Assumption 2 (Low Polynomial Degree)**.** *The potential outcomes model has polynomial degree at most β, i.e. there exist coefficients $\{c_{i,\mathcal{S}}\}_{i\in[n],\mathcal{S}\subseteq[n]}$ such that for all i and $\mathbf{z}$,*

$$Y_i(\mathbf{z}) = \sum_{\mathcal{S}\subseteq\mathcal{N}_i,|\mathcal{S}|\leq\beta} c_{i,\mathcal{S}} \cdot \mathbb{I}\big(\mathcal{S} \text{ treated}\big) = \sum_{\mathcal{S}\subseteq\mathcal{N}_i,|\mathcal{S}|\leq\beta} c_{i,\mathcal{S}} \prod_{j\in\mathcal{S}} z_j. \tag{2.2}$$

We interpret the parameter $c_{i,\mathcal{S}}$ as the effect that treating all individuals in $\mathcal{S}$ has on the outcome of individual i. The coefficient $c_{i,\emptyset}$ represents individual i's outcome when everyone is assigned to control (i.e. their *baseline outcome*); this is unaffected by the treatment assignment. In the case of a singleton set $\mathcal{S} = \{j\}$, we will use shorthand $c_{ij} = c_{i,\{j\}}$. It follows that the total treatment effect is the sum of all $c_{i,\mathcal{S}}$ for nonempty subsets $\mathcal{S}$, i.e. TTE $= \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{\mathcal{S}\subseteq\mathcal{N}_i\\1\leq|\mathcal{S}|\leq\beta}} c_{i,\mathcal{S}}$.

The number of unknown parameters in this model are $\sum_{i\in[n]}\sum_{k=0}^{\beta}\binom{|\mathcal{N}_i|}{k}$, which scales as nd^{β}. When $\beta = 1$, the network effects resulting from treated neighbors is additive, and is also equivalent to the the heterogeneous linear outcomes model in [25]. This low degree assumption will not generally admit threshold models or saturation models, both of which would require the degree of Y_i to be $|\mathcal{N}_i|$.

An example in which the polynomial degree may be smaller than the neighborhood size would be a setting in which an individual's neighborhood can be further partitioned into smaller subcommunities: colleagues, university friends, high school friends, family, etc. Each subcommunity could have an additive affect on the individuals' outcome, but there may be nontrivial interactions among the treatments of individuals in the subcommunities. The polynomial degree would be bounded by the size of the largest subcommunity, which could be significantly smaller than the full neighborhood.

We let $Y_{\max}$ denote an upper bound on the absolute treatment effects for each individual, i.e.

$$Y_{\max} := \max_{i\in[n]} \textstyle\sum_{\mathcal{S}\subseteq\mathcal{N}_i,|\mathcal{S}|\leq\beta} |c_{i,\mathcal{S}}|.$$

It follows that the magnitude of the outcomes $Y_i(\mathbf{z})$ are bounded by $Y_{\max}$ for any treatment vector $\mathbf{z}$. We let L_j denote the absolute effect or influence that individual j has on the population outcomes,

$$L_j := \textstyle\sum_{i:j\in\mathcal{N}_i}\sum_{\mathcal{S}\subseteq\mathcal{N}_i,|\mathcal{S}|\leq\beta,j\in\mathcal{S}} |c_{i,\mathcal{S}}|.$$

Our boundedness assumption and the finiteness of our network imply the boundedness of the L_j. We denote the upper bound on the absolute effect or influence of any individual by $L_{\max} := \max_j \big\{L_j\big\}$.

Randomized Experiment Design. As it may be costly and/or detrimental to expose the entire population to treatment, we wish to estimate the total treatment effect after treating only a small random subset of individuals. In particular we assume that there may be an experimental budget that limits the proportion of individuals who may be treated. We will focus on two standard randomized designs. In Bernoulli design, a treatment vector $\mathbf{z}$ is obtained by independently sampling each coordinate from a Bernoulli(p) distribution, so that the probability that a subset of individuals $\mathcal{S}$ are all treated is $p^{|\mathcal{S}|}$. We assume that $p > \frac{1}{n}$ so that at least one individual is treated in expectation. In completely randomized design, a treatment vector $\mathbf{z}$ is obtained by uniformly sampling a subset of k individuals to treat for some fixed k. Here, the probability that a subset of individuals $\mathcal{S}$ are all treated is

$$\prod_{i=0}^{|\mathcal{S}|-1} \frac{k-i}{n-i} =: \left[\tfrac{k}{n}\right]^{|\mathcal{S}|}. \tag{2.3}$$

Throughout the paper, we utilize a *staggered rollout* experimental design. Treatment is assigned to individuals in T stages throughout the experiment. Overall, the individuals' outcomes are measured $T+1$ times: a baseline measurement before treatment, as well as a measurement after each treatment round. We'll use $\mathbf{z}^t$ to denote the vector of treatment assignment in round t, and assume that each entry z_i^t is monotone increasing with t (individuals cannot be un-treated). Notably, this monotonicity requirement introduces significant correlation between the treatment vectors.

3 Graph Agnostic Estimators under Staggered Rollout Design

To motivate the design of our estimators, we begin with a high-level view of estimating the total treatment effect. When we have no information about the underlying causal network, we do not

know how much of each individual's neighborhood is treated, so have no systematic way to predict what their potential outcome would be if the entire population were treated. However, we can aggregate the average of the individuals' outcomes to obtain a meaningful statistic. Consider the expected population average outcomes where the expectation is taken over the distribution of treatment vectors $\mathbf{z}$ sampled from a parameterized class of distributions $\mathcal{D}_x$, where $\mathcal{D}_0$ refers to the distribution that deterministically assigns all individuals to control, and $\mathcal{D}_1$ refers to the distribution that deterministically assigns all individuals to treatment. Consider the underlying expected outcome function $F_{\mathcal{D}} : [0,1] \to \mathbb{R}$ given by

$$F_{\mathcal{D}}(x) = \mathbb{E}\Big[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} Y_i(\mathbf{z})\Big]$$

where the expectation is taken over the distribution of treatment vectors $\mathbf{z} \sim \mathcal{D}_x$. By construction, the TTE is exactly $F(1) - F(0)$.

If we can implement a staggered rollout design where at stage t of the experiment, the marginal distribution of the treatment vector is $\mathcal{D}_{x_t}$, the observed average outcomes collected in the experiment at stage t would give noisy estimates of $F_{\mathcal{D}}(x_t)$. Under this framing, our goal is to use these measurements to extrapolate the value of $F(1)$. This provides a general framework for utilizing staggered rollout design to simplify estimation of the total treatment effect.

The simplest class of distributions we can consider is the Bernoulli(p) randomized design, in which each individual is independently assigned to treatment or control with probability p. For a degree-β polynomial potential outcomes model, the expected outcome function under this design is polynomial in the treatment probability p:

$$F_B(p) = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z})\Big] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{\mathcal{S}\subseteq\mathcal{N}_i\\|\mathcal{S}|\le\beta}} c_{i,\mathcal{S}}\cdot\mathbb{E}\Big[\prod_{j\in\mathcal{S}} z_j\Big] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{\mathcal{S}\subseteq\mathcal{N}_i\\|\mathcal{S}|\le\beta}} c_{i,\mathcal{S}}\cdot p^{|\mathcal{S}|}.$$

To implement a staggered rollout Bernoulli design with treatment probabilities $p_1 < p_2 < \ldots < p_T$ we independently sample $u_i \sim U[0,1]$ for each individual i. Then, for each $t \in [T]$, we define treatment vector $\mathbf{z}^t$ with $\mathbf{z}_i^t = \mathbb{I}\big(u_i \le p_t\big)$. This both ensures that the marginal distribution of the treatment vector at stage t is equivalent to the Bernoulli(p_t) randomized design, and that the treatment assignments are monotone over the rounds.

Alternatively, we can consider a completely randomized design (CRD) in which we fix a number of treated individuals k, and sample a subset of k individuals uniformly at random among all size k subsets in the population. For a degree-β polynomial potential outcomes model, the expected outcome function under this design is polynomial in the treated fraction k/n:

$$F_C(\tfrac{k}{n}) = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z})\Big] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{\mathcal{S}\subseteq\mathcal{N}_i\\|\mathcal{S}|\le\beta}} c_{i,\mathcal{S}}\cdot\mathbb{E}\Big[\prod_{j\in\mathcal{S}} z_j\Big] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{\mathcal{S}\subseteq\mathcal{N}_i\\|\mathcal{S}|\le\beta}} c_{i,\mathcal{S}}\cdot\Big[\tfrac{k}{n}\Big]^{|\mathcal{S}|}.$$

To implement a complete staggered rollout design, we sample a treatment vector from CRD(k_1) at stage 1, and at stage $t > 1$, we sample a treatment vector from CRD($k_t - k_{t-1}$) out of the remaining untreated individuals. The marginal distribution of the treatment vector at state t will be equivalent to the completely randomized design with parameter k_t.

To construct our estimators, we will make use of the Lagrange interpolation formula.

Definition 1 (Lagrange Interpolation)**.** *Given a dataset $\big\{(x_t, y_t)\big\}_{t=0}^{T}$ with distinct x-coordinates, the unique polynomial F of degree at most T with $F(x_t) = y_t$ for each t is given by*

$$F(x) = \sum_{t=0}^{T} \ell_{t,\mathbf{x}}(x)\cdot y_t, \qquad \ell_{t,\mathbf{x}}(x) = \prod_{\substack{s=0\\s\neq t}}^{T} \frac{x - x_s}{x_t - x_s}.$$

To estimate TTE, we require estimates of $F_B(x)$ or $F_C(x)$ at $x \in [0,1]$. As both F_B and F_C have degree at most β, they can be recovered from $\beta+1$ observations, requiring $T = \beta$ rounds of treatment. Given treatment targets $\mathbf{x} = (x_0, x_1, \ldots x_T)$ with realized treatment schedule $\{\mathbf{z}_t \sim \mathcal{D}_{x_t}\}$, we can

utilize Lagrange interpolation to derive the following polynomial interpolation (PI) estimator:

$$\widehat{\mathrm{TTE}}_{\mathrm{PI}}(\mathbf{x}) := \begin{cases} \sum_{t=0}^{T} \left(\ell_{t,\mathbf{x}}(1) - \ell_{t,\mathbf{x}}(0)\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z}^t)\right) & x_0 < x_1 < \ldots < x_T, \\ 0 & x_t = x_{t-1} \text{ for some } t \in [T]. \end{cases} \tag{3.1}$$

The separation into cases ensures that the Lagrange coefficients are well-defined. We assume that the degree β is known such that the experimenter can select $T = \beta$. We also assume that $\mathbf{x}$ is monotone, and define $\Delta_{\mathbf{x}} = \min_{t=1..m}\left\{x_t - x_{t-1}\right\}$. We can apply this estimator in both the Bernoulli and completely randomized design settings.

Theorem 2. *Consider a potential outcomes model with degree β. Under a staggered rollout Bernoulli design with distinct treatment probabilities $\mathbf{p} = (p_0 = 0, \ldots, p_\beta)$, the estimator $\widehat{\mathrm{TTE}}_{\mathrm{PI}}(\mathbf{p})$ is unbiased with variance $O\left(\frac{d^2\beta^2}{n} Y_{\max}^2 \Delta_{\mathbf{p}}^{-2\beta}\right)$.*

Theorem 3. *Consider a potential outcomes model with degree β. Under a staggered rollout completely randomized design with distinct treatment counts $\mathbf{k} = (k_0 = 0, k_1, \ldots, k_\beta)$, the* TTE *estimator $\widehat{\mathrm{TTE}}_{\mathrm{PI}}(\mathbf{k}/n)$ is unbiased with variance $O\left(\beta^2\, Y_{\max}^2 \left(\frac{d^2}{n} + \frac{\beta^2}{k_1}\right) \cdot \left(\frac{n}{\Delta_{\mathbf{k}}}\right)^{2\beta}\right)$.*

Proofs of both of these theorems are given in Appendix A. A key technical piece of the analysis is handling the strong correlation in the observations across measurements due to the monotonicity enforced by the staggered rollout design. In the case of a linear potential outcomes model, we can strengthen both of these variance bounds, which match the results from [25]

Corollary 4. *For a linear potential outcomes model:*

- *The estimator $\widehat{\mathrm{TTE}}_{\mathrm{PI}}(\mathbf{p})$ under* Bernoulli$(0, p)$ *design has variance at most $\frac{1-p}{np} \cdot L_{\max}^2$.*
- *The estimator $\widehat{\mathrm{TTE}}_{\mathrm{PI}}(\mathbf{k}/n)$ under* CRD$(0, k)$ *has variance at most $\frac{n-k}{nk} \cdot L_{\max}^2$.*

Observe that the Bernoulli estimator does not incorporate any information about the realized treatments. Notably, it does not account for the number of treated individuals. While this binomial random variable concentrates around its mean (especially for large values of n), it fails to account for significant deviations from this mean. Since this information is available at the time of estimation, it can be incorporated into an estimator. We let $\hat{\mathbf{k}} = (\hat{k}_0 = 0, \hat{k}_1, \ldots, \hat{k}_\beta)$ be the realized number of treated individuals at each time step, and consider the estimator $\widehat{\mathrm{TTE}}_{\mathrm{PI}}(\hat{\mathbf{k}}/n)$.

Theorem 5. *Consider a potential outcomes model with degree β. Under a staggered rollout Bernoulli design with treatment probabilities $\mathbf{p} = (p_0 = 0, \ldots, p_\beta)$, $\widehat{\mathrm{TTE}}_{\mathrm{PI}}(\hat{\mathbf{k}}/n)$ has bias decaying exponentially in n and variance $O\left(\beta^2\, Y_{\max}^2 \left(\frac{d^2}{n} + \frac{\beta^2}{p_1 n}\right) \cdot \Delta_{\mathbf{p}}^{-2\beta}\right)$.*

A proof is given in Appendix A. For large n, the performance of these three estimators will converge to each other. While our theoretical variance bound in Theorem 5 does not show improvement upon that from Theorem 2, our experimental results illustrate empirical improvements of this estimator.

Discussion. The above sections illustrate a natural relationship between the complexity of the model (i.e. its degree β) and the complexity of the randomized design and corresponding estimator; we require $\beta + 1$ outcome measurements in order to construct an unbiased estimator. Intuitively, each of these measurements allows us to quantify one "degree" of the network effects. A natural question is whether we continue to see improvements in the estimator when we increase the number of estimates beyond $\beta + 1$. Note that we restrict our attention to unbiased estimators, as we desire the asymptotic reduction in mean-squared error as the population grows large. We may thus assess the quality of an estimator by its variance.

In the linear setting, we show that these extra measurements do not help to reduce variance. In fact, we'll argue that the unbiased estimator with minimum variance is the one that ignores all but its first and last observations and then performs polynomial interpolation on these endpoints. We record this result in Theorem 6. A proof is given in Appendix B.

Theorem 6. *Suppose that the potential outcomes model is linear, and a staggered rollout Bernoulli design is implemented with a set of $T + 1$ distinct treatment probabilities $p_0 < p_1 < \ldots < p_T$. Then,*

the unbiased estimator for TTE *of the form*

$$\widehat{\text{TTE}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\alpha_t Y_i(\mathbf{z}^t)$$

that minimizes variance has $\alpha_T = \alpha_0 = \frac{-1}{p_T - p_0}$, $\alpha_T = \frac{1}{p_T - p_0}$ *and* $\alpha_1, \ldots, \alpha_{T-1} = 0$.

On one hand, such a result seems surprising: having more observations seems like it would only lead to a stronger estimator. However, what is overlooked is that there is strong correlation in the different measurements due to the monotonicity of treatments enforced in the staggered rollout design, such that the information in the first and last measurements contain all the useful information one could construct from the intermediate measurements.

Another question of interest is how one should determine the degree β if it is not known in advance. Even if we have many measurements, it may not always be wise to increase the degree of the interpolant, as this increases the magnitude of its slope outside of the interpolating region $[0, p]$. When the expected number of treated individuals np is small relative to the population size n (so $p \ll 1$), the value of the interpolant at 1 will be highly sensitive to any deviation of the later measurements from their expectation. On the other hand, choosing to fit a low degree polynomial may lead to bias if the underlying network effects exhibit higher order interactions. An interesting future direction is how to optimally perform model selection to balance between bias and variance.

4 Experiments

We provide simulations on synthetic data to illustrate the performance of our estimators relative to existing estimators. For a population of n individuals, we generate random directed networks of n nodes using a configuration model with in-degrees distributed as a power law with exponent 2.5, and out-degrees evenly shared among individuals. For degree β, we construct the following potential outcomes model:

$$Y_i(\mathbf{z}) = c_{i,\emptyset} + \sum_{j=1}^{n}\tilde{c}_{ij}z_j + \sum_{\ell=2}^{\beta}\left(\frac{\sum_{j\in\mathcal{N}_i}\tilde{c}_{ij}z_j}{\sum_{j\in\mathcal{N}_i}\tilde{c}_{ij}}\right)^{\ell}, \tag{4.1}$$

where $c_{i,\emptyset} \sim U[0,1]$, $\tilde{c}_{ii} \sim U[0,1]$, and for $i \neq j$, $\tilde{c}_{ij} = v_j|\mathcal{N}_i|/\sum_{k:(k,j)\in E}|\mathcal{N}_k|$ for $v_j \sim U[0,r]$, where r denotes a hyperparameter that governs the relative magnitude of the network effects relative to the direct effects. Essentially v_j represents the magnitude of individual j's influence, which is then shared among its out-neighbors proportional to their in-degrees.

Other Algorithms. We benchmark our proposed estimators against least squares regression and difference-in-mean estimators. As these estimators don't utilized the staggered rollout design, we evaluate them on the measurements taken at the last stage, T, of the experiment. We will use $\mathbf{z}$ to denote the treatment vector at time T (suppressing the superscript). As a network sampled from a configuration model does not exhibit clustering, the solutions that propose cluster randomized designs perform poorly, and thus we omit them from the experiments.

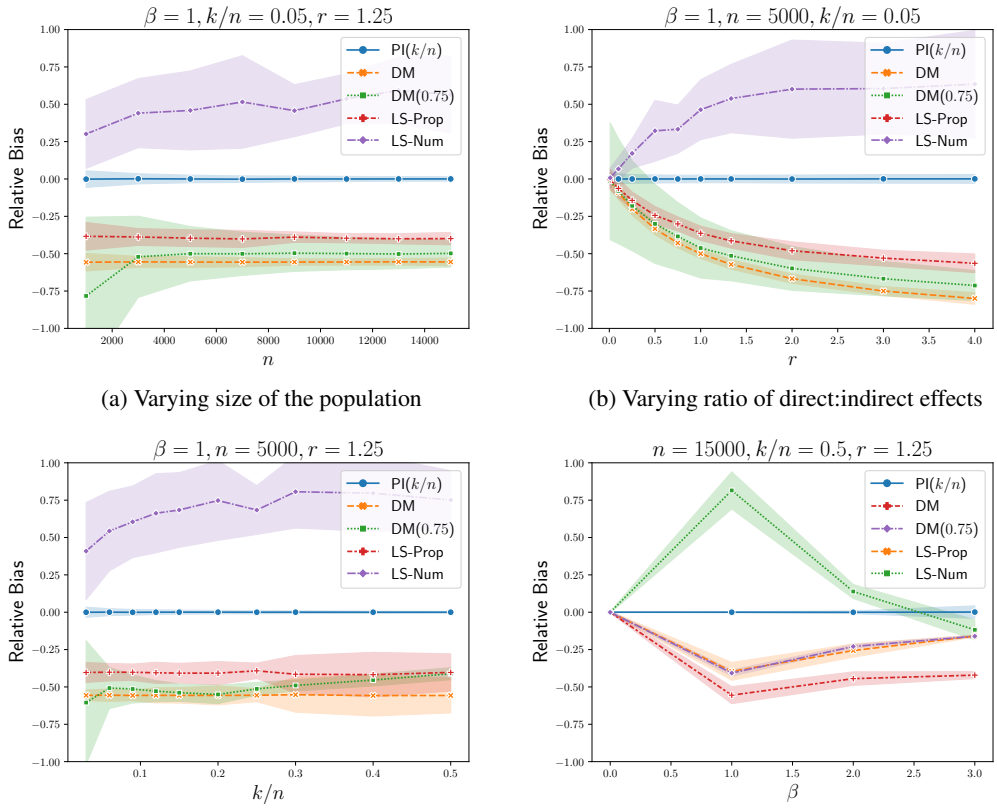
The standard difference in means estimator is the difference between the average outcome of individuals assigned to treatment and the average outcome of individuals assigned to control, given by

$$\widehat{\text{TTE}}_{\text{DM}} = \frac{\sum_{i\in[n]} z_i Y_i(\mathbf{z})}{\sum_{i\in[n]} z_i} - \frac{\sum_{i\in[n]}(1-z_i)Y_i(\mathbf{z})}{\sum_{i\in[n]}(1-z_i)}. \tag{4.2}$$

This estimator is biased under the presence of network interference. Note that $\widehat{\text{TTE}}_{\text{DM}}$ does not take into account any information about each individual's neighborhood.

A modification of the difference in means estimator incorporates knowledge of the number of treated neighbors of each individual. Let U_i denote the number of individuals in $\mathcal{N}_i \setminus \{i\}$ assigned to treatment, and let $\tilde{U}_i$ denote the number of neighbors individuals in $\mathcal{N}_i \setminus \{i\}$ assigned to control. This estimator is given by

$$\widehat{\text{TTE}}_{\text{DM}(\lambda)} = \frac{\sum_{i\in[n]} z_i \mathbb{I}(U_i \geq \lambda)Y_i(\mathbf{z})}{\sum_{i\in[n]} z_i \mathbb{I}(U_i \geq \lambda)} - \frac{\sum_{i\in[n]}(1-z_i)\mathbb{I}(\tilde{U}_i \geq \lambda)Y_i(\mathbf{z})}{\sum_{i\in[n]}(1-z_i)\mathbb{I}(\tilde{U}_i \geq \lambda)}, \tag{4.3}$$



(a) Varying size of the population

(b) Varying ratio of direct:indirect effects

(c) Varying treatment budget

(d) Varying the model degree

Figure 1: Four graphs visualizing the performance of various TTE estimators as various parameters are adjusted. The height of each graph depicts the experimental relative bias of the estimator and the shaded width depicts the experimental standard deviation.

for some user-defined tolerance $\lambda \in [0, 1]$. This estimator only counts an individual's outcome if at least λ of the individual's neighborhood is assigned to the same treatment as the individual itself. In our experiments, we set $\lambda = 0.75$.

Finally we compare against least squares regression models of degree β, which posit that the potential outcomes model can be described as

$$Y_i(\mathbf{z}) = g(z_i, \bar{z}_i) = \left(\rho + \sum_{k=1}^{\beta} \gamma_k X_i^k\right) + z_i\left(\tilde{\rho} + \sum_{k=1}^{\beta-1} \tilde{\gamma}_k X_i^k\right), \tag{4.4}$$

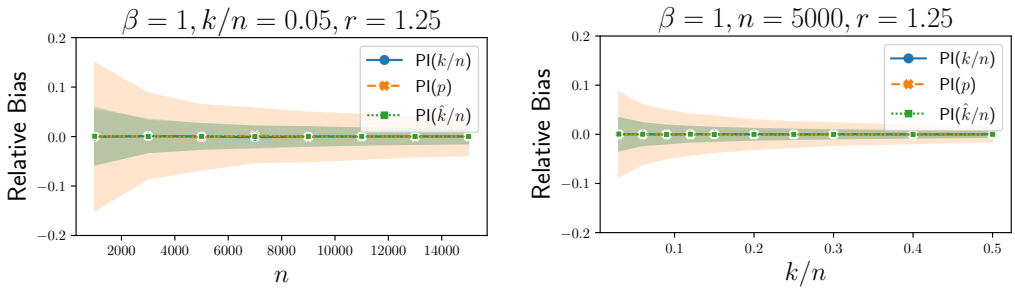
for some covariate X_i. In the two variations we consider, we set X_i equal to either the number of treated neighbors or the proportion of treated neighbors, where we do not include i itself. The two sets of coefficients $(\rho, \gamma_1, \ldots \gamma_\beta)$ and $(\rho, \tilde{\gamma}_1, \ldots \tilde{\gamma}_\beta)$ allow for the model to be different when i is treated vs not treated, and the second summation only goes until $\beta - 1$ since we want to only allow degree β interactions. The total number of coefficients in the model is $2\beta + 1$. Least squares regression finds the set of coefficients that minimizes the least squares predictive error on the dataset, which consists of $\{z_i, X_i, Y_i(\mathbf{z})\}_{i \in [n]}$. The estimated coefficients define an estimate for the function $\hat{g}$. For the variation which uses the number of treated neighbors as the covariates, setting $X_i = U_i$, the estimate is given by

$$\widehat{\text{TTE}}_{\text{LS-Num}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{g}(1, |\mathcal{N}_i| - 1) - \hat{g}(0, 0)). \tag{4.5}$$

For the variation which uses the proportion of treated neighbors as the covariates, setting $X_i = U_i/(|\mathcal{N}_i| - 1)$, the estimate is given by

$$\widehat{\text{TTE}}_{\text{LS-Prop}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{g}(1, 1) - \hat{g}(0, 0)). \tag{4.6}$$

As completely randomized design is more balanced than Bernoulli randomized design, we evaluate all the benchmark algorithms under a completely randomized design.



(a) Varying size of the population

(b) Varying treatment budget

Figure 2: Two graphs visualizing the performance of our proposed TTE estimators as the size of the population (n) or treatment budget (k/n) is varied. The height of each graph depicts the experimental relative bias of the estimator and the shaded width depicts the experimental standard deviation. The blue and the green plots essentially overlap.

Results and Discussion. For each population size n, we sample G networks from the distribution described above. For each configuration of parameters in the experiment, we sample N treatment schedules $\{\mathbf{z}^0, \ldots, \mathbf{z}^\beta\}$ from our parameterized distribution class (Bernoulli or CRD) compute the TTE using each estimator. For each estimator, we plot the relative bias of the TTE estimates averaged over the results from these GN samples and normalized by the magnitude of the TTE. The width of the shading in the figures depicts the standard deviation across the GN estimates. We ran all experiments on a Linux-based machine with 20 CPU(s) and 10 cores. The experiments for the linear setting took 8.3 minutes and the experiments with varying polynomial degree took 4.6 minutes.

In Figure 1, we visualize the effect of four network/estimator parameters on the quality of each of the five TTE estimators (the four described above, and our CRD estimator with treatment targets $k_t = \frac{tk}{\beta}$). Specifically, we consider the effects of the population size (n), the maximum proportion of treated individuals (k/n), the degree of the potential outcomes model (β), and the ratio between the network and direct effects (r). Each of the plots fixes three of these parameters and varies the fourth. Specific settings of the parameters are listed on each plot.

In Figure 1, our estimator (in blue) is unbiased as expected and the variance decreases as n and k/n increases. However, the other estimators remain significantly biased, with higher variances than ours, regardless of treatment budget or population size. As the ratio r increases the network effects become more significant relative to the direct effect, and thus the bias of other estimators also increases. As a sanity check, when the ratio is close to 0, all estimators are unbiased as there are no network effects.

In Figure 2, we compare the variants of our estimator, evaluating $\widehat{\text{TTE}}_{\text{PI}}(\mathbf{k}/n)$ under CRD and evaluating $\widehat{\text{TTE}}_{\text{PI}}(\mathbf{p})$ and $\widehat{\text{TTE}}_{\text{PI}}(\hat{\mathbf{k}}/n)$ under Bernoulli($\mathbf{p}$) randomized design, where $p_t = tp/\beta$ and $\hat{\mathbf{k}}$ is the vector of realized treatment counts. The estimators $\widehat{\text{TTE}}_{\text{PI}}(\mathbf{k}/n)$ and $\widehat{\text{TTE}}_{\text{PI}}(\hat{\mathbf{k}}/n)$ perform nearly identically. $\widehat{\text{TTE}}_{\text{PI}}(\hat{\mathbf{k}}/n)$ has lower variance than $\widehat{\text{TTE}}_{\text{PI}}(\mathbf{p})$, which is intuitive as it performs polynomial interpolation on the realized treatment fraction rather than the expected treatment fraction. We include additional experiments for higher degree models in Appendix C.

5 Conclusion

We propose a new approach for causal inference under network interference which performs significantly better than existing approaches without requiring knowledge of the graph. In particular, the additional measurements from a staggered rollout design enable us to reduce the task of estimating total treatment effect to that of polynomial interpolation. We show that under a flexible class of low degree polynomial potential outcomes our estimator is unbiased with variance scaling as $O(1/n)$. Future directions include how to optimally perform model selection when β is unknown, and allowing for the potential outcomes model to incorporate time dependence. The staggered rollout design framework has implications towards estimation under other model classes beyond polynomial, such as sublinear or monotone functions, under which one may be able to construct bounds on TTE.

References

[1] 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.

[2] Eric Auerbach and Max Tabord-Meehan. The local approach to causal inference under network interference. Technical report, 2021.

[3] Falco J Bargagli-Stoffi, Costanza Tortù, and Laura Forastiere. Heterogeneous treatment and spillover effects under clustered network interference. Technical report, 2020.

[4] Guillaume W Basse and Edoardo M Airoldi. Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858, 2018.

[5] 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. URL https://proceedings.mlr.press/v115/bhattacharya20a.html.

[6] 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.

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

[8] 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.

[9] 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.

[10] Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103:832–842, 2008. URL https://EconPapers.repec.org/RePEc:bes:jnlasa:v:103:y:2008:m:june:p:832-842.

[11] Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. *arXiv preprint arXiv:2007.13302*, 2020.

[12] Wenrui Li, Daniel L Sussman, and Eric D Kolaczyk. Causal inference under network interference with noise. Technical report, 2021.

[13] 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. doi: 10.1080/01621459.2013.844698. URL https://doi.org/10.1080/01621459.2013.844698. PMID: 24659836.

[14] Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1), 2013.

[15] 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. doi: https://doi.org/10.1111/rssc.12170. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12170.

[16] Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007. doi: 10.1198/016214506000001112. URL https://doi.org/10.1198/016214506000001112.

[17] Michael E Sobel. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476):1398–1407, 2006. doi: 10.1198/016214506000000636. URL https://doi.org/10.1198/016214506000000636.

[18] Daniel L Sussman and Edoardo M Airoldi. Elements of estimation theory for causal effects in the presence of network interference. Technical report, 2017.

[19] 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. doi: 10.1177/0962280210386779. URL https://doi.org/10.1177/0962280210386779. PMID: 21068053.

[20] Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International Conference on Machine Learning*, pages 1489–1497, 2013.

[21] Johan Ugander and Hao Yin. Randomized graph cluster randomization. *arXiv preprint arXiv:2009.02297*, 2020.

[22] 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.

[23] Tyler J. VanderWeele, Eric J. Tchetgen Tchetgen, and M. Elizabeth Halloran. Interference and sensitivity analysis. *Statist. Sci.*, 29(4):687–706, 11 2014. doi: 10.1214/14-STS479. URL https://doi.org/10.1214/14-STS479.

[24] Davide Viviano. Experimental design under network interference. Technical report, 2020.

[25] Christina Lee Yu, Edoardo Airoldi, Christian Borgs, and Jennifer Chayes. Graph agnostic randomized experimental design under heterogeneous linear network interference, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the end of section 3.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This is implied by the limitations of our work–because causal inference is applied to societal problems, problems such as model misspecification can have negative consequences. We comment on this at the end of section 3.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are included in the appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] see Section 4 for details on how we got the results. Link to code and data to reproduce results will be included in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 1. The shading indicates experimental standard deviation over all samples.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the results and discussion paragraph in Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] All the code we used will be accessible via URL in the supplementary material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]