Network Synthetic Interventions: A Causal Framework for Panel Data Under Network Interference

Anish Agarwal[1], Sarah H. Cen*[2], Devavrat Shah[2], and Christina Lee Yu[3]

[1]Industrial Engineering and Operations Research, Columbia University
[2]Electrical Engineering and Computer Science, Massachusetts Institute of Technology
[3]Operations Research and Information Engineering, Cornell University

Abstract

We propose a generalization of the synthetic controls and synthetic interventions methodology to incorporate network interference. We consider the estimation of unit-specific potential outcomes from panel data in the presence of spillover across units 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, *Network Synthetic Interventions* (NSI), and show that it consistently estimates the mean outcomes for a unit under an arbitrary set of counterfactual treatments for the network. We further establish that the estimator is asymptotically normal. We furnish two validity tests for whether the NSI estimator reliably generalizes to produce accurate counterfactual estimates. We provide a novel graph-based experiment design that guarantees the NSI estimator produces accurate counterfactual estimates, and also analyze the sample complexity of the proposed design. We conclude with simulations that corroborate our theoretical findings.

1 Introduction

There is growing interest in the identification and estimation of causal effects in the context of spillover on networks, in which the outcomes of a unit are affected by the treatments assigned to other units, known as the unit's "neighbors." Here, a unit could be an individual, customer cohort, or region, and correspondingly, treatments could be recommendations, discounts, or legislation. For example, whether an individual gets COVID-19 is a function of not only the individual's vaccination status but also the vaccination status of that individual's social network. In the setting of e-commerce, the number of goods sold of a particular product is a function of not only whether that product gets a discount, but the discount level of other products that are substitutes or complements of it. That is, there is *network interference*.

In this work, we focus on network inference with *panel data*, a ubiquitous manner in which data is structured, where we collect multiple measurements of different units, and each unit can undergo a different sequence of treatments. See Figure 1 for an example of panel data and the type of causal question we are interested in. Causal inference with panel data has recently received significant attention, and a popular class of estimators in such settings are known as *matching estimators*, where one represents the outcomes of one unit as some combination of other units to answer counterfactual questions. Such estimators have been very popular in practice due to their

*Correspondence to Sarah H. Cen at shcen@mit.edu.

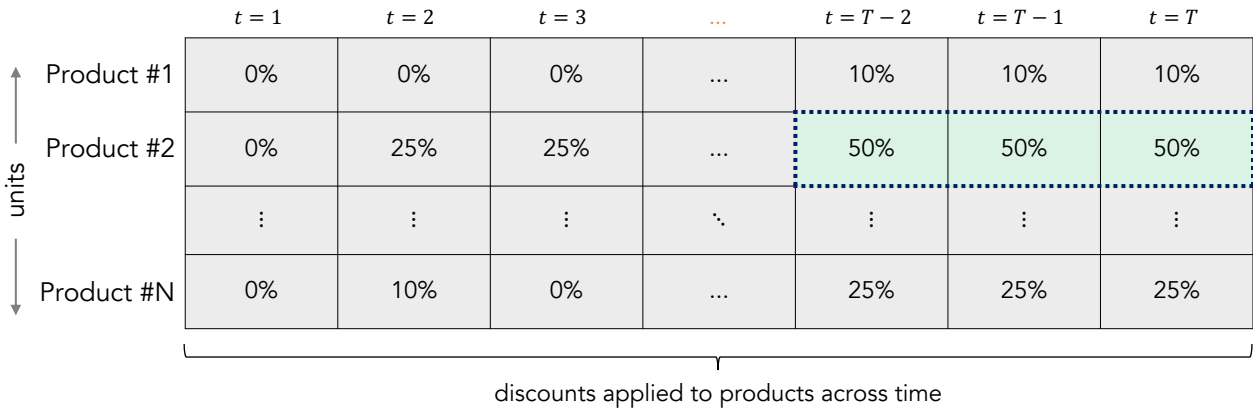


Figure 1: Panel data setting illustrated via an online retail example. Each row corresponds to a product (or unit). Each column corresponds to a week (or measurement). A discount (the treatment) is applied to each product each week. In this work, we ask questions of the form: Given every product's sales numbers across time and under various discounts, what would the sales numbers (i.e., potential outcomes) have been for a specific unit, like Product #2, under a different set of end-of-year discounts?

flexibility and simplicity in addition to the fact that they provide valid causal estimates under unobserved confounding with appropriate assumptions. Some examples of matching estimators with panel data include Difference-in-Differences (DiD) (Bertrand et al. 2004), Synthetic Controls (SC) (Abadie 2021), and variants thereof. However, such matching estimators rely on the Stable Unit Treatment Value Assumption (SUTVA), which implies that there is no spillover across units, i.e., the treatment applied to one unit does not affect the outcomes of other units. Failing to account for spillovers can lead to biased estimates.

We propose a novel latent factor model—which is a generalization of models studied in the panel data literature—that accounts for network interference. Given this model, we establish an identification result where the counterfactual potential outcome for a given unit and its neighbors can be written as a linear combination of the observed outcomes of a carefully selected set of other units. This identification result leads to a natural estimator, which we call *Network Synthetic Interventions (NSI)*, a simple two-step procedure, that estimates the mean counterfactual potential outcome for a given unit. We then show that, given our latent factor model, the NSI estimator is finite-sample consistent and asymptotically normal under suitable conditions. NSI and our analysis of it can be viewed as a generalization of the Synthetic Interventions (Agarwal et al. 2020b) and, in turn, Synthetic Controls frameworks to account for network interference.

We furnish two validity tests that verify whether the treatment assignment pattern and the observed data have enough variation such that valid counterfactual estimates can be produced. Motivated by these tests, we provide a novel graph-based experiment design.

To explain the efficacy of the experiment design and the NSI estimator, we consider the setting of a regular network graph with degree $d \geq 2$. We show that the proposed experiment design requires only $O(d^3)$ training samples in order to guarantee that the training data has enough

variation such that it is possible to generalize to a given target counterfactual treatment. Further, NSI obtains an estimate within error ε with high probability under the proposed experiment design when $O\big(\mathsf{poly}(d)/\varepsilon^4\big)$ training samples per unit are available. This is a significant improvement over the $O(\mathsf{exp}(d)/\varepsilon^2)$ training samples that a naive procedure would require.

We conclude with simulations showing that NSI is robust to spillovers under which existing estimators are biased.

1.1 Related work

The literature on causal inference with network interference or spillover effect has mostly considered the setting of a single measurement per unit, whether in the setting of a randomized experiment or an 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 identifiable (Manski 2013, Aronow et al. 2017, Basse and Airoldi 2018a, Karwa and Airoldi 2018). Subsequently, various models have been proposed in the literature that impose restrictions on the exposure functions (Manski 2013, Aronow et al. 2017, Viviano 2020, Auerbach and Tabord-Meehan 2021, Li et al. 2021), interference neighborhoods (Ugander et al. 2013, Bargagli-Stoffi et al. 2020, Sussman and Airoldi 2017a, Bhattacharya et al. 2020), parametric structure (Toulis and Kao 2013, Basse and Airoldi 2018b, Cai et al. 2015, Gui et al. 2015, Eckles et al. 2017), two-sided platforms (Johari et al. 2022, Bajari et al. 2021) or a combination of these, each leading to a different solution concept. A comprehensive review on network interference models is given by De Paula (2017). 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, or additivity of interference effects (Sussman and Airoldi 2017a, Yu et al. 2022, Cortez et al. 2022a,b).

Distinct to our work is that we consider a *panel data* setting in which there are multiple measurements (e.g., a time series) for each unit. The potential outcomes function is thus also dependent on both the unit and the measurement. Additionally, we allow for the estimation of unit-specific 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. Previous work has focused on causal estimands that capture population-level effects, such as the average direct treatment effect (the average difference in outcomes if only one unit and none of its neighbors get treated (Basse and Airoldi 2018b, Jagadeesan et al. 2020, Sävje et al. 2021, Sussman and Airoldi 2017a, Leung 2019, Ma and Tresp 2021)) and the average total treatment effect (the average difference in outcomes if all units get treated versus if they do not (Ugander et al. 2013, Eckles et al. 2017, Chin 2019, Yu et al. 2022, Cortez et al. 2022a,b)). Alternately there has been some literature that focuses on hypothesis testing for the presence of network interference (Aronow 2012, Bowers et al. 2013, Athey et al. 2018, Pouget-Abadie et al. 2017, Saveski et al. 2017); these results do not immediately extend to estimation as they are based on randomization inference with a fixed network size, and focus on testing the sharp null hypotheses.

While a majority of the literature focuses on randomized experiments, there is a growing interest in the literature to account for network interference when analyzing observational studies. The existing literature generally assumes partial interference, where the network consists of many disconnected sub-communities (Tchetgen and VanderWeele 2012, Perez-Heydrich et al. 2014, Liu et al. 2016, DiTraglia et al. 2020, Vazquez-Bare 2022). Without this strong clustering condition, other works impose strong parametric assumptions on the potential outcomes function, assuming that the potential outcomes only depend on a known statistic of the neighborhood treatment, e.g. the number or fraction of treated (Verbitsky-Savitz and Raudenbush 2012, Chin 2019, Ogburn et al.

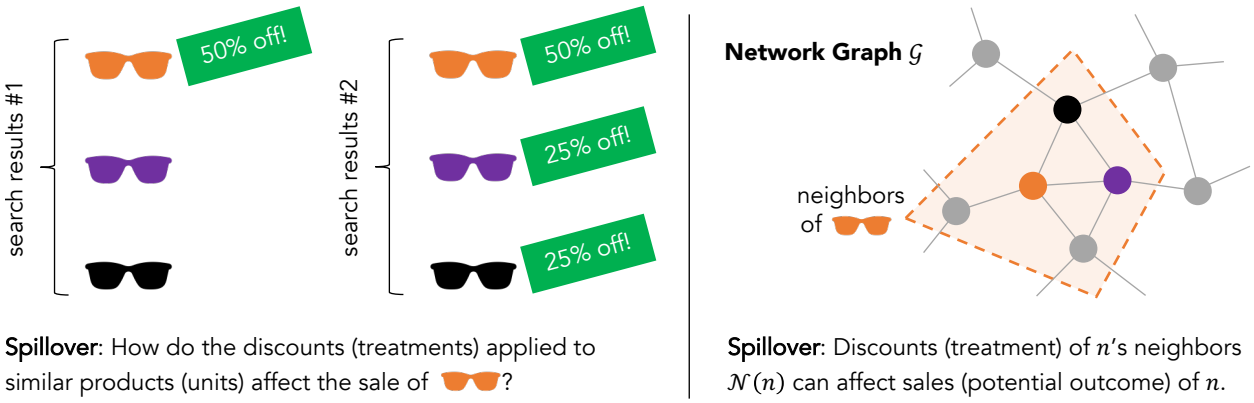


Figure 2: Example of spillover effects and how they can be captured via a graph network. On the left, suppose that an online retailer presents similar products alongside one another. Then, the sale of one product (e.g., orange sunglasses) is affected by the discounts applied to similar products; in this case, other sunglasses. On the right, spillover is often modeled via a network graph $\mathcal{G}$, in which the treatments applied to the neighbors $\mathcal{N}(n)$ of a unit n may affect the potential outcomes of n.

2017). This reduces estimation to a regression task under requirements of sufficient diversity in the treatments. Belloni et al. (2022) also consider a setting in which the exposure mapping is known but allow the "radius" of interference to vary across units, then learn this radius from data to devise a doubly robust estimator. Forastiere et al. (2021) consider a general exposure mapping model alongside an inverse propensity weighted estimator, but the estimator has high variance when the exposure mapping is complex. De Paula et al. (2018) and De Paula et al. (2019) derive identification conditions when the observational panel data contains no information about the social ties (i.e., network). Further, building on recent works in panel data (Agarwal et al. 2020b, 2021a), we allow for unobserved confounding in treatment assignment as long as there exist low-rank latent factors that mediate the unobserved confounding, i.e., there is "selection on latent factors".

2 Setup & Model

We begin with some notation. Let $[X] := \{1, \ldots, X\}$ for any positive integer X. For vector $\mathbf{a} \in [D]^N$ and set $S \subseteq [N]$, let $\mathbf{a}_S \in [D]^{|S|}$ denote the vector containing the elements of $\mathbf{a}$ indexed by S and $a_i \in [D]$ denote the i-th element of $\mathbf{a}$. Let $\mathbb{I}_x$ denote the $x \times x$ identity matrix and $\otimes$ denote the Kronecker product. Let $\text{Ind}(\cdot)$ denote the indicator function. Let $\|\cdot\|_{\psi_2}$ denote the Orlicz norm. Let O_p denote a probabilistic version of big-O notation and $\tilde{\Omega}$ denote the variation on big-Ω notation that ignores logarithmic terms (see Appendix A for precise definitions). For sets of indices $S_1 \subseteq [m_1]$ and $S_2 \subseteq [m_2]$ and a matrix $\Pi \in \mathbb{R}^{m_1 \times m_2}$, let $\Pi[S_1, S_2] \in \mathbb{R}^{|S_1| \times |S_2|}$ denote the submatrix corresponding to the rows indexed by S_1 and columns index by S_2. We use ":" as a shorthand for all indices such that $\Pi[:, S_2] \in \mathbb{R}^{m_1 \times |S_2|}$ and $\Pi[S_1, :] \in \mathbb{R}^{|S_1| \times m_2}$. Let $\mathcal{X}^*$ denote the $*$-product space, where its length is not pre-determined. Let Π^+ denote the pseudo-inverse of Π.

2.1 Setup

Consider $N \geq 1$ units, $D \geq 1$ treatments, and $T \geq 1$ measurements of interest. We denote the potential outcome for a given unit n and measurement t by the real-valued random variable $Y_{t,n}^{(\mathbf{a})}$, where $\mathbf{a} \in [D]^N$ denotes the vector of treatments over *all* N units. This definition allows for spillover

effects because the potential outcome for a given unit is a function of the treatment assignment of all units. To model spillover across units, we use a network graph. Let $\mathcal{G} = ([N], \mathcal{E})$ denote a graph over the N units, where $\mathcal{E} \subseteq [N] \times [N]$ denotes the edges of the graph. Throughout, we assume that $\mathcal{G}$ is *fixed* and *known*. Let $\mathcal{N}(n)$ denote the neighbors of unit $n \in [N]$ with respect to $\mathcal{G}$ such that $j \in \mathcal{N}(n) \iff (j, n) \in \mathcal{E}$. For simplicity of notation, let self-edges be included, i.e., $(n, n) \in \mathcal{E}$ for all $n \in [N]$. We assume that the network graph $\mathcal{G}$ captures spillover effects in the following way.

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

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

where $\mathbf{a}_{\mathcal{N}(n)} \in [D]^{|\mathcal{N}(n)|}$ 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 $j \in [N] \setminus \mathcal{N}(n)$.

See Figure 2 for an example of spillover and its network representation. Several prior works on network interference also assume SNTVA, e.g., as the *Neighborhood Interference Assumption (NIA)* (Sussman and Airoldi 2017b). It can be viewed as a particular instantiation of exposure mappings, as defined by Aronow and Samii (2017), and effective treatment functions (e.g., under the constant treatment response assumption) (Manski 2013).

Remark 1. *SNTVA only captures first-order spillover effects, i.e., assumes that the potential outcome of unit n is only affected by the treatments of its* immediate *neighbors. One could capture higher-order spillover effects by adding edges to $\mathcal{G}$. The trade-off is that, as the number of edges in $\mathcal{G}$ increases, the estimation bounds for the NSI estimator in Section 4 get correspondingly weaker.*

Remark 2. *Although we assume $\mathcal{G}$ is an undirected graph, our results can be adapted for directed graphs by changing the definition of $\mathcal{N}(i)$. When $\mathcal{G}$ is directed, $j \in \mathcal{N}(i)$ if and only if $(j, i) \in \mathcal{E}$.*

2.2 Network latent-factor model

In this section, we introduce the model that we use to develop our estimator and formal results.

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

$$Y_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})} = \langle \mathbf{u}_{n,n}, \mathbf{w}_{t,a_n} \rangle + \sum_{j \in \mathcal{N}(n) \setminus n} \langle \mathbf{u}_{j,n}, \mathbf{w}_{t,a_j} \rangle + \epsilon_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{1}$$

where $\mathbf{u}_{\cdot,\cdot} \in \mathbb{R}^r$ and $\mathbf{w}_{\cdot,\cdot} \in \mathbb{R}^r$ represent latent (unobserved) factors; $\epsilon_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})}$ represents additive, idiosyncratic shocks, and r is the "rank" or model complexity. Further, we assume that $\mathbb{E}\big[\epsilon_{t,i}^{(\mathbf{a}_{\mathcal{N}(i)})} \,|\, LF\big] = 0$, where $LF := \big\{\mathbf{u}_{j,i}, \mathbf{w}_{t,a} : i, j \in [N],\, t \in [T],\ \text{and } a \in [D]\big\}$.

We make several remarks. First, we note that Assumption 2 automatically satisfies Assumption 1. Second, the latent factor $\mathbf{u}_{j,n}$ captures the effect in the potential outcome $Y_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})}$ due to the interaction between node n and its neighbour j; analogously $\mathbf{w}_{t,a_j}$ captures the effect due to the treatment that neighbor j receives (i.e., a_j) for measurement t. Specifically, their effect is captured

through the inner product $\langle \mathbf{u}_{j,n}, \mathbf{w}_{t,a_j} \rangle$. In this sense, the spillover effect of different neighbors in (1) is additive. Lastly, (1) can be equivalently written as

$$Y_{t,n}^{(\mathbf{a})} = Y_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})} = \left\langle \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}} \right\rangle + \epsilon_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{2}$$

where

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

Here, $\mathcal{N}_i(n)$ refers to i-th neighbor of n. (2) is reminiscent of classical interactive fixed effects models studied in the literature. Indeed, we can think of $\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} \in \mathbb{R}^{r|\mathcal{N}(n)|}$ and $\tilde{\mathbf{w}}_{t,a_{\mathcal{N}(n)}} \in \mathbb{R}^{r|\mathcal{N}(n)|}$ as the *network-adjusted* latent factors and $r|\mathcal{N}(n)| \in \mathbb{N}_{>0}$ as denoting the *network-adjusted* "rank" (note that $r|\mathcal{N}(n)|$ is actually an upper bound on the model's rank, but we will refer to it as the network-adjusted rank for convenience).

2.3 Examples of latent-factor model

We discuss how examples of latent factor models previously studied in the literature are captured by the model we propose in Assumption 2. Further, we discuss how additive non-linear latent factor models can be approximated by the linear additive model we propose.

Example 1. *Consider a setting with* no spillover effects*, i.e., $\mathcal{N}(n) = \{n\}$ for all $n \in [N]$. Then, the latent factor model in (1) reduces to*

$$Y_{t,n}^{(\mathbf{a})} = Y_{t,n}^{(a_n)} = \langle \mathbf{u}_{n,n}, \mathbf{w}_{t,a_n} \rangle + \epsilon_{t,n}^{(a_n)}, \tag{3}$$

This recovers the model considered in (Agarwal et al. 2020b). As explained in (Agarwal et al. 2020b), this also captures the models considered in (Abadie 2021) and (Arkhangelsky et al. 2019).

Example 2. *There are several prior works that assume that network interference is additive. For instance, consider the model proposed by Yu et al. (2022) in which $D = 2$, i.e., the treatments are binary, denoted by $\{0, 1\}$, and*

$$Y_n^{(\mathbf{a})} = Y_n^{(\mathbf{a}_{\mathcal{N}(n)})} = u_{0,n} + u_{n,n}a_n + \sum_{j \in \mathcal{N}(n) \setminus n} u_{j,n}a_j + \epsilon_n^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{4}$$

where $u_{0,n}, u_{n,n}, u_{j,n}, \epsilon_n^{(\mathbf{a}_{\mathcal{N}(n)})} \in \mathbb{R}$. One can verify that (4) can be recovered from (1) by taking $r = 1$; $T = 1$ (i.e., no index t); $w_a = a$; and there is an auxiliary node 0 for which $a_0 = 1$ and $0 \in \mathcal{N}(n)$ for all $n \in [N]$.

That is, both our model (1) and (4) assume that spillover is additive, and in order to exploit the structure across measurements t that exists in panel data, we extend (4) by: (i) allowing for multiple measurements t, and (ii) assuming that $u_{j,n}a_j$ in (4) has the measurement-dependent latent-factor representation $\langle \mathbf{u}_{j,n}, \mathbf{w}_{t,a_j} \rangle$. These repeated measurements are exactly what lets us create personalized counterfactual trajectories per units and implicitly correct for unobserved confounding.

Example 3. *Consider a setting where network interference is additive but the effect of the latent factors is* non-linear*. Precisely, consider the following variation of* (1)*:*

$$Y_{t,n}^{(\mathbf{a})} = Y_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})} = h(\mathbf{u}_{n,n}, \mathbf{w}_{t,a_n}) + \sum_{j \in \mathcal{N}(n) \setminus n} g(\mathbf{u}_{j,n}, \mathbf{w}_{t,a_j}) + \epsilon_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{5}$$

where $h, g : \mathbb{R}^r \times \mathbb{R}^r \to \mathbb{R}$ are potentially non-linear functions. If the latent factors take value in a bounded domain, say $\mathcal{C} \subset \mathbb{R}^r$, and h, g are Lipschitz continuous (or more generally smooth*), then it can be argued that (see Theorem 1 by Shah et al. (2020) for example) for any given $\delta > 0$, there is some $r' = r'(\delta)$ large enough and choice of functions $\{\phi_k, \psi_k, \phi'_k, \psi'_k : \mathbb{R}^r \to \mathbb{R},\ k \le r'\}$ such that*

$$\left| h(\mathbf{u}, \mathbf{w}) - \sum_{k=1}^{r'} \phi_k(\mathbf{u}) \psi_k(\mathbf{w}) \right| \le \delta,$$

$$\left| g(\mathbf{u}, \mathbf{w}) - \sum_{k=1}^{r'} \phi'_k(\mathbf{u}) \psi'_k(\mathbf{w}) \right| \le \delta,$$

for all $\mathbf{u}, \mathbf{w} \in \mathcal{C}$. Then, by setting $\tilde{\mathbf{u}} = [\phi_k(\mathbf{u}) : k \le r']$, $\tilde{\mathbf{w}} = [\psi_k(\mathbf{u}) : k \le r']$, $\tilde{\mathbf{u}}' = [\phi'_k(\mathbf{u}) : k \le r']$, $\tilde{\mathbf{w}}' = [\psi'_k(\mathbf{u}) : k \le r']$, it follows that (5) is pointwise δ-approximated as a linear latent factor model as given in (1), with $\mathbf{u}, \mathbf{w}$ appropriately replaced by $\tilde{\mathbf{u}}, \tilde{\mathbf{w}}, \tilde{\mathbf{u}}'$, and $\tilde{\mathbf{w}}'$.

2.4 Target causal estimand

Recall that we consider the panel data setting in which we observe T measurements (e.g., a time series) for every unit. Let $a_n^t \in [D]$ denote the treatment assignment for unit n at measurement t; $\mathbf{a}^t \in [D]^N$ denote the vector of treatment assignments for all N units at t; and $A := [\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^T] \in [D]^{N \times T}$ denote the sequence of treatment assignments across all N units and T measurements. Note that the treatment assignments A are observed, and the potential outcome $Y_{t,n}^{(\mathbf{a}^t)}$ is observed for every unit $n \in [N]$ and measurement $t \in [T]$. We denote the *observed* outcomes for unit n at measurement t by $Z_{t,n} = Y_{t,n}^{(\mathbf{a}^t)}$ for all $t \in [T]$ and the matrix of observations by $Z \in \mathbb{R}^{T \times N}$.

To define the target causal estimand of interest, let $\mathcal{T}_{\text{pr}} \subseteq [T]$ refer to a subset of measurements for which we would like to make counterfactual predictions; let $T_{\text{pr}} = |\mathcal{T}_{\text{pr}}| \le T$. To simplify notation, we assume without loss of generality that the treatment assignments are fixed across the measurements in $\mathcal{T}_{\text{pr}}$, i.e., $A[:, t] = \mathbf{a}^{\text{pr}} \in [D]^N$ for all $t \in \mathcal{T}_{\text{pr}}$ (see Remark 3).

For any given unit n and target treatment assignment $\tilde{\mathbf{a}} \in [D]^N$, our goal is to estimate the individual potential outcome averaged over the prediction period:

$$\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \mathbb{E}\Big[Y_{t,n}^{(\tilde{\mathbf{a}}_{\mathcal{N}(n)})} \mid LF \Big], \tag{6}$$

using observations Z, where we condition on the latent factors, LF. Under Assumption 1, the outcome of unit n depends only on the treatments applied to $\mathcal{N}(n)$, i.e., on $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ rather than the entire $\tilde{\mathbf{a}}$. See Figure 3 for an illustration of our target causal estimand and observation pattern.

Note that our results are given for any $T_{\text{pr}} \ge 1$. That is, we show identifiability and finite-sample consistency even for point estimates (i.e., for $T_{\text{pr}} = 1$).

Remark 3. *Our assumption $A[:, t] = \mathbf{a}^{pr} \in [D]^N$ for all $t \in \mathcal{T}_{pr}$ is without loss of generality. First, our work does not allow for spillover across measurements, i.e., $Y_{t,n}^{(\mathbf{a})}$ does not depend on treatments*

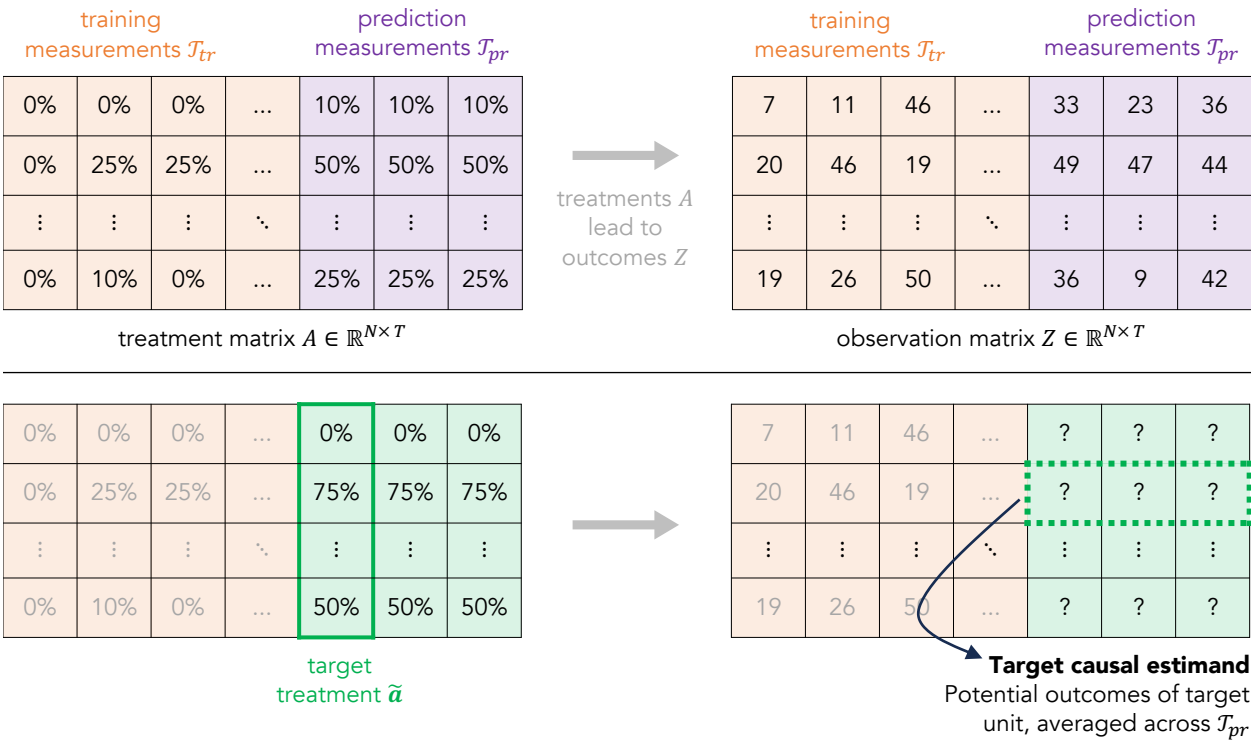


Figure 3: Illustration of our setup and target causal estimand. On the top-left is an example $N \times T$ treatment matrix A, split into the training and prediction measurements $\mathcal{T}_{\text{tr}}$ and $\mathcal{T}_{\text{pr}}$. On the top-right is the corresponding observation matrix Z, i.e., the (n, t)-th element of Z is the outcome of unit n at measurement t under treatments $\mathbf{a}^t$. The bottom-left gives the counterfactual treatment matrix, i.e., the treatments during $\mathcal{T}_{\text{tr}}$ remain intact and the counterfactual prediction treatment is given by $\tilde{\mathbf{a}}$. Lastly, on the bottom-right are the potential outcomes of interest under $\tilde{\mathbf{a}}$. The target causal estimand is the average of potential outcomes of a specific unit under $\tilde{\mathbf{a}}$ across $\mathcal{T}_{\text{pr}}$.

other than those assigned at t. We can therefore extract measurements in $\mathcal{T}_{pr}$ that share the same treatment $\mathbf{a}^{pr}$, i.e., we can redefine the prediction set as $\mathcal{T}'_{pr}$ so that $\mathbf{a}^t = \mathbf{a}^{pr}$ for all $t \in \mathcal{T}'_{pr}$. We can repeat this for all unique prediction treatments and apply NSI separately to each. Further, we note that our consistency and normality results allow for $T_{pr} = 1$, and so our results go through even if we have a different target prediction treatment for every measurement in $\mathcal{T}_{pr}$.

3 Network Synthetic Interventions (NSI) Estimator

We now describe our estimator for the estimand of interest (6), which we term *Network Synthetic Intervention* (NSI). It can be seen as a natural extension of the Synthetic Interventions (SI) estimator (Agarwal et al. 2020b), which is itself a generalization of Synthetic Controls (SC) (Abadie 2021) estimator, to settings in which there is *network interference*. For the remainder of this work, we fix the unit n and counterfactual treatment assignment $\tilde{\mathbf{a}}$ of interest.

3.1 Donor set

To define the NSI estimator, we introduce some necessary concepts. First, let $\mathcal{T}_{\text{tr}} \subset [T]$ denote a subset of the measurements known as *training* measurements. Without loss of generality, let $\mathcal{T}_{\text{tr}} := \{1, 2, \ldots, T_{\text{tr}}\}$, $\mathcal{T}_{\text{pr}} := \{T_{\text{tr}} + 1, \ldots, T\}$, $T_{\text{tr}} := |\mathcal{T}_{\text{tr}}|$, and $T_{\text{pr}} := |\mathcal{T}_{\text{pr}}|$. We note that $\mathcal{T}_{\text{tr}}$ does not need to be $[T] \backslash \mathcal{T}_{\text{pr}}$ but we keep it as such to simplify the exposition. Recall that $A \in [D]^{N \times T}$ denotes the treatment assignments to the various units over time. Let $A^{\text{tr}} := A[:, \mathcal{T}_{\text{tr}}]$ and $A^{\text{tr}}_n := A[\mathcal{N}(n), \mathcal{T}_{\text{tr}}]$. Next, we introduce the notion of a "donor set."

Definition 1 (Donors)**.** *For a given unit $n \in [N]$ and counterfactual treatment assignment $\tilde{\mathbf{a}}_{\mathcal{N}(n)} \in [D]^{|\mathcal{N}(n)|}$, we consider $i \in [N] \setminus \{n\}$ a "donor unit" if the following conditions hold:*

1. *$|\mathcal{N}(i)| = |\mathcal{N}(n)|$, i.e., donor unit i has the same number of neighbors as unit n.*
2. *There exists a permutation $\pi_i : [\mathcal{N}(i)] \to [\mathcal{N}(i)]$ such that:*
   (a) *$A[\pi_i(\mathcal{N}(i)), \mathcal{T}_{tr}] = A[\mathcal{N}(n), \mathcal{T}_{tr}]$, i.e., the training treatment assignment of donor unit i and its neighbors match that of unit n and its neighbors, once permuted by π_i.*
   (b) *$\mathbf{a}^{pr}_{\pi_i(\mathcal{N}(i))} = \tilde{\mathbf{a}}_{\mathcal{N}(n)}$, i.e., the prediction treatment assignment of donor unit i and its neighbors matches the target counterfactual treatment assignment $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$, once permuted by π_i.*

For the remainder of this work, we fix the unit n and counterfactual treatments $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest and let $\mathcal{I}^{(n)} \subset [N] \setminus \{n\}$ denote the corresponding set of donors. One can think of the donor set $\mathcal{I}^{(n)}$ as units whose *observed* outcomes can be used to estimate the unobserved *potential* outcome of unit n under the counterfactual treatments of interest.

3.2 NSI Estimation procedure

Recall that $Z \in \mathbb{R}^{T \times N}$ denotes the matrix of observations. We define $\mathbf{z}_{\text{tr},n} := Z[\mathcal{T}_{\text{tr}}, n] \in \mathbb{R}^{T_{\text{tr}}}$, $Z_{\text{tr},\mathcal{I}^{(n)}} := Z[\mathcal{T}_{\text{tr}}, \mathcal{I}^{(n)}] \in \mathbb{R}^{T_{\text{tr}} \times |\mathcal{I}^{(n)}|}$, and $Z_{\text{pr},\mathcal{I}^{(n)}} := Z[\mathcal{T}_{\text{pr}}, \mathcal{I}^{(n)}] \in \mathbb{R}^{T_{\text{pr}} \times |\mathcal{I}^{(n)}|}$. NSI takes in one hyperparameter $\kappa \in [\min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)]$ and proceeds in two steps, as follows.

1. Point estimate. Let $\{(\hat{s}_\ell, \hat{\boldsymbol{\mu}}_\ell, \hat{\boldsymbol{\nu}}_\ell)\}_{\ell=1}^{\min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)}$ denote the set of singular values, left singular vectors, and right singular vectors for the observed matrix $Z_{\text{tr},\mathcal{I}^{(n)}}$, where $\hat{s}_1 \geq \hat{s}_2 \geq \ldots \geq \hat{s}_{\min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)} \geq 0$. The NSI estimator produced a point estimate as follows:

$$\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\text{pr}}} \mathbf{1}^T Z_{\text{pr},\mathcal{I}^{(n)}} \; \hat{\mathbb{E}}[Z_{\text{tr},\mathcal{I}^{(n)}} | LF, A]^+ \mathbf{z}_{\text{tr},n}. \tag{7}$$

For the given hyperparameter κ, $\hat{\mathbb{E}}[Z_{\text{tr},\mathcal{I}^{(n)}} | LF, A]^+ = \sum_{\ell=1}^{\kappa} \frac{1}{\hat{s}_\ell} \hat{\boldsymbol{\nu}}_\ell \hat{\boldsymbol{\mu}}_\ell^\top$, can be viewed as an estimate of $\mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}} | LF, A]$ that is obtained via hard singular value thresholding, where only the top κ components are preserved. This estimator, which begins with singular value thresholding, has been shown to be equivalent to principal component regression (Agarwal et al. 2021b, 2020a).

2. Confidence interval. Let $\hat{\boldsymbol{\alpha}} = \hat{\mathbb{E}}[Z_{\text{tr},\mathcal{I}^{(n)}} | LF, A]^+ \mathbf{z}_{\text{tr},n}$. Then, the CI-percent confidence interval can be constructed as:

$$\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) \in \left[\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) \pm \frac{\Phi^{-1}(\text{CI}/100) \hat{\sigma} \, \|\hat{\boldsymbol{\alpha}}\|_2}{\sqrt{T_{\text{pr}}}} \right],$$

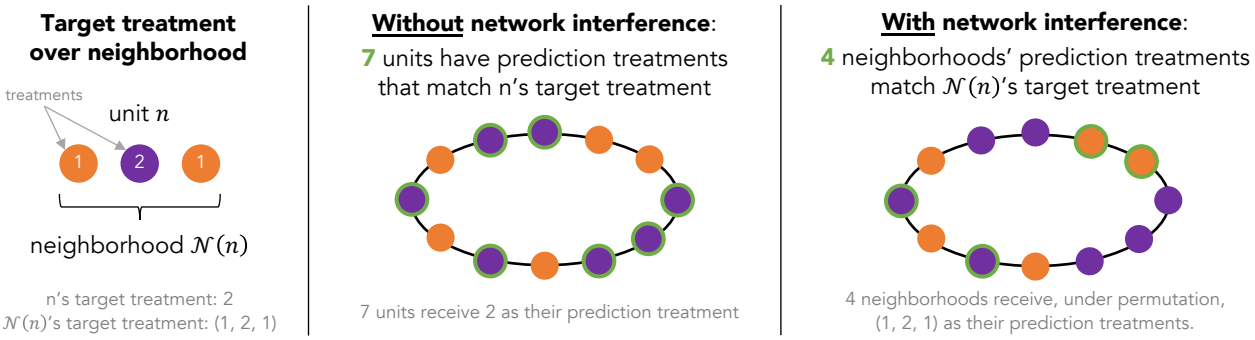


Figure 4: One of the main differences between the NSI estimator and previous estimators (such as SC and SI) is the choice of donors. To illustrate this point, consider a unit n whose target treatment is 2, as given in the left panel. Suppose that $\mathcal{G}$ is a ring graph and the prediction treatments are assigned as given in the middle and right panels. Then, under SC and SI, there are 7 units (with green borders) whose prediction treatments match unit n's target treatment (middle panel). Under NSI, however, the donor requirements are stricter. Specifically, NSI looks at the target treatment $(1, 2, 1)$ across all neighbors $\mathcal{N}(n)$, as given in the left panel. The only units j that could be considered as potential donors must have neighborhood $\mathcal{N}(j)$ that receive prediction treatments $(1, 2, 1)$ subject to permutation (recall Definition 1). As shown on the right, only 4 units (with green borders) meet this requirement.

where Φ denotes the cumulative distribution function (CDF) of the standard normal distribution, Φ^{-1} is the inverse CDF, and

$$\hat{\sigma}^2 = \frac{1}{T_{\text{tr}}} \left\| \mathbf{z}_{\text{tr},n} - Z_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2,$$

which can be interpreted as the in-sample prediction error of the NSI estimator.

3.3 Discussion of NSI

We briefly provide intuition for the NSI estimator, then compare it to the traditional Synthetic Control (Abadie 2021) and Synthetic Interventions (Agarwal et al. 2020b) estimators.

NSI linearly combines donor outcomes. NSI begins by finding units, called "donors," whose outcomes can be used to estimate the potential outcomes of unit n. In Section 4, under suitable assumptions, we establish that the expected potential outcome of unit n can be expressed as a linear combination of the expected outcomes of the donor units, i.e.,

$$\mathbb{E}[Y_{t,n}^{(\tilde{\mathbf{a}})}] = \sum_{j \in \mathcal{I}^{(n)}} \alpha_j \cdot \mathbb{E}[Y_{t,j}^{(\mathbf{a})}], \tag{8}$$

where $\alpha_j \in \mathbb{R}$ (note that α_j can be negative). NSI can be viewed as a method for estimating the coefficients $\{\alpha_j\}$. More precisely, recall that $\hat{\boldsymbol{\alpha}} = \hat{\mathbb{E}}[Z_{\text{tr},\mathcal{I}^{(n)}} | LF, A]^+ \mathbf{z}_{\text{tr},n}$. Then, the NSI estimator (7) can be rewritten as

$$\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \sum_{j \in \mathcal{I}^{(n)}} \hat{\alpha}_j Z[t, j],$$

which is precisely what would follow from expressing the target causal estimand (6) using (8).

Comparing NSI to SI & SC: Choosing donors appropriately. NSI is a generalization of SI and SC. The key difference between SC/SI and NSI is the choice of donor units. In SC/SI, a valid donor unit only needs to undergo the same training and prediction treatments as the training and target prediction treatments of n. In NSI, there are more stringent requirements on donors, as given by Definition 1 and illustrated in Figure 4. These more stringent requirements on how donors are chosen are how NSI removes the bias that SI and SC suffer from when there is spillover.

Under the appropriate choice of donors, the way that the linear model (8) is learned can depend on modeling assumptions. NSI uses principal component regression (PCR), which is motivated by Agarwal et al. (2020b)). However, other estimators, such as convex regression (Abadie 2021) and variants thereof can also be used. In Section 4, we detail the conditions under which PCR produces consistent and asymptotically normal estimates.

4 Formal results

In this section, we provide formal results for the NSI estimator. We characterize conditions under which $\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ can be identified. We then establish that NSI provides a consistent estimate of $\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ and its estimation error is asymptotically normal, justifying the confidence interval given in Step 3 of Section 3.2. As before, we restrict our attention to a specific unit n and target counterfactual treatment assignment $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$. All proofs are given in Appendices B-C.

4.1 Identification Result

We now discuss the key assumptions we make about the intervention assignments A. We begin with an assumption on the treatment assignment.

Assumption 3 (Conditional exogeneity)**.** *For all $n \in [N]$, $t \in [T]$, and $\mathbf{a} \in [D]^N$, we have that $Y_{n,t}^{(\mathbf{a}_{\mathcal{N}(n)})} \perp A \,|\, LF$.*

Given Assumption 2, this conditional independence is equivalent to assuming that $\epsilon_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})} \perp A \,|\, LF$. Similar conditions of "selection on latent factors" have been considered in the literature (see (Agarwal et al. 2020b) and discussion therein). In this work, we analogously require "selection on network-adjusted latent factors." We make two additional assumptions, as follows.

Assumption 4 (Linear span inclusion)**.** *Given a unit $n \in [N]$ and counterfactual treatments $\tilde{\mathbf{a}}_{\mathcal{N}(n)} \in [D]^{|\mathcal{N}(n)|}$ of interest, consider the donor set $\mathcal{I}^{(n)}$. We assume the treatment assignment A is such that $\mathcal{I}^{(n)}$ is non-empty and that $\tilde{\mathbf{u}}_{n,\mathcal{N}(n)}$ lies in the linear span of $\{\tilde{\mathbf{u}}_{i,\pi_i(\mathcal{N}(i))}\}_{i\in\mathcal{I}^{(n)}}$, where $\{\pi_i\}_{i\in\mathcal{I}^{(n)}}$ is defined in Definition 1. That is, there exists $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{I}^{(n)}|}$ such that*

$$\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} = \sum_{i\in\mathcal{I}^{(n)}} \lambda_i \tilde{\mathbf{u}}_{i,\pi_k(\mathcal{N}(i))}.$$

Assumption 5 (Subspace inclusion)**.** *Assume that the rowspace of $\mathbb{E}\big[Z_{pr,\mathcal{I}^{(n)}} \big| LF, A\big]$ lies within the rowspace of $\mathbb{E}\big[Z_{tr,\mathcal{I}^{(n)}} \big| LF, A\big]$.*

We discuss Assumptions 4-5 in Sections 4.3 and 5. We show that NSI's confidence interval indicates the degree to which Assumption 4 holds, and we provide a way to test for Assumption 5 in Section 5.

Before stating our identification result, we first recall identifiability for an arbitrary estimation problem of interest. In the definition below, $P_\theta \in \mathcal{P}$ refers to data distribution under model parameters θ, i.e., P_θ describes how the data behaves under model θ.

Definition 2 (Identifiability). *Let $\theta \in \Theta$ denote the ground-truth model parameters and $\mathcal{P} = \{P_{\theta'} : \theta' \in \Theta\}$ denote the set of possible data distributions. Let $f : \Theta \to \mathbb{R}$ denote the estimand of interest and $P_\theta \in \mathcal{P}$ denote the data generating distribution parameterized by θ. Then, $f(\theta)$ is* identifiable *if there exists a function $g : \mathcal{P} \to \mathbb{R}$ such that $f(\theta) = g(P_\theta)$, i.e., the estimand can be written as a function of the data distribution.*

Identifiability implies that if $f(\theta) \neq f(\theta')$, then $g(P_\theta) \neq g(P_{\theta'})$; otherwise, $f(\theta) = g(P_\theta) = g(P_{\theta'}) = f(\theta')$. In other words, the estimand is identifiable if we can compute it *exactly* when given access to the full data distribution, which is necessary for estimation from (noisy) data to be possible.

In our setting, $\theta = LF$ denotes the latent factors and $f(\theta) = \text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ is the target causal estimand, which we note can be written solely as a function of the latent factors. Recall that the quantities $(A, \mathcal{G}, \mathcal{T}_{\text{tr}}, \mathcal{T}_{\text{pr}})$ are observed and known. Let P_θ denote the joint distribution over the matrices of observed outcomes $Z[\mathcal{T}_{\text{tr}}, :]$ and $Z[\mathcal{T}_{\text{pr}}, :]$. Note that, by Assumption 2, the distribution of $Z[\mathcal{T}_{\text{tr}}, :]$ and $Z[\mathcal{T}_{\text{pr}}, :]$ is given by the latent factors LF, the treatment assignment A, and the random variables $\epsilon_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})}$. We now show that $\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ is identifiable under (1) and Assumptions 2-5.

Theorem 1 (Identification). *If Assumptions 1-5 hold, then*

$$IPO(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{pr}} \mathbf{1}_{T_{pr}}^T \, \mathbb{E}[Z_{pr,\mathcal{I}^{(n)}} \mid LF, A] \, \mathbb{E}[Z_{tr,\mathcal{I}^{(n)}} | LF, A]^+ \, \mathbb{E}[\mathbf{z}_{tr,n} | LF, A], \tag{9}$$

where $\mathbf{1}_{T_{pr}}$ is the all ones vector of length T_{pr}, and the set $\mathcal{I}^{(n)}$ is defined in Definition 1. This implies that $IPO(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ is identifiable by Definition 2.

Under Definition 2, g is given by (9). Note only the first moments of $P_\theta = P_{LF}$ are needed. NSI estimates $\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ by replacing the expectations in (9) with the corresponding empirically observed quantities and smoothing out the pseudoinverse using hard singular value thresholding as given by (7). For the purposes of the analysis, we denote

$$\boldsymbol{\alpha} = \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}} | LF, A]^+ \mathbb{E}[\mathbf{z}_{\text{tr},n} | LF, A]. \tag{10}$$

4.2 Consistency and asymptotic normality

Next, we give conditions under which the NSI estimator achieves finite-sample consistency and asymptotic normality. Let $r_{\text{tr}} \in [r|\mathcal{N}(n)|]$ be the rank of $\mathbb{E}\big[Z_{\text{tr},\mathcal{I}^{(n)}} \big| LF, A\big]$, $s_1 \geq \ldots \geq s_{r_{\text{tr}}} \geq 0$ denote its singular values, and $R_{\text{tr}} \in \mathbb{R}^{|\mathcal{I}^{(n)}| \times r_{\text{tr}}}$ denote its right singular vectors.

Assumption 6 (Sub-Gaussian noise). *Conditioned on LF, we assume that, for all $i \in [N]$, $t \in [T]$, and $\mathbf{a} \in [D]^N$, $\epsilon_{t,i}^{(\mathbf{a}_{\mathcal{N}(i)})}$ are independent, sub-Gaussian random variables with $Var(\epsilon_{t,i}^{(\mathbf{a}_{\mathcal{N}(i)})} | LF) = \sigma^2$ and that $\big\|\epsilon_{t,i}^{(\mathbf{a}_{\mathcal{N}(i)})} \big| LF\big\|_{\psi_2} \leq \xi\sigma$ for some constant $\xi > 0$.*

Assumption 7 (Boundedness). *We assume that $\mathbb{E}\big[Y_{t,i}^{(\mathbf{a}_{\mathcal{N}(j)})} \big| LF\big] \in [-1, 1]$ for all $i \in [N]$, $t \in [T]$.*

Assumption 8 (Well-balanced spectrum). *For parameters $\xi', \xi'' > 0$, we assume $s_{r_{tr}}/s_1 \geq \xi'$ and $\big\|\mathbb{E}\big[Z_{tr,\mathcal{I}^{(n)}} \big| LF, A\big]\big\|_F^2 \geq \xi'' T_{tr} |\mathcal{I}^{(n)}|$, where $\mathcal{I}^{(n)}$ is defined in Definition 1.*

In Section 6, we prove that in the setting of d-regular graphs and where the latent factors are sampled independently from a Gaussian distribution, the parameters ξ' and ξ'' are inverse polynomials of r and d.

Assumption 9 (Sufficient number of components)**.** *We assume that $\kappa = r_{tr}$, where κ is defined in Section 3 and $r_{tr} \leq r|\mathcal{N}(n)|$.*

The following results establish that NSI is consistent and asymptotically normal.

Theorem 2 (Finite-sample consistency)**.** *Let Assumptions 1-9 hold. Then,*

$$\left|\widehat{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - IPO(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})\right| = O_P\left(\log(T_{tr}|\mathcal{I}^{(n)}|)\left(\frac{r_{tr}^{3/4}}{(\xi''')^{3/2}T_{tr}^{1/4}} + \frac{r_{tr}^2}{(\xi''')^4}\max\left(\frac{1}{\sqrt{T_{tr}}}, \frac{1}{\sqrt{|\mathcal{I}^{(n)}|}}, \frac{\sqrt{|\mathcal{I}^{(n)}|}}{T_{tr}^{3/2}}\right)\right)\right),$$

where $\xi''' = \xi'\sqrt{\xi''}$ and ξ', ξ'' are defined in Assumption 8.

Theorem 2 indicates that, under the stated conditions, NSI is consistent. Specifically, for fixed r, d, and $r_{\text{tr}} \leq r(d+1)$, the estimation error of NSI approaches 0 as the number of training measurements T_{tr} and number of donors $|\mathcal{I}^{(n)}|$ grow if $T_{\text{tr}} = \omega(|\mathcal{I}^{(n)}|^{1/3})$. Importantly, the number of prediction measurements T_{pr} need not grow in order for NSI's estimation error to decay to 0.

Let $\Delta = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$, the estimation error of learning the linear weights that represent the outcomes of the target units in terms of the donor units (see (10)). The following result establishes a general result that as long as Δ is decaying sufficiently quickly for any linear estimator (it does not have to be via principal component regression as we do), the NSI estimator is asymptotically normal. While Theorem 2 allows NSI to produce the point estimate in Step 1 of Section 3, Theorem 3 justifies the confidence interval provided in Step 2.

Theorem 3 (Asymptotic normality)**.** *Suppose Assumptions 1-9 hold. Suppose*

$$\|\Delta\|_2 = o_P\left(\min\left(\frac{\sigma\|\boldsymbol{\alpha}\|_2}{\sqrt{T_{pr}|\mathcal{I}^{(n)}|}}, \sqrt{\frac{\|\boldsymbol{\alpha}\|_2}{\sigma}}\right)\right).$$

Then, conditioned on LF and A,

$$\frac{\sqrt{T_{pr}}}{\sigma\|\boldsymbol{\alpha}\|_2}\left(IPO(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \widehat{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})\right) \xrightarrow{d} \mathcal{N}(0,1),$$

as $T_{tr}, T_{pr}, |\mathcal{I}^{(n)}| \to \infty$. Moreover, the $\hat{\sigma}$ used to construct the NSI confidence interval in Step 3 of Section 3.2 satisfies:

$$|\hat{\sigma}^2 - \sigma^2| = O_p\left(\frac{r_{tr}}{\sqrt{T_{tr}}} + \frac{r_{tr}^2\log(T_{tr}|\mathcal{I}^{(n)}|)}{(\xi''')^4\min(T_{tr}, |\mathcal{I}^{(n)}|)}\right),$$

where $\xi''' = \xi'\sqrt{\xi''}$ and ξ', ξ'' are defined in Assumption 8.

4.3 Assumptions 4, 5, and 9

The key enabling conditions for Theorems 1-3 are Assumptions 4, 5, and 9. In this section, we discuss these assumptions further.

Assumption 4. Recall from Section 3.3 that NSI can be interpreted as linearly combining the potential outcomes of donors under appropriately chosen coefficients, denoted by $\hat{\boldsymbol{\alpha}}$. The key

enabling condition that makes linearly combining donor outcomes valid under our model (1) is Assumption 4. The extent to which Assumption 4 holds can be examined in two ways.

First, recall that $\hat{\sigma}^2 = \frac{1}{T_{\text{tr}}} \left\| \mathbf{z}_{\text{tr},n} - Z_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2$ is a measure of how well NSI's linear fit explains the training data. Recall further that the coefficients $\hat{\boldsymbol{\alpha}}$ are estimates of the coefficients $\boldsymbol{\alpha}$ that are used to combine donor outcomes. As such, $\hat{\sigma}^2$ can be viewed as a statistic for Assumption 4, where a large $\hat{\sigma}^2$ suggests Assumption 4 does not hold. Since NSI's confidence interval scales with $\hat{\sigma}$ (see Step 2 of 3.2), how well Assumption 4 holds is captured by NSI's confidence interval.

Second, since Assumption 4 requires that unit n's $\tilde{\mathbf{u}}$-latent factor is contained in the span of the donors' $\tilde{\mathbf{u}}$-latent factors and $\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} \in \mathbb{R}^{r|\mathcal{N}(n)|}$, at least $r|\mathcal{N}(n)|$ donors are needed. Suppose, for example, that all units' $\tilde{\mathbf{u}}$-latent factors are drawn i.i.d. from a multivariate Gaussian. Then, Assumption 4 holds almost surely if and only if there are at least $r|\mathcal{N}(n)|$ donors (cf. Lemma 10). Given an estimate $\bar{r}$ of r, one can therefore perform a simple sanity check that $|\mathcal{I}^{(n)}| \geq \bar{r}|\mathcal{N}(n)|$.

Assumption 5. This condition ensures that the linear coefficients that NSI learns from the training observations generalize to the prediction task. In Section 5, we provide two validity tests to verify whether Assumption 5 holds, i.e., whether the observations are sufficiently rich such that a generalizable model can be learned. To motivate the need for such tests, below we provide a simple example where Assumption 5 does not hold. Suppose that $D = 2$ (the treatments are binary). Let

$$B^{\text{tr},n} = \Big[\text{Ind}(A[\mathcal{N}(n), \mathcal{T}_{\text{tr}}] = 1), \quad \text{Ind}(A[\mathcal{N}(n), \mathcal{T}_{\text{tr}}] = 2) \Big] \in \{0,1\}^{|\mathcal{N}(n)| \times 2T_{\text{tr}}}.$$

Intuitively, $B_{\text{tr},n}$ is an indicator matrix that tracks the training treatment assignment over $\mathcal{N}(n)$.

Proposition 4. *Suppose Assumption 2 holds. Unless colrank$(B^{tr,n}) = |\mathcal{N}(n)|$, there exist latent factors LF and target treatment assignments under which Assumption 5 cannot hold.*

Proposition 4 shows that the diversity of treatment assignments (as captured by $B^{\text{tr},n}$) affects the feasibility of Assumption 5. We unpack this relationship in detail in the next section. Before doing so, we present a negative example in which Assumption 5 does not hold.

Example 4. *Suppose $\mathbf{a}^t = \mathbf{1}_N$ for all $t \in \mathcal{T}_{tr}$. As such, $B^{tr,n} = [[1]_{|\mathcal{N}(n)| \times T_{tr}}, [0]_{|\mathcal{N}(n)| \times T_{tr}}]$ and colrank$(B^{tr,n}) = 1 < |\mathcal{N}(n)|$. One can show that Assumption 5 does* not *hold unless $\tilde{\mathbf{a}}_{\mathcal{N}(n)} = \mathbf{1}$ or $\mathbf{2}$. The reason Assumption 5 does not hold is that all of n's neighbors have only been observed under the* same *treatment. As such, there is no way for NSI to estimate the potential outcome of n under $\tilde{\mathbf{a}}^\top_{\mathcal{N}(n)} = [2,1,1,\ldots]$, for example, where only the first neighbor is treated. It is impossible for NSI (or any estimator, for that matter) to disentangle the spillover of the first neighbor from that of any other neighbor because the training measurements only contain data in which all neighbors receive the same treatment. The validity tests in Section 5 provide a way of testing for whether the treatment assignment and the observations during the training period are rich enough.*

Assumption 9. This condition requires that the number of components used by NSI, given by κ matches the rank r_{tr} of $\mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]$. Since $\mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]$ is unknown in practice, one must estimate r_{tr}, which can be done by applying an elbow point (or knee point) method to the spectrum of the observed matrix $Z_{\text{tr},\mathcal{I}^{(n)}}$ (Zack et al. 1977, Satopaa et al. 2011). There are other heuristics for setting κ, such as the universal thresholding method given in (Chatterjee 2015). Alternatively, suppose we have an estimate $\bar{r}$ of the model "rank" r, defined in Section 2. By our model (1), r_{tr} is upper bounded by $r|\mathcal{N}(n)|$, which suggests that one can use the heuristic $\kappa = \bar{r}|\mathcal{N}(n)|$. It also suggests that one should always set κ to be at least $|\mathcal{N}(n)|$, assuming that $r \geq 1$.

5 Validity tests

We present two validity tests for Assumption 5, one of the key enabling assumptions of Theorems 2 and 3. The first test can be performed *before* any data is collected. It tests for whether the treatment assignment in the training period is diverse enough relative to the target treatment assignment in the prediction period. The second test can be performed only *after* the data is collected and, as such, is a relatively stronger test. Proofs for this section can be found in Appendix D.

5.1 Validity test #1: Pre-Data Collection

The first test can be run *before* the prediction samples are collected.

TrainingTreatmentTest. This test takes in one hyperparameter $\bar{r} \in \mathbb{N}_{>0}$, which is an estimate of the model "rank" r (see Section 2.2). If one does not have a good estimate, $\bar{r}$ can also be an upper bound on r. In order to run this test, we first define several "masking" matrices. For a given treatment $a \in [D]$, let $B^{\text{tr}}(a) \in \{0,1\}^{N \times T_{\text{tr}}}$ and $\mathbf{b}^{\text{pr}}(a) \in \{0,1\}^{N}$ be defined such that their (i,t)-th elements are $B^{\text{tr}}_{it}(a) = \text{Ind}(A^{\text{tr}}_{it} = a)$ and $\tilde{b}^{\text{pr}}_{i}(a) = \text{Ind}(\tilde{a}_i = a)$. That is, the (i,t)-th entry of $B^{\text{tr}}(a)$ is 1 if and only if unit i at measurement t receives treatment a under the training treatments A^{tr}. Similarly, the i-th entry of $\tilde{\mathbf{b}}^{\text{pr}}(a)$ is 1 if and only if unit i is assigned the target counterfactual treatment a under $\tilde{\mathbf{a}}$. Let B^{tr} and $\tilde{B}^{\text{pr}}$ be the concatenated matrices across different treatments:

$$B^{\text{tr}} = [B^{\text{tr}}(1), B^{\text{tr}}(2), \ldots, B^{\text{tr}}(D)] \in \{0,1\}^{N \times T_{\text{tr}} D},$$
$$\tilde{B}^{\text{pr}} = [\tilde{\mathbf{b}}^{\text{pr}}(1), \tilde{\mathbf{b}}^{\text{pr}}(2), \ldots, \tilde{\mathbf{b}}^{\text{pr}}(D)] \in \{0,1\}^{N \times D}.$$

For the hyperparameter $\bar{r}$, the NSI estimator passes the TrainingTreatmentTest if

1. columnspace$(\tilde{B}^{\text{pr}}[\mathcal{N}(n), :]) \subseteq$ columnspace$(B^{\text{tr}}[\mathcal{N}(n), :])$, and
2. for every $t \in \mathcal{T}_{\text{tr}}$, the treatment $A[\mathcal{N}(n), t]$ is repeated at least $\bar{r}D$ times in training;

otherwise, it fails.

Connecting the TrainingTreatmentTest to Assumption 5. The following result formalizes the relationship between the test and Assumption 5 under a natural data generating process.

Proposition 5. *Suppose Assumption 2 holds and $\bar{r} = r$. Suppose that $u_{k,n,\ell} \overset{i.i.d.}{\sim} p_u$ and $w_{t,a,\ell} \overset{i.i.d.}{\sim} p_w$ for all $k, n \in [N]$, $t \in [T]$, $a \in [D]$, and $\ell \in [r]$, where p_u and p_w are continuous and non-degenerate (i.e., the support of p_u or p_w does not have dimension less than r). If* TrainingTreatmentTest *is passed, Assumption 5 holds almost surely for any n and target treatment $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest.*

As such, TrainingTreatmentTest tests whether Assumption 5 can hold under the training treatments for the given target treatment of interest. Intuitively, it requires that the training treatments are sufficiently diverse relative to $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$. In Section 6, we provide an experiment design that guarantees TrainingTreatmentTest is passed for any n and $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$. Note that the i.i.d. condition in Proposition 5 can be relaxed, as long as the latent factors are always drawn from non-degenerate distributions. The condition on latent factors ensures that there is enough variation across units and time such that we can isolate the role that training treatment assignments plays in Assumption 5 from the role that latent factors play.

5.2 Validity test #2: Post-Data Collection

We now furnish a data-driven check for Assumption 5 that we call the SubspaceInclusionTest. This test can be run only *after* the training and prediction samples are collected as opposed to the TrainingTreatmentTest test, which can be run beforehand.

SubspaceInclusionTest. The test takes in three hyperparameters: κ, κ', and γ. Note that we overload κ (which also appears in Section 3) because both instances refer to an estimate of the rank of $\mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}]$. Similarly, let κ' denote the estimated rank of $\mathbb{E}[Z_{\text{pr},\mathcal{I}^{(n)}}]$, respectively. (Refer to Appendix A for various approaches to selecting parameters κ and κ'.) The third $\gamma \in (0,1)$ is a tunable parameter, where a smaller γ results in a stricter test.

Let $\hat{R}_{\text{tr}} \in \mathbb{R}^{|\mathcal{I}^{(n)}| \times \kappa}$ and $\hat{R}_{\text{pr}} \in \mathbb{R}^{|\mathcal{I}^{(n)}| \times \kappa'}$ denote the matrices constructed from the top κ right singular vectors of $Z_{\text{tr},\mathcal{I}^{(n)}}$ and the top κ' right singular vectors of $Z_{\text{pr},\mathcal{I}^{(n)}}$, respectively. Let

$$\hat{\beta} = \left\| (\mathbb{I}_{|\mathcal{I}^{(n)}|} - \hat{R}_{\text{tr}} \hat{R}_{\text{tr}}^\top) \hat{R}_{\text{pr}} \right\|_F^2.$$

Then, the NSI estimator passes the SubspaceInclusionTest if $\hat{\beta} \leq (1-\gamma)\kappa'$; otherwise, it fails.

SubspaceInclusionTest is a data-driven check for Assumption 5. Let R_{pr} and R_{tr} denote the matrices constructed from the right singular vectors of $\mathbb{E}\big[Z_{\text{tr},\mathcal{I}^{(n)}} \big| LF, A\big]$ and $\mathbb{E}\big[Z_{\text{pr},\mathcal{I}^{(n)}} \big| LF, A\big]$, respectively. Then, Assumption 5 can equivalently be stated as requiring that columnspace$(R_{\text{pr}}) \subseteq$ columnspace(R_{tr}). Although one cannot directly test for Assumption 5 since both $\mathbb{E}\big[Z_{\text{tr},\mathcal{I}^{(n)}} \big| LF, A\big]$ and $\mathbb{E}\big[Z_{\text{pr},\mathcal{I}^{(n)}} \big| LF, A\big]$ are not observable due to noise, we now show that SubspaceInclusionTest is a sample-based test for Assumption 5 using $\hat{R}_{\text{pr}}$ and $\hat{R}_{\text{tr}}$. Recall that SubspaceInclusionTest fails if $\hat{\beta} = \left\| (\mathbb{I} - \hat{R}_{\text{tr}} \hat{R}_{\text{tr}}^\top) \hat{R}_{\text{pr}} \right\|_F^2 \geq (1-\gamma)\kappa'$, where $\hat{R}_{\text{tr}}$ and $\hat{R}_{\text{pr}}$ contain the top κ and κ' right singular vectors of $Z_{\text{tr},\mathcal{I}^{(n)}}$ and $Z_{\text{pr},\mathcal{I}^{(n)}}$, respectively. This can be viewed as a test for Assumption 5 since smaller values of $\hat{\beta}$ indicate the extent to which columnspace$(\hat{R}_{\text{pr}}) \subseteq$ columnspace$(\hat{R}_{\text{tr}})$. Indeed, suppose that $R_{\text{tr}} = \hat{R}_{\text{tr}}$, $R_{\text{pr}} = \hat{R}_{\text{pr}}$, and Assumption 5 holds; then, $\hat{\beta} = 0$. As the span of $\hat{R}_{\text{pr}}$ moves outside of the span of $\hat{R}_{\text{tr}}$, $\hat{\beta}$ increases. Since $\hat{\beta} = \left\| (\mathbb{I} - R_{\text{tr}} R_{\text{tr}}^\top) R_{\text{pr}} \right\|_F^2$ is always upper bounded by r_{pr}, which is estimated by κ', we use the threshold $(1-\gamma)\kappa'$ such that the test fails if $\hat{\beta} \geq (1-\gamma)\kappa'$. A formal analysis of this test remains important future work.

Remark 4. *The equivalence between Assumption 5 and columnspace$(R_{pr}) \subseteq$ columnspace(R_{tr}) implies that* LatentFactorTest *would supersede* TrainingTreatmentTest *if R_{tr} and R_{pr} are known exactly and the prediction samples are already collected. However,* TrainingTreatmentTest *remains useful for two reasons. First, it can be run even before measurements are collected as it only requires the treatment assignment pattern. Second,* LatentFactorTest *requires estimating R_{tr} and R_{pr}, which* TrainingTreatmentTest *does not.*

6 NSI's sample complexity: Experiment design

In this section, we propose an experiment design based on graph coloring, under which we can precisely answer the question of how should T, N scale to enable the estimation of IPO$(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ within $\varepsilon \in (0,1)$? Proofs for this section can be found in Appendix E.

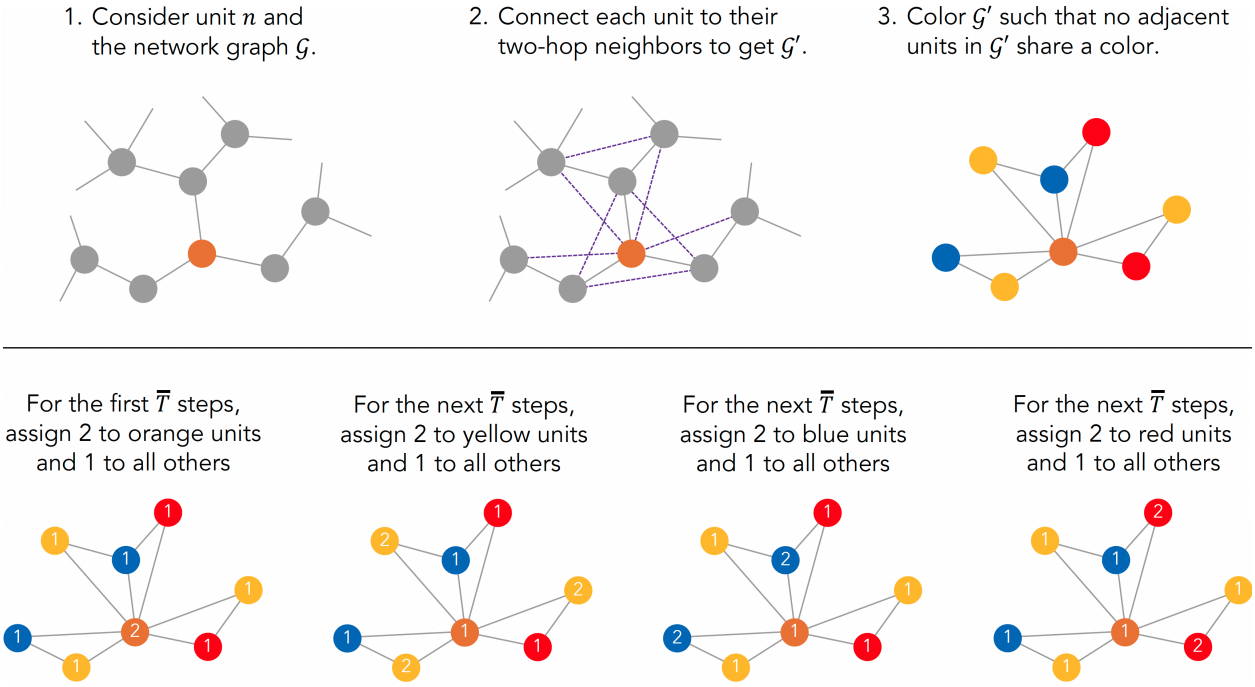


Figure 5: Illustration of experiment design. Consider a unit n and network graph $\mathcal{G}$ (top left). The experiment design generates the two-hop graph $\mathcal{G}'$ by connecting every unit to its two-hop neighbors, which translates to adding edges, as given by the purple dotted lines (top center). Next, color $\mathcal{G}'$ such that no units that are adjacent in $\mathcal{G}'$ share a color (top right). This coloring is used to generate the training treatment assignments. Specifically, consider $D = 2$. Then, during each $\bar{T}$ training measurements, every unit receives the control treatment 1, except units of a specific color.

6.1 Graph coloring-based experiment design

We begin by describing the experiment design. The design assumes access to a subroutine $\textsc{TwoHopColoring}(\mathcal{G})$ that outputs $\textsc{NumColors}, \textsc{Coloring}$ and proceeds in two steps: (1) Construct the graph $\mathcal{G}' = ([N], \mathcal{E}')$ such that $(i,j) \in \mathcal{E} \implies (i,j) \in \mathcal{E}'$ and $(i,j),(j,k) \in \mathcal{E} \implies (i,k) \in \mathcal{E}'$. That is, $\mathcal{G}'$ is constructed by taking $\mathcal{G}$ and adding edges between every node and its two-hop neighbors. (2) Perform a coloring on $\mathcal{G}'$ by labeling the vertices in a graph such that no two adjacent vertices receive the same color, greedily adding colors whenever an existing color cannot be used. (Under a color, vertices of the same color form an independent set of $\mathcal{G}'$.) Let $\textsc{NumColors}$ denote the number of colors required to color $\mathcal{G}'$. Let $\textsc{Coloring} \in [\textsc{NumColors}]^N$ denote the colors assigned to each node (or unit). As before, let $\bar{r} \in \mathbb{N}_{>0}$ denote an estimate (or, alternatively, an upper bound) of the model "rank" r. Then, the experiment design procedure proceeds as follows.

Step 1. Let $\textsc{NumColors}, \textsc{Coloring} = \textsc{TwoHopColoring}(\mathcal{G})$.

Step 2. Divide the colors into $T' = \lceil \frac{\textsc{NumColors}}{D-1} \rceil$ disjoint sets $\{\textsc{Colors}_\ell : \ell = 0, 1, \ldots, T'-1\}$ such that $\textsc{Colors}_1$ contains the first $D-1$ colors, $\textsc{Colors}_2$ contains the next $D-1$ colors, and so on.

Step 3. Then, for $\ell = 0, 1, \ldots, T' - 1$, let $\mathbf{c}^\ell \in [D]^N$ denote a treatment vector such that

$$a_i^\ell = \begin{cases} \textsc{Coloring}_i \bmod (D-1) + 2, & \text{if } \textsc{Coloring}_i \in \textsc{Colors}_\ell, \\ 1, & \text{otherwise.} \end{cases}$$

The intuition behind $\mathbf{c}^\ell$ is as follows. Since $\textsc{Colors}_\ell$ contains at most $D-1$ colors, each color in $\textsc{Colors}_\ell$ can be associated with a different treatment in $\{2, 3, \ldots, D\}$. Units with one of those colors receive the corresponding treatment. That is, any unit i for which $\textsc{Coloring}_i \in \textsc{Colors}_\ell$ is assigned the corresponding treatment in $\{2, 3, \ldots, D\}$. All other units receive treatment 1.

Step 4. Let $\mathcal{T}_{\text{tr}}$ be divided into disjoint sets $\{\mathcal{T}_{\text{tr}}^\ell : \ell = 0, 1, \ldots, T' - 1\}$, each of length $\bar{T} \geq \bar{r}D$ such that $T_{\text{tr}} = \bar{T}T'$. Then, let $A^{\text{tr}} \in [D]^{N \times T_{\text{tr}}}$ be defined such that, $A^{\text{tr}}[:, t] = \mathbf{c}^\ell \quad \forall t \in \mathcal{T}_{\text{tr}}^\ell$ for all $\ell = 0, 1, \ldots, T' - 1$. For the remainder of this work, we assume $\bar{T} = \bar{r}D$ unless otherwise stated.

Step 5. Let the prediction treatment of each unit be assigned i.i.d. uniformly at random from the D possible treatments, i.e., $a_i^{\text{pr}} \overset{\text{i.i.d.}}{\sim} \text{Unif}(D)$ for all $i \in [N]$.

Discussion of experiment design. The experiment design described above uses the graph coloring over the two-hop version of $\mathcal{G}$ to assign treatments. The training measurements are divided into T' disjoint sets—that we will call "periods"—denoted by $\mathcal{T}_{\text{tr}}^\ell$ for $\ell = 0, 1, \ldots, T' - 1$. The treatment assignment during each period remains constant, i.e., for any ℓ, $\mathbf{a}^t = \mathbf{c}^\ell$ for all $t \in \mathcal{T}_{\text{tr}}^\ell$.

The experiment design ensures several important properties hold. First, all nodes that receive the same color also receive the same treatment at any $t \in \mathcal{T}_{\text{tr}}$. Second, nodes that receive the same color have to be at least three hops away from one another, which ensures that, for any neighborhood $\mathcal{N}(n)$ of an arbitrary node n, no two nodes receive the same non-control treatment at any given $t \in \mathcal{T}_{\text{tr}}$. Third, each node receives the control treatment 1 for every period, except for one period during which it receives a non-control treatment. The non-control treatment that a node receives is given by $\textsc{Coloring}_i \bmod (D-1) + 2$, where $\textsc{Coloring}_i$ denotes the color assigned to unit i. Lastly, each period has length $\bar{T} \geq \bar{r}D$. We show that these properties are important to proving Lemma 6 in the next section. See Jensen and Toft (2011) for further information on graph colorings.

6.2 Theoretical guarantees on experiment design

We present two results. The first establishes that the graph-theoretic experiment design described in Section 6.1 guarantees that TrainingTreatmentTest is passed.

Lemma 6. *Suppose Assumption 2 holds. If the training treatments A^{tr} are assigned as in Section 6.1, then* TrainingTreatmentTest *is passed for any unit n and treatments $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$.*

Importantly, Lemma 6 gives a guarantee for *any* unit n and counterfactual treatments $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest, i.e., for all ND^d possible estimands of interest. That is, the experiment design in Section 6.1 can be used to ensure that TrainingTreatmentTest is passed for any target estimand of interest. Moreover, the experiment design can be applied to any graph of interest (and any valid two-hop coloring, as described in Section 6.1). In Appendix E.3, we discuss how one can adapt our experiment design to be less stringent if one is interested in a *specific* choice of n and $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$.

The next result shows that the number of training measurements required under the graph-theoretic experiment design is $O(d^2)$, where d denotes the maximum degree of $\mathcal{G}$.

Lemma 7. *The number of training measurements required by the experimental procedure in Section 6.1 is* $T_{tr} = \bar{r}D\lceil \frac{\textsc{NumColors}}{D-1} \rceil$, *where* $\textsc{NumColors} \leq Degree(\mathcal{G}') + 1$. *As a result,* $T_{tr} \leq \frac{\bar{r}D(d^2+D)}{D-1}$.

By Lemma 7, the experiment design requires $T_{\text{tr}} = O(d^2)$ training measurements. Since $r|\mathcal{N}(n)| = rd$ donors are needed for a given n and $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest (as discussed in Section 4.3), this result shows that one needs $O(d^3)$ training samples under the experiment design to guarantee that TrainingTreatmentTest passes for a given unit and target treatment of interest. That is, with $O(d^3)$ training samples, there is enough variation in the training data such that it is possible to generalize to the training data to a given target counterfactual treatment of interest.

6.3 Data requirement: Generative example

In this section, we examine how much data is needed for the NSI estimator to get within ε accuracy, using regular graphs as an illustrative setting.

Assumption 10. *$\mathcal{G}$ is a d-regular graph.*

Assumption 11. *Assume that each* $u_{j,i,k} \overset{i.i.d.}{\sim} Unif(-\frac{1}{\sqrt{rd}}, \frac{1}{\sqrt{rd}})$ *for all* $j, i \in [N]$ *and* $k \in [r]$. *Further, each* $w_{t,a,k} \overset{i.i.d.}{\sim} Unif(-\frac{1}{\sqrt{rd}}, \frac{1}{\sqrt{rd}})$ *for all* $t \in [T]$, $a \in [D]$, *and* $k \in [r]$.

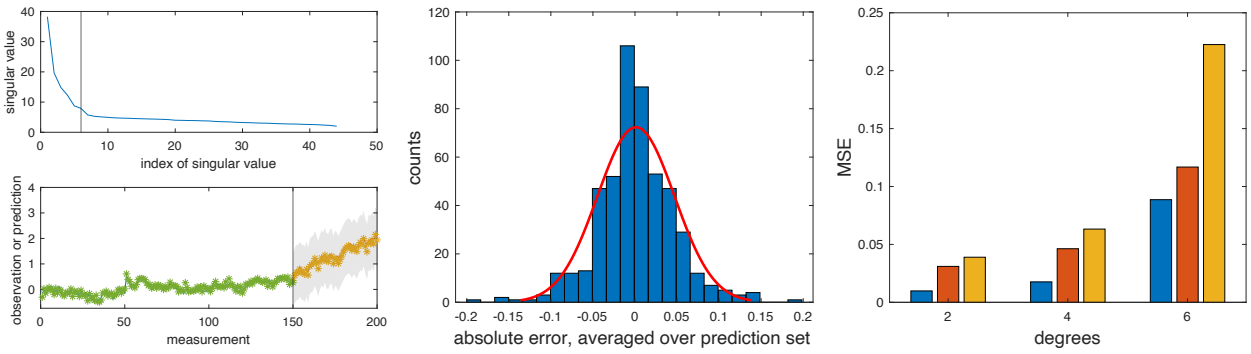
Proposition 8. *Suppose* $\bar{r} = r$. *Suppose Assumptions 2, 3, 6, 7, 9, 10, and 11, hold. Suppose that there are at least* $\frac{rD(d^2+D)}{D-1}$ *training measurements assigned according to the experiment design in Section 6 and* $N = \Omega\left(r^2 d^2 D^{2d+2}\right)$ *units. Then, there is a set of units* E *such that* $|E| = N - \Theta(\sqrt{N})$ *and a method of choosing donors (i.e., units that satisfy Definition 1) such that, for all* $n \in E$,

$$\left|\widehat{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - IPO(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})\right| = O_P\left(r^6 d^{10} \log\left(\frac{T_{tr}N}{D^{d+1}}\right) \max\left(\frac{1}{T_{tr}^{1/4}}, \frac{D^{(d+1)/2}}{N^{1/4}}, \frac{N^{1/4}}{D^{(d+1)/2} T_{tr}^{3/2}}\right)\right),$$

For a d-regular graph, Proposition 8 suggests that to achieve error of order ε with high probability, NSI needs $T_{\text{tr}} = \tilde{\Omega}\left(\frac{r^{24} d^{40}}{\varepsilon^4}\right)$. To understand whether the sample complexity is reasonable, consider a naive alternative. For a given unit n, there are $D^{|\mathcal{N}(n)|}$ possible counterfactual treatments that could be applied to the neighborhood $\mathcal{N}(n)$. Suppose that we do not impose any structure on the potential outcomes, i.e., we do not assume (1). Then, to learn how unit n behaves under every possible neighborhood treatment, one would naively need at least one observation per $D^{|\mathcal{N}(n)|} \leq D^{d+1}$ possible treatments. This naive approach would require $\Omega(D^d)$ samples in order to estimate the potential outcome for a given n and any $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest. Moreover, to achieve an error of ε, at least $\frac{1}{\varepsilon^2}$ samples are needed per treatment, which implies $T_{\text{tr}} = \Omega\left(\frac{D^d}{\varepsilon^2}\right)$ under a naive approach.

7 Simulations

In this section, we present simulation results illustrating the behavior of the NSI estimator and compare it to two related estimators. Let $\mathcal{G}$ be a regular graph with degree d, and let the treatments be binary, i.e., $D = 2$. In each of the experiments below, we indicate the graph degree. Let the training treatments be assigned according to the experiment design in Section 6. Further



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

Figure 6: Simulation results illustrating the performance of the NSI estimator. (a) considers a specific unit n and counterfactual $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest. The top plots the spectrum produced in Step 1 of NSI (see Section 3.2). The bottom visualizes the pointwise estimates produced by NSI for all measurements t in asterisks $*$ with the 95 percent confidence interval in gray. The ground truth potential outcomes are given by the solid lines. (b) plots the residuals ($\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$) across 500 simulations. (c) plots the MSE across different hyperparameters. Each group of bars gives the MSE for regular graphs of degree 2, 4, and 6, as indicated on the x-axis. Within each group of bars, the left (blue) bars are for $N = 1000$, $T_{\text{tr}} = 100$, and $T_{\text{pr}} = 50$; the middle (red) bars for $N = 1000$ and $T_{\text{tr}} = T_{\text{pr}} = 50$; and the right (yellow) bars for $N = 500$ and $T_{\text{tr}} = T_{\text{pr}} = 50$.

experiments and details are given in Appendix F.

Predictions. Note that the NSI estimator (7) can be adapted to produce pointwise estimates

$$\widehat{\mathbb{E}}\left[Y_{t,n}^{(\tilde{\mathbf{a}}_{\mathcal{N}(n)})}\right] = Z[t, \mathcal{I}^{(n)}]\hat{\mathbb{E}}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]^{+}\mathbf{z}_{\text{tr},n},$$

such that $\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\widehat{\mathbb{E}}\left[Y_{t,n}^{(\tilde{\mathbf{a}}_{\mathcal{N}(n)})}\right]$. Under this observation, Figure 6(a) shows an example of the pointwise estimates given by NSI for an example unit n. Consider the bottom plot. The solid line gives the ground truth potential outcomes for unit n across measurements $t \in [200]$. The pointwise estimates produced by NSI are marked by asterisks $*$, with the 95 percent confidence interval in gray. The measurements to the left of the vertical line (i.e., in blue and green) correspond to the training set $\mathcal{T}_{\text{tr}}$ while those to the right (i.e., in red and orange) correspond to the prediction set $\mathcal{T}_{\text{pr}}$. The top plot gives the spectrum $\{\hat{s}_\ell\}_{\ell=1}^{q}$ produced in Step 1 of Section 3.2, where the vertical line marks the hard singular value threshold κ that is used in Step 1. In Figure 6(a), $\mathcal{G}$ is a ring graph ($d = 2$) with $N = 1000$ units, $\epsilon_{\tau,i}^{(\mathbf{c}_{\mathcal{N}(i)})} \sim \mathcal{N}(0, 0.1)$, and $r = 2$.

As shown in the bottom plot of Figure 6(a), the predictions closely match the ground-truth values. As shown on top, 6 components are used to construct the estimates. Since the network-adjusted rank is 6 ($r = 2$ and $|\mathcal{N}(n)| = 3$), the fact that NSI uses 6 components explains why its estimates are fairly accurate. We provide similar plots for other units and target treatments in Appendix F.

Consistency and asymptotic normality. Figure 6(b) verifies that the NSI estimates are consistent and asymptotically normal. Specifically, we let $\mathcal{G}$ be a ring graph (i.e., $d = 2$) with $N = 1000$ units, $\epsilon_{\tau,i}^{(\mathbf{c}_{\mathcal{N}(i)})} \sim \mathcal{N}(0, 0.1)$, and $r = 2$. For each simulation, we randomly generate the

potential outcomes of all units under (1) (see Appendix F for details). We run 500 simulations, then compute the NSI residuals $(\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}))$ for all units $n \in [N]$ and across all possible counterfactual treatments for each n. That is, we use NSI to estimate $\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ for $\tilde{\mathbf{a}}_{\mathcal{N}(n)} = (1, 0, 0)$, $\tilde{\mathbf{a}}_{\mathcal{N}(n)} = (0, 1, 0)$, $\tilde{\mathbf{a}}_{\mathcal{N}(n)} = (1, 1, 0)$, and so on. Figure 6(b) gives a histogram of the NSI residuals. A Gaussian distribution is fit to the residuals and given by the red line.

MSE trends. Figure 6(c) summarizes the performance of NSI across different parameters. The performance is given by the mean-squared error (MSE) across the prediction measurements $\mathcal{T}_{\text{pr}}$, averaged across 50 units. Each group of bars gives the MSE for regular graphs of degree 2, 4, and 6, as indicated on the x-axis. Within each group of bars, the left (blue) bars are for $N = 1000$, $T_{\text{tr}} = 100$, and $T_{\text{pr}} = 50$; the middle (red) bars for $N = 1000$ and $T_{\text{tr}} = T_{\text{pr}} = 50$; and the right (yellow) bars for $N = 500$ and $T_{\text{tr}} = T_{\text{pr}} = 50$. Each bar is the average of 200 simulations with $\epsilon_{\tau,i}^{(\mathbf{c}_{\mathcal{N}(i)})} \sim \mathcal{N}(0, 0.1)$, and $r = 2$. As expected, the MSE increases with degree (because, holding N fixed, a higher degree leads to fewer valid donors), fewer nodes (which also leads to fewer valid donors), and fewer training measurements.

Comparing to other estimators. We also compare the NSI estimator to two others: the SI estimator (Agarwal et al. 2020b) and a baseline estimator. The SI estimator is similar to NSI, but SI assumes that there is no spillover and therefore does not account for network interference. 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 F). We compare the estimators for a ring graph under the same parameters as those used in Figure 6(b) 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, **(0.1174, 0.8735)**, **(0.2310, 0.8149)**, and **(3.398, -2.957)**. Both the NSI and baseline estimators use donor sets that contain, on average, 41 units. The SI estimator uses donor sets with, on average, 166 units. As such, even though the SI estimator has more donors, the performance of NSI is better than that of SI, which is better than that of the baseline estimator.

8 Conclusion and future work

There is rising interest in estimating unit-specific potential outcomes. In this work, we consider the estimation of unit-specific potential outcomes in the presence of spillover, i.e., the treatment assigned to one unit affects the outcome of another unit. We focus on the panel data setting and model spillover as network interference.

As our main contribution, we provide an estimator that we call Network Synthetic Interventions (NSI). In addition to producing point estimates, NSI provides confidence intervals. We show that, under a low-rank latent-factor model and suitable conditions, the NSI estimates are consistent and asymptotically normal. We provide two validity tests that determine whether key conditions hold. We find that obtaining good estimates under spillover requires that the data is rich enough. To this end, we provide an experiment design.

There are many paths for future work. Although the method that we provide comes with strong performance guarantees, it has strict data requirements, as discussed in Section 4.3. One path for future work would be to explore whether these requirements can be relaxed. Second, we explore unit-specific potential outcome estimation. If such fine-grained estimates are not needed, one could examine whether NSI and its data requirements can be improved for coarser estimands of interest. Finally, one compelling path for future work would be to test NSI on real-world datasets.

Acknowledgments

The authors gratefully acknowledge funding from the MIT-IBM project on Causal Representation, the National Science Foundation (NSF) grant CNS-1955997, and the Air Force Research Laboratory (AFOSR) grant FA9550-23-1-0301.

References

Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425.

Agarwal, A., Agarwal, A., and Vijaykumar, S. (2023a). Synthetic Combinations: A Causal Inference Framework for Combinatorial Interventions. *arXiv preprint arXiv:2303.14226*.

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2021a). Causal Matrix Completion. *arXiv preprint arXiv:2109.15154*.

Agarwal, A., Shah, D., and Shen, D. (2020a). On Principal Component Regression in a High-dimensional Error-in-variables Setting. *arXiv preprint arXiv:2010.14449*.

Agarwal, A., Shah, D., and Shen, D. (2020b). Synthetic Interventions. *arXiv preprint arXiv:2006.07691v4*.

Agarwal, A., Shah, D., and Shen, D. (2023b). Synthetic A/B Testing Using Synthetic Interventions. *arXiv preprint arXiv:2006.07691*.

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021b). On Robustness of Principal Component Regression. *Journal of the American Statistical Association*, 116(536):1731–1745.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019). Synthetic Difference in Differences. Technical report, National Bureau of Economic Research.

Aronow, P. M. (2012). A General Method for Detecting Interference Between Units in Randomized Experiments. *Sociological Methods & Research*, 41(1):3–16.

Aronow, P. M. and Samii, C. (2017). Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment.

Aronow, P. M., Samii, C., et al. (2017). Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact P-values for Network Interference. *Journal of the American Statistical Association*, 113(521):230–240.

Auerbach, E. and Tabord-Meehan, M. (2021). The Local Approach to Causal Inference Under Network Interference. Technical report.

Bajari, P., Burdick, B., Imbens, G. W., Masoero, L., McQueen, J., Richardson, T., and Rosen, I. M. (2021). Multiple Randomization Designs. *arXiv preprint arXiv:2112.13495*.

Bargagli-Stoffi, F. J., Tortù, C., and Forastiere, L. (2020). Heterogeneous Treatment and Spillover Effects Under Clustered Network Interference. Technical report.

Basse, G. W. and Airoldi, E. M. (2018a). Limitations of Design-based Causal Inference and A/B Testing Under Arbitrary and Network Interference. *Sociological Methodology*, 48(1):136–151.

Basse, G. W. and Airoldi, E. M. (2018b). Model-assisted Design of Experiments in the Presence of Network-correlated Outcomes. *Biometrika*, 105(4):849–858.

Belloni, A., Fang, F., and Volfovsky, A. (2022). Neighborhood Adaptive Estimators for Causal Inference Under Network Interference. *arXiv preprint arXiv:2212.03683*.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-differences Estimates? *The Quarterly journal of economics*, 119(1):249–275.

Bhattacharya, R., Malinsky, D., and Shpitser, I. (2020). Causal Inference Under Interference And Network Uncertainty. In Adams, R. P. and Gogate, V., editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 1028–1038. PMLR.

Bowers, J., Fredrickson, M. M., and Panagopoulos, C. (2013). Reasoning about Interference Between Units: A General Framework. *Political Analysis*, 21(1):97–124.

Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social Networks and the Decision to Insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Chatterjee, S. (2015). Matrix Estimation by Universal Singular Value Thresholding. *The Annals of Statistics*, 43(1):177–214.

Chin, A. (2019). Regression Adjustments for Estimating the Global Treatment Effect in Experiments with Interference. *Journal of Causal Inference*, 7(2).

Cortez, M., Eichhorn, M., and Yu, C. L. (2022a). Exploiting Neighborhood Interference with Low Order Interactions Under Unit Randomized Design. *arXiv preprint arXiv:2208.05553*.

Cortez, M., Eichhorn, M., and Yu, C. L. (2022b). Staggered Rollout Designs Enable Causal Inference Under Interference Without Network Knowledge. *arXiv preprint arXiv:2205.14552*.

De Paula, A. (2017). Econometrics of Network Models. In *Advances in economics and econometrics: Theory and applications, eleventh world congress*, pages 268–323. Cambridge University Press Cambridge.

De Paula, A., Rasul, I., and Souza, P. (2018). Recovering Social Networks from Panel Data: Identification, Simulations and an Application.

De Paula, Á., Rasul, I., and Souza, P. (2019). Identifying Network Ties from Panel Data: Theory and an Application to Tax Competition. *arXiv preprint arXiv:1910.07452*.

DiTraglia, F. J., Garcia-Jimeno, C., O'Keeffe-O'Donovan, R., and Sanchez-Becerra, A. (2020). Identifying Causal Effects in Experiments with Spillovers and Non-compliance. *arXiv preprint arXiv:2011.07051*.

Eckles, D., Karrer, B., and Ugander, J. (2017). Design and Analysis of Experiments in Networks: Reducing Bias from Interference. *Journal of Causal Inference*, 5(1).

Forastiere, L., Airoldi, E. M., and Mealli, F. (2021). Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks. *Journal of the American Statistical Association*, 116(534):901–918.

Gui, H., Xu, Y., Bhasin, A., and Han, J. (2015). Network A/B Testing: From Sampling to Estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pages 399–409. International World Wide Web Conferences Steering Committee.

Jagadeesan, R., Pillai, N. S., and Volfovsky, A. (2020). Designs for Estimating the Treatment Effect in Networks with Interference. *The Annals of Statistics*, 48(2):679–712.

Jensen, T. R. and Toft, B. (2011). *Graph Coloring Problems*. John Wiley & Sons.

Johari, R., Li, H., Liskovich, I., and Weintraub, G. Y. (2022). Experimental Design in Two-sided Platforms: An Analysis of Bias. *Management Science*.

Karwa, V. and Airoldi, E. M. (2018). A Systematic Investigation of Classical Causal Inference Strategies Under Mis-specification Due to Network Interference. Technical report.

Leung, M. P. (2019). Causal Inference Under Approximate Neighborhood Interference. Technical report.

Li, W., Sussman, D. L., and Kolaczyk, E. D. (2021). Causal Inference Under Network Interference with Noise. Technical report.

Liu, L., Hudgens, M. G., and Becker-Dreps, S. (2016). On Inverse Probability-weighted Estimators in the Presence of Interference. *Biometrika*, 103(4):829–842.

Ma, Y. and Tresp, V. (2021). Causal Inference Under Networked Interference and Intervention Policy Enhancement. In *International Conference on Artificial Intelligence and Statistics*, pages 3700–3708. PMLR.

Manski, C. F. (2013). Identification of Treatment Response with Social Interactions. *The Econometrics Journal*, 16(1).

Matoušek, J. (2008). On Variants of the Johnson–lindenstrauss Lemma. *Random Structures & Algorithms*, 33(2):142–156.

Ogburn, E. L., Sofrygin, O., Diaz, I., and Van der Laan, M. J. (2017). Causal Inference for Social Network Data. Technical report.

Perez-Heydrich, C., Hudgens, M. G., Halloran, M. E., Clemens, J. D., Ali, M., and Emch, M. E. (2014). Assessing Effects of Cholera Vaccination in the Presence of Interference. *Biometrics*, 70(3):731–741.

Pouget-Abadie, J., Saveski, M., Saint-Jacques, G., Duan, W., Xu, Y., Ghosh, S., and Airoldi, E. M. (2017). Testing for Arbitrary Interference on Experimentation Platforms. Technical report.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a "kneedle"' in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Saveski, M., Pouget-Abadie, J., Saint-Jacques, G., Duan, W., Ghosh, S., Xu, Y., and Airoldi, E. M. (2017). 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.

Sävje, F., Aronow, P. M., and Hudgens, M. G. (2021). Average Treatment Effects in the Presence of Unknown Interference. *The Annals of Statistics*, 49(2):673–701.

Shah, D., Song, D., Xu, Z., and Yang, Y. (2020). Sample Efficient Reinforcement Learning Via Low-rank Matrix Estimation. *arXiv preprint arXiv:2006.06135*.

Sussman, D. L. and Airoldi, E. M. (2017a). Elements of Estimation Theory for Causal Effects in the Presence of Network Interference. Technical report.

Sussman, D. L. and Airoldi, E. M. (2017b). Elements of Estimation Theory for Causal Effects in the Presence of Network Interference. *arXiv preprint arXiv:1702.03578*.

Tchetgen, E. J. T. and VanderWeele, T. J. (2012). On Causal Inference in the Presence of Interference. *Statistical Methods in Medical Research*, 21(1):55–75. PMID: 21068053.

Toulis, P. and Kao, E. (2013). Estimation of Causal Peer Influence Effects. In *International Conference on Machine Learning*, pages 1489–1497.

Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph Cluster Randomization: Network Exposure to Multiple Universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337. ACM.

Vazquez-Bare, G. (2022). Identification and Estimation of Spillover Effects in Randomized Experiments. *Journal of Econometrics*.

Verbitsky-Savitz, N. and Raudenbush, S. W. (2012). Causal Inference Under Interference in Spatial Settings: a Case Study Evaluating Community Policing Program in Chicago. *Epidemiologic Methods*, 1(1):107–130.

Viviano, D. (2020). Experimental Design Under Network Interference. Technical report.

Yu, C. L., Airoldi, E. M., Borgs, C., and Chayes, J. T. (2022). Estimating Total Treatment Effect in Randomized Experiments with Unknown Network Structure. *arXiv preprint arXiv:2205.12803*.

Zack, G. W., Rogers, W. E., and Latt, S. A. (1977). Automatic Measurement of Sister Chromatid Exchange Frequency. *Journal of Histochemistry & Cytochemistry*, 25(7):741–753.

Appendix

A Preliminaries

Let $[X] := \{1, \ldots, X\}$ for any positive integer X. For any treatment vector $\mathbf{a} \in [D]^N$ and some set $S \subseteq [N]$, let $\mathbf{a}_S \in [D]^{|S|}$ denote the vector containing the elements of $\mathbf{a}$ indexed by S. Similarly, let $a_i \in [D]$ denote the i-th element of $\mathbf{a}$. Let $\mathbb{I}_x$ denote the $x \times x$ identity matrix and $\otimes$ denote the Kronecker product. Let $\text{Ind}(\cdot)$ denote the indicator function. Let $\|\cdot\|_{\psi_2}$ denote the Orlicz norm. Let O_p denote 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." Similarly, let o_P denote the probabilistic version of little-o notation. Formally, for any sequence of random variables X_n, $X_n = o_P(1)$ if and only if $X_n \xrightarrow{p} 0$. Let $\tilde{\Omega}$ denote the variation on big-Ω notation that ignores logarithmic terms such that $\Omega(a(n)\log^k(n)) = \tilde{\Omega}(a(n))$. For two sets of indices $S_1 \subseteq [m_1]$ and $S_2 \subseteq [m_2]$ as well as a matrix $\Pi \in \mathbb{R}^{m_1 \times m_2}$, let $\Pi[S_1, S_2] \in \mathbb{R}^{|S_1| \times |S_2|}$ denote the submatrix of Π corresponding to the rows indexed by S_1 and columns index by S_2. We use ":" as a shorthand for the entire set of indices such that $\Pi[:, S_2] \in \mathbb{R}^{m_1 \times |S_2|}$ and $\Pi[S_1, :] \in \mathbb{R}^{|S_1| \times m_2}$. Let $\mathcal{X}^*$ denote the $*$-product space $\mathcal{X} \times \mathcal{X} \times \ldots \times \mathcal{X}$, where the length of the product is not pre-determined. Lastly, let Π^+ denote the pseudo-inverse of Π.

Remark 5. *The estimation procedure for NSI requires the use of a singular value thresholding (SVT) method. Given a list of singular values (also known as a spectrum) $(s_1, s_2, \ldots, s_X)$ where $s_1 \geq s_2 \geq \ldots \geq s_X$, an SVT method determines a threshold $r_{SVT} \leq X$. The singular values $(s_1, s_2, \ldots, s_{r_{SVT}})$ are preserved, and the remaining are discarded. SVT is often used to "de-noise" a matrix using its singular values. That is, one reconstructs a de-noised matrix by keeping the top r_{SVT} components of that matrix and attributing the remaining components to noise. In this way, one can think of r_{SVT} as the matrix's estimated rank.*

There are several popular methods for SVT. One could, for instance, use a universal SVT method, such as that given in (Chatterjee 2015). There are also popular methods for what is known as elbow (or knee) point detection (Zack et al. 1977, Satopaa et al. 2011), which look for the point of maximum curvature along a monotonic curve.

B Helper lemmas

Lemma 9. *Let $X \in \mathbb{R}^{m_1 \times m_2}$ be a random matrix where $X_{i,j} \overset{i.i.d.}{\sim} p_x$ is drawn from a continuous, non-degenerate distribution p_x. Then, rank$(X) = \min(m_1, m_2)$ almost surely.*

Proof. Let $V_j = \text{span}(\{\mathbf{x}_k : k = 1, \ldots, j\})$, where $\mathbf{x}_k$ denotes the k-th column of X. Since V_1 is a one-dimensional subspace and $\mathbf{x}_k$ consists of i.i.d. non-degenerate, continuous random variables,

$$P(\mathbf{x}_2 \in V_1) = 0.$$

In other words, with probability 1, V_2 is a two-dimensional subspace. By induction, V_j is a j-dimensional subspace almost surely as long as $j \leq m_1$. For any $j \geq m_1$, V_j is an m_1-dimensional subspace. Therefore, $rank(X) = \min(m_1, m_2)$ almost surely. □

Lemma 10. *Suppose Assumptions 1, 2, and 11 hold. Then, Assumption 4 holds almost surely if there are at least $r|\mathcal{N}(n)|$ donors.*

Proof. Consider a unit n. Recall that

$$\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} = [\mathbf{u}_{\mathcal{N}_1(n),n}^\top, \mathbf{u}_{\mathcal{N}_2(n),n}^\top, \ldots, \mathbf{u}_{\mathcal{N}_{|\mathcal{N}(n)|}(n),n}^\top]^\top \in \mathbb{R}^{r|\mathcal{N}(n)|}.$$

Further, recall that linear span inclusion (Assumption 4) requires that

$$\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} = \sum_{j \in \mathcal{I}^{(n)}} \lambda_j \tilde{\mathbf{u}}_{j,\pi_j(\mathcal{N}(j))}, \tag{11}$$

for some $\{\lambda_j\}_{j \in \mathcal{I}^{(n)}}$, where π_j is defined in Definition 1. To show that linear span inclusion holds, suppose that we construct a matrix $U = [\tilde{\mathbf{u}}_{j,\pi_j(\mathcal{N}(j))} : j \in \mathcal{I}^{(n)}] \in \mathbb{R}^{r|\mathcal{N}(n)| \times |\mathcal{I}^{(n)}|}$. By Lemma 9, $\text{rank}(U) = \min(r|\mathcal{N}(n)|, |\mathcal{I}^{(n)}|)$ almost surely. Since $|\mathcal{I}^{(n)}| \geq r|\mathcal{N}(n)|$, $\text{rank}(U) = r|\mathcal{N}(n)|$, which implies that (11) and therefore that Assumption 4 holds almost surely. □

Lemma 11. *Consider $\mathcal{G}$. Suppose that $\mathcal{G}$ contains N' disjoint clusters, each of size M. Suppose that every unit is assigned a treatment $a \in [D]$ independently and uniformly at random. Let $\mathbf{s}_i \in [D]^M$ denote the (ordered) sequence of treatments for cluster $i \in [N']$. Let $\mathbf{s}_0$ denote a reference sequence. Let B denote the number of clusters for which the cluster's treatments $\mathbf{s}_i$ match the reference treatments $\mathbf{s}_0$, i.e., $s_i = s_0$. Then,*

$$P\left(B \geq \frac{\chi N'}{D^M}\right) \leq \exp\left(\frac{-2(\chi-1)^2 N'}{D^{2M}}\right),$$

for any $\chi > 1$, and

$$P\left(B \leq \frac{\chi N'}{D^M}\right) \leq \exp\left(\frac{-2(\chi-1)^2 N'}{D^{2M}}\right),$$

for any $\chi < 1$.

Proof. Let there be N' clusters, each of size M. Let $\mathbf{s}_i \in [D]^M$ denote the (ordered) sequence of treatments for cluster $i \in [N']$. Let $\mathbf{s}_0$ denote a reference sequence.

Let $\mathbf{b} = (s_1 = s_0, s_2 = s_0, \ldots, s_{N'} = s_0) \in \{0,1\}^{N'}$. Intuitively, $\mathbf{b}$ is a binary vector, where entry b_i indicates whether s_i matches the reference sequence s_0. Let $B = \sum_{i=1}^{N'} b_i$, i.e., the number of clusters for which the cluster's treatments $\mathbf{s}_i$ match the reference treatments $\mathbf{s}_0$.

Under the setup (in particular, that units are assigned treatments independently and uniformly at random, and the sequences s_i are over disjoint clusters), $\mathbf{b}$ is a sequence of i.i.d. Bernoulli random variables and $\mathbb{E}[B] = \frac{N'}{D^M}$. Then, by Hoeffding's inequality,

$$\begin{aligned} P\left(B \geq \chi \mathbb{E}B\right) &= P\left(B - \mathbb{E}B \geq (\chi-1)\mathbb{E}B\right) \\ &\leq \exp\left(\frac{-2(\chi-1)^2(\mathbb{E}B)^2}{N'}\right) \\ &= \exp\left(\frac{-2(\chi-1)^2 N'}{D^{2M}}\right), \end{aligned}$$

for $\chi > 1$. □

Lemma 12. *Suppose Assumption 10 holds. For every unit $i \in [N]$, fix an ordering over the neighborhood $\mathcal{N}(i)$, and let C_i denote the (ordered) colors assigned to $\mathcal{N}(i)$. We say that a unit i*

is a "coloring donor" for an ego-unit $n \neq i$ if $C_i = C_n$. Then, there are at least $N - \Theta(\sqrt{N})$ units with at least $\sqrt{N}$ coloring donors.

Proof. First, note that every unit has d neighbors by Assumption 10. In addition, it is well known that a coloring of $\mathcal{G}'$ (as defined in Section 6.1) requires at most $d^2 + 1$ colors. Therefore, there are $\rho(d) = (d^2+1)^{d+1}$ ways to color each neighborhood. Let $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_{\rho(d)}$ denote the possible (ordered) colorings and $N_k = |\{i : C_i = \mathcal{O}_k\}|$ denote the number of units for which the (ordered) neighborhood is colored according to $\mathcal{O}_k$. Let $\mathcal{O}^- = \{k : N_k < \sqrt{N}\}$, i.e., $\mathcal{O}^-$ is formed by removing all colorings that match fewer than $\sqrt{N}$ neighborhoods.

Since at most $\rho(d)\sqrt{N}$ units have colors $\mathcal{O}^-$ and all remaining units have at least $\sqrt{N}$ coloring donors by definition of $\mathcal{O}^-$, there are at least $N - \rho(d)\sqrt{N} = N - \Theta(\sqrt{N})$ units with at least $\sqrt{N}$ coloring donors. □

Lemma 13. *Consider two matrices $X_1 \in \mathbb{R}^{m_1 \times m_2}$ and $X_2 \in \mathbb{R}^{m_1' \times m_2}$. Then, rowspace$(X_1) \not\subseteq$ rowspace(X_2) if and only if there exists a vector $\mathbf{v} \neq \mathbf{0}_{m_1}$ such that $X_2 X_1^\top \mathbf{v} = \mathbf{0}_{m_1'}$ and $X_1^\top \mathbf{v} \neq \mathbf{0}_{m_2}$.*

Proof. rowspace$(X_1) \not\subseteq$ rowspace(X_2) if and only if there exists a vector $\mathbf{v} \neq \mathbf{0}_{m_1}$ such that $X_1^\top \mathbf{v} \neq \mathbf{0}_{m_2}$ is in the null space of $X_2^\top$, which is equivalent to $X_2 X_1^\top \mathbf{v} = \mathbf{0}_{m_1'}$. □

Notation and definitions. For Lemmas 14-19, we suppress the conditioning on LF and A. Let $\boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}} = [\epsilon_{t,j}^{(\mathbf{a}^t_{\mathcal{N}(j)})} : j \in \mathcal{I}^{(n)}] \in \mathbb{R}^{|\mathcal{I}^{(n)}|}$. Let $\boldsymbol{\epsilon}_{\mathrm{tr},n} = [\epsilon_{t,n}^{(\mathbf{a}^t_{\mathcal{N}(n)})} : t \in \mathcal{T}_{\mathrm{tr}}] \in \mathbb{R}^{T_{\mathrm{tr}}}$. Let $\Delta = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is defined below Theorem 1. Recall that R_{tr} denotes the matrix containing the right singular vectors of $\mathbb{E}\big[Z_{\mathrm{tr},\mathcal{I}^{(n)}} \big| LF, A\big]$. Let $Z^{r_{\mathrm{tr}}}_{\mathrm{tr},\mathcal{I}^{(n)}} = \sum_{\ell=1}^{r_{\mathrm{tr}}} \hat{s}_\ell \hat{\boldsymbol{\mu}}_\ell \hat{\boldsymbol{\nu}}_\ell^\top = \bar{L}_{\mathrm{tr}} \bar{\Sigma}_{\mathrm{tr}} \bar{R}_{\mathrm{tr}}^\top$, where $\hat{s}_\ell$, $\hat{\boldsymbol{\mu}}_\ell$, and $\hat{\boldsymbol{\nu}}_\ell$ are defined in Section 3.2. Let $\mathcal{P} = R_{\mathrm{tr}} R_{\mathrm{tr}}^\top$ and $\bar{\mathcal{P}} = \bar{R}_{\mathrm{tr}} \bar{R}_{\mathrm{tr}}^\top$. Let $\bar{\mathcal{Q}} = \bar{L}_{\mathrm{tr}} \bar{L}_{\mathrm{tr}}^\top$.

Lemma 14 (Adapted from Agarwal et al. (2023b))**.** *Consider the setup of Theorem 2. Then,*

$$\|\mathcal{P}\Delta\|_2 = O_P\left(\frac{\sqrt{r_{tr}}}{\xi''' T_{tr}^{1/4}\sqrt{|\mathcal{I}^{(n)}|}} + \frac{r_{tr}^{3/2}\sqrt{\log\left(T_{tr}|\mathcal{I}^{(n)}|\right)}}{(\xi''')^{5/2}\sqrt{|\mathcal{I}^{(n)}|}\min\left(\sqrt{T_{tr}}, \sqrt{|\mathcal{I}^{(n)}|}\right)} + \frac{r_{tr}^{2}\sqrt{\log\left(T_{tr}|\mathcal{I}^{(n)}|\right)}}{(\xi''')^{4}\min\left(T_{tr}^{3/2}, |\mathcal{I}^{(n)}|^{3/2}\right)}\right),$$

where ξ''' is defined in Theorem 2 and depends only on r and d. Furthermore,

$$\left\|Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}}\hat{\boldsymbol{\alpha}} - \mathbb{E}\mathbf{z}_{tr,n}\right\|_2^2 \leq \left\|Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}} - \mathbb{E}Z_{tr,\mathcal{I}^{(n)}}\right\|_{2,\infty}^2 \|\boldsymbol{\alpha}\|_1^2 + 2\left\langle Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}}\Delta, \boldsymbol{\epsilon}_{tr,n}\right\rangle.$$

Lemma 15 (Adapted from Agarwal et al. (2023b))**.** *Let x_t be a sequence of independent, zero-mean sub-Gaussian random variables with variance $\bar{\sigma}^2$. Then, $\frac{1}{H}\sum_{t=1}^{H}\gamma_t = O_P(\bar{\sigma}^2/\sqrt{H})$.*

Lemma 16 (Adapted from Agarwal et al. (2023b))**.** *Let Assumptions 2-3 and 6-9 hold. Then,*

$$\left\|Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}} - \mathbb{E}Z_{tr,\mathcal{I}^{(n)}}\right\|_{2,\infty} = O_P\left(\frac{1}{\xi'''}\left(\frac{\sqrt{r_{tr}T_{tr}\log(T_{tr}|\mathcal{I}^{(n)}|)}}{\min(\sqrt{T_{tr}}, |\mathcal{I}^{(n)}|)}\right)\right),$$

where ξ''' is defined in Theorem 2 and depends only on r and d.

Lemma 17 (Adapted from Agarwal et al. (2023b))**.** *Let Assumptions 2-8 hold. Then, $Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}}\hat{\boldsymbol{\alpha}} = \bar{\mathcal{Q}}(\mathbb{E}\mathbf{z}_{tr,n} + \boldsymbol{\epsilon}_{tr,n})$ and, for any $\chi > 0$,*

$$P\left(\left\langle\bar{\mathcal{Q}}\boldsymbol{\epsilon}_{tr,n}, \boldsymbol{\epsilon}_{tr,n}\right\rangle \geq \sigma^2 r_{tr} + \chi\right) \leq \exp\left(-\bar{\xi}\left(\frac{\chi^2}{\sigma^4 r_{tr}}, \frac{\chi}{\sigma^2}\right)\right),$$

for some universal constant $\bar{\xi} > 0$. *Moreover, given* $Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}}$,

$$\left\langle Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}}\Delta, \boldsymbol{\epsilon}_{tr,n}\right\rangle = O_P\left(r_{tr} + \sqrt{T_{tr}} + \left\|Z^{r_{tr}}_{tr,\mathcal{I}^{(n)}} - \mathbb{E}Z_{tr,\mathcal{I}^{(n)}}\right\|_{2,\infty}\|\boldsymbol{\alpha}\|_1\right),$$

with respect to the randomness in $\boldsymbol{\epsilon}_{tr,n}$.

Lemma 18 (Adapted from Agarwal et al. (2023b))**.** *Let* $\mathbf{x} \in \mathbb{R}^m$ *be a random variable with independent, zero-mean sub-Gaussian random coordinates with* $\|x_i\|_{\psi_2} \leq K$ *for every* $i \in [m]$. *Let* $\mathbf{x}' \in \mathbb{R}^m$ *be another random variable that satisfies* $\|\mathbf{x}'\|_2 \leq K'$. *Then, for any* $\chi \geq 0$,

$$P\left(\left|\sum_{i=1}^m x_i' x_i\right| \geq \chi\right) \leq 2\exp\left(-\frac{\bar{\xi}\chi^2}{(KK')^2}\right).$$

Lemma 19 (Adapted from Agarwal et al. (2023b))**.** *For a given unit* $n \in [N]$ *and counterfactual treatment* $\tilde{\mathbf{a}}$ *of interest, suppose Assumptions 2-9 hold. Then, conditioned on* LF *and* A,

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2 = O_P\left(\frac{\sqrt{\log(T_{tr}|\mathcal{I}^{(n)}|)}}{(\xi''')^{3/2}}\left(\frac{r_{tr}^{3/4}}{T_{tr}^{1/4}|\mathcal{I}^{(n)}|^{1/2}} + \frac{r_{tr}^{3/2}}{(\xi''')^{3/2}\min(\sqrt{T_{tr}}, \sqrt{|\mathcal{I}^{(n)}|})}\right)\right).$$

Remark 6. *Lemmas 14-19 are adapted from Agarwal et al. (2020b). One can check that the assumptions for these lemmas hold in our setting. In particular, the main difference between the assumptions in our work and in Agarwal et al. (2020b) is the definition of the "donor set."*

To see how this affects the assumptions, first note that observation pattern in this work allows for any sequence of treatments during the training period (referred to as the "pre-intervention" period in Agarwal et al. (2020b)). In contrast, in Agarwal et al. (2020b), the treatment must be constant across the training measurements, and it is assumed that all units are under treatment 0 *during the pre-intervention (i.e., training) period. This difference is reflected in the assumptions via the donor set. Once we adjust the choice of donor set (Definition 1) to suit the network interference setting, the assumptions in Agarwal et al. (2020b) can be mapped to ours.*

Second, note that the model in (Agarwal et al. 2020b) is given by (in their notation)

$$Y^{(d)}_{tn} = \left\langle u^{(d)}_t, v_n\right\rangle + \varepsilon^{(d)}_{tn}, \tag{12}$$

where $u^{(d)}_t, v_n \in \mathbb{R}^r$ *are latent factors;* $\varepsilon^{(d)}_{tn}$ *is a zero-mean, independent noise term; and* $Y^{(d)}_{tn}$ *is the potential outcome of interest. Recall from* (2) *that our model is given by (in our notation)*

$$Y^{(\mathbf{a}_{\mathcal{N}(n)})}_{t,n} = \left\langle \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}, \tilde{\mathbf{w}}_{t,\mathbf{a}_{\mathcal{N}(n)}}\right\rangle + \epsilon^{(\mathbf{a}_{\mathcal{N}(n)})}_{t,n}, \tag{13}$$

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;* $\epsilon^{(\mathbf{a}_{\mathcal{N}(n)})}_{t,n}$ *is a zero-mean, independent noise term; and* $Y^{(\mathbf{a}_{\mathcal{N}(n)})}_{t,n}$ *is the potential outcome of interest. As such, our setup model is analogous to the model used by Agarwal et al. (2020b), with a change of notation.*

We now go through the assumptions one-by-one. As we saw above, Assumption 2 is equivalent to Assumption 2 in (Agarwal et al. 2020b), with a change of notation. Furthermore, Assumption 1 is automatically satisfied when Assumption 2 holds. Assumptions 2 and 3 together give Assumption 3 of (Agarwal et al. 2020b). Similarly, Assumptions 6 and 7 map one-to-one to Assumptions 5 and 6 of (Agarwal et al. 2020b) under the change of notation. Assumptions 4 and 5 also map one-

to-one to Assumptions 4 and 8 under the new definition of a donor set, as given by Definition 1. Assumption 8 is slightly different than Assumption 7 of (Agarwal et al. 2020b) in that the constants in this work can depend on model rank r or maximum degree d of $\mathcal{G}$. In (Agarwal et al. 2020b), this change is mainly reflected via Equations (41) and (54). In both, the right-hand side should be multiplied by a factor of $1/\xi'''$, where ξ''' is defined in Theorem 2. Both these changes are reflected in our Lemmas 14 and 16 above.

Lastly, note that (Agarwal et al. 2020b) analyze the coefficients $\boldsymbol{\alpha}$ as well as $\boldsymbol{\alpha}_\perp = R_{tr}R_{tr}^\top\boldsymbol{\alpha}$. In our work, our proofs only utilize $\boldsymbol{\alpha}$ because $\mathbb{E}[Z_{tr,\mathcal{I}^{(n)}}|LF,A]^+ = R_{tr}\Sigma_{tr}^{-1}L_{tr}$, which implies that $\boldsymbol{\alpha}_\perp = R_{tr}R_{tr}^\top R_{tr}\Sigma_{tr}^{-1}L_{tr}\mathbb{E}[\mathbf{z}_{tr,n}|LF,A] = \mathbb{E}[Z_{tr,\mathcal{I}^{(n)}}|LF,A]^+\mathbb{E}[\mathbf{z}_{tr,n}|LF,A] = \boldsymbol{\alpha}$.

Lemma 20 (Adapted from Lemma 19 by Agarwal et al. (2023a))**.** *Suppose that Assumptions 2, 4, 7, and 8 hold. Then, $\|\boldsymbol{\alpha}\|_2 \le \frac{\sqrt{r}}{\xi'\sqrt{\xi''|\mathcal{I}^{(n)}|}}$, where ξ' and ξ'' are defined in Assumption 8 and depend only on the model rank r and maximum degree d of $\mathcal{G}$.*

Remark 7. *Lemma 20 is adapted from Lemma 19 of (Agarwal et al. 2023a). It is easy to verify that the setup is identical, with Assumptions 2 (which automatically satisfies Assumption 1), 4, 7, and 8 map to Assumptions 1, 3b, 6, and 9, respectively. Note that the proof of Lemma 19 only requires Assumption 3a (and not Assumption 3b). There is one important difference, which is that Assumption 8's constants ξ' and ξ'' can depend on r and d. In Agarwal et al. (2023a), this simply means translates to a factor of $\frac{1}{\xi'\sqrt{\xi''}}$, where ξ' and ξ'' are defined in Assumption 8, as reflected in Lemma 20.*

Lemma 21 (Adapted from Theorem 3.1 by Matoušek (2008))**.** *Let $R \in \mathbb{R}^{d_1\times d_2}$. Let $R_{ij} \overset{i.i.d.}{\sim} p_R$, where $\mathbb{E}[R_{ij}] = 0$, $var(R_{ij}) = 1$, and p_R is a sub-Gaussian distribution. Let $\eta \in (0, 1/2]$, $\delta \in (0,1)$, $d = C\eta^{-2}\log(2/\delta)$, where C is a constant that depends on p_R. Then, with probability at least $1-\delta$,*

$$(1-\eta)\|\mathbf{x}\|_2 \le \|R\mathbf{x}\|_2 \le (1+\eta)\|\mathbf{x}\|_2,$$

for all $\mathbf{x} \in \mathbb{R}^{d_2}$.

Lemma 22. *Suppose Assumption 2, 3, 10, and 11 hold. Then, under the experiment design in Section 6.1, Assumption 8 holds with high probability.*

Proof. In this proof, we abbreviate $\mathcal{I}^{(n)}$ to $\mathcal{I}$.

Decomposing $\mathbb{E}[Z_{\mathbf{tr},\mathcal{I}}|LF,A]$. Let $\tilde{\mathcal{N}}(j)$ denote $\pi_j(\mathcal{N}(j))$, where π_j is specified in Definition 1, i.e., $\tilde{\mathcal{N}}(j)$ corresponds to the permuted neighborhood of donor j, where the permutation is fixed under Definition 1. In the remainder of this proof, we use the decomposition $\mathbb{E}[Z_{\mathrm{tr},\mathcal{I}}|LF,A] = \widetilde{W}U_{\mathcal{I}}$, where

$$\widetilde{W} = \left[\tilde{\mathbf{w}}_{t,A[\mathcal{N}(n),t]}^\top : t \in \mathcal{T}_{\mathrm{tr}}\right] \in \mathbb{R}^{T_{\mathrm{tr}}\times r(d+1)}, \tag{14}$$

$$U_{\mathcal{I}} = \left[\mathbf{u}_{\tilde{\mathcal{N}}_j(\mathcal{I}_k),\mathcal{I}_k} : j \le |\mathcal{N}(n)|, k \le |\mathcal{I}|\right] \in \mathbb{R}^{r(d+1)\times|\mathcal{I}|}. \tag{15}$$

Reducing the problem to upper bounding the condition number of $\widetilde{W}$. By Assumption 11, the variance of $u_{j,i,k}$ and $w_{t,a,k}$ is $\frac{1}{3r(d+1)}$. Applying Lemma 21,

$$(1-\eta)\left\|\widetilde{W}^\top\mathbf{x}\right\|_2 \le \sqrt{\frac{3r(d+1)}{|\mathcal{I}|}}\left\|U_{\mathcal{I}}^\top\widetilde{W}^\top\mathbf{x}\right\|_2 \le (1+\eta)\left\|\widetilde{W}^\top\mathbf{x}\right\|_2,$$

for all $\mathbf{x} \in \mathbb{R}^{T_{\text{tr}}}$ with probability at least $1 - 2\exp\left(-\frac{|\mathcal{I}|\eta^2}{C}\right)$ for $\eta \in (0, 1/2]$ and $|\mathcal{I}| > \frac{C \log 2}{\eta^2}$.

Let $\phi(X)$ denote the condition number of matrix X, i.e., the ratio of the largest to r_X-th largest singular values of X, where r_X denotes the rank of X. Let $\mathcal{B}_b$ denote the unit ball in $\mathbb{R}^b$. This implies that

$$\begin{aligned}
(\phi(\widetilde{W}U))^2 &= \frac{\max_{\mathbf{x}\in\mathcal{B}_{T_{\text{tr}}}} \left\| U_{\mathcal{I}}^\top \widetilde{W}^\top \mathbf{x} \right\|_2}{\min_{\mathbf{x}\in\mathcal{B}_{T_{\text{tr}}}} \left\| U_{\mathcal{I}}^\top \widetilde{W}^\top \mathbf{x} \right\|_2} \\
&\leq \frac{\max_{\mathbf{x}\in\mathcal{B}_{T_{\text{tr}}}} (1+\eta)\sqrt{\frac{|\mathcal{I}|}{3r(d+1)}} \left\| \widetilde{W}^\top \mathbf{x} \right\|_2}{\min_{\mathbf{x}\in\mathcal{B}_{T_{\text{tr}}}} (1-\eta)\sqrt{\frac{|\mathcal{I}|}{3r(d+1)}} \left\| \widetilde{W}^\top \mathbf{x} \right\|_2} \\
&= \frac{(1+\eta)\max_{\mathbf{x}\in\mathcal{B}_{T_{\text{tr}}}} \left\| \widetilde{W}^\top \mathbf{x} \right\|_2}{(1-\eta)\min_{\mathbf{x}\in\mathcal{B}_{T_{\text{tr}}}} \left\| \widetilde{W}^\top \mathbf{x} \right\|_2} \\
&\leq \frac{3}{2}(\phi(\widetilde{W}))^2.
\end{aligned}$$

Therefore, upper bounding the condition number of $\mathbb{E}[Z_{\text{tr},\mathcal{I}}|LF, A]$ comes down to upper bounding the condition number of $\widetilde{W}$.

Bounding the condition number of $\widetilde{W}$. The condition number of $\widetilde{W}$ is given by

$$\sqrt{\frac{\max_{\mathbf{x}\in\mathcal{B}_{r(d+1)}} \mathbf{x}^\top \widetilde{W}^\top \widetilde{W} \mathbf{x}}{\min_{\mathbf{x}\in\mathcal{B}_{r(d+1)}} \mathbf{x}^\top \widetilde{W}^\top \widetilde{W} \mathbf{x}}}, \tag{16}$$

where we once again take max and min over $\mathbf{x}$ for which $\|\mathbf{x}\|_2 = 1$. We therefore study $\widetilde{W}^\top \widetilde{W}$. Note that $\widetilde{W}^\top \widetilde{W}$ can be split into $(d+1) \times (d+1)$ block matrices, where each block matrix is $r \times r$. Let the (j,k)-th block matrix be denoted by $X^{j,k} \in \mathbb{R}^{r\times r}$. By the definition of $\widetilde{W}$ and $\widetilde{w}_{\cdot,\cdot}$, the (j,k)-th block matrix can be written as:

$$X^{j,k} = \sum_{t\in\mathcal{T}_{\text{tr}}} \mathbf{w}_{t,A[\mathcal{N}_j(n),t]} \mathbf{w}_{t,A[\mathcal{N}_k(n),t]}^\top.$$

For the remainder of the proof, we assume $D = 2$ for ease of exposition. However, it is easy to show that our results extend for $D > 2$.

Now, note that, under the experiment design in Section 6, three facts hold true if $j \neq k$:

1. j and k receive the treatment 1 at the same time for exactly $T_{\text{tr}} - 2\bar{T}$ measurements; and
2. j receives a non-control treatment and k receives treatment 1 for exactly $\bar{T}$ time steps; and
3. k receives a non-control treatment and j receives treatment 1 for exactly $\bar{T}$ time steps.

Let $\mathcal{T}_j$ denote the measurements for which j receives a non-control treatment and $\mathcal{T}_k$ denote the measurements for which k receives a non-control treatment. Note that $\mathcal{T}_j$ and $\mathcal{T}_k$ are disjoint by the experiment design in Section 6.1. Note further that $|\mathcal{T}_j| = |\mathcal{T}_k| = \bar{T}$.

Then, if $j \neq k$,

$$X^{j,k} = \sum_{t \in \mathcal{T}_{\text{tr}} \setminus \{\mathcal{T}_j \cup \mathcal{T}_k\}} \mathbf{w}_{t,0}\mathbf{w}_{t,0}^\top + \sum_{t \in \mathcal{T}_j} \mathbf{w}_{t,1}\mathbf{w}_{t,0}^\top + \sum_{t \in \mathcal{T}_k} \mathbf{w}_{t,0}\mathbf{w}_{t,1}^\top,$$

We now make use of two facts. First, since $\mathbf{w}_{\cdot,\cdot}$ is bounded, it is sub-Gaussian. Second, $\mathbb{E}[\mathbf{w}_{t,0}\mathbf{w}_{t,1}^\top] = \mathbb{E}[\mathbf{w}_{t,1}\mathbf{w}_{t,0}^\top] = 0$. Third, $\mathbb{E}[\mathbf{w}_{t,0}\mathbf{w}_{t,0}^\top] = \mathbb{E}[\mathbf{w}_{t,1}\mathbf{w}_{t,1}^\top] = \text{Cov}(\mathbf{w}_{\cdot,\cdot}) = \frac{1}{3r(d+1)}\mathbb{I}_{r\times r}$. As such, we can characterize $X^{j,j}$ and $X^{j,k}$ in a high-probability sense, as follows:

1. If $j = k$, $X^{j,k} = \Theta_P\left(\frac{T_{\text{tr}}}{3r(d+1)}\mathbb{I}_{r\times r}\right)$.
2. If $j \neq k$, $\sum_{t \in \mathcal{T}_{\text{tr}} \setminus \{\mathcal{T}_j \cup \mathcal{T}_k\}} \mathbf{w}_{t,0}\mathbf{w}_{t,0}^\top = \Theta_P\left(\frac{T_{\text{tr}} - 2\bar{T}}{3r(d+1)}\mathbb{I}_{r\times r}\right)$.
3. If $j \neq k$, the two right-hand sums are $\sum_{t \in \mathcal{T}_j} \mathbf{w}_{t,1}\mathbf{w}_{t,0}^\top + \sum_{t \in \mathcal{T}_k} \mathbf{w}_{t,0}\mathbf{w}_{t,1}^\top = \Theta_P\left([0]_{r\times r}\right)$.

Therefore,

$$\widetilde{W}^\top\widetilde{W} = \Theta_P\left(\frac{2\bar{T}}{3r(d+1)}\mathbb{I}_{r(d+1)\times r(d+1)} + \frac{T_{\text{tr}} - 2\bar{T}}{3r(d+1)}\begin{bmatrix}\mathbb{I}_{r\times r} & \mathbb{I}_{r\times r} & \cdots \\ \mathbb{I}_{r\times r} & \mathbb{I}_{r\times r} & \cdots \\ \vdots & \vdots & \ddots\end{bmatrix}\right), \tag{17}$$

and

$$\mathbf{x}^\top\widetilde{W}^\top\widetilde{W}\mathbf{x} = \Theta_P\left(\frac{2\bar{T}}{3r(d+1)}\left\|\mathbf{x}\right\|_2^2 + \frac{T_{\text{tr}} - 2\bar{T}}{3r(d+1)}\left\|\sum_{k=1}^{d+1}\mathbf{x}_k\right\|_2^2\right) \tag{18}$$

$$= \Theta_P\left(\frac{2\bar{T}}{3r(d+1)} + \frac{T_{\text{tr}} - 2\bar{T}}{3r(d+1)}\left\|\sum_{k=1}^{d+1}\mathbf{x}_k\right\|_2^2\right), \tag{19}$$

where $\mathbf{x}^\top = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_{d+1}^\top]$, and the second equality follows from the fact that we restrict our attention to $\mathbf{x}$ for which $\left\|\mathbf{x}\right\|_2 = 1$. Note that $\left\|\sum_{k=1}^{d+1}\mathbf{x}_k\right\|_2^2 = \sum_{\ell=1}^{r}(\sum_{k=1}^{d+1}x_{k,\ell})^2 \leq (\sum_{\ell=1}^{r}|\sum_{k=1}^{d+1}x_{k,\ell}|)^2 \leq (\sum_{\ell=1}^{r}\sum_{k=1}^{d+1}|x_{k,\ell}|)^2 \leq \left\|\mathbf{x}\right\|_1^2 \leq r(d+1)\left\|\mathbf{x}\right\|_2^2$. Therefore,

$$\begin{aligned}
\phi(\widetilde{W}) &= \sqrt{\frac{\max_{\mathbf{x}\in\mathcal{B}_{r(d+1)}}\mathbf{x}^\top\widetilde{W}^\top\widetilde{W}\mathbf{x}}{\min_{\mathbf{x}\in\mathcal{B}_{r(d+1)}}\mathbf{x}^\top\widetilde{W}^\top\widetilde{W}\mathbf{x}}} \\
&= O_P\left(\sqrt{1 + \frac{r(d+1)(T_{\text{tr}} - 2\bar{T})}{2\bar{T}}}\right) \\
&= O_P\left(\sqrt{1 + \frac{r(d+1)(d^2-1)}{2}}\right) \\
&= O_P\left(\sqrt{1 + \frac{r(d+1)^3}{2}}\right) \\
&= O_P\left(\sqrt{1 + 4rd^3}\right),
\end{aligned}$$

where the second inequality follows from the fact that there are at most d^2+1 sets of $\bar{T}$ that make up $\mathcal{T}_{\text{tr}}$ under the experiment design in Section 6.1. Therefore, the requirement on the condition number in Assumption 8 holds with $\xi' = (1+4rd^3)^{-1/2}$.

Requirement on Frobenius norm. It remains to show that the second requirement of Assumption 8 holds. To do so, note that

$$\begin{aligned}
\|\mathbb{E}[Z_{\text{tr},\mathcal{I}}|LF,A]\|_2^2 &= \sum_{t,j}(\mathbb{E}[Z_{\text{tr},\mathcal{I}}|LF,A]_{tj})^2 \\
&= \sum_{t,j}(\tilde{\mathbf{w}}_{t,A[\mathcal{N}(n),t]}^\top \tilde{\mathbf{u}}_{j,\mathcal{N}(j)})^2 \\
&= \sum_{t,j}\left(\sum_{b=1}^{d+1}\mathbf{w}_{t,A[\mathcal{N}_b(n),t]}^\top \mathbf{u}_{\mathcal{N}_b(j),j}\right)^2 \\
&= \sum_{t,j}\left(\sum_{a\in[D]}\mathbf{w}_{t,a}^\top \sum_{b=1}^{d+1}\mathbf{u}_{\mathcal{N}_b(j),j}\text{Ind}(A[\mathcal{N}_b(n),t]=a)\right)^2. \qquad (20)
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{a\in[D]}\mathbf{w}_{t,a}^\top \sum_{b=1}^{d+1}\mathbf{u}_{\mathcal{N}_b(j),j}\text{Ind}(A[\mathcal{N}_b(n),t]=a)\right)^2\right] \\
&\geq \sum_{a\in[D]}\mathbb{E}\left[\left(\mathbf{w}_{t,a}^\top \sum_{b=1}^{d+1}\mathbf{u}_{\mathcal{N}_b(j),j}\text{Ind}(A[\mathcal{N}_b(n),t]=a)\right)^2\right] \\
&= \sum_{a\in[D]}\mathbb{E}\left[\left(\sum_{k=1}^{r} w_{t,a,k}\sum_{b=1}^{d+1}u_{\mathcal{N}_b(j),j,k}\text{Ind}(A[\mathcal{N}_b(n),t]=a)\right)^2\right] \\
&\geq \sum_{a\in[D]}\sum_{k=1}^{r}\mathbb{E}\left[\left(w_{t,a,k}\sum_{b=1}^{d+1}u_{\mathcal{N}_b(j),j,k}\text{Ind}(A[\mathcal{N}_b(n),t]=a)\right)^2\right] \\
&= \sum_{a\in[D]}\sum_{k=1}^{r}\mathbb{E}\left[w_{t,a,k}^2\left(\sum_{b=1}^{d+1}u_{\mathcal{N}_b(j),j,k}\text{Ind}(A[\mathcal{N}_b(n),t]=a)\right)^2\right] \\
&\geq \sum_{a\in[D]}\sum_{k=1}^{r}\mathbb{E}\left[w_{t,a,k}^2\sum_{b=1}^{d+1}u_{\mathcal{N}_b(j),j,k}^2\text{Ind}(A[\mathcal{N}_b(n),t]=a)^2\right] \\
&= \sum_{a\in[D]}\sum_{k=1}^{r}\mathbb{E}\left[w_{t,a,k}^2\right]\sum_{b=1}^{d+1}\mathbb{E}\left[u_{\mathcal{N}_b(j),j,k}^2\right]\text{Ind}(A[\mathcal{N}_b(n),t]=a) \\
&= \frac{1}{9r^2(d+1)^2}\sum_{a\in[D]}\sum_{k=1}^{r}\sum_{b=1}^{d+1}\text{Ind}(A[\mathcal{N}_b(n),t]=a) \\
&= \frac{1}{9r(d+1)}. \qquad (21)
\end{aligned}$$

Since $w_{t,a,k}$ and $u_{\ell,j,k}$ are bounded, $w_{t,a,k}^2$ and $u_{\ell,j,k}^2$ are sub-Gaussian. As such, combining (20) and (21) gives

$$\mathbb{E}[Z_{\mathrm{tr},\mathcal{I}}|LF, A] = \omega_P \left(\frac{T_{\mathrm{tr}}|\mathcal{I}|}{9r(d+1)}\right),$$

which confirms that Assumption 8 holds with high probability for $\xi'' = (9r(d+1))^{-1}$. □

C Proofs for Section 4

C.1 Proof of Theorem 1

Proof. Recall from Definition 2 that identifiability requires that the estimand $f(\theta)$ can be written as a function $g(P_\theta)$ of the data distribution P_θ, where θ are the unknown model parameters. In our setting, the unknown model parameters are the latent factors, thus $\theta = LF$. The observed dataset consists of the matrices of observed outcomes $Z[\mathcal{T}_{\mathrm{tr}}, :]$ and $Z[\mathcal{T}_{\mathrm{pr}}, :]$, whose joint distribution, denoted P_θ, is both a function of the unknown parameter θ and the known and fixed parameters $(A, G, \mathcal{T}_{\mathrm{tr}}, \mathcal{T}_{\mathrm{pr}})$.

Our estimand $f(\theta) = \mathrm{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\mathrm{pr}}} \sum_{t \in \mathcal{T}_{\mathrm{pr}}} \mathbb{E}\left[Y_{t,n}^{(\tilde{\mathbf{a}}_{\mathcal{N}(n)})}\right]$. The claim in Theorem 1 is that $\mathrm{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ is identifiable as we can write it as a function of expectations of the data, as given by

$$f(\theta) := \mathrm{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\mathrm{pr}}} \mathbf{1}^T \, \mathbb{E}[Z_{\mathrm{pr},\mathcal{I}^{(n)}} \,|\, LF, A] \, \mathbb{E}[Z_{\mathrm{tr},\mathcal{I}^{(n)}}|LF, A]^+ \, \mathbb{E}[\mathbf{z}_{\mathrm{tr},n}|LF, A] =: g(P_\theta).$$

To show this claim, we first define some additional notation. Let $U_{\mathcal{I}^{(n)}}$ be a $r|\mathcal{N}(n)| \times |\mathcal{I}^{(n)}|$ matrix where the j-th column of $U_{\mathcal{I}^{(n)}}$ corresponds to the network-adjusted latent factor associated to the j-th donor, i.e. $\tilde{\mathbf{u}}_{i,\mathcal{N}(v)}$ where unit i is the j-th donor in $\mathcal{I}^{(n)}$. Let W_{tr} be a $r|\mathcal{N}(n)| \times T_{\mathrm{tr}}$ matrix where the j-th column of W_{tr} corresponds to the network adjusted latent factor and the *applied* treatment associated to the j-th measurement in the training period $\mathcal{T}_{\mathrm{tr}}$, i.e. $\tilde{\mathbf{w}}_{t,a^t_{\mathcal{N}(n)}}$ where t is the j-th measurement in $\mathcal{T}_{\mathrm{tr}}$. Similarly let W_{pr} be a $r|\mathcal{N}(n)| \times T_{\mathrm{pr}}$ matrix where the j-th column of W_{pr} corresponds to the network adjusted latent factor associated to the j-th measurement and the *counterfactual* treatment $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ in the prediction period $\mathcal{T}_{\mathrm{pr}}$.

By conditional exogeneity as stated in Assumption 3 and the latent factor model as stated in Assumption 2, it follows that for any $t, i, \tilde{\mathbf{a}}$,

$$\mathbb{E}[Y_{t,i}^{(\tilde{\mathbf{a}})} |\, LF, A] = \langle \tilde{\mathbf{w}}_{t,\tilde{\mathbf{a}}_{\mathcal{N}(i)}}, \tilde{\mathbf{u}}_{i,\mathcal{N}(i)} \rangle.$$

As a result, we can write the target estimand as

$$\mathrm{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) := \frac{1}{T_{\mathrm{pr}}} \sum_{t \in \mathcal{T}_{\mathrm{pr}}} \mathbb{E}\left[Y_{t,n}^{(\tilde{\mathbf{a}}_{\mathcal{N}(n)})}\right] = \frac{1}{T_{\mathrm{pr}}} \mathbf{1}^T \, W_{\mathrm{pr}}^T \tilde{\mathbf{u}}_{n,\mathcal{N}(n)}.$$

By the linear span property as stated in Assumption 4, there must exist a vector $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{I}^{(n)}|}$ such that $\tilde{\mathbf{u}}_{n,\mathcal{N}(n)} = U_{\mathcal{I}^{(n)}} \boldsymbol{\lambda}$. Along with the latent factor model decomposition and the condition that donors must share the same applied treatment as n in the training period, it also follows that

$\mathbb{E}[\mathbf{z}_{\text{tr},n}|LF, A] = \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]\boldsymbol{\lambda}$. By substitution,

$$\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\text{pr}}}\mathbf{1}^T\ W_{\text{pr}}^T\ U_{\mathcal{I}^{(n)}}\boldsymbol{\lambda} = \frac{1}{T_{\text{pr}}}\mathbf{1}^T\ \mathbb{E}[Z_{\text{pr},\mathcal{I}^{(n)}}\,|\,LF, A]\boldsymbol{\lambda},$$

where the latter equality follows from the latent factor model, conditional exogeneity, and the construction of the donor set, which enforces that the applied treatments to the donors during the prediction period must match the counterfactual treatment. By the subspace inclusion property as stated in Assumption 5, there must exist a matrix $\Gamma \in \mathbb{R}^{T_{\text{pr}} \times T_{\text{tr}}}$ such that $\mathbb{E}[Z_{\text{pr},\mathcal{I}^{(n)}}\,|\,LF, A] = \Gamma\ \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]$, such that by substitution

$$\begin{aligned}
\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) &= \frac{1}{T_{\text{pr}}}\mathbf{1}^T\ \Gamma\ \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]\boldsymbol{\lambda}, \\
&\overset{\text{(a)}}{=} \frac{1}{T_{\text{pr}}}\mathbf{1}^T\ \Gamma\ \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]\ \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]^+\ \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]\boldsymbol{\lambda}, \\
&\overset{\text{(b)}}{=} \frac{1}{T_{\text{pr}}}\mathbf{1}^T\ \mathbb{E}[Z_{\text{pr},\mathcal{I}^{(n)}}\,|\,LF, A]\ \mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]^+\ \mathbb{E}[\mathbf{z}_{\text{tr},n}|LF, A] =: g(P_\theta).
\end{aligned}$$

where (a) follows from the property of pseudoinverses, and (b) follows from the construction of Γ and $\boldsymbol{\lambda}$ from the linear span and subspace inclusion properties. □

C.2 Proof of Theorem 2

Proof. In this proof, we suppress the conditioning on LF and A. Let $\Delta = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is defined in Section 4.1. Let $\boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}} = \left[\epsilon_{t,j}^{(\mathbf{a}^t_{\mathcal{N}(j)})} : j \in \mathcal{I}^{(n)}\right] \in \mathbb{R}^{|\mathcal{I}^{(n)}|}$. Lastly, recall that R_{pr} and R_{tr} denote the matrices containing the right singular vectors of $\mathbb{E}\big[Z_{\text{pr},\mathcal{I}^{(n)}}\big|\,LF, A\big]$ and $\mathbb{E}\big[Z_{\text{tr},\mathcal{I}^{(n)}}\big|\,LF, A\big]$, respectively.

By (6), (7), the definition of $\boldsymbol{\alpha}$ in Section 4.1 and the definition of $\hat{\boldsymbol{\alpha}}$ in Section 4.1,

$$\begin{aligned}
&\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) \\
&\qquad = \frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left(\left\langle Z[t, \mathcal{I}^{(n)}], \hat{\boldsymbol{\alpha}}\right\rangle - \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \boldsymbol{\alpha}\right\rangle\right) \\
&\qquad = \frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left(\left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \Delta\right\rangle - \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \hat{\boldsymbol{\alpha}}\right\rangle + \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}] + \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \hat{\boldsymbol{\alpha}}\right\rangle\right) \\
&\qquad = \frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left(\left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \Delta\right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \hat{\boldsymbol{\alpha}}\right\rangle\right) \\
&\qquad = \frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left(\left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \Delta\right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}\right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta\right\rangle\right),
\end{aligned}$$

where the second equality follows from $Z[t, \mathcal{I}^{(n)}] = \mathbb{E}Z[t, \mathcal{I}^{(n)}] + \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}$.

Let $\mathcal{P} = R_{\text{tr}} R_{\text{tr}}^\top$. By Assumption 5, $R_{\text{pr}} = \mathcal{P} R_{\text{pr}}$, which implies $\mathbb{E}[Z_{\text{pr},\mathcal{I}^{(n)}}] = \mathbb{E}[Z_{\text{pr},\mathcal{I}^{(n)}}]\mathcal{P}$. Therefore,

$$\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left(\left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha} \right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta \right\rangle \right). \tag{22}$$

The three terms on the right-hand side can be bounded using Lemmas 14, 15, and 19, as follows.

Bounding the three terms in (22). To bound the first term in (22), observe that

$$\begin{aligned} \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle &\leq \left\| \mathbb{E}Z[t, \mathcal{I}^{(n)}] \right\|_2 \|\mathcal{P}\Delta\|_2 \leq \|\mathcal{P}\Delta\|_2 \sqrt{|\mathcal{I}^{(n)}|} \\ &\implies \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle \leq \|\mathcal{P}\Delta\|_2 \sqrt{|\mathcal{I}^{(n)}|}, \end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality from Assumption 7. One can then upper bound $\|\mathcal{P}\Delta\|_2$ using Lemma 14 to get

$$\begin{aligned} &\frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle \\ &\qquad = O_P \left(\frac{\sqrt{r_{\text{tr}}}}{\xi''' T_{\text{tr}}^{1/4}} + \frac{r_{\text{tr}}^{3/2} \sqrt{\log\left(T_{\text{tr}} |\mathcal{I}^{(n)}|\right)}}{(\xi''')^{5/2} \min\left(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|}\right)} + \frac{r_{\text{tr}}^2 \sqrt{|\mathcal{I}^{(n)}| \log\left(T_{\text{tr}} |\mathcal{I}^{(n)}|\right)}}{(\xi''')^4 \min\left(T_0^{3/2}, N_d^{3/2}\right)} \right), \end{aligned}$$

where ξ''' is defined in Theorem 2.

To bound the second term in (22), observe that

$$\mathbb{E}\left[\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha} \right\rangle \right] = 0, \qquad \text{Var}(\boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}) = \sigma^2 \|\boldsymbol{\alpha}\|_2^2,$$

for all $t \in \mathcal{T}_{\text{pr}}$ by by Assumptions 2 and 6. Furthermore, Assumption 6 gives that $\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha} \right\rangle$ are independent across t. By Lemmas 15 and 20,

$$\frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha} \right\rangle = O_P\left(\frac{\|\boldsymbol{\alpha}\|_2}{\sqrt{T_{\text{pr}}}} \right) = O_P\left(\frac{\sqrt{r_{\text{tr}}}}{\xi' \sqrt{\xi'' T_{\text{pr}} |\mathcal{I}^{(n)}|}} \right).$$

Lastly, to bound the third term in (22), we define the following events

$$\begin{aligned} E_1 &= \left\{ \|\Delta\|_2 = O\left(\frac{\sqrt{\log(T_{\text{tr}}|\mathcal{I}^{(n)}|)}}{(\xi''')^{3/2}} \left(\frac{r_{\text{tr}}^{3/4}}{T_{\text{tr}}^{1/4} |\mathcal{I}^{(n)}|^{1/2}} + \frac{r_{\text{tr}}^{3/2}}{(\xi''')^{3/2} \min(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|})} \right) \right) \right\}, \\ E_2 &= \left\{ \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta \right\rangle \right. \\ &\qquad \left. = O\left(\frac{\sqrt{\log(T_{\text{tr}}|\mathcal{I}^{(n)}|)}}{\sqrt{T_{\text{pr}}} (\xi''')^{3/2}} \left(\frac{r_{\text{tr}}^{3/4}}{T_{\text{tr}}^{1/4} |\mathcal{I}^{(n)}|^{1/2}} + \frac{r_{\text{tr}}^{3/2}}{(\xi''')^{3/2} \min(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|})} \right) \right) \right\}. \end{aligned}$$

Noting that $\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta \right\rangle$ are independent across t, Lemmas 19 and 15 imply that E_1 and $E_2|E_1$ occur with high probability, which also implies that E_2 occurs with high probability and therefore that

$$\frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta \right\rangle = O_P\left(\frac{\sqrt{\log(T_{\text{tr}}|\mathcal{I}^{(n)}|)}}{\sqrt{T_{\text{pr}}}(\xi''')^{3/2}} \left(\frac{r_{\text{tr}}^{3/4}}{T_{\text{tr}}^{1/4}|\mathcal{I}^{(n)}|^{1/2}} + \frac{r_{\text{tr}}^{3/2}}{(\xi''')^{3/2}\min(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|})} \right) \right).$$

Note that we can safely assume that $\xi''' < 1$ because if there exists a $\xi''' \geq 1$, letting ξ''' take some value less than $1/\xi'''$ will always satisfy Assumption 8. Together, the three bounds and the observation above give the theorem result. □

C.3 Proof of Theorem 3

Proof. In this proof, we suppress the conditioning on LF and A. Let $\Delta = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is defined below Theorem 1. Let $\boldsymbol{\epsilon}_{\text{tr},n} = [\epsilon_{\tau,n}^{(\mathbf{a}_{\mathcal{N}(n)}^{\tau})} : \tau \in \mathcal{T}_{\text{tr}}]$. Let $\boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}} = [\epsilon_{t,j}^{(\mathbf{a}_{\mathcal{N}(j)}^{t})} : j \in \mathcal{I}^{(n)}] \in \mathbb{R}^{|\mathcal{I}^{(n)}|}$. Recall that R_{tr} denotes the matrix containing the right singular vectors of $\mathbb{E}\big[Z_{\text{tr},\mathcal{I}^{(n)}} \big| LF, A\big]$. Let $Z_{\text{tr},\mathcal{I}^{(n)}}^{r_{\text{tr}}} = \sum_{\ell=1}^{r_{\text{tr}}} \hat{s}_\ell \hat{\boldsymbol{\mu}}_\ell \hat{\boldsymbol{\nu}}_\ell^\top = \bar{L}_{\text{tr}} \bar{\Sigma}_{\text{tr}} \bar{R}_{\text{tr}}^\top$, where $\hat{s}_\ell$, $\hat{\boldsymbol{\mu}}_\ell$, and $\hat{\boldsymbol{\nu}}_\ell$ are defined in Section 3.2. Let $\mathcal{P} = R_{\text{tr}} R_{\text{tr}}^\top$ and $\bar{\mathcal{Q}} = \bar{L}_{\text{tr}} \bar{L}_{\text{tr}}^\top$.

Asymptotic normality. We begin by establishing that, conditioned on LF and A,

$$\frac{\sqrt{T_{\text{tr}}}}{\sigma \|\boldsymbol{\alpha}\|_2} \left(\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) \right) \xrightarrow{d} \mathcal{N}(0,1).$$

We use an equation from the proof of Theorem 2. By (22), we have

$$\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) = \frac{1}{T_{\text{pr}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left(\left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha} \right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta \right\rangle \right) \quad (23)$$

We now analyze each of the three terms on the right-hand side of (23).

To characterize the first term in (23), observe that

$$\left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle \leq \left\| \mathbb{E}Z[t, \mathcal{I}^{(n)}] \right\|_2 \|\mathcal{P}\Delta\|_2 \leq \|\mathcal{P}\Delta\|_2 \sqrt{|\mathcal{I}^{(n)}|}$$

$$\implies \frac{1}{\sigma \|\boldsymbol{\alpha}\|_2 \sqrt{T_{\text{pr}}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle \leq \|\mathcal{P}\Delta\|_2 \frac{\sqrt{T_{\text{pr}}|\mathcal{I}^{(n)}|}}{\sigma \|\boldsymbol{\alpha}\|_2},$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality from Assumption 7. Then,

$$\sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle \leq \sum_{t \in \mathcal{T}_{\text{pr}}} \|\mathcal{P}\Delta\|_1 \leq \sum_{t \in \mathcal{T}_{\text{pr}}} \|\mathcal{P}\Delta\|_2 \sqrt{|\mathcal{I}^{(n)}|} \leq \|\mathcal{P}\Delta\|_2 T_{\text{pr}} \sqrt{|\mathcal{I}^{(n)}|}.$$

Therefore,

$$\frac{1}{\sigma \|\boldsymbol{\alpha}\|_2 \sqrt{T_{\text{pr}}}} \sum_{t \in \mathcal{T}_{\text{pr}}} \left\langle \mathbb{E}Z[t, \mathcal{I}^{(n)}], \mathcal{P}\Delta \right\rangle \leq \frac{\|\mathcal{P}\Delta\|_2 \sqrt{T_{\text{pr}}|\mathcal{I}^{(n)}|}}{\sigma \|\boldsymbol{\alpha}\|_2} = o_P(1),$$

under the condition $\|\Delta\|_2 = o_P\left(\frac{\sigma\|\boldsymbol{\alpha}\|_2}{\sqrt{T_{\text{pr}}|\mathcal{I}^{(n)}|}}\right)$, as given in the theorem statement.

To characterize the second term in (23), observe that

$$\mathbb{E}\left[\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}\right\rangle\right] = 0, \qquad \text{Var}(\boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}) = \sigma^2 \|\boldsymbol{\alpha}\|_2^2,$$

for all $t \in \mathcal{T}_{\text{pr}}$ by Assumptions 2 and 6. By the Lindelberg-Lévy Central Limit Theorem,

$$\sqrt{T_{\text{pr}}}\left(\frac{\frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}\right\rangle}{\sqrt{\text{Var}(\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}\right\rangle)}}\right) \xrightarrow{d} \mathcal{N}(0,1) \tag{24}$$

$$\implies \frac{1}{\sigma\|\boldsymbol{\alpha}\|_2\sqrt{T_{\text{pr}}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}\right\rangle \xrightarrow{d} \mathcal{N}(0,1) \tag{25}$$

To characterize the third term in (23), note that $\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta\right\rangle$ are independent across t and mean-zero for $t \in \mathcal{T}_{\text{pr}}$. We define two events:

$$E_1 = \left\{\|\Delta\|_2 = o\left(\sqrt{\frac{\|\boldsymbol{\alpha}\|_2}{\sigma}}\right)\right\},$$

$$E_2 = \left\{\frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta\right\rangle = o\left(\frac{\|\boldsymbol{\alpha}\|_2\,\sigma}{\sqrt{T_{\text{pr}}}}\right)\right\}$$

By the theorem statement, E_1 holds with high probability. Moreover, Lemma 15 implies that $E_2|E_1$ holds with high probability since, conditioned on E_1, $\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta\right\rangle$ is sub-Gaussian with variance upper bounded by $\|\boldsymbol{\alpha}\|_2\,\sigma$. Therefore, E_2 holds with high probability, which implies that

$$\frac{1}{\sigma\|\boldsymbol{\alpha}\|_2\sqrt{T_{\text{pr}}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta\right\rangle = o_P(1),$$

Combining all three terms,

$$\frac{\sqrt{T_{\text{pr}}}}{\sigma\|\boldsymbol{\alpha}\|_2}\left(\frac{1}{T_{\text{pr}}}\sum_{t\in\mathcal{T}_{\text{pr}}}\left(\left\langle \mathbb{E}Z[t,\mathcal{I}^{(n)}], \mathcal{P}\Delta\right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \boldsymbol{\alpha}\right\rangle + \left\langle \boldsymbol{\epsilon}_{t,\mathcal{I}^{(n)}}, \Delta\right\rangle\right)\right) \xrightarrow{d} \mathcal{N}(0,1) \tag{26}$$

$$\implies \frac{\sqrt{T_{\text{pr}}}}{\sigma\|\boldsymbol{\alpha}\|_2}\left(\widehat{\text{IPO}}(n,\tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \text{IPO}(n,\tilde{\mathbf{a}}_{\mathcal{N}(n)})\right) \xrightarrow{d} \mathcal{N}(0,1), \tag{27}$$

as stated in the theorem.

Convergence of variance. By the definition of $\hat{\boldsymbol{\alpha}}$ in Section 3.2 and Assumption 9, we know that $Z_{\text{tr},\mathcal{I}^{(n)}}\hat{\boldsymbol{\alpha}} = Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}}\hat{\boldsymbol{\alpha}}$. Then by the definition of $\hat{\sigma}^2$ in Section 3.2,

$$|\hat{\sigma}^2 - \sigma^2| = \left|\frac{1}{T_{\text{tr}}}\left\|\mathbf{z}_{\text{tr},n} - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}}\hat{\boldsymbol{\alpha}}\right\|_2^2 - \sigma^2\right|$$

$$= \left| \frac{1}{T_{\text{tr}}} \left\| \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2 + \left(\frac{1}{T_{\text{tr}}} \left\| \boldsymbol{\epsilon}_{\text{tr},n} \right\|_2^2 - \sigma^2 \right) + \frac{2}{T_{\text{tr}}} \left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\rangle \right|$$
$$\leq \frac{1}{T_{\text{tr}}} \left\| \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2 + \left| \frac{1}{T_{\text{tr}}} \left\| \boldsymbol{\epsilon}_{\text{tr},n} \right\|_2^2 - \sigma^2 \right| + \frac{2}{T_{\text{tr}}} \left| \left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\rangle \right|. \tag{28}$$

We upper bound these three terms next.

The first term of (28) can be upper bounded using Lemmas 14, 16, and 17. First, by Lemma 14,

$$\frac{1}{T_{\text{tr}}} \left\| \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2 \leq \frac{1}{T_{\text{tr}}} \left\| Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} - \mathbb{E} Z_{\text{tr},\mathcal{I}^{(n)}} \right\|_{2,\infty}^2 \left\| \boldsymbol{\alpha} \right\|_1^2 + \frac{2}{T_{\text{tr}}} \left\langle Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \Delta, \boldsymbol{\epsilon}_{\text{tr},n} \right\rangle,$$

which by Lemma 17,

$$= O_P \left(\frac{1}{T_{\text{tr}}} \left\| Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} - \mathbb{E} Z_{\text{tr},\mathcal{I}^{(n)}} \right\|_{2,\infty}^2 \left\| \boldsymbol{\alpha} \right\|_1^2 + \frac{2 r_{\text{tr}}}{T_{\text{tr}}} + \frac{2}{\sqrt{T_{\text{tr}}}} + \frac{2}{T_{\text{tr}}} \left\| Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} - \mathbb{E} Z_{\text{tr},\mathcal{I}^{(n)}} \right\|_{2,\infty} \left\| \boldsymbol{\alpha} \right\|_1 \right).$$

By Lemma 16,

$$\frac{1}{T_{\text{tr}}} \left\| \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2$$
$$= O_P \left(\frac{\left\| \boldsymbol{\alpha} \right\|_1^2 r_{\text{tr}} \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}{(\xi''')^2 \min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)} + \frac{r_{\text{tr}}}{T_{\text{tr}}} + \frac{1}{\sqrt{T_{\text{tr}}}} + \frac{\left\| \boldsymbol{\alpha} \right\|_1 \sqrt{r_{\text{tr}} \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}}{\xi''' \sqrt{T_{\text{tr}}} \min(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|})} \right),$$

where ξ''' is defined in Theorem 3. Note that $\left\| \boldsymbol{\alpha} \right\|_1 \leq \sqrt{|\mathcal{I}^{(n)}|} \left\| \boldsymbol{\alpha} \right\|_2$. Therefore, by Lemma 20,

$$\begin{aligned}
\frac{1}{T_{\text{tr}}} \left\| \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2 &= O_P \left(\frac{\left\| \boldsymbol{\alpha} \right\|_2^2 |\mathcal{I}^{(n)}| r_{\text{tr}} \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}{(\xi''')^2 \min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)} \right. \\
&\qquad \left. + \frac{r_{\text{tr}}}{T_{\text{tr}}} + \frac{1}{\sqrt{T_{\text{tr}}}} + \frac{\left\| \boldsymbol{\alpha} \right\|_2 \sqrt{r_{\text{tr}} |\mathcal{I}^{(n)}| \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}}{\xi''' \sqrt{T_{\text{tr}}} \min(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|})} \right) \\
&= O_P \left(\frac{r r_{\text{tr}} \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}{(\xi''')^4 \min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)} \right. \\
&\qquad \left. + \frac{r_{\text{tr}}}{T_{\text{tr}}} + \frac{1}{\sqrt{T_{\text{tr}}}} + \frac{\sqrt{r r_{\text{tr}} \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}}{(\xi''')^2 \sqrt{T_{\text{tr}}} \min(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|})} \right) \\
&= O_P \left(\frac{r_{\text{tr}}^2 \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}{(\xi''')^4 \min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)} \right. \\
&\qquad \left. + \frac{r_{\text{tr}}}{T_{\text{tr}}} + \frac{1}{\sqrt{T_{\text{tr}}}} + \frac{\sqrt{r_{\text{tr}}^2 \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}}{(\xi''')^2 \sqrt{T_{\text{tr}}} \min(\sqrt{T_{\text{tr}}}, \sqrt{|\mathcal{I}^{(n)}|})} \right),
\end{aligned}$$

which, after grouping terms, implies that

$$\frac{1}{T_{\text{tr}}} \left\| \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z^{r_{\text{tr}}}_{\text{tr},\mathcal{I}^{(n)}} \hat{\boldsymbol{\alpha}} \right\|_2^2 = O_P \left(\frac{r_{\text{tr}}^2 \log(T_{\text{tr}} |\mathcal{I}^{(n)}|)}{(\xi''')^4 \min(T_{\text{tr}}, |\mathcal{I}^{(n)}|)} + \frac{r_{\text{tr}}}{\sqrt{T_{\text{tr}}}} \right). \tag{29}$$

The second term in (28) can be bounded using Assumption 6 and Lemma 15 to obtain

$$\left|\frac{1}{T_{\text{tr}}}\left\|\boldsymbol{\epsilon}_{\text{tr},n}\right\|_2^2 - \sigma^2\right| = O_P\left(T_{\text{tr}}^{-1/2}\right). \tag{30}$$

The third term in (28) can be bounded using Lemma 17 to obtain

$$\left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z_{\text{tr},\mathcal{I}^{(n)}}^{r_{\text{tr}}}\hat{\boldsymbol{\alpha}}\right\rangle = \left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \mathbb{E}[\mathbf{z}_{\text{tr},n}]\right\rangle - \left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \bar{\mathcal{Q}}\mathbb{E}[\mathbf{z}_{\text{tr},n}]\right\rangle - \left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \bar{\mathcal{Q}}\boldsymbol{\epsilon}_{\text{tr},n}\right\rangle. \tag{31}$$

Note that $\left\|\bar{\mathcal{Q}}\right\|_{op} \leq 1$ and, by Assumption 7, $\left\|\bar{\mathcal{Q}}\mathbb{E}[\mathbf{z}_{\text{tr},n}]\right\|_2 \leq \left\|\mathbb{E}[\mathbf{z}_{\text{tr},n}]\right\|_2 \leq \sqrt{T_{\text{tr}}}$. By Lemma 18 and Assumption 6, for any $\eta > 0$,

$$P(\langle \boldsymbol{\epsilon}_{\text{tr},n}, \mathbb{E}[\mathbf{z}_{\text{tr},n}] \geq \xi\rangle) \leq \exp\left(-\frac{\xi\eta^2}{T_{\text{tr}}\sigma^2}\right),$$
$$P(\left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \bar{\mathcal{Q}}\mathbb{E}[\mathbf{z}_{\text{tr},n}] \geq \xi\right\rangle) \leq \exp\left(-\frac{\xi\eta^2}{T_{\text{tr}}\sigma^2}\right),$$

which imply that

$$\langle \boldsymbol{\epsilon}_{\text{tr},n}, \mathbb{E}[\mathbf{z}_{\text{tr},n}]\rangle = O_P(\sqrt{T_{\text{tr}}}),$$
$$\left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \bar{\mathcal{Q}}\mathbb{E}[\mathbf{z}_{\text{tr},n}]\right\rangle = O_P(\sqrt{T_{\text{tr}}}).$$

From Lemma 17, for any $\eta > 0$,

$$P(\left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \bar{\mathcal{Q}}\boldsymbol{\epsilon}_{\text{tr},n}\right\rangle \geq \sigma^2 r_{\text{tr}} + \eta) \leq \exp\left(-\xi \min\left(\frac{\eta^2}{\sigma^4 r_{\text{tr}}}, \frac{\eta}{\sigma^2}\right)\right) \implies \left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \bar{\mathcal{Q}}\boldsymbol{\epsilon}_{\text{tr},n}\right\rangle = O_P(r_{\text{tr}}).$$

Therefore, by (31),

$$\frac{2}{T_{\text{tr}}}\left|\left\langle \boldsymbol{\epsilon}_{\text{tr},n}, \mathbb{E}[\mathbf{z}_{\text{tr},n}] - Z_{\text{tr},\mathcal{I}^{(n)}}^{r_{\text{tr}}}\hat{\boldsymbol{\alpha}}\right\rangle\right| = O_P\left(\frac{1}{\sqrt{T_{\text{tr}}}} + \frac{r_{\text{tr}}}{T_{\text{tr}}}\right) \tag{32}$$

Together (29), (30), and (32) give the desired result. □

C.4 Proof of Proposition 4

In order to prove Proposition 4, we prove another result that subsumes Proposition 4.

We first introduce some notation. First, we assume that $D = 2$ for ease of exposition. The proof can be straightforwardly extended for $D > 2$. Consider a unit $n \in [N]$ and counterfactual, prediction treatments of interest $\tilde{\mathbf{a}}_{\mathcal{N}(n)} \in \{1, 2\}^{|\mathcal{N}(n)|}$. We use $\mathcal{I}$ as a shorthand for $\mathcal{I}^{(n)}$ and let $\mathcal{I}_j$ refer to the j-th donor in the donor set $\mathcal{I}$.

Recall that $B^{\text{tr}}(a) \in \{0, 1\}^{N \times T_{\text{tr}}}$ and $\mathbf{b}^{\text{pr}}(a) \in \{0, 1\}^N$ are defined such that

$$B_{it}^{\text{tr}}(a) = \text{Ind}(A_{it}^{\text{tr}} = a) \qquad \text{and} \qquad \tilde{b}_i^{\text{pr}}(a) = \text{Ind}(\tilde{a}_i = a),$$

and B^{tr} and $\tilde{B}^{\text{pr}}$ be the concatenated matrices across different treatments, i.e.,

$$B^{\text{tr}} = [B^{\text{tr}}(1), B^{\text{tr}}(2), \ldots, B^{\text{tr}}(D)] \in \{0, 1\}^{N \times T_{\text{tr}} D},$$
$$\tilde{B}^{\text{pr}} = [\tilde{\mathbf{b}}^{\text{pr}}(1), \tilde{\mathbf{b}}^{\text{pr}}(2), \ldots, \tilde{\mathbf{b}}^{\text{pr}}(D)] \in \{0, 1\}^{N \times D}.$$

Finally, without loss of generality, let us re-order the training measurements such that the treatment assignments over $\mathcal{N}(n)$ are grouped together, i.e.,

$$A[\mathcal{N}(n), \mathcal{T}_{\text{tr}}] = [\mathbf{c}^1_{\mathcal{N}(n)}, \ldots, \mathbf{c}^1_{\mathcal{N}(n)} \,\big|\, \mathbf{c}^2_{\mathcal{N}(n)}, \ldots, \mathbf{c}^2_{\mathcal{N}(n)} \,\big|\, \ldots \,\big|\, \mathbf{c}^K_{\mathcal{N}(n)}, \ldots, \mathbf{c}^K_{\mathcal{N}(n)}],$$

where K is the number of distinct training treatment vectors. Let $\mathcal{T}_1$ denote the first T_1 measurements such that $A[\mathcal{N}(n), \tau] = \mathbf{c}^1_{\mathcal{N}(n)}$ for all $\tau \in \mathcal{T}_1$, $\mathcal{T}_2$ denote the next T_2 measurements such that $A[\mathcal{N}(n), \tau] = \mathbf{c}^2_{\mathcal{N}(n)}$ for all $\tau \in \mathcal{T}_2$, and so on through $\mathcal{T}_K$.

Matrix representation of $\mathbb{E}[Z_{\mathbf{tr},\mathcal{I}}|LF, A]$. Recall that for unit $n \in [N]$, measurement $t \in [T]$, and treatments $\mathbf{a} \in [D]_0^N$,

$$Y_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})} = \sum_{k \in \mathcal{N}(n)} \langle \mathbf{u}_{k,n}, \mathbf{w}_{t,a_k} \rangle + \epsilon_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})}, \tag{33}$$

where $n \in \mathcal{N}(n)$. Under $D = 2$,

$$\begin{aligned} Y_{t,n}^{(\mathbf{a}_{\mathcal{N}(n)})} - \epsilon_{tn}^{(\mathbf{a}_{\mathcal{N}(n)})} &= \sum_{k \in \mathcal{N}(n)} \text{Ind}(a_k = 1) \mathbf{u}_{k,n}^\top \mathbf{w}_{t,1} + \sum_{k \in \mathcal{N}(n)} \text{Ind}(a_k = 2) \mathbf{u}_{k,n}^\top \mathbf{w}_{t,2} \\ &= \left[\textstyle\sum_{k \in \mathcal{N}(n)} \text{Ind}(a_k = 1) \mathbf{u}_{k,n}^\top \quad , \quad \sum_{k \in \mathcal{N}(n)} \text{Ind}(a_k = 2) \mathbf{u}_{k,n}^\top \right] \begin{bmatrix} \mathbf{w}_{t,1} \\ \mathbf{w}_{t,2} \end{bmatrix}, \end{aligned} \tag{34}$$

We will use this decomposition to rewrite $\mathbb{E}[Z_{\text{tr},\mathcal{I}}|LF, A]$ as a product of matrices.

First, recall that

$$Z_{\text{tr},\mathcal{I}} = \left[Z_{t,\mathcal{I}_j} : t \in \mathcal{T}_{\text{tr}}, j \le |\mathcal{I}| \right] \in \mathbb{R}^{T_{\text{tr}} \times |\mathcal{I}|}. \tag{35}$$

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

$$U_{\mathcal{I}} = \left[\mathbf{u}_{\tilde{\mathcal{N}}_j(\mathcal{I}_k), \mathcal{I}_k} : j \le |\mathcal{N}(n)|, k \le |\mathcal{I}| \right] \in \mathbb{R}^{r|\mathcal{N}(n)| \times |\mathcal{I}|},$$

$$H_{\text{tr}}^\ell = \begin{bmatrix} \text{Ind}(c^\ell_{\mathcal{N}_1(n)} = 1), & \text{Ind}(c^\ell_{\mathcal{N}_2(n)} = 1), & \ldots, & \text{Ind}(c^\ell_{\mathcal{N}_{|\mathcal{N}(n)|}(n)} = 1) \\ \text{Ind}(c^\ell_{\mathcal{N}_1(n)} = 2), & \text{Ind}(c^\ell_{\mathcal{N}_2(n)} = 2), & \ldots, & \text{Ind}(c^\ell_{\mathcal{N}_{|\mathcal{N}(n)|}(n)} = 2) \end{bmatrix} \in \{0,1\}^{2 \times |\mathcal{N}(n)|}$$

Let $H_{\text{tr}} \in \{0,1\}^{2K \times |\mathcal{N}(n)|}$ be constructed by stacking $H_{\text{tr}}^1, H_{\text{tr}}^2, \ldots, H_{\text{tr}}^K$ on top of one another. Let

$$W^j = [(\mathbf{w}_{\tau,1}^\top, \mathbf{w}_{\tau,2}^\top) : \tau \in \mathcal{T}_j] \in \mathbb{R}^{T_j \times 2r},$$

and $W_{\text{tr}} \in \mathbb{R}^{T_{\text{tr}} \times 2rK}$ be the block-diagonal matrix with matrices $W^1, W^2, \ldots, W^K$ along the diagonal.

Then, by the decomposition in (34) and Definition 1,

$$\mathbb{E}[Z_{\text{tr},\mathcal{I}}|LF, A] = W_{\text{tr}} \left(H_{\text{tr}} \otimes \mathbb{I}_r \right) U_{\mathcal{I}}. \tag{36}$$

Matrix representation of $\mathbb{E}[Z_{\mathbf{pr},\mathcal{I}}|LF, A]$. Using the same reasoning as above, one can write

$$\mathbb{E}[Z_{\text{pr},\mathcal{I}}|LF, A] = W_{\text{pr}} \left(H_{\text{pr}} \otimes \mathbb{I}_r \right) U_{\mathcal{I}}, \tag{37}$$

where

$$H_{\text{pr}} = \begin{bmatrix} \text{Ind}(a^{\text{pr}}_{\mathcal{N}_1(n)} = 1) & \text{Ind}(a^{\text{pr}}_{\mathcal{N}_2(n)} = 1) & \dots & \text{Ind}(a^{\text{pr}}_{\mathcal{N}_{|\mathcal{N}(n)|}(n)} = 1) \\ \text{Ind}(a^{\text{pr}}_{\mathcal{N}_1(n)} = 2) & \text{Ind}(a^{\text{pr}}_{\mathcal{N}_2(n)} = 2) & \dots & \text{Ind}(a^{\text{pr}}_{\mathcal{N}_{|\mathcal{N}(n)|}(n)} = 2) \end{bmatrix} \otimes \mathbf{1}_{T_{\text{pr}}} \in \{0,1\}^{2T_{\text{pr}} \times |\mathcal{N}(n)|},$$

$$W^{\tau}_{\text{pr}} = [\mathbf{w}^{\top}_{\tau,1}, \mathbf{w}^{\top}_{\tau,2}] \in \mathbb{R}^{1 \times 2r},$$

and $W_{\text{pr}} \in \mathbb{R}^{T_{\text{pr}} \times 2rT_{\text{pr}}}$ denote the block-diagonal matrix with $W^{T_{\text{tr}}+1}_{\text{pr}}, W^{T_{\text{tr}}+2}_{\text{pr}}, \dots, W^{T}_{\text{pr}}$ along the diagonal.

Recall that Assumption 5 is that the rowspace of $\mathbb{E}[Z_{\text{pr},\mathcal{I}^{(n)}}|LF, A]$ is contained within the rowspace of $\mathbb{E}[Z_{\text{tr},\mathcal{I}^{(n)}}|LF, A]$.

Lemma 23. *Suppose Assumption 2 holds. If*

$$\textit{columnspace}(\tilde{B}^{pr}[\mathcal{N}(n), :]) \not\subseteq \textit{columnspace}(B^{tr}[\mathcal{N}(n), :]),$$

then there exist latent factors LF under which Assumption 5 cannot hold.

Proof. Our goal is to show that there exist latent factors LF such that, if columnspace($\tilde{B}^{\text{pr}}[\mathcal{N}(n), :]$) $\not\subseteq$ columnspace($B^{\text{tr}}[\mathcal{N}(n), :]$), then Assumption 5 does not hold. Since rowspace(H_{pr}) = rowspace($\tilde{B}^{\text{pr}}[\mathcal{N}(n), :]^{\top}$) and rowspace($H^{\top}_{\text{tr}}$) = rowspace($\tilde{B}^{\text{tr}}[\mathcal{N}(n), :]^{\top}$), columnspace($\tilde{B}^{\text{pr}}[\mathcal{N}(n), :]$) $\not\subseteq$ columnspace($B^{\text{tr}}[\mathcal{N}(n), :]$) is equivalent to rowspace(H_{pr}) $\not\subseteq$ rowspace(H_{tr}), which is equivalent to rowspace($H_{\text{pr}} \otimes \mathbb{I}_r$) $\not\subseteq$ rowspace($H_{\text{tr}} \otimes \mathbb{I}_r$).

By Lemma 13, there exists a vector $\mathbf{v} \neq \mathbf{0}_{2rT_{\text{pr}}}$ such that $(H_{\text{pr}} \otimes \mathbb{I}_r)^{\top}\mathbf{v} \neq \mathbf{0}_{r|\mathcal{N}(n)|}$ and

$$(H_{\text{tr}} \otimes \mathbb{I}_r)(H_{\text{pr}} \otimes \mathbb{I}_r)^{\top}\mathbf{v} = \mathbf{0}_{2K}. \tag{38}$$

Since $(H_{\text{pr}} \otimes \mathbb{I}_r)^{\top}\mathbf{v} \neq \mathbf{0}_{r|\mathcal{N}(n)|}$, there must exist $\mathbf{u}$-latent factors such that, for the same $\mathbf{v}$, $U^{\top}_{\mathcal{I}}(H_{\text{pr}} \otimes \mathbb{I}_r)^{\top}\mathbf{v} \neq \mathbf{0}_{|\mathcal{I}|}$. Suppose that $U_{\mathcal{I}}$ reflects these latent factors. Then, (38) implies

$$(H_{\text{tr}} \otimes \mathbb{I}_r)U_{\mathcal{I}}U^{\top}_{\mathcal{I}}(H_{\text{pr}} \otimes \mathbb{I}_r)^{\top}\mathbf{v} = \mathbf{0}_{2K}. \tag{39}$$

Let $\mathbf{v}' = (W_{\text{pr}}W^{\top}_{\text{pr}})^{-1}W_{\text{pr}}\mathbf{v}$. Let the $\mathbf{w}$-latent factors be defined such that $\mathbf{v}' \neq \mathbf{0}_{T_{\text{pr}}}$. Then, (39) implies

$$\begin{aligned} & W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r)U_{\mathcal{I}}U^{\top}_{\mathcal{I}}(H_{\text{pr}} \otimes \mathbb{I}_r)^{\top}\mathbf{v} = \mathbf{0}_{2K}, \\ \implies & W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r)U_{\mathcal{I}}U^{\top}_{\mathcal{I}}(H_{\text{pr}} \otimes \mathbb{I}_r)^{\top}W^{\top}_{\text{pr}}\mathbf{v}' = \mathbf{0}_{2K}. \end{aligned} \tag{40}$$

By Lemma 13, (40) implies that rowspace($W_{\text{pr}}(H_{\text{pr}} \otimes \mathbb{I}_r)U_{\mathcal{I}}$) $\not\subseteq$ rowspace($W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r)U_{\mathcal{I}}$), which implies that rowspace($\mathbb{E}[Z_{\text{pr},\mathcal{I}}|LF, A]$) $\not\subseteq$ rowspace($\mathbb{E}[Z_{\text{tr},\mathcal{I}}|LF, A]$). Therefore, then Assumption 5 does not hold, as claimed. □

Proposition 4 follows immediately from Lemma 23. First note that $B^{\text{tr},n} = B^{\text{tr}}[\mathcal{N}(n), :]$. Second, if colrank($B^{\text{tr},n}$) < $|\mathcal{N}(n)|$, then there exists a $\tilde{\mathbf{a}}$ such that columnspace($\tilde{B}^{\text{pr}}[\mathcal{N}(n), :]$) $\not\subseteq$ columnspace($B^{\text{tr}}[\mathcal{N}(n), :]$). By Lemma 23, if colrank($B^{\text{tr},n}$) < $|\mathcal{N}(n)|$, then there exists a target treatment $\tilde{\mathbf{a}}$ and latent factors LF such that Assumption 5 does not hold, as claimed.

D Proofs for Section 5

D.1 Proof of Proposition 5

Proof. The subspace inclusion assumption (SIA), or Assumption 5, requires that

$$\text{rowspace}(\mathbb{E}[Z_{\text{pr},\mathcal{I}}|LF, A]) \subseteq \text{rowspace}(\mathbb{E}[Z_{\text{tr},\mathcal{I}}|LF, A]).$$

By (36) and (37), this is equivalent to requiring that, for the given n and $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ and for every $i \in [T_{\text{pr}}]$, there exists some $\boldsymbol{\phi} \in \mathbb{R}^{T_{\text{tr}}}$ such that

$$\mathbf{e}_i^\top W_{\text{pr}}(H_{\text{pr}} \otimes \mathbb{I}_r)U_{\mathcal{I}} = \boldsymbol{\phi}^\top W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r)U_{\mathcal{I}}. \tag{41}$$

Under the assumption on latent factors, as long as $|\mathcal{I}^{(n)}| \geq r|\mathcal{N}(n)|$, then (41) holds if and only if there exists some $\boldsymbol{\phi} \in \mathbb{R}^{T_{\text{tr}}}$ such that

$$\mathbf{e}_i^\top W_{\text{pr}}(H_{\text{pr}} \otimes \mathbb{I}_r) = \boldsymbol{\phi}^\top W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r). \tag{42}$$

Therefore, subspace inclusion requires that $\text{rowspace}(W_{\text{pr}}(H_{\text{pr}} \otimes \mathbb{I}_r)) \subseteq \text{rowspace}(W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r))$.

To conclude the proof, we use several facts. First, $\text{rowspace}(W_{\text{pr}}(H_{\text{pr}} \otimes \mathbb{I}_r)) \subseteq \text{rowspace}(H_{\text{pr}} \otimes \mathbb{I}_r)$. Second, by Lemma 9, each W^j is has linearly independent columns almost surely (since $T_j \geq 2r$ by the second condition of TrainingTreatmentTest) and W_{tr} therefore also has linearly independent columns almost surely. As such, $\text{rowspace}(W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r)) = \text{rowspace}(H_{\text{tr}} \otimes \mathbb{I}_r)$ almost surely.

Therefore,

$$\begin{aligned}
&\text{rowspace}(H_{\text{pr}}) \subseteq \text{rowspace}(H_{\text{tr}}) \\
\iff &\text{rowspace}(H_{\text{pr}} \otimes \mathbb{I}_r) \subseteq \text{rowspace}(H_{\text{tr}} \otimes \mathbb{I}_r) \\
\implies &\text{rowspace}(W_{\text{pr}}(H_{\text{pr}} \otimes \mathbb{I}_r)) \subseteq \text{rowspace}(H_{\text{tr}} \otimes \mathbb{I}_r) \\
\iff &\text{rowspace}(W_{\text{pr}}(H_{\text{pr}} \otimes \mathbb{I}_r)) \subseteq \text{rowspace}(W_{\text{tr}}(H_{\text{tr}} \otimes \mathbb{I}_r)),
\end{aligned}$$

i.e., subspace inclusion holds if $\text{rowspace}(H_{\text{pr}}) \subseteq \text{rowspace}(H_{\text{tr}})$. This condition is equivalent to the first condition in TrainingTreatmentTest because because $H_{\text{pr}}^\top = \tilde{B}^{\text{pr}}[\mathcal{N}(n), :]$ and $\text{columnspace}(H_{\text{tr}}^\top) = \text{columnspace}(B^{\text{tr}}[\mathcal{N}(n), :])$. As such, given the two conditions in TrainingTreatmentTest, Assumption 5 holds almost surely. □

E Proofs for Section 6

E.1 Proof of Lemma 6

Recall that, for a given treatment $a \in [D]$, $B^{\text{tr}}(a) \in \{0,1\}^{N \times T_{\text{tr}}}$ and $\mathbf{b}^{\text{pr}}(a) \in \{0,1\}^N$ are defined such that their (i,t)-th elements are given by

$$B_{it}^{\text{tr}}(a) = \text{Ind}(A_{it}^{\text{tr}} = a) \qquad \text{and} \qquad \tilde{b}_i^{\text{pr}}(a) = \text{Ind}(\tilde{a}_i = a).$$

That is, the (i,t)-th entry of $B^{\text{tr}}(a)$ is 1 if and only if unit i at measurement t receives treatment a under the training treatments A^{tr}. Similarly, the i-th entry of $\tilde{\mathbf{b}}^{\text{pr}}(a)$ is 1 if and only if unit i is assigned counterfactual treatment a under $\tilde{\mathbf{a}}$. Further recall that

$$B^{\text{tr}} = [B^{\text{tr}}(1), B^{\text{tr}}(2), \ldots, B^{\text{tr}}(D)] \in \{0,1\}^{N \times T_{\text{tr}} D},$$

$$\tilde{B}^{\text{pr}} = [\tilde{\mathbf{b}}^{\text{pr}}(1), \tilde{\mathbf{b}}^{\text{pr}}(2), \ldots, \tilde{\mathbf{b}}^{\text{pr}}(D)] \in \{0,1\}^{N \times D}.$$

Before proving Lemma 6, we first introduce a lemma.

Lemma 24. *Under the experiment design in Section 6.1, $B^{tr}[i,:] = B^{tr}[j,:]$ if and only if units i and j have been assigned the same color.*

Proof. Observe that Step 3 in the experiment design identifies which units received certain colors and assigns those units a treatment other than 1. That is, for a given iteration ℓ in the for loop, every unit is assigned treatment 1 except for units of certain colors. Moreover, Step 3 never examines the same color twice. Therefore, since each unit i has only one color, $c_i^\ell \neq 1$ for exactly one value of ℓ, whose value is determined by the color given to unit i. One can conclude that $A^{\text{tr}}[i,:] = A^{\text{tr}}[j,:]$ if and only if unit i and unit j receive the same color. From the definitions of B^{tr}, it therefore follows that $B^{\text{tr}}[i,:] = B^{\text{tr}}[j,:]$ if and only if units i and j have been assigned the same color. □

We now provide a proof of Lemma 6.

Proof. We show that the two conditions in TrainingTreatmentTest hold when the training treatments A^{tr} are assigned as described in Section 6.1.

First requirement of TrainingTreatmentTest. We begin by proving that

$$\text{columnspace}(\tilde{B}^{\text{pr}}[\mathcal{N}(n),:]) \subseteq \text{columnspace}(B^{\text{tr}}[\mathcal{N}(n),:]),$$

for all possible $\tilde{B}^{\text{pr}}[\mathcal{N}(n),:]$ when A^{tr} is generated using the experiment design in Section 6.1. It suffices to prove that $B^{\text{tr}}[\mathcal{N}(n),:]$ has full row-rank. We make use of the following three facts.

First, the proposed procedure ensures that no two units in the same neighborhood ever receive the same color. This is guaranteed under TwoHopColoring, which returns a coloring on the graph $\mathcal{G}'$ that is created by connecting every node in $\mathcal{G}$ to its immediate *and* its two-hop neighbors. As such, for any unit i, no two units in its neighborhood $\mathcal{N}(i)$ share the same color because any two units in $\mathcal{N}(i)$ must be within each others' two-hop neighborhoods.

Second, $T_{\text{tr}} D \geq |\mathcal{N}(n)|$, i.e., there are at least as many columns in B^{tr} as there are rows. To see why, observe that, under the proposed procedure, $T_{\text{tr}} = T' \bar{r} D = \lceil \frac{\text{NumColors}}{D-1} \rceil \bar{r} D \geq \lceil \frac{|\mathcal{N}(n)|}{D-1} \rceil \bar{r} D$, where the inequality follows from the first fact, i.e., that no two units in $\mathcal{N}(n)$ share the same color.

Third, by Lemma 24, $B^{\text{tr}}[i,:] = B^{\text{tr}}[j,:]$ if and only if units i and j have been assigned the same color. However, by the first fact above, this cannot occur when $i, j \in \mathcal{N}(n)$. As such, $B^{\text{tr}}[\mathcal{N}(n),:]$ has $|\mathcal{N}(n)|$ distinct rows. Since $B^{\text{tr}}[\mathcal{N}(n),:]$ is a binary matrix *and* $B^{\text{tr}}[\mathcal{N}(n),:]$ has at least as many columns as rows, these rows must be linearly independent. In other words, $B^{\text{tr}}[\mathcal{N}(n),:]$ has full row-rank, as we sought to prove.

Second requirement of TrainingTreatmentTest. The second requirement holds by Step 4 of the experiment design in Section 6.1. □

E.2 Proof of Lemma 7

Proof. Recall that T_{tr} denotes the width of A^{tr}. Under the procedure in Section 6.1, the width of A^{tr} is given by $T' \bar{r} D$, where $T' = \lceil \frac{\text{NumColors}}{D-1} \rceil$. Therefore, $T_{\text{tr}} = \bar{r} D \lceil \frac{\text{NumColors}}{D-1} \rceil$. $T_{\text{tr}} \leq \frac{\bar{r} D (d^2 + D)}{D-1}$ follows from the facts that (i) any graph $\mathcal{G}''$ can be trivially colored using $\text{Degree}(\mathcal{G}'') + 1$ colors and (ii) the degree of $\mathcal{G}'$ is upper bounded by $d(d-1)$ by the definition of $\mathcal{G}'$ in Section 6 □

E.3 Tailoring experiment design to counterfactual treatments of interest

Recall that, by Lemma 6, the experimental design procedure in Section 6.1 produces a treatment schedule that passes TrainingTreatmentTest for *any* n and $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest under Assumptions 2 and 3 and $\bar{r} = r$. One can alternately tailor the treatment schedule to a specific n and $\tilde{\mathbf{a}}_{\mathcal{N}(n)}$ of interest. One need only ensure that columnspace$(\tilde{B}^{\mathrm{pr}}[\mathcal{N}(n), :]) \subseteq$ columnspace$(B^{\mathrm{tr}}[\mathcal{N}(n), :])$ is satisfied, as required by TrainingTreatmentTest. Such a modification is desirable because it might require fewer training samples.

To make this modification, all that would change in the procedure in Section 6 is the method TwoHopColoring. Specifically, one would form $\mathcal{G}'$ by connecting each unit i not to *every* one of its immediate and two-hop neighbors, but only those units j in the two-hop neighborhood for which $\tilde{a}_i \neq \tilde{a}_j$. That is, $(j, i) \in \mathcal{E}'$ if $((j, i) \in \mathcal{E}) \cup (\exists k \in [N] \setminus \{i, j\} : (j, k), (k, i) \in \mathcal{E})$ and $\tilde{a}_i \neq \tilde{a}_j$.

E.4 Proof of Proposition 8

Proof. Our goal is to apply Theorem 2. However, the conditions and assumptions of Theorem 2 differ from those used in Proposition 8. We proceed by showing that, under the assumptions in Proposition 8, we can recover the assumptions of Theorem 2 and obtain more precise estimates on the number of training samples T_{tr} and units N needed for finite-sample consistency.

We begin by characterizing the number of donors an ego-unit has under the conditions in Proposition 8. We then show that Assumptions 4 and 5 hold under the proposition conditions.

Number of donors. There are three requirements for a donor, as given in Definition 1. The first requirement is that a donor has the same number of neighbors as n. Since $\mathcal{G}$ is a d-regular graph, this requirement is automatically satisfied for all possible units.

The second requirement for a unit k to be a donor for unit n is that there exists a permutation π_k such that $A[\pi_k(\mathcal{N}(k)), \mathcal{T}_{\mathrm{tr}}] = A[\mathcal{N}(n), \mathcal{T}_{\mathrm{tr}}]$. By the experiment design procedure in Section 6.1, this second requirement is satisfied as long as n and k are assigned the same colors. By Lemma 12, there are at least $N - \Theta(\sqrt{N})$ ego-units for which there are at least $\sqrt{N}$ units that satisfy the second requirement. Let this set of ego-units be denoted by E.

Therefore, at least $\sqrt{N}$ units satisfy the first and second requirements of a donor unit. Suppose that these units are sub-sampled before checking whether they meet the third requirement of Definition 1. Specifically, suppose that exactly $\sqrt{N}$ of them are randomly chosen and the rest discarded. (This subsampling method is what we refer to as the "method of choosing donors" in the proposition statement.) Recall further that the third requirement for a unit k to be a donor is that $\mathbf{a}^{\mathrm{pr}}_{\pi_k(\mathcal{N}(k))} = \tilde{\mathbf{a}}_{\mathcal{N}(n)}$. By Lemma 11 and the subsampling condition, the number of units that satisfy the third requirement of Definition 1 for any ego-unit in $n \in E$ (and therefore are considered "donors" for n) is:

$$|\mathcal{I}^{(n)}| = \Theta\left(\frac{\sqrt{N}}{D^{d+1}}\right), \tag{43}$$

with high probability. We can therefore replace $|\mathcal{I}^{(n)}|$ in Theorem 2 with $\sqrt{N}/D^{d+1}$, noting that this substitution holds with high probability for $N - \Theta(\sqrt{N})$ ego-units, as stated in Proposition 8.

Assumption 4. Assumption 4 is required in Theorem 2. By Lemma 10 and Assumption 11, Assumption 4 holds if there are at least $r|\mathcal{N}(n)|$ donors. In a d-regular graph, this translates to needing at least $r(d+1)$ donors. Therefore, by (43), we require that

$$\begin{aligned} \frac{\sqrt{N}}{D^{d+1}} &= \Omega(r(d+1)) \\ \implies N &= \Omega\left(r^2 d^2 D^{2d+2}\right), \end{aligned}$$

for Assumption 4 to hold, which matches the condition stated in Proposition 8.

Combining results. By Assumption 11, $r = \bar{r}$, and Lemma 6, Assumption 5 holds. Therefore, both Assumptions 4 and 5 hold under the conditions given in Proposition 8. By Assumption 11, Assumption 7 holds. By Lemma 22, Assumption 8 holds. Furthermore, the condition in Proposition 8 that there are at least $\frac{rD(d^2+D)}{D-1}$ comes from the fact that the experiment design in Section 6.1 can always be carried out with at least $\frac{rD(d^2+D)}{D-1}$ training measurements. Combining these results with Theorem 2 gives

$$\begin{aligned} &\left|\widehat{\mathrm{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \mathrm{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})\right| \\ &\quad = O_P\left(\log\left(\frac{T_{\mathrm{tr}} N}{D^{d+1}}\right)\left(\frac{r_{\mathrm{tr}}^{3/4}}{(\xi''')^{3/2} T_{\mathrm{tr}}^{1/4}} + \frac{r_{\mathrm{tr}}^2}{(\xi''')^4}\max\left(\frac{1}{\sqrt{T_{\mathrm{tr}}}}, \frac{\sqrt{D^{d+1}}}{N^{1/4}}, \frac{N^{1/4}}{D^{(d+1)/2} T_{\mathrm{tr}}^{3/2}}\right)\right)\right). \end{aligned}$$

Note that, by Lemma 22, $\xi' = (1 + 4rd^3)^{-1/2}$ and $\xi'' = (9r(d+1))^{-1}$. Grouping terms and noting that $r_{\mathrm{tr}} \leq r(d+1)$ gives the result. □

F Simulation details

In this section, we provide full details behind the simulations produced in Section 7 and provide additional plots.

Setting. Let $\mathcal{G}$ be a regular graph with degree d, and let the treatments be binary, i.e., $D = 2$. In each of the experiments below, we will indicate the graph degree.

At the start of each simulation, the latent factors $\mathbf{u}_{k,n}$ and $\mathbf{w}_{1,a}$ are drawn uniformly at random from $\left[-\frac{1}{\sqrt{r(d+1)}}, \frac{1}{\sqrt{r(d+1)}}\right]^r$. The latent factors $\mathbf{w}_{\tau,a}$ are generated as random walk for $\tau > 1$, where each random step of the random walk is also drawn uniformly at random from $\left[-\frac{1}{\sqrt{r(d+1)}}, \frac{1}{\sqrt{r(d+1)}}\right]^r$.

Our experiments use a simple donor-finding algorithm. In particular, instead of searching for donors over all possible permutations π_j, as defined in Definition 1, we fix an ordering of units and restrict ourselves to the identity permutation $\pi_j(i) = i$. In this way, the number of donors reported in our experiments is lower than the actual number of available donors.

Predictions. Figure 6(a) shows an example of the estimates that NSI produces, where $\mathcal{G}$ is a ring graph ($d = 2$) with $N = 1000$ units, $\epsilon_{\tau,i}^{(\mathbf{c}_{\mathcal{N}(i)})} \sim \mathcal{N}(0, 0.1)$, $r = 2$, $T_{\mathrm{tr}} = 150$, and $T_{\mathrm{pr}} = 50$. Let the training treatments be assigned according to the experiment design in Section 6.

The prediction treatments $\mathbf{a}^{\text{pr}}$ for $\tau \in \mathcal{T}_{\text{pr}}$ are drawn uniformly at random from $[D]^N$. The plot is generated for a given target treatment $\tilde{\mathbf{a}}$ of interest. As discussed in Section 2, we assume that the prediction and target treatments are constant across $\mathcal{T}_{\text{pr}}$. Consider the bottom plot and a specific unit n. The solid line gives the ground truth potential outcomes for unit n across measurements $t \in [200]$. The estimates produced by NSI are marked by asterisks $*$, with the 95 percent confidence interval in gray. The measurements to the left of the vertical line (i.e., in blue and green) correspond to the training set $\mathcal{T}_{\text{tr}}$ while those to the right (i.e., in red and orange) correspond to the prediction set $\mathcal{T}_{\text{pr}}$. The top plot gives the spectrum $\{\hat{s}_\ell\}_{\ell=1}^{q}$ produced in Step 1 of Section 3.2, where the vertical line marks the singular value threshold κ that is chosen in Step 1. In all of our experiments, κ is chosen using a knee-point (otherwise known as elbow-point) method. As shown in the bottom plot, the predictions closely match the ground-truth values. As shown on top, 6 components are used to construct the estimates. Since the network-adjusted rank is 6 (the product of $r = 2$ and $|\mathcal{N}(n)| = 3$), that NSI uses 6 components explains why its estimates are fairly accurate.

Further examples of the estimates NSI produces are given at the end of this section.

Consistency and asymptotic normality. Figure 6(b) verifies that the NSI estimates are consistent and asymptotically normal. Specifically, we let $\mathcal{G}$ be a ring graph (i.e., $d = 2$) with $N = 1000$ units, $\epsilon_{\tau,i}^{(\mathbf{c}_{\mathcal{N}(i)})} \sim \mathcal{N}(0, 0.1)$, $r = 2$, $T_{\text{tr}} = 150$, and $T_{\text{pr}} = 50$. For each simulation, we randomly generate the latent factors in the same way as described above for Figure 6(a). We ran 500 simulations, then computed the NSI residuals $(\widehat{\text{IPO}}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}) - \text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)}))$ for 50 units in $[N]$ and across all possible counterfactual treatments for each unit. By all possible counterfactual treatments, we used NSI to estimate $\text{IPO}(n, \tilde{\mathbf{a}}_{\mathcal{N}(n)})$ for $\tilde{\mathbf{a}}_{\mathcal{N}(n)} = (1, 0, 0)$, $\tilde{\mathbf{a}}_{\mathcal{N}(n)} = (0, 1, 0)$, $\tilde{\mathbf{a}}_{\mathcal{N}(n)} = (1, 1, 0)$, and so on. The rest of setup is identical to that used for Figure 6(a).

Figure 6(b) gives a histogram of the NSI residuals. A Gaussian distribution is fit to the residuals and given by the red line. This result verifies the consistency and asymptotic normality of NSI.

MSE trends. Figure 6(c) summarizes the performance of NSI across different parameters. The performance is given by the mean-squared error (MSE) across the prediction measurements $\mathcal{T}_{\text{pr}}$, averaged across 50 units. Each group of bars gives the MSE for regular graphs of degree 2, 4, 6, and 8, as indicated on the x-axis. Within each group of bars, the left (blue) bars are for $N = 1000$, $T_{\text{tr}} = 100$, $T_{\text{pr}} = 50$; the middle (red) bars for $N = 1000$ and $T_{\text{tr}} = T_{\text{pr}} = 50$; and the right (yellow) bars for $N = 500$ and $T_{\text{tr}} = T_{\text{pr}} = 50$. Each bar is the average of 200 simulations with $\epsilon_{\tau,i}^{(\mathbf{c}_{\mathcal{N}(i)})} \sim \mathcal{N}(0, 0.1)$, and $r = 2$. The training treatments $\mathbf{a}^{\tau} = \mathbf{a}^{\text{tr}}$ for $\tau \in \mathcal{T}_{\text{tr}}$ are assigned randomly and remain constant across $\mathcal{T}_{\text{tr}}$. The prediction treatments are also generated randomly and remain constant across $\mathcal{T}_{\text{pr}}$. In the experiments for Figure 6(c), we compute the MSE for the synthetic control setting, that is, $\tilde{\mathbf{a}} = \mathbf{a}^{\text{tr}}$ for all $\tau \in \mathcal{T}_{\text{pr}}$. Note that studying the synthetic control setting does not bias the MSE, as the method we propose is agnostic to the counterfactual treatment of interest as long as TrainingTreatmentTest is passed. We use the synthetic control setting to simplify the computation, as TrainingTreatmentTest is always passed under synthetic control.

As expected, the MSE typically increases with degree, fewer nodes, and less training time.

Comparing to other estimators. We also compare the NSI estimator to two others: the SI estimator (Agarwal et al. 2020b) and a baseline estimator. The SI estimator is a method similar to NSI, but SI assumes that there is no spillover and therefore does not account for network interference. 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. We compare the

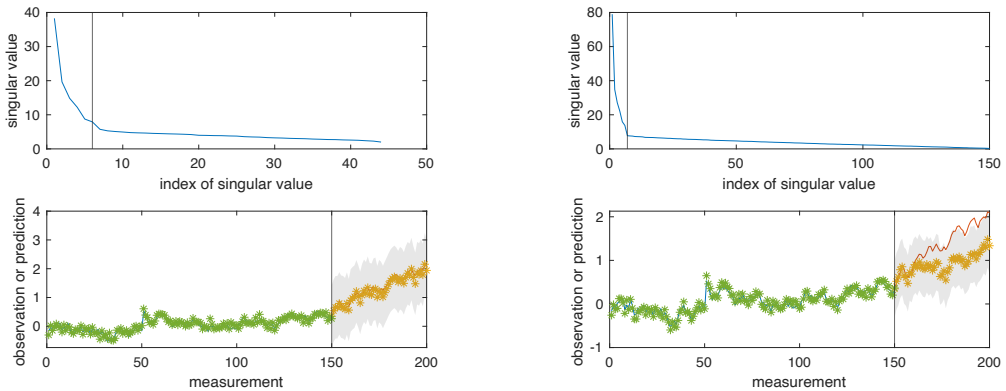
estimators for a ring graph under the same parameters as those used in Figure 6(b) averaging across 200 simulations, 50 units, and all possible counterfactual treatments.

The NSI numbers are given for $\kappa \geq d+1 = 3$ (i.e., estimates for which the elbow points are lower than 3 are removed). This heuristic is consistent with Theorems 2-3, which hold when $\kappa = r_{\text{tr}}$. That is, the NSI estimates are consistent and asymptotically normal when the number of components is at least r_{tr}. Since we do not know r_{tr} *a priori*, we can lower bound it and discard NSI estimates that are produced using fewer components than the lower bound. From (1), $r_{\text{tr}} \leq r|\mathcal{N}(n)|$, which gives a lower bound $r_{\text{tr}} \geq |\mathcal{N}(n)|$ as long as $r \geq 1$. Therefore, we can discard NSI estimates that are produced using fewer than $d+1$ components in a d-regular graph. Similarly, the SI estimates that are produced using fewer than 1 component are also discarded since r is lower bounded by 1. This is built on precisely the same intuition as that given for NSI's $d+1$ lower bound; the only difference is that the effective degree for SI is 0 because SI ignores network interference. No donors are discarded for the baseline estimator.

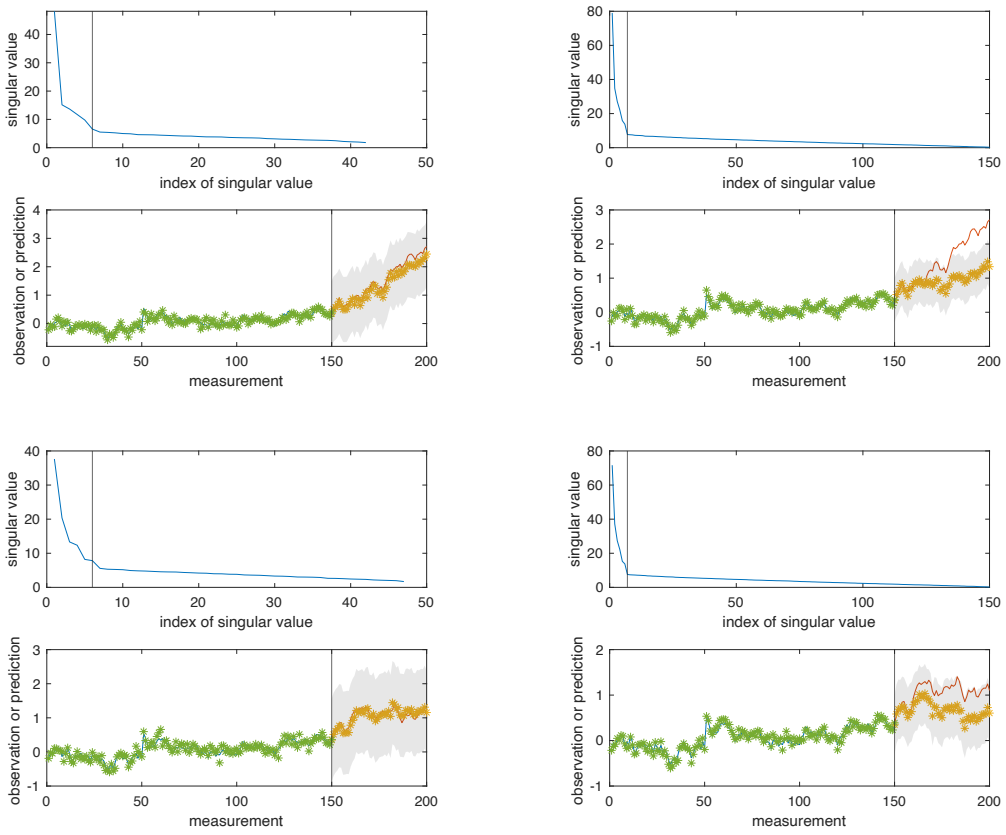
The MSEs and R-squared values for the NSI estimator, SI estimator, and baseline estimators are, respectively, **(0.1174, 0.8735)**, **(0.2310, 0.8149)**, and **(3.398, -2.957)**. Both the NSI and baseline estimators use donor sets that contain, on average, 41 units. The SI estimator uses donor sets with, on average, 166 units. As such, even though the SI estimator has more donors, the performance of NSI is better than that of SI, which is better than that of the baseline estimator.

Additional simulations. Below, we illustrate the results produced by NSI and SI. The setup is the same as that given in Figure 6(a). For all the plots below, the unit of interest and simulation is fixed. The counterfactual treatment of interest varies across the rows. In each row, the left plot gives the results for NSI, and the right plot gives the results for SI. Recall that SI and NSI differ in that NSI accounts for spillover effects while SI does not.

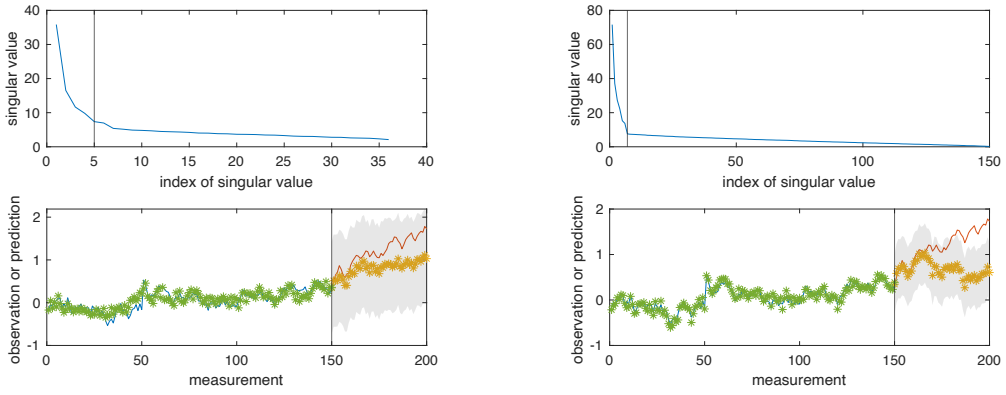
We can make several observations from the two plots directly below. First, the number of donors is greater for SI than NSI (as can be seen by the range of the x-axis of the top plot on the left versus that of the top plot on the right). Second, SI typically uses more components to construct its estimates as well (as can be seen by the number of components to the left of the vertical lines of the top plots). Both these trends hold true across the examples. Third, both NSI and SI perform well across the training set. However, SI performs poorly across the prediction set, indicating that it suffers in the presence of spillover. Even so, the confidence interval for SI is smaller than that for NSI, i.e., SI is overconfident in its estimates. Fourth, the number of components used by NSI (as marked the vertical line in the top-left plot) is 6, which matches the network-adjusted rank of 6 (the product of $r = 2$ and $|\mathcal{N}(n)| = 3$ for a ring graph) and suggest that NSI would perform well. This is confirmed by the fact that the estimates are close to the ground truth.



The next two pairs of plots show similar results. NSI performs well compared to SI, which is overconfident in its estimates. The number of components used by NSI is 6 in both cases, which suggests that it will produce good estimates.



The following two plots illustrate an instance for which NSI performs poorly. Indeed, the number of components used by NSI is only 5. As such, one would not expect NSI to do well. However, SI is still overconfident in its estimates (as there are ground truth values lie outside the gray area) whereas NSI's confidence interval covers the ground-truth potential outcomes.



It is worth noting that in many cases (including the one directly above), there seems to be leftover spectral energy beyond the κ chosen by the knee point method. As such, the NSI estimates could be improved by letting $\kappa = 6$. The fact that $\kappa = 5$ is due to the automated knee point

method that we utilize, and it motivates the incorporation of human oversight—that, when κ is too small and there is leftover spectral energy, κ can be increased. Generally speaking, one can increase κ until the training MSE is small, then apply that κ to the predictions.