A Causal Inference Framework for Network Interference with Panel Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a framework for causal inference with panel data in the presence of network interference and unobserved confounding. Key to our approach is a novel latent factor model that takes into account network interference and generalizes the factor models typically used in panel data settings. We propose an estimator–the Network Synthetic Interventions estimator—and show that it consistently estimates the counterfactual outcomes for a unit under an arbitrary set of treatments, if certain observation patterns hold in the data. We corroborate our theoretical findings with simulations. In doing so, our framework extends the Synthetic Control and Synthetic Interventions methods to incorporate network interference.

1 Introduction

There is growing interest in the identification and estimation of causal effects in the context of networks, in which the outcomes of a unit (e.g., an individual, customer cohort, or region) are affected by the treatments (e.g., recommendations, discounts, or legislation) assigned to other units, known as the unit's "neighbors". For example, whether an individual (i.e., the unit) gets COVID-19 (i.e., the outcome) is a function of not only the individual's vaccination status (i.e., the treatment), but also the vaccination status of that individual's social network. That is, there is *network interference*.

The majority of works on causal inference under network interference consider the setting of a single measurement or dataset, whether collected from a randomized experiment or observational study. It is known that estimating any desired causal estimand under arbitrary interference is impossible, as the model is not identifiable [26, 4, 8, 21]. As a result, prior works impose additional structure through assumptions on exposure functions [26, 4, 39, 6, 23], interference neighborhoods [36, 7, 32, 10], parametric structure [35, 9, 12, 19, 17], or a combination of these, each leading to a different solution concept. In this work, we focus on network interference that is additive across the neighbors, referred to in the literature as the joint assumptions of neighborhood interference, additivity of main effects, and additivity of interference effects [32, 40, 14, 15].

Distinct to our work is that we consider a *panel data* setting in which there are multiple measurements for each unit, as arises when units are observed across time. Additionally, we allow for estimation of counterfactuals under *multiple treatments*, whereas the existing literature has largely focused on binary treatments. Key to our approach is a novel latent factor model that takes into account network interference and is a generalization of the factor models typically used in panel data settings. Although adding time to our analysis might appear to introduce complexity, we show that being able to measure potential outcomes across time actually enables the inference of *unit-specific causal effects* as long as the dataset is "sufficiently rich" (specifically, as long as there is sufficient diversity in the observed treatments). Estimating unit-specific causal effects is typically not feasible in the single measurement setup unless one imposes strong parametric model assumptions on the potential outcomes function. As a result, previous work has focused on causal estimands that capture population-wide effects,

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

such as the average direct treatment effect (the average difference in outcomes if only one unit and none of its neighbours get treated [9, 20, 31, 32, 22, 25]) and the average total treatment effect (the average difference in outcomes if all units get treated versus if they do not [36, 17, 13, 40, 14, 15]). Further, building on recent works in panel data [2], we allow for unobserved confounding in treatment assignment as long as there is selection on latent factors.

2 Problem Statement

Setup. Consider a setting with $N \geq 1$ units, $D \geq 1$ treatments, and $T \geq 1$ measurements of interest. Unless otherwise stated, we index units with $n \in [N]$, measurements with $t \in [T]$, and treatments with $a \in [D]_0$.[1] Let $G = ([N], \mathcal{E}) \in \mathcal{G}$ denote a graph over the N units, where $\mathcal{E} \subset [N] \times [N]$ denotes the edges of the graph. Throughout, we shall assume $\mathcal{G}$ to be *fixed* and *observed*. Let $\mathcal{N}(n)$ denote the neighbors of unit $n \in [N]$ with respect to $\mathcal{G}$ such that $j \in \mathcal{N}(n) \iff (j, n) \in \mathcal{E}$.[2] Under network interference, the potential outcome for a given unit n and measurement t is a real-valued random variable denoted by $Y_{tn}^{(\mathbf{a})}$, where $\mathbf{a} \in [D]_0^N$ denotes the treatments over all N units. We impose the following additional structure on the potential outcomes.

Assumption 1 (Network SUTVA)**.** *The potential outcome of measurement $t \in [T]$ for unit $n \in [N]$ under treatments $\mathbf{a} \in [D]_0^N$ is given by*

$$Y_{tn}^{(\mathbf{a})} = Y_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})},$$

where $\mathbf{a}_{\mathcal{N}(n)} \in [D]_0^{|\mathcal{N}(n)|}$ denotes the treatments assigned to the units in n's neighborhood $\mathcal{N}(n)$ for measurement t. That is, the potential outcome of unit n depends on its neighbors' treatments but does not depend the treatment of any other unit $j \in [N] \setminus \mathcal{N}(n)$.

Several prior works on network interference also assume Network SUTVA, e.g., as the *Neighborhood Interference Assumption (NIA)* [33].

Observation pattern. In this work, let the measurement index t denote time. Let the T measurements be partitioned into two sets. Let $\mathcal{T}_{\text{tr}} \subset [T]$ denote the training period and $\mathcal{T}_{\text{pr}} \subset [T]$ denote the prediction period, where $\mathcal{T}_{\text{tr}} \cap \mathcal{T}_{\text{pr}} = \emptyset$, $T_{\text{tr}} = |\mathcal{T}_{\text{tr}}|$, and $T_{\text{pr}} = |\mathcal{T}_{\text{pr}}|$. Without loss of generality, let $\mathcal{T}_{\text{tr}} := \{1, 2, \ldots, T_{\text{tr}}\}$ and $\mathcal{T}_{\text{pr}} := \{T - T_{\text{pr}} + 1, \ldots, T\}$. Let $\mathbf{a}^t \in [D]_0^N$ denote the treatment vector assigned at time $t \in [T]$. Let

$$A^{\text{tr}} = \left[\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^{T_{\text{tr}}}\right] \in [D]_0^{N \times T_{\text{tr}}},$$
$$A^{\text{pr}} = \left[\mathbf{a}^{T - T_{\text{pr}} + 1}, \mathbf{a}^{T - T_{\text{pr}} + 2}, \ldots, \mathbf{a}^T\right] \in [D]_0^{N \times T_{\text{pr}}},$$

denote the training and prediction treatment sequences, respectively. We assume that we observe every unit at all $t \in \mathcal{T}_{\text{tr}} \cup \mathcal{T}_{\text{pr}}$ under treatments sequences A^{tr} and A^{pr}.

We denote the observation for unit n at time t as $Y_{tn} = Y_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})}$ for all $t \in \mathcal{T}_{\text{tr}} \cup \mathcal{T}_{\text{pr}}$.

Target causal parameter. Our goal is to estimate counterfactuals for a given unit during $\mathcal{T}_{\text{pr}}$. Specifically, for unit $n \in [N]$, let

$$\tilde{A}_n^{\text{pr}} = \left[\tilde{\mathbf{a}}_{\mathcal{N}(n)}^{T - T_{\text{pr}} + 1}, \tilde{\mathbf{a}}_{\mathcal{N}(n)}^{T - T_{\text{pr}} + 2}, \ldots, \tilde{\mathbf{a}}_{\mathcal{N}(n)}^{T}\right] \in [D]_0^{|\mathcal{N}(n)| \times T_{\text{pr}}},$$

denote the sequence of counterfactual treatments of interest for unit n. We are interested in estimating the following causal parameter:

$$\theta_n^{(\tilde{A}_n^{\text{pr}})} = \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \mathbb{E}\left[Y_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})}\right]. \tag{1}$$

That is, we seek to estimate the expected potential outcome of unit n, averaged over $\mathcal{T}_{\text{pr}}$, if unit n's neighborhood $\mathcal{N}(n)$ undergo the treatment sequence $\tilde{A}_{\mathcal{N}(n)}^{\text{pr}}$.

[1] Let $[X]_0 = \{0, 1, \ldots, X - 1\}$ and $[X] = \{1, \ldots, X\}$ for any positive integer X.

[2] For simplicity of notation, we include self-edges: $(n, n) \in \mathcal{N}(n)$ for all $n \in [N]$.

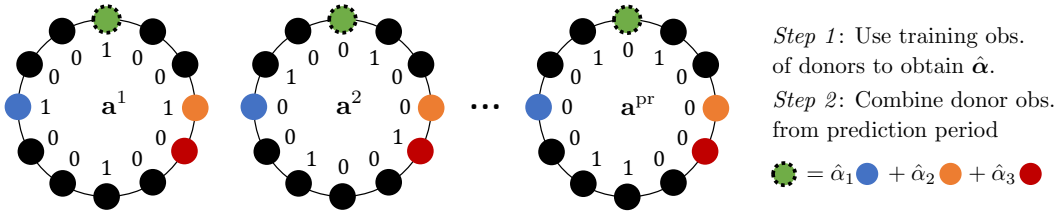


Figure 1: Visualization of the NSI estimator (Section 3). Consider a ring graph with 12 units and binary treatments. Suppose that $\mathbf{a}^t = \mathbf{a}^{\text{pr}}$ for all $t \in \mathcal{T}_{\text{pr}}$ and, similarly, the counterfactual treatments of interest $\tilde{\mathbf{a}}^t_{\mathcal{N}(n)} = (0, 0, 0)$ for all $t \in \mathcal{T}_{\text{pr}}$. Then, if n is given by the top unit (in green with the dotted black outline), the donor set is given by the blue, orange, and red units (see Definition 1). As illustrated on the right-hand side, the NSI estimator first runs principal component regression over the donors to obtain $\hat{\boldsymbol{\alpha}}$. As the second step, the NSI estimator linearly combines the observations of each donor using the coefficients $\hat{\boldsymbol{\alpha}}$ to produce estimates of n's potential outcomes.

3 The Network Synthetic Interventions (NSI) Estimator

We describe a simple estimator, which we term the *Network Synthetic Intervention* (NSI) estimator. It is a natural extension of the Synthetic Interventions estimator [2] in the presence of network interference. Below, we describe the estimator formally, with a caricature example in Figure 1.

NSI Estimator. Consider the causal parameter $\theta_n^{(\tilde{A}^{\text{pr}}_n)}$ of interest, as given in (1). Let

$$A^{\text{tr}}_n = \left[\mathbf{a}^1_{\mathcal{N}(n)}, \mathbf{a}^2_{\mathcal{N}(n)}, \ldots, \mathbf{a}^{T_{\text{tr}}}_{\mathcal{N}(n)}\right] \in [D]_0^{|\mathcal{N}(n)| \times T_{\text{tr}}},$$
$$\mathbf{z}_{\text{tr},n} = [Y_{nt} : t \in \mathcal{T}_{\text{tr}}] \in \mathbb{R}^{T_{\text{tr}}}.$$

Before presenting the estimation procedure, we define the useful notion of a "donor set".

Definition 1 (Donor set)**.** *For any $N' \leq N$, consider sequence of training treatments $C^{tr} = [\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^{T_{tr}}] \in [D]_0^{N' \times T_{tr}}$ and prediction treatments $C^{pr} = [\mathbf{c}^{T-T_{pr}+1}, \mathbf{c}^{T-T_{pr}+2}, \ldots, \mathbf{c}^T] \in [D]_0^{N' \times T_{pr}}$. Let $\mathcal{I}^{(C^{tr}, C^{pr}, N')} \subset [N]$ denote a set of "donor units" such that for all $j \in \mathcal{I}^{(C^{tr}, C^{pr}, N')}$,*

1. *$|\mathcal{N}(j)| = N'$, and*
2. *there exists a way π_j to permute $\mathcal{N}(j)$ such that: $\mathbf{a}^t_{\pi_j(\mathcal{N}(j))} = \mathbf{c}^t$ for all $t \in \mathcal{T}_{tr} \cup \mathcal{T}_{pr}$.*

To estimate $\theta_n^{(\tilde{A}^{\text{pr}}_n)}$, the relevant donor set is turns out to be $\mathcal{I}^{(A^{\text{tr}}_n, \tilde{A}^{\text{pr}}_n, |\mathcal{N}(n)|)} \subset [N]$. For simplicity, we use the shorthand $\mathcal{I}^n := \mathcal{I}^{(A^{\text{tr}}_n, \tilde{A}^{\text{pr}}_n, |\mathcal{N}(n)|)}$. Let the donors' training observations be given by

$$Z_{\text{tr},\mathcal{I}^n} = [Y_{jt} : t \in \mathcal{T}_{\text{tr}}\,,\ j \in \mathcal{I}^n] \in \mathbb{R}^{T_{\text{tr}} \times |\mathcal{I}^n|}. \tag{2}$$

Then, estimation proceeds in a two-step procedure with a parameter κ.[3]

Step 1: Principal component regression. Perform a singular value decomposition (SVD) of $Z_{\text{tr},\mathcal{I}^n}$ to obtain $Z_{\text{tr},\mathcal{I}^n} = \sum_{\ell=1}^{q_{\text{tr}}} \hat{s}_\ell \hat{\boldsymbol{\mu}}_\ell \hat{\boldsymbol{\nu}}_\ell^\top$. Using parameter $\kappa \leq q_{\text{tr}}$, compute

$$\hat{\boldsymbol{\alpha}} = \sum_{\ell=1}^{\kappa} \hat{s}_\ell^{-1} \hat{\boldsymbol{\nu}}_\ell \hat{\boldsymbol{\mu}}_\ell^\top \mathbf{z}_{\text{tr},n} \in \mathbb{R}^{|\mathcal{I}^n|}.$$

Step 2: Estimator. Using $\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_j : j \in \mathcal{I}^n]$[4], construct the estimate

$$\widehat{\mathbb{E}}\left[Y_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})}\right] = \sum_{j \in \mathcal{I}^n} \hat{\alpha}_j Y_{tj}, \quad \text{for all } t \in \mathcal{T}_{\text{pr}}, \tag{3}$$

and, accordingly,

$$\hat{\theta}_n^{(\tilde{A}^{\text{pr}}_n)} = \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \widehat{\mathbb{E}}\left[Y_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})}\right]. \tag{4}$$

[3] κ can be selected in a data-driven manner. Due to limitation of space, it is not discussed here.

[4] For simplicity, we abuse notation and let $\hat{\alpha}_j$ denote the element associated with donor $j \in \mathcal{I}^n$.

4 Formal Analysis: Model and Results

In this section, we provide a formal analysis of the NSI estimator. We start by presenting a model for network interference. This is followed by formal results for identification and finite sample analysis.

4.1 Model

We now introduce the model that we use to develop our formal results. We note that this model, given in Assumption 2 below, satisfies Assumption 1.

Assumption 2. *Let the potential outcome of measurement $t \in [T]$ for unit $n \in [N]$ under graph $G \in \mathcal{G}$ if assigned treatments $\mathbf{a} \in [D]_0^N$ be given by:*

$$Y_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} = \sum_{k \in \mathcal{N}(n)} \langle \mathbf{u}_{k,n}, \mathbf{w}_{t,a_k} \rangle + \epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{5}$$

where $\mathbf{u}_{\cdot,\cdot} \in \mathbb{R}^r$ and $\mathbf{w}_{\cdot,\cdot} \in \mathbb{R}^r$ represent latent (unobserved) factors; $\epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}$ is an additive, zero-mean, independent (or idiosyncratic) noise term; and r is the rank or model complexity.

Intuitively, the potential outcome of unit n with neighbors $\mathcal{N}(n)$ at time t is determined by two factors: (a) the effect of the treatment assigned to unit n and (b) the spillover effects from the treatments assigned to n's neighbors. Since $n \in \mathcal{N}(n)$, both effects are captured in the summation in (5). Note that (5) can be written as

$$Y_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} = \left\langle \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}} \right\rangle + \epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{6}$$

where

$$\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} = [\mathbf{u}_{\mathcal{N}_1(n),n}^\top, \mathbf{u}_{\mathcal{N}_2(n),n}^\top, \ldots, \mathbf{u}_{\mathcal{N}_{|\mathcal{N}(n)|}(n),n}^\top]^\top,$$
$$\tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}} = [\mathbf{w}_{t,a_{\mathcal{N}_1(n)}}^\top, \mathbf{w}_{t,a_{\mathcal{N}_2(n)}}^\top, \ldots, \mathbf{w}_{t,a_{\mathcal{N}_{|\mathcal{N}(n)|}(n)}}^\top]^\top.$$

Here, $\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} \in \mathbb{R}^{r|\mathcal{N}(n)|}$ and $\tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}} \in \mathbb{R}^{r|\mathcal{N}(n)|}$ are the *network-adjusted* latent factors, and $r|\mathcal{N}(n)| \in \mathbb{N}_{>0}$ denotes the *network-adjusted* "rank".

4.2 Formal results

In this section, we present an identification result for (1) under (6), then establish finite-sample consistency of the NSI estimator. We restrict our attention to a specific unit $n \in [N]$ and counterfactual treatments $\tilde{A}_n^{\text{pr}} \in [D]_0^{|\mathcal{N}(n)|}$ of interest. The proofs are relegated to Appendix B.

We begin with some notation and assumptions. Let $\mathcal{O}$ and LF be given by

$$\mathcal{O} = \left\{ (j, t, \mathbf{a}) : Y_{tj}^{(\mathbf{a}_{\mathcal{N}(j)})} \text{ is observed} \right\} \subset [N] \times [T] \times [D]_0^N,$$
$$LF = \left\{ \mathbf{u}_{k,j}, \mathbf{w}_{t,a} : k, j \in [N], t \in [T], \text{ and } a \in [D]_0 \right\}.$$

Assumption 3 (Conditional exogeneity)**.** *We assume that $\mathbb{E}\left[\epsilon_{tj}^{(\mathbf{a}_{\mathcal{N}(j)})} \,|\, LF\right] = 0$ and $\epsilon_{tj}^{(\mathbf{a}_{\mathcal{N}(j)})} \perp \mathcal{O} \,|\, LF$ for all $j \in [N]$, $t \in [T]$, and $\mathbf{a} \in [D]_0^N$.*

Assumption 4 (Linear span inclusion)**.** *Given a unit $n \in [N]$ and sequence of counterfactual, treatments $\tilde{A}_n^{pr} \in [D]_0^{|\mathcal{N}(n)|}$ of interest, consider the donor set $\mathcal{I}^n$. We assume that $\mathcal{I}^n$ is non-empty and that there exists $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{I}^n|}$ such that*

$$\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} = \sum_{j \in \mathcal{I}^n} \alpha_j \tilde{\mathbf{u}}_{j,\pi_j(\mathcal{N}(j))},$$

where π_j is defined in Definition 1.

Together, Assumptions 2-3 imply that $Y_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} \perp \mathcal{O} \,|\, LF$, which is analogous to requiring "selection on network-adjusted latent factors": that, conditioning on all latent factors, the treatment assignments are independent of the potential outcome. This requirement is analogous to "selection on latent factors" in [2]. While the treatment assignment is allowed to depend on the latent factors, Assumption 4 requires that the treatment assignment is "diverse" enough that the target unit's latent factor lies in the linear span of the donor units. We now state the identification result.

Theorem 1 (Identification)**.** *Consider a unit $n \in [N]$ and sequence of counterfactual treatments $\tilde{A}_n^{pr} \in [D]_0^{|\mathcal{N}(n)|}$ of interest. Suppose that Assumptions 1-4 hold. Let $\boldsymbol{\alpha}$ denote the coefficients from Assumption 4 for the donor set $\mathcal{I}^n$, where $\mathcal{I}^n$ is defined in Section 3. Then,*

$$\mathbb{E}\left[Y_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})} \,\middle|\, LF\right] = \sum_{j \in \mathcal{I}^n} \alpha_j \mathbb{E}[Y_{tj} \mid LF, \mathcal{O}] \quad \textit{and} \quad \theta_n^{(\tilde{A}_n^{pr})} = \frac{1}{T_{pr}} \sum_{t \in \mathcal{T}_{pr}} \sum_{j \in \mathcal{I}^n} \alpha_j \mathbb{E}[Y_{tj} \mid LF, \mathcal{O}].$$

Theorem 1 implies that estimating (1) comes down to acquiring good estimates of $\boldsymbol{\alpha}$. Estimating $\boldsymbol{\alpha}$ using observational data is precisely what the NSI estimator does. Next, we give conditions under which the NSI estimator achieves finite-sample consistency.

To that end, let $M = |\mathcal{I}^n|$ and $Z_{\text{post},\mathcal{I}^n} = [Y_{tj} : t \in \mathcal{T}_{\text{pr}}, j \in \mathcal{I}^n] \in \mathbb{R}^{T_{\text{pr}} \times M}$. Recall $Z_{\text{tr},\mathcal{I}^n}$ from (2) and let $r_{\text{tr}} \in [r|\mathcal{N}(n)|]$ be the rank of $\mathbb{E}\big[Z_{\text{tr},\mathcal{I}^n} \,|\, LF, \mathcal{O}\big]$, $s_1 \geq \ldots \geq s_{r_{\text{tr}}} > 0$ denote its singular values, and $R_{\text{tr}} \in \mathbb{R}^{M \times r_{\text{tr}}}$ denote its right singular vectors. Let $\boldsymbol{\alpha}_\perp = R_{\text{tr}} R_{\text{tr}}^\top \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is defined in Assumption 4. Finally, let $\|\cdot\|_{\psi_2}$ denote the Orlicz norm and O_p denote a probabilistic version of big-O notation.

Assumption 5 (Sub-Gaussian noise)**.** *Assume that $\left\|\epsilon_{tj}^{(\mathbf{a}_{\mathcal{N}(j)})} \,|\, LF, \mathcal{O}\right\|_{\psi_2} \leq c\bar{\sigma}$ for some constant $c > 0$ and for all $j \in [N]$, $t \in [T]$, and $\mathbf{a} \in [D]_0^N$.*

Assumption 6 (Boundedness)**.** $\mathbb{E}\big[Y_{tj}^{(\mathbf{a}_{\mathcal{N}(j)})} \,\big|\, LF, \mathcal{O}\big] \in [-1, 1]$ *for all $j \in [N]$, $t \in [T]$, and $\mathbf{a} \in [D]_0^N$.*

Assumption 7 (Well-balanced spectrum)**.** *For universal constants $c', c'' > 0$, assume $s_{r_{tr}}/s_1 \geq c'$ and $\left\|\mathbb{E}\big[Z_{tr,\mathcal{I}^n} \,\big|\, LF, \mathcal{O}\big]\right\|_F^2 \geq c'' T_{tr} |\mathcal{I}^n|$, where $\mathcal{I}^n$ is defined in Definition 1.*

Assumption 8 (Subspace inclusion)**.** *Assume that the row-space of $\mathbb{E}\big[Z_{post,\mathcal{I}^n} \,\big|\, LF, \mathcal{O}\big]$ lies within the row-space of $\mathbb{E}\big[Z_{tr,\mathcal{I}^n} \,\big|\, LF, \mathcal{O}\big]$.*

Theorem 2 (Finite-sample consistency)**.** *Let Assumptions 1-8 hold and $\kappa = r_{tr}$. Then,*

$$\left|\hat{\theta}_n^{(\tilde{A}_n^{pr})} - \theta_n^{(\tilde{A}_n^{pr})}\right| = O_P\left(\frac{\sqrt{r_{tr}}}{T_{tr}^{1/4}} + \frac{\|\boldsymbol{\alpha}_\perp\|_2}{\sqrt{T_{pr}}} + \frac{\|\boldsymbol{\alpha}_\perp\|_1 \, r_{tr}^{3/2} \sqrt{\log(T_{tr} M)}}{\min\left(\sqrt{T_{tr}}, \sqrt{M}\right)} \,\middle|\, LF, \mathcal{O}\right),$$

where we assume $\|\boldsymbol{\alpha}_\perp\|_2 \geq c'''$ for a universal constant $c''' > 0$.

4.3 Subspace Inclusion and Implications for Network-Aware Experiment Design

The key enabling condition for finite-sample consistency of the NSI estimator (Theorem 2) is Assumption 8, i.e., the subspace inclusion assumption (SIA). Below, we show that SIA implies that the training treatments A_n^{tr} must be diverse enough with respect to the prediction treatments of interest $\tilde{A}_n^{\text{pr}}$. In terms of experiment design, Propositions 3-4 suggest that the treatments assigned during the training period must be carefully designed.

To this end, consider a scenario where the treatments are binary such that $D = 2$ and the training period is split into L sub-periods, denoted by $\mathcal{T}_{\text{tr},1}$ through $\mathcal{T}_{\text{tr},L}$. During each sub-period, let the treatments assigned to each unit be constant, i.e., for all $\ell \in [L]$, $\mathbf{a}^t = \bar{\mathbf{a}}^\ell$ for all $t \in \mathcal{T}_{\text{tr},\ell}$. Let

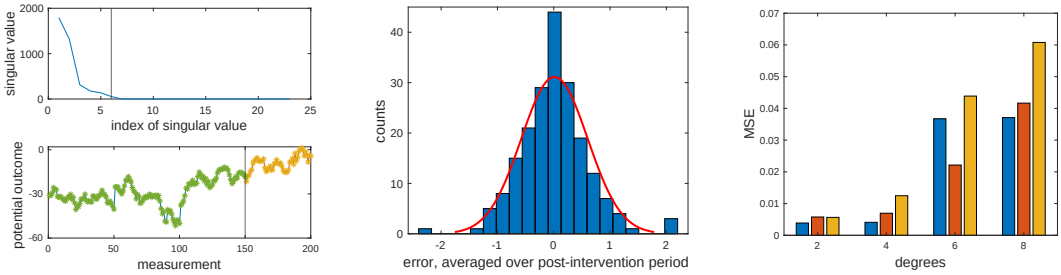
$$W_{\text{tr},\ell} = [\mathbf{w}_{t,a}^\top : t \in \mathcal{T}_{\text{tr},\ell}, a \in \{0,1\}] \in \mathbb{R}^{|\mathcal{T}_{\text{tr},\ell}| \times 2r},$$
$$B_{\text{tr}} = [1 - \bar{\mathbf{a}}^1_{\mathcal{N}(n)}, \bar{\mathbf{a}}^1_{\mathcal{N}(n)}, \ldots, 1 - \bar{\mathbf{a}}^L_{\mathcal{N}(n)}, \bar{\mathbf{a}}^L_{\mathcal{N}(n)}]^\top \in \{0,1\}^{2L \times |\mathcal{N}(n)|}.$$

Let $W_{\text{tr}} \in \mathbb{R}^{T_{\text{tr}} \times 2rL}$ be a block diagonal matrix, with $W_{\text{tr},1}$ through $W_{\text{tr},L}$ along the diagonal.

Proposition 3. *SIA holds for any $\tilde{A}_n^{pr}$ if W_{tr} and B_{tr} have linearly independent columns.*

Proposition 4. *Suppose W_{tr} has linearly independent columns. Then, SIA holds for any $\{\mathbf{u}_{k,j} : k, j \in [N]\}$ if $\tilde{\mathbf{a}}^t_{\mathcal{N}(n)}$ and $1 - \tilde{\mathbf{a}}^t_{\mathcal{N}(n)}$ are in the rowspace of B_{tr} for all $t \in \mathcal{T}_{pr}$.*

Recall that the latent factors are, by definition, unobserved. As such, W_{tr} is also unobserved, and it is not possible to verify that W_{tr} has linearly independent columns, as required in Propositions 3-4. However, as an example that, suppose $\mathbf{w}_{t,a}$ are sampled i.i.d. from a multivariate Gaussian and $|T_{\text{tr},\ell}| \geq 2r$ for all sub-periods $\ell \in [L]$. Then, with high probability, W_{tr} has linearly independent columns. As for B_{tr}, consider the following illustrative examples.



(a) NSI estimates (b) Residuals of NSI estimator (c) MSE for regular graphs

Figure 2: Simulation results for NSI estimator under $\epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} \sim \mathcal{N}(0, 0.1)$, $D = 2$, and $r = 2$.

Example 1. *Suppose $L = 1$, i.e., $\mathbf{a}^t = \bar{\mathbf{a}}^1$ for all $t \in \mathcal{T}_{tr}$. Then, $B_{tr} = [1 - \bar{\mathbf{a}}^1, \bar{\mathbf{a}}^1]^\top$. Suppose that the treatment assignment during the prediction period is not equal to $\bar{\mathbf{a}}^1$, but we are interested in estimating what* would have happened *if all units had remained under $\bar{\mathbf{a}}^1$, i.e., if $\tilde{\mathbf{a}}^t = \bar{\mathbf{a}}^1$ for all $t \in \mathcal{T}_{pr}$. Under this setup, $\tilde{\mathbf{a}}^t, 1 - \tilde{\mathbf{a}}^t \in B_{tr}$, as required by Proposition 4. This setup could be viewed as Synthetic Control for Panel Data [1] under network interference.*

Example 2. *Suppose $L = 1$ and $\bar{\mathbf{a}}^1 = \mathbf{0}_N$. As such, $B_{tr} = [\mathbf{0}, \mathbf{1}]^\top$. Then, unless all units in $\mathcal{N}(n)$ receive the same treatment as one another under $\tilde{\mathbf{a}}^t$, neither $\tilde{\mathbf{a}}^t$ nor $1 - \tilde{\mathbf{a}}^t$ are in the rowspace of B_{tr}.*

Example 3. *Suppose that $L = |\mathcal{N}(n)|$. Suppose that during each sub-period $\mathcal{T}_{tr,\ell}$, a single distinct unit in $\mathcal{N}(n)$ is assigned treatment 1 and all other are assigned 0. Suppose that which unit in $\mathcal{N}(n)$ is assigned treatment 1 rotates at each subsequent sub-period. Then, B_{tr} has linearly independent columns, as required in Proposition 3.*

5 Simulations

In this section, we present simulation results illustrating the behavior of the NSI estimator and compare it to two related estimators. Experimental details can be found in Appendix C.

In particular, we consider the following setting. Suppose G is a regular graph with degree d, and the treatments are binary. For simplicity, suppose that $\mathbf{a}^t = \mathbf{a}^{\text{pr}}$ for all $t \in \mathcal{T}_{\text{pr}}$ and, similarly, $\tilde{\mathbf{a}}^t = \tilde{\mathbf{a}}^{\text{pr}}$ for all $t \in \mathcal{T}_{\text{pr}}$, i.e., the prediction and counterfactual treatments are constant across $\mathcal{T}_{\text{pr}}$. Lastly, suppose that the training treatments are assigned as described in Section 4.3 with $L = d + 1$. More precisely, let the training period be divided into $d + 1$ sub-periods. During each of the sub-periods, 1 out of every $d + 1$ units receives treatment 1, and all others receive treatment 0. Each subsequent sub-period rotates which units are treated such that each unit is only treated during one of the sub-periods.

Under this setup, Fig. 2(a) shows an example of NSI estimates for the ring graph ($d = 2$) with $N = 400$. On top, it plots the spectrum $\{\hat{s}_\ell\}_{\ell=1}^{q_{\text{tr}}}$ produced in Step 1 of Section 3, where the vertical line marks κ. On bottom, it gives the NSI estimates, where the vertical line separates the training and prediction periods. The ground-truth values are given as lines, and the predictions are marked with $*$'s. As shown, the predictions closely match the ground-truth values. Under the same setup, Fig. 2(b) plots the histogram of NSI residuals (the difference between the estimated and ground-truth potential outcomes) of 200 simulations, verifying that the residuals are consistent. Fig. 2(c) gives the MSE across $d = 2, 4, 6$, and 8. The left (blue) bars are for $N = 100$ and $T_{\text{tr}} = T_{\text{pr}} = 100$; the middle (red) bars for $N = 100$ and $T_{\text{tr}} = T_{\text{pr}} = 50$; and the right (yellow) bars for $N = 500$ and $T_{\text{tr}} = T_{\text{pr}} = 50$. As expected, the MSE typically increases with degree, fewer nodes, and less training time.

We also compare the NSI estimator to two others: the SI estimator (which does not account for network interference) [2] and a baseline estimator. The baseline estimator finds donor units that satisfy Definition 1, then averages the donor units' observed outcomes. We compare the estimators for a ring graph (details given in Appendix C). The MSEs and R-squared values for the NSI estimator, SI estimator, and baseline estimators are, respectively, **(0.08013, 0.9994)**, **(53.10, 0.9101)**, and **(576.1, -1.389)**. Both the NSI and baseline estimators use donor sets that contained, on average, 16 units. The SI estimator used donor sets with, on average, 66 units.

References

[1] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

[2] Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2020.

[3] Peter M. Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.

[4] 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.

[5] 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.

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

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

[8] Guillaume W. Basse and Edoardo M. Airoldi. Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology*, 48(1):136–151, 2018.

[9] 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.

[10] 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.

[11] Jake Bowers, Mark M. Fredrickson, and Costas Panagopoulos. Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97–124, 2013.

[12] 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.

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

[14] Mayleen Cortez, Matthew Eichhorn, and Christina Lee Yu. Exploiting neighborhood interference with low order interactions under unit randomized design. *arXiv preprint arXiv:2208.05553*, 2022.

[15] Mayleen Cortez, Matthew Eichhorn, and Christina Lee Yu. Staggered rollout designs enable causal inference under interference without network knowledge. *arXiv preprint arXiv:2205.14552*, 2022.

[16] Francis J. DiTraglia, Camilo Garcia-Jimeno, Rossa O'Keeffe-O'Donovan, and Alejandro Sanchez-Becerra. Identifying causal effects in experiments with spillovers and non-compliance. *arXiv preprint arXiv:2011.07051*, 2020.

[17] 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.

[18] Laura Forastiere, Edoardo M Airoldi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918, 2021.

[19] 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.

[20] 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.

[21] Vishesh Karwa and Edoardo M Airoldi. A systematic investigation of classical causal inference strategies under mis-specification due to network interference. Technical report, 2018.

[22] Michael P Leung. Causal inference under approximate neighborhood interference. Technical report, 2019.

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

[24] Lan Liu, Michael G Hudgens, and Sylvia Becker-Dreps. On inverse probability-weighted estimators in the presence of interference. *Biometrika*, 103(4):829–842, 2016.

[25] 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.

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

[27] Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. Technical report, 2017.

[28] Carolina Perez-Heydrich, Michael G Hudgens, M Elizabeth Halloran, John D Clemens, Mohammad Ali, and Michael E Emch. Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741, 2014.

[29] Jean Pouget-Abadie, Martin Saveski, Guillaume Saint-Jacques, Weitao Duan, Ya Xu, Souvik Ghosh, and Edoardo Maria Airoldi. Testing for arbitrary interference on experimentation platforms. Technical report, 2017.

[30] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. 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.

[31] 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.

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

[33] Daniel L Sussman and Edoardo M Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*, 2017.

[34] 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.

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

[36] 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.

[37] Gonzalo Vazquez-Bare. Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, 2022.

[38] Natalya Verbitsky-Savitz and Stephen W Raudenbush. Causal inference under interference in spatial settings: a case study evaluating community policing program in chicago. *Epidemiologic Methods*, 1(1):107–130, 2012.

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

[40] Christina Lee Yu, Edoardo M Airoldi, Christian Borgs, and Jennifer T Chayes. Estimating total treatment effect in randomized experiments with unknown network structure. *arXiv preprint arXiv:2205.12803*, 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] We discuss the required conditions in Assumptions 1-8 as well as the real-world implications of the major assumptions in Sections 4.2-4.3
   (c) Did you discuss any potential negative societal impacts of your work? [No] We focus on providing a method to correct for biases introduced when network interference exists. We present the method as well as the conditions required for the method. Beyond the limitations, we do not discuss the societal impacts of the work.
   (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] See Appendix B.
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)? [N/A] We run simple simulations; they are not the main focus of this work. We provide all simulation details in Section 5 and Appendix C.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In place of error bars, we plot the distribution of residuals in Figure 2(b).
   (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 Appendix C.
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] We do not use existing assets.
   (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? [N/A]
   (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] We did not crowdsource or conduct research with human subjects.
   (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]

A Related Work

There has been great interest in studying causal inference in the presence of network interference. The majority of works consider the setting of a single measurement or dataset, whether collected from a randomized experiment or observational study. Under fully arbitrary interference, it has been shown that it is impossible to estimate any desired causal estimands as the model is not fully identifiable [26, 4, 8, 21]. As a result, there have been many proposed models that impose assumptions on exposure functions [26, 4, 39, 6, 23], interference neighborhoods [36, 7, 32, 10], parametric structure [35, 9, 12, 19, 17], or a combination of these, each leading to a different solution concept.

In this work we focus on network interference which is additive across the neighbors, referred to in the literature as the joint assumptions of neighborhood interference, additivity of main effects, and additivity of interference effects [32, 40, 14, 15]. However, distinct to our work is that we allow for multiple treatments, whereas the existing literature has largely focused on binary treatments. More importantly we consider a panel data setting in which we are given multiple measurements from each unit, as arises if we observe time series data from each unit. The potential outcomes function are thus also time dependent.

Additionally, previous work has focused on specific causal estimands that correspond to population wide averages, most notably the average direct treatment effect, which is the average difference in outcomes if only a unit and none of its neighbours get treated [9, 20, 31, 32, 22, 25], and the average total treatment effect, which is the average difference in outcomes if all units get treated versus if they do not [36, 17, 13, 40, 14, 15]. Alternately there has been some literature that focus on hypothesis testing for the presence of network interference [3, 11, 5, 29, 30]; these results do not immediately extend to estimation as they are based on randomization inference with a fixed network size and study testing sharp null hypotheses.

In contrast, in this work we obtain estimates for unit-specific causal effects. This is typically impossible in the single measurement setup unless one imposes strong parametric model assumptions on the potential outcomes function.

While a majority of the literature focuses on randomized experiment, there is growing interest as well to develop theory for accounting for network interference when analyzing observational studies. A majority of the literature assumes partial interference, where the network consists of many disconnected subcommunities [34, 28, 24, 16, 37]. Without this strong clustering condition, other works impose strong parametric assumptions on the potential outcomes function, assuming that the potential outcomes only depends on a known statistic of the neighborhood treatment, e.g. the number or fraction of treated [38, 13, 27]. This reduces estimation to a regression task under requirements of sufficient diversity in the treatments. [18] considers a general exposure mapping model alongside an inverse propensity weighted estimator, but the estimator has high variance when the exposure mapping is complex.

B Proofs

The notation O_p is a probabilistic version of big-O notation. Formally, for any sequence of random vectors X_n, $X_n = O_p(\chi_n)$ if, for any $\varepsilon > 0$, there exists constants c_ε and n_ε such that $P(\|X_n\|_2 > c_\varepsilon \chi_n) < \varepsilon$ for every $n \geq n_\varepsilon$. Equivalently, we say that X_n/χ_n is "uniformly tight" or "bounded in probability".

B.1 Proof of Theorem 1

Proof. Below, the symbol $\overset{AX}{=}$ and $\overset{DX}{=}$ imply that the equality follows from Assumption X and Definition X, respectively. Recall that $\mathcal{I}^n$ is shorthand for $\mathcal{I}^{(A^{\text{tr}}_n, \tilde{A}^{\text{pr}}_n, |\mathcal{N}(n)|)}$. Then, for $t \in \mathcal{T}_{\text{pr}}$,

$$\begin{aligned}
\mathbb{E}\left[Y_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})} \,\middle|\, LF \right] &\overset{A2}{=} \mathbb{E}\left[\left\langle \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\mathcal{N}(n)}} \right\rangle + \epsilon_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})} \,\middle|\, LF \right] \\
&\overset{A3}{=} \left\langle \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\mathcal{N}(n)}} \right\rangle \Big| LF \\
&= \left\langle \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\mathcal{N}(n)}} \right\rangle \,\Big|\, \{LF, \mathcal{O}\} \qquad (7)
\end{aligned}$$

$$
\begin{aligned}
&\overset{A4}{=} \left\langle \sum_{j \in \mathcal{I}^n} \alpha_j \tilde{\mathbf{u}}_{j,\pi_j(\mathcal{N}(j))}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\mathcal{N}(n)}} \right\rangle \Bigg| \{LF, \mathcal{O}\} \\
&= \sum_{j \in \mathcal{I}^n} \alpha_j \left\langle \tilde{\mathbf{u}}_{j,\pi_j(\mathcal{N}(j))}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\mathcal{N}(n)}} \right\rangle \Bigg| \{LF, \mathcal{O}\} \\
&\overset{D1}{=} \sum_{j \in \mathcal{I}^n} \alpha_j \left\langle \tilde{\mathbf{u}}_{j,\pi_j(\mathcal{N}(j))}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\pi_j(\mathcal{N}(j))}} \right\rangle \Bigg| \{LF, \mathcal{O}\} \\
&= \sum_{j \in \mathcal{I}^n} \alpha_j \left\langle \tilde{\mathbf{u}}_{j,\mathcal{N}(j)}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\mathcal{N}(j)}} \right\rangle \Bigg| \{LF, \mathcal{O}\} \qquad (8) \\
&\overset{A3}{=} \sum_{j \in \mathcal{I}^n} \alpha_j \mathbb{E}\left[\left\langle \tilde{\mathbf{u}}_{j,\mathcal{N}(j)}, \tilde{\mathbf{w}}_{t,\mathbf{a}^t_{\mathcal{N}(j)}} \right\rangle + \epsilon_{tj}^{(\mathbf{a}^t_{\mathcal{N}(j)})} \Bigg| \{LF, \mathcal{O}\} \right] \\
&\overset{A2}{=} \sum_{j \in \mathcal{I}^n} \alpha_j \mathbb{E}\left[Y_{tj}^{(\mathbf{a}^t_{\mathcal{N}(j)})} \Big| LF, \mathcal{O} \right] \\
&= \sum_{j \in \mathcal{I}^n} \alpha_j \mathbb{E}\left[Y_{tj} \Big| LF, \mathcal{O} \right], \qquad (9)
\end{aligned}
$$

where (7) follows from the fact that, conditioned on LF, the left-hand side is deterministic, which implies that event on which it is conditioned can be exchanged for $\{LF, \mathcal{O}\}$. Therefore,

$$
\begin{aligned}
\phi_n^{(\tilde{A}^{\text{pr}}_n)} &= \frac{1}{|\mathcal{T}_{\text{pr}}|} \sum_{t \in \mathcal{T}_{\text{pr}}} \mathbb{E}\left[Y_{tn}^{(\mathbf{a}^t_{\mathcal{N}(n)})} \Big| LF \right] \qquad (10) \\
&= \frac{1}{|\mathcal{T}_{\text{pr}}|} \sum_{t \in \mathcal{T}_{\text{pr}}} \sum_{j \in \mathcal{I}^n} \alpha_j \mathbb{E}\left[Y_{tj} \Big| LF, \mathcal{O} \right], \qquad (11)
\end{aligned}
$$

where the first equality follows from the definition of $\phi_n^{(\tilde{A}^{\text{pr}}_n)}$ and the second equality follows from (9). Note that Assumption 1 immediately holds from Assumption 2. □

B.2 Proof of Theorem 2

As indicated in the main text, Theorem 2 is adapted from Theorem 4.2 of [2]. Below, we explain how to adapt Theorem 4.2 for this work.

Model. The model in [2] is given by (in their notation)

$$
Y_{tn}^{(d)} = \left\langle u_t^{(d)}, v_n \right\rangle + \varepsilon_{tn}^{(d)}, \qquad (12)
$$

where $u_t^{(d)}, v_n \in \mathbb{R}^r$ are latent factors; $\varepsilon_{tn}^{(d)}$ is a zero-mean, independent noise term; and $Y_{tn}^{(d)}$ is the potential outcome of interest.

Recall from (6) that our model is given by (in our notation)

$$
Y_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} = \langle \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}} \rangle + \epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}, \qquad (13)
$$

where $\tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}} \in \mathbb{R}^{r|\mathcal{N}(n)|}$ are latent factors; $\epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}$ is a zero-mean, independent noise term; and $Y_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}$ is the potential outcome of interest.

As such, our setup model is analogous to the model used by [2], with a change of notation. Specifically, $\tilde{\mathbf{u}}_{n,\mathcal{N}(n)}$ in this work corresponds to $u_t^{(d)}$ in [2], $\tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}}$ to v_n, and $\epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}$ to $\varepsilon_{tn}^{(d)}$.

Assumptions of Theorem 4.2 in [2]. Given that our model (6) can be mapped to the model in [2], it remains to check whether the assumptions in Theorem 4.2 of [2] are satisfied by those in Theorem 2.

In particular, one of the main differences between our work and [2] is the observation pattern. In this work, the observation pattern is more general, allowing for any sequence of treatments during the

training and prediction periods (referred to as the "pre-intervention" and "post-intervention" periods in [2]). In [2], the treatment must be constant across each period, and it is assumed that all units are under treatment 0 during the pre-intervention (i.e., training) period. This difference only affects Theorem 2 via the donor set. In other words, once we adjust the choice of donor set (see Definition 1) to suit the network interference setting, Theorem 4.2 can be mapped directly to Theorem 2.

We now go through the assumptions one-by-one. As we saw above, Assumption 2 is equivalent to Assumption 2 in [2], with a change of notation. Furthermore, as discussed in Section 4, Assumption 1 is automatically satisfied when Assumption 2 holds. Assumptions 3-6 map one-to-one to Assumptions 3-6 of [2] under the change of notation. Lastly, Assumptions 7-8 also map one-to-one to Assumptions 7-8 under the new definition of a donor set, as given by Definition 1.

B.3 Proof of Proposition 3

Proof. Recall that for unit $n \in [N]$, measurement $t \in [T]$, and treatments $\mathbf{a} \in [A]_0^N$,

$$Y_{tn}^{(\mathbf{a})} = \sum_{k \in \mathcal{N}(n)} \langle \mathbf{u}_{k,n}, \mathbf{w}_{t,a_k} \rangle + \epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{14}$$

where $n \in \mathcal{N}(n)$. Because $D = 2$, $a_k \in \{0, 1\}$ for all $k \in [N]$. As such,

$$Y_{tn}^{(\mathbf{a})} - \epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} = \sum_{k \in \mathcal{N}(n)} \mathbf{1}(a_k = 0)\mathbf{u}_{k,n}^\top \mathbf{w}_{t,0} + \sum_{k \in \mathcal{N}(n)} \mathbf{1}(a_k = 1)\mathbf{u}_{k,n}^\top \mathbf{w}_{t,1} \tag{15}$$

$$= \begin{bmatrix} \sum_{k \in \mathcal{N}(n)} \mathbf{1}(a_k = 0)\mathbf{u}_{k,n}^\top & \sum_{k \in \mathcal{N}(n)} \mathbf{1}(a_k = 1)\mathbf{u}_{k,n}^\top \end{bmatrix} \begin{bmatrix} \mathbf{w}_{t,0} \\ \mathbf{w}_{t,1} \end{bmatrix} \tag{16}$$

$$= \begin{bmatrix} \sum_{k \in \mathcal{N}(n)} (1 - a_k)\mathbf{u}_{k,n}^\top & \sum_{k \in \mathcal{N}(n)} a_k \mathbf{u}_{k,n}^\top \end{bmatrix} \begin{bmatrix} \mathbf{w}_{t,0} \\ \mathbf{w}_{t,1} \end{bmatrix} \tag{17}$$

Given a unit $n \in [N]$ and sequence of counterfactual, prediction treatments of interest $\tilde{A}_n^{\text{pr}} \in \{0, 1\}^{|\mathcal{N}(n)| \times \mathcal{T}_{\text{pr}}}$. Recall that we use $\mathcal{I}^n$ as a shorthand for $\mathcal{I}^{(A_n^{\text{tr}}, \tilde{A}_n^{\text{pr}}, |\mathcal{N}(n)|)}$. Further, we let $\mathcal{I}_j^n$ refer to the j-th donor in the donor set $\mathcal{I}^n$.

Recall that:

$$Z_{\text{tr},\mathcal{I}^n} = \begin{bmatrix} Y_{1,\mathcal{I}_1^n} & Y_{1,\mathcal{I}_2^n} & \dots & Y_{1,\mathcal{I}_{|\mathcal{I}^n|}^n} \\ Y_{2,\mathcal{I}_1^n} & Y_{2,\mathcal{I}_2^n} & \dots & Y_{2,\mathcal{I}_{|\mathcal{I}^n|}^n} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{T_{\text{tr}},\mathcal{I}_1^n} & Y_{T_{\text{tr}},\mathcal{I}_2^n} & \dots & Y_{T_{\text{tr}},\mathcal{I}_{|\mathcal{I}^n|}^n} \end{bmatrix} \in \mathbb{R}^{T_{\text{tr}} \times |\mathcal{I}^n|}, \tag{18}$$

denotes the observations across all training periods, and

$$Z_{\text{pr},\mathcal{I}^n} = \begin{bmatrix} Y_{T-T_{\text{pr}}+1,\mathcal{I}_1^n} & Y_{T-T_{\text{pr}}+1,\mathcal{I}_2^n} & \dots & Y_{T-T_{\text{pr}}+1,\mathcal{I}_{|\mathcal{I}^n|}^n} \\ Y_{T-T_{\text{pr}}+2,\mathcal{I}_1^n} & Y_{T-T_{\text{pr}}+2,\mathcal{I}_2^n} & \dots & Y_{T-T_{\text{pr}}+2,\mathcal{I}_{|\mathcal{I}^n|}^n} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{T,\mathcal{I}_1^n} & Y_{T,\mathcal{I}_2^n} & \dots & Y_{T,\mathcal{I}_{|\mathcal{I}^n|}^n} \end{bmatrix} \in \mathbb{R}^{T_{\text{pr}} \times |\mathcal{I}^n|}, \tag{19}$$

denotes the observations during the prediction period.

Without loss of generality, we assume that the first of the L sub-periods occupies the first $T_{\text{tr},1}$ time steps of $\mathcal{T}_{\text{tr}}$, the second sub-period occupies the next $T_{\text{tr},2}$ time steps of $\mathcal{T}_{\text{tr}}$, and so on.

The **subspace inclusion assumption** (SIA) requires that $\text{rowspace}(Z_{\text{pr},\mathcal{I}^n}) \subset \text{rowspace}(Z_{\text{tr},\mathcal{I}^n})$.

Let $\tilde{\mathcal{N}}(j)$ denote the $\pi_j(\tilde{\mathcal{N}}(j))$, where π_j is specified in Definition 1, i.e., $\tilde{\mathcal{N}}(j)$ corresponds to the already-permuted neighborhood of donor j, where the permutation is fixed under Definition 1.

$$U_{\mathcal{I}^n} = \begin{bmatrix} \mathbf{u}_{\tilde{\mathcal{N}}_1(\mathcal{I}_1^n),\mathcal{I}_1^n} & \dots & \mathbf{u}_{\tilde{\mathcal{N}}_1(\mathcal{I}_{|\mathcal{I}^n|}^n),\mathcal{I}_{|\mathcal{I}^n|}^n} \\ \vdots & \ddots & \vdots \\ \mathbf{u}_{\pi_1(\tilde{\mathcal{N}}_{|\tilde{\mathcal{N}}(n)|}(\mathcal{I}_1^n)),\mathcal{I}_1^n} & \dots & \mathbf{u}_{\tilde{\mathcal{N}}_{|\tilde{\mathcal{N}}(n)|}(\mathcal{I}_{|\mathcal{I}^n|}^n),\mathcal{I}_{|\mathcal{I}^n|}^n} \end{bmatrix} \in \mathbb{R}^{r|\tilde{\mathcal{N}}(n)| \times |\mathcal{I}^n|}$$

We now define W_{tr} and B_{tr}. For ease of exposition, we express them for $L = 2$. Let $\mathbb{I}_r$ denote the $r \times r$ identity matrix. Using (16),

$$\mathbb{E}[Z_{\text{tr},\mathcal{I}^n}] = \underbrace{\begin{bmatrix} \mathbf{w}_{1,0}^\top & \mathbf{w}_{1,1}^\top & 0 & 0 \\ \mathbf{w}_{2,0}^\top & \mathbf{w}_{2,1}^\top & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \mathbf{w}_{T_{\text{tr},1}+1,0}^\top & \mathbf{w}_{T_{\text{tr},1}+1,1}^\top \\ 0 & 0 & \mathbf{w}_{T_{\text{tr},1}+2,0}^\top & \mathbf{w}_{T_{\text{tr},1}+2,1}^\top \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{W_{\text{tr}} \in \mathbb{R}^{T_{\text{tr}} \times 2rL}} \left(\underbrace{\begin{bmatrix} 1 - \bar{a}^1_{\mathcal{N}_1(n)} & 1 - \bar{a}^1_{\mathcal{N}_2(n)} & \dots \\ \bar{a}^1_{\mathcal{N}_1(n)} & \bar{a}^1_{\mathcal{N}_2(n)} & \dots \\ 1 - \bar{a}^2_{\mathcal{N}_1(n)} & 1 - \bar{a}^2_{\mathcal{N}_2(n)} & \dots \\ \bar{a}^2_{\mathcal{N}_1(n)} & \bar{a}^2_{\mathcal{N}_2(n)} & \dots \end{bmatrix}}_{B_{\text{tr}} \in \{0,1\}^{2L \times |\mathcal{N}(n)|}} \otimes \mathbb{I}_r \right) U_{\mathcal{I}^n}, \tag{20}$$

and, analogously,

$$\mathbb{E}[Z_{\text{pr},\mathcal{I}^n}] = \underbrace{\text{diag}_{t \in \mathcal{T}_{\text{pr}}} \left([\mathbf{w}_{t,0}^\top, \mathbf{w}_{t,1}^\top] \right)}_{W_{\text{pr}} \in \mathbb{R}^{T_{\text{pr}} \times 2rT_{\text{pr}}}} \left(\underbrace{\begin{bmatrix} 1 - a^{T-T_{\text{pr}}+2}_{\mathcal{N}_1(n)} & 1 - a^{T-T_{\text{pr}}+2}_{\mathcal{N}_2(n)} & \dots \\ a^{T-T_{\text{pr}}+2}_{\mathcal{N}_1(n)} & a^{T-T_{\text{pr}}+2}_{\mathcal{N}_2(n)} & \dots \\ 1 - a^{T-T_{\text{pr}}+2}_{\mathcal{N}_1(n)} & 1 - a^{T-T_{\text{pr}}+2}_{\mathcal{N}_2(n)} & \dots \\ a^{T-T_{\text{pr}}+2}_{\mathcal{N}_1(n)} & a^{T-T_{\text{pr}}+2}_{\mathcal{N}_2(n)} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}}_{B_{\text{pr}} \in \{0,1\}^{2T_{\text{pr}} \times |\mathcal{N}(n)|}} \otimes \mathbb{I}_r \right) U_{\mathcal{I}^n}. \tag{21}$$

Let $K_{\text{tr}} = W_{\text{tr}}(B_{\text{tr}} \otimes \mathbb{I}_r)$ and $K_{\text{pr}} = W_{\text{pr}}(B_{\text{pr}} \otimes \mathbb{I}_r)$.

Note that any matrix that has linear independent columns has full row space. Hence to complete the proof, is suffices to show that K_{tr} has linearly independent columns. Now if B_{tr} and W_{tr} have linearly independent columns, then it immediately implies that K_{tr} has linearly independent columns. □

B.4 Proof of Proposition 4

Proof. Below, we use the same notation as in the proof of Proposition 3.

Subspace inclusion effectively requires that, for every $i \in [T_{\text{pr}}]$ there exists some $\boldsymbol{\phi} \in \mathbb{R}^{T_{\text{tr}}}$ such that

$$\mathbf{e}_i^\top W_{\text{pr}}(B_{\text{pr}} \otimes \mathbb{I}_r) U_{\mathcal{I}^n} = \boldsymbol{\phi}^\top W_{\text{tr}}(B_{\text{tr}} \otimes \mathbb{I}_r) U_{\mathcal{I}^n}.$$

Therefore, subspace inclusion holds for any $U_{\mathcal{I}^n}$ if there exists some $\boldsymbol{\phi} \in \mathbb{R}^{T_{\text{tr}}}$ such that

$$\mathbf{e}_i^\top W_{\text{pr}}(B_{\text{pr}} \otimes \mathbb{I}_r) = \mathbf{e}_i^\top K_{\text{pr}} = \boldsymbol{\phi}^\top K_{\text{tr}} = \boldsymbol{\phi}^\top W_{\text{tr}}(B_{\text{tr}} \otimes \mathbb{I}_r). \tag{22}$$

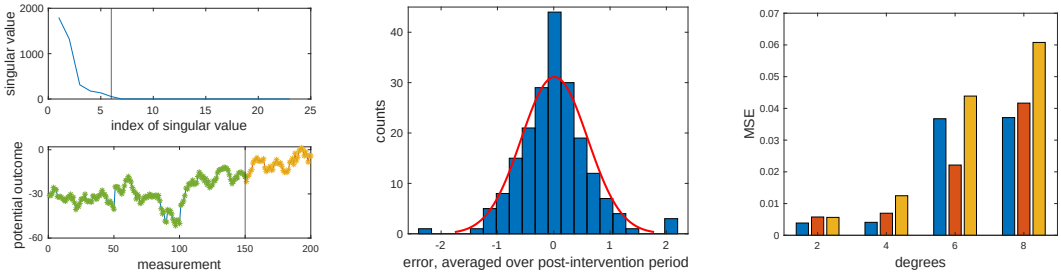
Therefore, by the second equality, subspace inclusion requires that $\text{rowspace}(K_{\text{pr}}) \subset \text{rowspace}(K_{\text{tr}})$. Note . Given that (i) $\text{rowspace}(K_{\text{pr}}) \subset \text{rowspace}(B_{\text{pr}} \otimes \mathbb{I}_r)$ and (ii) $\text{rowspace}(K_{\text{tr}}) = \text{rowspace}(B_{\text{tr}} \otimes \mathbb{I}_r)$ since W_{tr} has linearly independent columns, it suffices to show that $\text{rowspace}(B_{\text{pr}} \otimes \mathbb{I}_r) \subset \text{rowspace}(B_{\text{pr}} \otimes \mathbb{I}_r)$. This is equivalent to showing that $\text{rowspace}(B_{\text{pr}}) \subset \text{rowspace}(B_{\text{tr}})$.

Since the rows of B_{pr} are $\tilde{\mathbf{a}}^t_{\mathcal{N}(n)}$ and $1 - \tilde{\mathbf{a}}^t_{\mathcal{N}(n)}$ for all $t \in \mathcal{T}_{\text{pr}}$, $\text{rowspace}(B_{\text{pr}}) \subset \text{rowspace}(B_{\text{tr}})$ holds when $\tilde{\mathbf{a}}^t_{\mathcal{N}(n)}, 1 - \tilde{\mathbf{a}}^t_{\mathcal{N}(n)} \in \text{rowspace}(B_{\text{tr}})$ for all $t \in \mathcal{T}_{\text{pr}}$. □

C Simulations

Below, we re-present the results given in Section 5, providing additional simulation results.

All results are given for binary treatments, i.e., $D = 2$. The latent factors $\mathbf{u}_{k,n}$ and $\mathbf{w}_{0,a}$ are drawn from a standard random normal distribution, and $\mathbf{w}_{\cdot,a}$ are a Gaussian random walk. Our experiments use a simple donor-finding algorithm. In particular, instead of searching for donors over all possible



(a) NSI estimator predictions (b) Residuals of NSI estimator (c) MSE for regular graphs

Figure 3: Simulation results for NSI estimator under $\epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} \sim \mathcal{N}(0, 0.1)$. (a) Consider a ring graph with $N = 400$, $T_{\text{tr},\ell} = T_{\text{pr}} = 50$, $r = 2$, and $L = 3$. The top graph plots the spectrum $\{\hat{s}_\ell\}_{\ell=1}^{q_{\text{tr}}}$, and the vertical line marks κ from Section 3. The NSI estimates are plotted below, where the vertical line separates the training and prediction periods. The ground-truth values are given as lines, and the predictions are marked with $*$'s. (b) gives the histogram of residuals (the difference between the estimated and ground-truth potential outcomes) of 200 simulations, averaged over 50 units and all possible counterfactual treatments. The experimental parameters match those (a). (c) plots the MSE of the NSI estimator across regular graphs of different degrees. The left (blue) bars are for $N = 1000$ and $T_{\text{tr},\ell} = T_{\text{pr}} = 100$. The middle (red) bars are for $N = 1000$, $T_{\text{tr},\ell} = T_{\text{pr}} = 50$. The right (yellow) bars are for $N = 500$, $T_{\text{tr},\ell} = T_{\text{pr}} = 50$. All other parameters match those for (a).

permutations π_j, as defined in Definition 1, we fix an ordering of units (as described in Section 1) and restrict ourselves to the identity permutation $\pi_j(i) = i$.

Figure 2 shows results for the NSI estimator over a ring graph, such that the size of each neighborhood set is 3. For Fig. 2(a)-(b), we adopt the setup described in Section 4.3, where $\mathcal{T}_{\text{tr}}$ is divided into $L = 3$ sub-periods, each of length $T_{\text{tr},\ell} = 50$ and $\bar{\mathbf{a}}^1 = (1, 0, 0, 1, 0, 0, \ldots)$, $\bar{\mathbf{a}}^2 = (0, 1, 0, 0, 1, 0, \ldots)$, and $\bar{\mathbf{a}}^3 = (0, 0, 1, 0, 0, 1, \ldots)$. For Fig. 2(c), we study the Network Synthetic Control setting described in Example 1, where $L = 1$, $\bar{\mathbf{a}}^1 = \tilde{\mathbf{a}}^{\text{pr}} = \mathbf{0}_N$, $\mathbf{a}^{T-T_{\text{pr}}+1}$ is drawn uniformly at random, and $\mathbf{a}^t$ is constant across $\mathcal{T}_{\text{pr}}$.

We also compare the NSI estimator to two others: the SI estimator (which does not account for network interference) [2] and a baseline estimator. The baseline estimator finds donor units that satisfy Definition 1, then averages the donor units' observed outcomes. We compare the estimators for a ring graph under the same hyper-parameters as those used in Fig. 2(a), averaging across 200 simulations, 50 units, and all possible counterfactual treatments. The MSEs and R-squared values for the NSI estimator, SI estimator, and baseline estimators are, respectively, **(0.08013, 0.9994)**, **(53.10, 0.9101)**, and **(576.1, -1.389)**. Both the NSI and baseline estimators used donor sets that contained, on average, 16 units. The SI estimator used donor sets with, on average, 66 units. These results as well as those for Fig. 2(a)-(b) are given for $\kappa \geq 3r$, and the results for Fig. 2(c) are given for $\kappa \geq r$.

The simulations were simple. They were run on a local machine with a 2.3 GHz processor. The simulations were completed in under two hours.