Graph Neural Networks for Causal Inference Under Network Confounding

Mc ae P Leun[†] Pante s Loupos[‡]

Marc

Abstract. [illegible]

1 Introduction

Treatment assignment is said to be unconfounded if it is as good as random within subpopulations of observationally equivalent units. In settings where the stable unit treatment value assumption (SUTVA) is plausible, units with identical covariates are naturally considered observationally equivalent. However, when units are connected through a network, they may differ on other observed dimensions that may confound causal inference if SUTVA is violated and interference is mediated by the network. These dimensions include, for example, the number of type-$\mathbf{x}$ neighbors, the number of type-$\mathbf{x}$ neighbors with $\mathbf{m}$ neighbors of type $\mathbf{y}$, and so on through higher-order neighbors.

Existing formulations of unconfoundedness only utilize a small subset of these dimensions. For example, a common set of controls used in the literature is the vector consisting of own covariates, number of neighbors, and average covariates of neighbors. This choice may be difficult to justify in practice due to a lack of behavioral models of selection. Neighbor covariates may influence selection into treatment in more complex ways not adequately captured by the mean. Furthermore, this choice of controls implies no confounding from higher-order neighbors, which we show rules out economically interesting sources of interference in treatment selection, such as endogenous peer effects.

In this paper, we study estimation and inference for treatment and spillover effects under a fully nonparametric formulation of unconfoundedness motivated by a model of selection. To allow for peer effects, selection is governed by the reduced form of a simultaneous-equations model, which is a function of the entirety of $\boldsymbol{X}$, the matrix of all units' covariates, and $\boldsymbol{A}$, the network adjacency matrix. As a result, it is not generally possible to summarize confounding by a simple low-dimensional function of these objects. Our unconfoundedness condition therefore considers units observationally equivalent if they occupy identical positions in the network, meaning that they match on all observed neighborhood and higher-order neighborhood dimensions.

Existing methods that rule out complex forms of interference in selection may result in biased estimates of treatment and spillover effects. For example, consider the causal effect of vaccine adoption on illness. With peer effects in vaccine adoption, vaccinated individuals tend to have more vaccinated direct and indirect social contacts, and a simple comparison of adopters and nonadopters may overstate vaccine

GNNs for Network Confounding

utilize the model of approximate neighborhood interference (ANI) proposed by Leung (2022a), which posits that interference in the outcome stage decays with network path distance. Leung shows that ANI allows for endogenous peer effects but focuses on a setting with randomized assignment. In observational settings, it stands to reason that peer effects in selection may be a possibility. We theref

We provide conditions under which the doubly robust estimator is approximately normally distributed as the network size grows large. This type of result is well known for i.i.d. data (e.g. Farrell, 2018), but it is nontrivial to extend to our setting since we allow for a complex form of network dependence. For example, asymptotically linearizing the doubly robust estimator requires a new argument due to dependence, and application of an appropriate CLT requires verification of a high-level weak dependence condition under a nonparametric model with outcome and selection stages both governed by simultaneous-equations models. For inference, we utilize a network HAC estimator due to Kojevnikov et al. (2021) and propose a new bandwidth that adjusts for estimation error in the first-stage machine learners.

We substantiate the theory in a simulation study and empirical application to microfinance diffusion. The simulations demonstrate that the use of GNNs can substantially reduce bias relative to conventional choices of network controls even with shallow architectures. The empirical illustration revisits the microfinance diffusion application of He and Song (2024). We show how our estimands can capture complementary aspects of diffusion relative to their "average diffusion at the margin" measure. Our theoretical framework allows for more complex diffusion processes without requiring the econometrician to prespecify the maximum number of within-period rounds of diffusion. Finally, by including richer controls that account for network confounding, we find more attenuated diffusion effects.

1.2 Related Literature

There is a large literature on interference, much of which focuses on randomized control trials (e.g. Athey et al., 2018; Li and Wager, 2022; Toulis and Kao, 2013). We contribute to a growing recent literature on unconfoundedness, much of which operates in a partial interference setting where units are partitioned into disjoint groups with no interference across groups (e.g. Liu et al., 2019; Qu et al., 2022).

Studying a network interference setting, Veitch et al. (2019) propose to use "node embeddings" as network controls, which are learned functions of the graph. Since node embeddings can be obtained from a variety of methods, there remains the issue of justifying a particular choice of network controls. GNNs can be interpreted as a method of estimating node embeddings (see §3), and our behavioral model provides justification for their use. We defer to §2.1 a more detailed review of the literature

i's n o and the elements of the same set of K 1 as i's o neighbors. A unit i's is $n(i, 1$, the number of neighbors.

2 Setup

Let $\mathcal{N}_n$ $1, \dots, n$ be the set of units connected through the network A. Each unit i $\mathcal{N}_n$ is endowed with unobservables $(\varepsilon_i, \nu_i$ $\mathbb{R}^{d_\varepsilon}$ $\mathbb{R}^{d_\nu}$ and observables X_i $\mathbb{R}^{d_x}$. The model primitives determine outcomes and treatments according to

$$Y_i \quad g_n(i, D, X, A, \varepsilon \quad \text{and} \quad D_i \quad h_n(i, X, A, \nu\ , \tag{1}$$

respectively, where X $(X_i{}_{i=1}^{n}$ is the matrix with ith row equal to X_i'; Y, D, ε, and ν are similarly defined; and $(g_n, h_n{}_{n\in\mathbb{N}}$ is a sequence of function pairs such that each $g_n($ has range $\mathbb{R}$ and $h_n($ has range $0, 1$. The econometrician observes $(Y, D, X, A$. Our analysis treats $(A, X, \varepsilon, \nu$ as random, but the asymptotic theory in §4 conditions on $(X, A$ to avoid imposing additional assumptions on its dependence structure.

We view the timing of the model as follows. First, nature draws the primitives $(A, X, \varepsilon, \nu$. Next, units select into treatment, potentially based on other units' decisions, and $h_n($ is the reduced-form outcome of that process. Finally, $g_n($ is the reduced form of the subsequent process that generates outcomes. Because $g_n($ and $h_n($ may depend on the primitives of all units, the setup allows Y_i and D_i to be outcomes of simultaneous-equations models with endogenous peer effects, as shown in the next examples.

E a p 1 (Linear-in-Means) Consider the outcome model

$$Y_i \quad \alpha + \beta \frac{\check{r}_{j=1}^{n} A_{ij} Y_j}{\check{r}_{j=1}^{n} A_{ij}} + \frac{\check{r}_{j=1}^{n} A_{ij} Z_j'}{\check{r}_{j=1}^{n} A_{ij}} \gamma + Z_j' \delta + \varepsilon_i,$$

where Z_i $(D_i, X_i'\,'$ (Manski, 1993). The coefficient β captures endogenous peer denote the row-normalized adjacency matrix and $\mathbf{1}$ the n-dimensional vector of ones, if $\boldsymbol{A}$ is connected, the reduced form of the model can be written in matrix form as

$$\boldsymbol{Y} = \frac{\alpha}{1-\beta}\mathbf{1} + \boldsymbol{Z}\delta + \gamma\beta\sum_{k=0}^{\infty}\beta^k\tilde{\boldsymbol{A}}^{k+1}\boldsymbol{Z} + \sum_{k=0}^{\infty}\beta^k\tilde{\boldsymbol{A}}^k\boldsymbol{\varepsilon}.$$

This characterizes Y_i as a function $g_n(i,\boldsymbol{D},\boldsymbol{X},\boldsymbol{A},\boldsymbol{\varepsilon})$.

Example 2 (Binary Game) Consider the binary analog of Example 1 but for selection into treatment:

$$D_i = \mathbf{1}\left\{\alpha + \beta\frac{\sum_{j=1}^n A_{ij}D_j}{\sum_{j=1}^n A_{ij}} + \frac{\sum_{j=1}^n A_{ij}Z_j'}{\sum_{j=1}^n A_{ij}}\gamma + Z_i'\delta + \nu_i > 0\right\}. \tag{2}$$

Unlike Example 1, there may exist multiple equilibria. The equilibrium selection mechanism is a reduced-form mapping from the primitives $(\boldsymbol{X},\boldsymbol{A},\boldsymbol{\nu})$ to outcomes $\boldsymbol{D}$ and therefore characterizes D_i as a function $h_n(i,\boldsymbol{X},\boldsymbol{A},\boldsymbol{\nu})$. This formulation corresponds to a game of complete information. In a game of incomplete information, as modeled by Xu (2018) for instance, a unit i's information set is $(\nu_i,\boldsymbol{X},\boldsymbol{A})$. Here an analog of (2) holds with each D_j replaced with $\sigma_j(\boldsymbol{X},\boldsymbol{A})$, the equilibrium belief that $D_j = 1$. This characterizes D_i as a function $h_n(i,\boldsymbol{X},\boldsymbol{A},\nu_i)$.

Example 3 (Diffusion) He and Song (2024) study the following two-period diffusion model. Let D_i denote i's decision to adopt microfinance in period 0 and Y_i its decision in period 1. Their equations (2.4) and (3.6) posit that

$$Y_i = g_n(\boldsymbol{D}_{\mathcal{N}(i,K)},\varepsilon_i) \quad \text{and} \quad D_i = \mathbf{1}\{W_i'\gamma > \nu_i\},$$

where W_i is a known function of $(\boldsymbol{X},\boldsymbol{A})$ and K is the maximum distance that adoption decisions can diffuse through the network between periods 0 and 1. We provide a more detailed comparison of our models in §7.

Given specification (1), we define potential outcomes as

$$Y_i(\boldsymbol{d}) = g_n(i,\boldsymbol{d},\boldsymbol{X},\boldsymbol{A},\boldsymbol{\varepsilon}).$$

Confounding may arise first because $Y_i(\boldsymbol{d})$ is potentially correlated with D_i due to the high-dimensional observables $(\boldsymbol{X},\boldsymbol{A}$ and second because of dependence between unobservables that drive outcomes $\boldsymbol{\varepsilon}$ and those that drive selection $\boldsymbol{\nu}$. We restrict the second source of confounding.

Assu pt on 1 (Unconfoundedness) *o ny* **n** ℕ, $\boldsymbol{\varepsilon}$ $\boldsymbol{\nu}$ $\boldsymbol{X},\boldsymbol{A}$

As discussed below, unconfoundedness conditions used in the existing literature additionally limit the first source of confounding to known summary statistics of $(\boldsymbol{X},\boldsymbol{A}$. Ours is analogous to standard formulations of unconfoundedness under SUTVA (ε_i ν_i X_i) since we do not impose further restrictions on the nature of observed confounding.

Because the econometrician only observes a single network, a large-sample theory requires restrictions on interference in order to obtain some form of weak dependence. We next specify a nonparametric model of decaying interference that accommodates the previous examples. For any S $\mathcal{N}_n$, let $\boldsymbol{D}_S$ $(D_i$ $_{i\in S}$, and similarly define $\boldsymbol{X}_S$ and other such submatrices. Let $\boldsymbol{A}_S$

counterfactual s-neighborhood outcomes. The error from approximating the observed outcome with the s-neighborhood counterfactual is bounded by $\gamma_n($s , which decays with the neighborhood radius s. This formalizes the idea that Y_i is primarily determined by units relatively proximate to i, so that the further a unit is from i, the less it influences i's outcome. The second equation imposes the analogous requirement on D_i.

E a p For the linear-in-means model in Example 1, an argument similar to Proposition 1 of Leung (2022a) shows that (3) holds with $\sup_n \gamma_n(s \quad C|\beta|^s$ for some $C \quad 0$. For the binary game in Example 2, an argument similar to Proposition 2 of Leung (2022a) establishes (4) with $\sup_n \eta_n(s$ decaying at an exponential rate with s. Finally, for the He and Song (2024) diffusion model in Example 3, Y_i only depends on D through $D_{\mathcal{N}(i,K)}$, so (3) holds with $\gamma_n(s \quad c\,1\; s \quad K$ for some universal constant c. In their empirical application, they use own covariates as controls, so $W_i \quad X_i$, in which case (4) holds with $\eta_n(s \quad 0$ for all s.

2.1 Related Literature

The standard SUTVA model and unconfoundedness condition correspond to

$$Y_i \quad g(D_i, X_i, \varepsilon_i \quad \text{and} \quad \varepsilon_i \quad D_i \quad X_i. \tag{5}$$

To generalize this setup to allow for network interference, the typical approach in the existing literature is as follows. Define

$$T_i \quad f_n(i, D, A \quad \text{and} \quad W_i \quad q_n(i, X, A \tag{6}$$

where $f_n($ and $q_n($ are known vector-valued functions. The n
(Manski, 2013) or o u n (Aronow and Samii, 2017,t9(i)-0.34923.463(e)-297.001000cmBT 2T

o oo n n model and unconfoundedness condition

$$Y_i \quad g(T_i, W_i, \varepsilon_i \quad \text{and} \quad \varepsilon_i \quad T_i \quad W_i, \tag{7}$$

which is a direct generalization of (5) (Emmenegger et al., 2022; Forastiere et al., 2021; Ogburn et al., 2022). Here T_i entirely summarizes interference while W_i summarizes confounding.

Common examples of T_i and W_i are

$$T_i \quad \Big(D_i, \sum_{j=1}^{n} A_{ij} D_j \Big) \quad \text{and} \quad W_i \quad \Big(X_i, \sum_{j=1}^{n} A_{ij}, \frac{\sum_{j=1}^{n} A_{ij} X_j}{\sum_{j=1}^{n} A_{ij}} \Big). \tag{8}$$

This choice of T_i implies that Y_i depends on $\boldsymbol{D}$ only through two statistics: own treatment and the number of treated neighbors. Variation in the first component identifies a direct treatment effect and variation in the second a spillover effect. Like most exposure mappings used in the literature, this only depends on $\boldsymbol{D}_{\mathcal{N}(i,1)}$, so the outcome model (7) implies no interference beyond the 1-neighborhood. Likewise, this choice of W_i implies no confounding beyond 1-neighborhood covariates.

More generally, one could restrict the outcome model to depend only on the K-neighborhood treatments $\boldsymbol{D}_{\mathcal{N}(i,K)}$ for some fixed threshold K. As shown by Leung (2022a), this rules out economically interesting forms of interference such as endogenous peer effects, which motivates the ANI condition (3). Furthermore, (7) assumes the econometrician can correctly specify the summary statistic T_i in the outcome model, which may be difficult to justify (Sävje, 2024).

Whereas Leung (2022a) and Sävje (2024) focus on randomized experiments, we study observational data on economic agents that choose to select into treatment. It then becomes important to specify a behavioral model rationalizing the choice of controls W_i. Sánchez-Becerra (2022) is the first to provide such a model. Under neighborhood interference (7) and an exposure mapping similar to (8), he shows that it is sufficient to set $W_i \quad X_i$, that is, to solely control for own covariates. Since much of the literature utilizes controls such as (8), this raises the question of what model of selection justifies their use or more broadly the use of "network controls" that depend more generally on $\boldsymbol{X}$ and $\boldsymbol{A}$.

Our model (1) provides an answer. The presence of complex interference in both the outcome and treatment stages induces selection on $(\boldsymbol{X}, \boldsymbol{A}$, so that it is generally

with and without at least one treated neighbor, which captures a spillover effect. For $\mathbf{t}$ $(1,0$ and $\mathbf{t}'$ $(0,0$, it compares the average outcomes of treated and untreated units with no treated neighbors, which captures a treatment effect. For overlap, we need to exclude units with zero degree since a treated neighbor occurs with probability zero for such units. This is accomplished by choosing $\mathcal{M}_\mathbf{n}$ to be the subset of units whose degree $\mathbf{n}(\mathbf{i},1$ $|\mathcal{N}(\mathbf{i},1\ |$ lies in some desired set excluding zero. That is, choose some $\mathbf{\Gamma}$ $\mathbb{R}_{+|}$ 0 and

$$\mathcal{M}_\mathbf{n} \quad \mathbf{i}$$

$$(\qquad _{|}\, ^{+} \qquad | \qquad (\quad | \quad | \quad |$$

can be a complex function of $(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\nu}$, and both $\mathbf{T_i}$ and $\mathbf{Y_i}$ can be complex functions of $\boldsymbol{D}$, which makes it difficult to characterize the dependence structure necessary for the application of a central limit theorem without additional structure.

Identification results in Leung (2024) provide conditions under which $\boldsymbol{\tau}(\mathbf{t}, \mathbf{t}'$ has a causal interpretation. The focus of our paper is estimation, so we provide only a brief discussion here. Under the neighborhood interference model (7), $\boldsymbol{\tau}(\mathbf{t}, \mathbf{t}'$ has a transparent causal interpretation. In settings where (7) fails to hold, Leung (2024) shows that $\boldsymbol{\tau}(\mathbf{t}, \mathbf{t}'$ retains a causal interpretation under restrictions on interference either in potential outcomes or selection into treatment. For example, suppose treatment adoption follows a nonparametric game of incomplete information where $\boldsymbol{\nu}_\mathbf{i}$ is unit $\mathbf{i}$'s private information, so that $\mathbf{D_i} \quad \mathbf{h_n}(\mathbf{i}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\nu}_\mathbf{i}$ (see Example 2). If private information is independent across units conditional on $(\boldsymbol{X}, \boldsymbol{A}$, as is typically assumed in structural analyses of the model (e.g. Lin and Vella, 2021; Xu, 2018), then by Theorem 1 of Leung (2024) $\boldsymbol{\tau}(\mathbf{t}, \mathbf{t}'$ can be written as a non-negatively weighted average of certain unit-level effects.

Returning to the vaccine adoption example in §1, recall that,AduAexAerelct where

$$\hat{\tau}_i(t, t' \quad \frac{1 \; T_i \quad t \; (Y_i \quad \hat{\mu}_t(i, X, A}{\hat{p}_t(i, X, A} + \hat{\mu}_t(i, X, A$$

$$\frac{1 \; T_i \quad t' \; (Y_i \quad \hat{\mu}_{t^1}(i, X, A}{\hat{p}_{t^1}(i, X, A} \quad \hat{\mu}_{t^1}(i, X, A \; .$$

t To estimate

3.1 Architecture

The standard GNN architecture consists of nested, paramete

aggregation no longer holds when the support of $\mathbf{X}_\mathbf{i}$ is uncountable, so using multiple aggregators can result in more powerful architectures (Corso et al., 2020).

For an example of $\Gamma($, let $\boldsymbol{\mu}($ $,\boldsymbol{\sigma}($ $,\Sigma($ $,\min($ $,$ and $\max($ be respectively the mean, standard deviation, sum, min, and max functions, defined component-wise for multisets of vectors. Then setting $\Gamma($ $\quad\Gamma_1($ for

$$\Gamma_1(\qquad \acute{} \; \boldsymbol{\mu}(\quad \boldsymbol{\sigma}(\quad \Sigma(\quad \min(\quad \max(\quad \bar{}$$

results in an architecture utilizing five aggregation functions.

The authors combine multiple aggregators with "degree scalars" that multiply each aggregation function by a function of the size of the multiset input $\mathbf{n}(\mathbf{i},1$. The simplest example is the identity scalar, which maps any multiset to unity. This trivially multiplies each aggregation function in $\Gamma_1($ above, but it is useful to consider non-identity scalars. Let $| \, |$ be the function that takes as input a multiset and outputs its size. Corso et al. (2020) define logarithmic amplification and attenuation scalers

$$\mathbf{S}(\, ,\boldsymbol{\alpha} \quad \hat{} \; \frac{\log(| \; | \; +1}{\boldsymbol{\delta}} \; \dot{}^{\,\boldsymbol{\alpha}} \, , \quad \boldsymbol{\delta} \quad \frac{1}{\mathbf{n}} \, \ddot{\mathbf{y}}_{\mathbf{i}=1} \log \, \tilde{} \, \ddot{\mathbf{y}}_{\mathbf{j}=1} \mathbf{A}_{\mathbf{ij}} \; +1 \; \grave{} \, , \quad \boldsymbol{\alpha} \quad | \quad 1,1 \, .$$

The choice of $\boldsymbol{\alpha}$ defines whether the scalar "amplifies" ($\boldsymbol{\alpha}$ $\quad 1$) or "attenuates" ($\boldsymbol{\alpha}$ $\quad 1$) the aggregation function, and $\boldsymbol{\alpha}$ $\quad 0$ is the identity scalar. The purpose of the logarithm is to prevent small changes in degree from amplifying gradients in an exponential manner with each successive GNN layer. Thus, an aggregation function that augments $\Gamma_1($ with logarithmic amplification and attenuation is

$$\Gamma_2(\qquad \acute{} \; \mathbf{S}($$

function □(:

$$\hat{\mathbf{f}}_{\text{GNN}} \qquad \underset{\mathbf{f}\in\mathcal{F}_{\text{GNN}}(\mathbf{L})}{\text{argmin}} \; {}_{\mathbf{i}\in}^{\ddot{\mathbf{y}}}$$

As discussed in §3.1, any $\mathbf{f}$ $\mathcal{F}_{\mathrm{GNN}}(\mathbf{L}$ is invariant, so by using GNNs to esti-that does not depend on i, so evaluating i's propensity score is now only a matter of supplying the correct i-specific inputs $(\pi_{\mathsf{i}}(X\ ,\pi_{\mathsf{i}}(A\ \ .$

We close this section with a result demonstrating that invariance is an extremely weak requirement. In particular, it holds under minimal exchangeability conditions on the structural primitives.

ropos t on ^{1} *u o o ny* n $\mathbb{N}$ n *u on* π,

$$\mathsf{f_n}(\mathsf{i}, D, A \qquad \mathsf{f_n}(\pi(\mathsf{i}\ , \pi(D\ , \pi(A\ \ ,$$

$$\mathsf{g_n}(\mathsf{i}, D, X, A, \varepsilon \qquad \mathsf{g_n}(\pi(\mathsf{i}\ , \pi(D\ , \pi(X\ , \pi(A\ , \pi(\varepsilon\ \ , \qquad n$$

conditions imposed below. Define

$$\phi_{t,t^1}(i \quad \frac{1\ T_i \quad t\ (Y_i \quad \mu_t(i, X, A}{p_t(i, X, A} + \mu_t(i, X, A$$

$$\frac{1\ T_i \quad t'\ (Y_i \quad \mu_{t^1}(i, X, A}{p_{t^1}(i, X, A} \quad \mu_{t^1}(i, X, A \quad \tau(t, t' ,$$

whose average over $i \quad \mathcal{M}_n$ is the doubly robust moment condition, and let

$$\sigma_n^2 \quad \mathrm{Var} \quad \tilde{} \quad \frac{1}{\overline{m_n}} \ddot{y}_{i\in\mathcal{M}_n} \phi_{t,t^1}(i \quad X, A \quad .$$

Assu ptỵ on (Moments) M n p 4 u o n
n $\mathbb{N}$, i $\mathcal{N}_n$, n d $0, 1$ ^{n}, $\mathbf{E}||Y_i(d\ |^p$ X, A M
$|\underline{\pi}, \pi$ $(0, 1$ u $\hat{p}_t(i, X, A\ , p_t(i, X, A$ $|\underline{\pi}, \pi$ n

The next assumption is used to show that $\boldsymbol{\phi}_{\mathbf{t},\mathbf{t}'}(\mathbf{i} \quad {}^{\mathbf{n}}_{\mathbf{i}=1}$ is $\boldsymbol{\psi}$-dependent (see Definition C.1, which is due to Kojevnikov et al., 2021) to apply m728(o)-313.551(a)-.823-5.1.(o)-313.551(a)-.823-5.1.(

Parts (a) and (b) are used to establish that $\phi_{t,t'}(i \quad {}_{i=1}^{n}$ is ψ-dependent. Part (b) is a Lipschitz condition that holds if potential outcomes are uniformly bounded. In particular we can take $\Lambda_n(s \quad 2M$ where M is the uniform bound on the ranges strengthens Assumption 5(b) but only mildly since we nonparametrically estimate both nuisance functions. Since it does not require uniform convergence, it is more readily verifiable for machine learning estimators. Part (c) is Assumption 4.1(iii) of Kojevnikov et al. (2021). Parts (d)–(f) correspond to Assumptions 7(b)–(d) of Leung (2022a), which are used to characterize the bias of the variance estimator. We discuss verification of (c)–(f) in §B.2; the derivations there show that (f) is closely related to (c).

or n $\boldsymbol{\phi_{t,t^1}(i}$ n $\boldsymbol{\tau(t,t'}$ n n $on\ o$ $\boldsymbol{\phi_{t,t^1}(i}$ $\boldsymbol{\tau_i(t,t'}$
$\mathbf{E}|\mathbf{Y_i}$ $\mathbf{T_i}$ $\mathbf{t}, \boldsymbol{X},$

case $\hat{\sigma}^2$ would be asymptotically conservative. This can be formalized under additional weak dependence conditions on the superpopulation as in §A of Leung (2022b).

5 Approximate Sparsity

As discussed in §3, the number of layers $\mathbf{L}$ in a GNN determines its ,
the neighborhood ($\boldsymbol{X}_{\mathcal{N}(\mathbf{i},\mathbf{L})}$

$(W_i{}_{i=1}^{n}$ and ν $(\nu_i{}_{i=1}^{n}$, define

$$V_i(W,\nu;\theta \quad \alpha + \beta\frac{\sum_{j=1}^{n} A_{ij}W_j}{\sum_{j=1}^{n} A_{ij}} + \delta\frac{\sum_{j=1}^{n} A_{ij}X_j}{\sum_{j=1}^{n} A_{ij}} + \gamma X_j + \nu_i + \frac{\sum_{j=1}^{n} A_{ij}\nu_j}{\sum_{j=1}^{n} A_{ij}}$$

where θ $(\alpha,\beta,\delta,\gamma$. We generate $Y_i{}_{i=1}^{n}$ from the linear-in-means model, where Y_i $V_i(Y,\varepsilon;\theta_y$ and θ_y $(0.5, 0.8, 10, \; 1$. We generate $D_i{}_{i=1}^{n}$ according to Example 2, so that D_i 1 $V_i(D,\nu;\theta_d$ 0 with θ_d $(\; 0.5, 1.5, 1, \; 1$. The equilibrium selection mechanism is myopic best-response dynamics starting from the initial condition $D_i^0{}_{i=1}^{n}$ for D_i^0 1 $V_i(0,\nu;\theta_d$ 0 .

The design induces a greater degree of dependence than what our assumptions allow. The error term $\nu_i + \sum_{j=1}^{n} A_{ij}\nu_j \quad \sum_{j=1}^{n} A_{ij}$ is not conditionally independent across units unlike what Assumption 6(a) requires. Also, back-of-the-envelope calculations indicate that peer effects are sufficiently large in magnitude that Assumption 6(d) is violated.

We use the estimand in Example 5 whose true value is zero. About 57 percent of units select into treatment, so the effective sample size used to estimate the outcome regressions is around n 2 since $\mathrm{E}[Y_i \quad T_i \quad t, X, A$ is estimated only with observations for which T_i t. We report results for n $1000, 2000, 4000$.

6.2 Nonparametric Estimators

The GNNs use the PNA architecture in Example 9 with aggregator $\Gamma_2($ defined in the example and L $1,$

(2022) and Forastiere et al. (2021). For these, we estimate the nuisance functions using GLMs (logistic and linear regression) with polynomial sieves of order 1, 2, or 3. Recall that a GNN with $\mathbf{L}$ 1 corresponds to a receptive field that only encompasses the ego's 1-neighborhood. This is the same as the implied receptive field of the GLM estimators.

Table 1: Simulation results for random geometric graph

	$L = 1$			$L = 2$			$L = 3$		
n	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]									
H									
$\hat{\tau}(1,0)$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
CI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible] CI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
W $\hat{\tau}(1,0)$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
W CI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
W	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible] CI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

[illegible] si u ations, he esti and is $\tau(1,0) = 0$, treated $\approx$

A choice of L 2 is not unusual in the literature. Zhou et al. (2021) compute the prediction error of GNNs on several different datasets with L $2, 4, 8, \dots$ and find that L 2 has the best performance across several architectures. The fact that GNN performance often fails to improve (and indeed can worsen) with larger L is well known in the GNN literature, and we survey different expl study or later.

7.1 Comparison with He and Song (2024)

For the first (second) choice, $\boldsymbol{\tau}(\mathbf{t},\mathbf{t}'$ measures the effect of going from 0 to 1 (more than 1) adopting neighbor(s). This sheds light on a different dimension of diffusion relative to the ADM. Whereas the ADM measures how many others are affected by the ego's adoption, our estimands quantify the effect of having multiple adopting neighbors on the ego's adoption. We find below that having multiple adopting neighbors has a much larger effect than having only one.

As previously stated, He and Song (2024) define the treatment in two ways. One is a binary indicator for having a leader in the household, the idea being that all leaders were initially informed about microfinance and told to spread the word. However, not all leaders adopted in the first trimester, which perhaps

To compute the estimates, we concatenate the village networks into a single adjacency matrix of size $\mathbf{n}$ 4413. For the GNN and GLM estimates, we trim observations with propensity scores outside of $[0.01, 0.99$. Standard errors are obtained from the network HAC variance estimator defined in §2.3.

Table 3: Exposure Mapping $\mathbf{T}_{\mathbf{i}}^{(1)}$

	ADM	GNN			GLM		
		Layer	Layer	Layer	Order	Order	Order
Leader case							
$\mathbf{G}_{ee}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$\mathbf{G}_{sc}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$\mathbf{G}_{all}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Leader adopter case							
$\mathbf{G}_{ee}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$\mathbf{G}_{sc}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$\mathbf{G}_{all}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Adopter case							
$\mathbf{G}_{ee}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$\mathbf{G}_{sc}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$\mathbf{G}_{all}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

$\mathbf{n}$ = 4413

the leader case.

First consider the estimand using $\mathbf{T}_i^{(1)}$, which contrasts microfinance adoption rates for units with 1 versus 0 initially adopting neighbors. The GNN results are consistent across $\mathbf{L}$. For the leader case, we obtain precise zeros for almost all estimates, including the ADM. For the leader-adopter case, the GNN estimates are substantially smaller in magnitude than the ADM with an effect size of at most 10 percentage points compared to the smallest ADM estimate of 20 percentage points. This may be attributed to the use of richer network controls. For the adopter case, the contrast is even starker. Our estimates are an order of magnitude smaller than the corresponding ADM estimates. The GLM estimates are typically slightly smaller than the GNN estimates except for the order-3 polynomials, which are outliers in terms of magnitude.

The estimand using $\mathbf{T}_i^{(2)}$ contrasts units with 2 +versus 0 initially adopting neighbors. The estimates for the leader case are similar to those of $\mathbf{T}_i^{(1)}$. We find sizeable effects for the leader-adopter case, almost of the same order as ADM, but the robustness of the result is partly tempered by the large amount of trimming discussed below. The adopter case sees estimates of around 20 percentage points, whereas the ADM is double or triple that. Once again the GLM estimates are often slightly smaller relative to the GNN estimates, except for the order-3 polynomials.

The number of observations trimmed for $\mathbf{T}_i^{(1)}$ is negligible in the leader and adopter cases. In the leader-adopter case, more units are trimmed since fewer units are leader-adopters, but trimming never drops more than 200 observations. The story is quite different for $\mathbf{T}_i^{(2)}$, as reported in Table 5. The first three columns of the table report the number of observations for which the number of initially adopting neighbors $\mathbf{N}_i \quad \sum_{j=1}^{n} \mathbf{A}_{ij} \mathbf{D}_j$

in the lasso literature, which posit that a high-dimensional regression function is well-approximated by a function of a relatively small number of covariates.

A Additional Results on GNNs

Primitive conditions for Assumption 5 appear to be beyond the scope of the existing GNN literature, but we provide some potentially useful intermediate results. Consider the problem of establishing a rate of convergence for the propensity score:

$$\frac{1}{m_n}\sum_{i\in\mathcal{M}_n}\left(\hat{p}_t(i,\boldsymbol{X},\boldsymbol{A}) - p_t(i,\boldsymbol{X},\boldsymbol{A})\right)^2 = o_p(n^{-1/2}).$$

Under network approximate sparsity (Definition 1), the problem simplifies to showing

$$\frac{1}{m_n}\sum_{i\in\mathcal{M}_n}\left(\hat{p}_t(i,\boldsymbol{X},\boldsymbol{A}) - p_t(i,\boldsymbol{X}_{\mathcal{N}(i,L)},\boldsymbol{A}_{\mathcal{N}(i,L)})\right)^2 = o_p(n^{-1/2}). \tag{A.1}$$

Since $\hat{p}_t(i,\boldsymbol{X},\boldsymbol{A})$ is an L-layer GNN, which only uses information from $(\boldsymbol{X}_{\mathcal{N}(i,L)},\boldsymbol{A}_{\mathcal{N}(i,L)})$, this should well approximate $p_t(i,\boldsymbol{X}_{\mathcal{N}(i,L)},\boldsymbol{A}_{\mathcal{N}(i,L)})$ under appropriate conditions, so (A.1) should be more feasible to verify directly.

Farrell et al. (2021) provide a bound analogous to (A.1) for MLPs, which, were it applicable to our setting, would be of the form

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_t(i,\boldsymbol{X},\boldsymbol{A}) - p_t(i,\boldsymbol{X}_{\mathcal{N}(i,L)},\boldsymbol{A}_{\mathcal{N}(i,L)})\right)^2 \leq C\left(\frac{WL\log R}{n}\log n + \frac{\log\log n + \gamma}{n} + \epsilon^2\right) \tag{A.2}$$

with probability at least $1 - e^{-\gamma}$. Here W is the number of GNN parameters, C is a constant that does not depend on n, R depends on the architecture through the number of hidden neurons, and ϵ is the function approximation error, a measure of the ability of the neural network to approximate any function in a desired class. Establishing a corresponding result for GNNs requires an analog of Lemma 6 of Farrell et al. (2021), which is a bound on the pseudo-dimension of the GNN class, and concentration inequalities for ψ-dependent data. Jegelka (2022) surveys the few available complexity and generalization bounds for GNNs. These are not sufficiently general for our setup and only apply to settings where the sample consists of many independent networks.

To use a bound of the form (A.2) to verify Assumption 5, we require knowledge of how □ varies with key aspects of the architecture, such as $\mathbf{W}, \mathbf{R}, \mathbf{L}, \mathbf{n}$. As a first step toward obtaining such a result, it is necessary to characterize the function class that GNNs can approximate. Our next result, which draws heavily from existing results in the GNN literature, shows that an additional shape restriction on the function class beyond invariance (§3.3) is required.

A.1 WL Function Class

MLPs can approximate any measurable function (Hornik et al., 1989), so given the discussion in §3.3, a natural question is whether GNNs can approximate any measurable, n n function of graph-structured inputs. In other words, is it enough to require invariance (and regularity conditions), or are stronger restrictions on the function class necessary? For reasons related to the graph isomorphism problem, it turns out stronger restrictions are necessary. We next motivate the need for such restrictions and then state our function approximation result.

Chen et al. (2019) show that, for a function class such as GNNs to approximate any invariant function, some element of the class must be able to separate any pair of non-isomorphic graphs. By "separate," we mean that for any non-isomorphic "labeled graphs" $(\boldsymbol{X}, \boldsymbol{A}$ $,(\boldsymbol{X}', \boldsymbol{A}'$, the function $\mathbf{f}$ satisfies $\mathbf{f}(\boldsymbol{X}, \boldsymbol{A}$ $\neq$ $\mathbf{f}(\boldsymbol{X}', \boldsymbol{A}'$. Hence, a function with separating power of this sort solves the graph isomorphism problem, a problem for which no known polynomial-time solution exists (Kobler et al., 2012; Morris et al., 2021). Since GNNs can be computed in polynomial time, this suggests that approximating any invariant function is too demanding of a requirement.

To define the subclass of invariant functions that GNNs can approximate, we need to take a detour and discuss graph isomorphism tests. The subclass will be defined by a weaker graph separation criterion than solving the graph isomorphism problem, in particular one defined by the W n W . This is a (generally imperfect) test for graph isomorphism on which almost all practical graph isomorphism solvers are based (Morris et al., 2021).

Given a labeled graph $(\boldsymbol{X}, \boldsymbol{A}$, the WL test outputs a graph coloring (a vector of labels for each unit) according to the following recursive procedure, whose definition

$\mathbf{H}^2$ such that

$$(\mathbf{h}, \mathbf{h}' \quad \rho(\mathcal{F} \quad \text{if and only if} \quad \mathbf{f}(\mathbf{h} \quad \mathbf{f}(\mathbf{h}' \quad \text{for all} \quad \mathbf{f} \quad \mathcal{F}.$$

For any two sets of functions $\mathcal{E}, \mathcal{F}$ with domain $\mathbf{H}$, we say that $\mathcal{E}$ is *o* *n* $\mathcal{F}$ if $\rho(\mathcal{F} \quad \rho(\mathcal{E}$.

This is essentially Definition 2 of Azizian and Lelarge (2021). Intuitively, if $\mathcal{E}$ is at most as separating as $\mathcal{F}$, the latter is more complex in the sense that some function in $\mathcal{F}$ can separate weakly more elements of $\mathbf{H}$ than any function in $\mathcal{E}$.

Let $\mathbf{f}_{\mathrm{L},\mathbf{L}}$ denote the function of $(\boldsymbol{X}, \boldsymbol{A}$ with range $\Sigma^{\mathbf{n}}$ that outputs the vector of node colorings from the WL test run for $\mathbf{L}$ iterations. Let $\mathcal{C}(\mathbf{H}$ be the set of continuous functions with domain $\mathbf{H}$. For any $\mathbf{L} \quad \mathbb{N}$, define the *W* *un* *on*

$$\mathcal{F}_{\mathrm{L}}(\mathbf{L} \quad \mathbf{f} \quad \mathcal{C}(\mathbf{H} : \rho(\mathbf{f}_{\mathrm{L},\mathbf{L}} \quad \rho(\mathbf{f} \quad .$$

This is the set of continuous functions of $(\boldsymbol{X}, \boldsymbol{A}$ that are at most as separating as the WL test with $\mathbf{L}$ iterations.

The next result says that $\mathbf{p}_{\mathbf{t}}($ and $\boldsymbol{\mu}_{\mathbf{t}}($ can be approximated by

In other words, any function in the class $\mathcal{F}\ {}_{\mathrm{L}}($**L** can be approximated by **L**-layer GNNs in $\mathcal{F}_{\mathrm{GNN}}\ ($**L** . The result is a consequence of a Stone-Weierstrauss theorem due to Azizian and Lelarge (2021) and a version of the Morris et al. (2019) and Xu et al. (2018) result on the equivalent separation power of GNNs and the WL test. The proof is given below.

The result is essentially Theorem 4 of Azizian and Lelarge (

$\mathbf{T}$ in place of $\mathbf{L}$) and add an additional MLP layer that implements their equation (26). Here we use the additional MLP layer added to the output of our architecture (see the paragraph prior to the statement of Theorem A.1). In particular, for any $\mathbf{f}$ $\mathcal{F}_{\text{GNN}}$ ($\mathbf{L}$, consider the mapping

$$(\boldsymbol{X}, \boldsymbol{A} \qquad \tilde{} \; \underbrace{\text{ÿ}_{i=1} \mathbf{f}(\mathbf{i}, \boldsymbol{X}, \boldsymbol{A} \mathbf{, \ldots,} \; \text{ÿ}_{i=1} \mathbf{f}(\mathbf{i}, \boldsymbol{X}, \boldsymbol{A}}_{\mathbf{n}\text{ ti es}} \qquad \mathbb{R}^{\mathbf{n}}$$

(their (26) in our notation) corresponds to adding a linear output layer $\mathbf{L}$ $+1$ that is implementable by an MLP of the form $\boldsymbol{\varphi}_{\mathbf{L}+1}(\mathbf{h}_{\mathbf{i}}^{(\mathbf{L})}\mathbf{,}\ \mathbf{h}_{\mathbf{j}}^{(\mathbf{L})} : \mathbf{j} \quad \mathcal{N}_{\mathbf{n}}$. The mapping remains an element of $\mathcal{F}_{\text{GNN}}$ ($\mathbf{L}$, which completes the argument for (A.5).

By Theorems VIII.1 and VIII.4 of Grohe (2021), which use finiteness of the support of $\mathbf{X}_{\mathbf{i}}$, $\boldsymbol{\rho}(\mathcal{F}_{\text{GNN}}$ ($\mathbf{L}$ $\qquad \boldsymbol{\rho}($ $\mathbf{f}$ $_{\text{L},\mathbf{L}}$. That is, $\mathbf{L}$-layer GNNs have the same separation power as the WL test run for $\mathbf{L}$ iterations. ■

A.2 Disadvantages of Depth

The receptive field is the main consideration when selecting $\mathbf{L}$, but Theorem A.1 provides a second consideration, which is imposing a weaker implicit shape restriction. It shows that, for GNNs to approximate a target function well, the target must satisfy a shape restriction stronger than invariance, namely that it is at most as separating as the WL test with $\mathbf{L}$ iterations. The larger the choice of $\mathbf{L}$, the weaker the shape restriction imposed. However, there are several reasons why shallow architectures remain preferable.

ow r turns to ` pt A natural question is how many iterations are required for the WL test to converge for a given graph, which corresponds to the choice of $\mathbf{L}$ for which the shape restriction is weakest. Unfortunately, the answer is not generally known, being determined by the topology of the input graph in a complex manner. However, there is a range of results bounding the number of iterations required for convergence. For instance, Kiefer and McKay (2020) construct graphs for which the WL test requires $\mathbf{n}$ $\quad 1$ iterations to converge, so such graphs require $\mathbf{n}$ $\quad 1$ layers to obtain the weakest shape restriction. This makes the estimation problem extremely high-dimensional, requiring substantially more layers than what is typically required

than the standard architecture (13) (Dwivedi et al., 2022). These disadvantages may partly explain the common use in practice of the standard architecture with few layers.

B Verifying §8 Assumptions

Leung (2022a), §A, verifies analogs of Assumptions 6(d) and 7(c) from an older working paper version of Kojevnikov et al. (2021). This section repeats the exercise for Assumptions 6(d) and 7(c) and (d). We assume throughout that max $\mathbf{y}_{\mathbf{n}}(\mathbf{s}\ 2\ ,\boldsymbol{\psi}_{\mathbf{n}}(\mathbf{s}$ exp($\mathbf{c}(1\quad 4\ \mathbf{p}\ ^{-1}\mathbf{s}$ for some $\mathbf{c}\quad 0$ and $\mathbf{p}$ in Assumption 4(a). As in Leung (2022a), we say a sequence of networks exhibits polynomial neighborhood growth if

$$\sup_{\mathbf{n}} \max_{\mathbf{i}\in\mathcal{N}_n} |\mathcal{N}_{\boldsymbol{A}}(\mathbf{i},\mathbf{s}\ | \qquad \mathbf{C}\mathbf{s}^{\mathbf{d}}$$

for some $\mathbf{C}\quad 0, \mathbf{d} \succ 1$

and $D'_{\mathbf{B}}$ $(\mathbf{D}'_{\mathbf{j}}$ $_{\mathbf{j}\in\mathbf{B}}$ for any $\mathbf{B}$ $\mathcal{N}_{\mathbf{n}}$. Using the first equality of (C.3),

$$
\begin{aligned}
\mathbf{p_t}(\mathbf{i}, X, A \quad & \mathrm{P}\,(\mathbf{D}'_{\mathbf{i}} + (\mathbf{D}_{\mathbf{i}} \quad \mathbf{D}'_{\mathbf{i}} \quad | \mathbf{a}, \mathbf{b}\ , \ \mathbf{V}'_{\mathbf{i}} + (\mathbf{V}_{\mathbf{i}} \quad \mathbf{V}'_{\mathbf{i}} \quad | \boldsymbol{\alpha}, \boldsymbol{\beta} \quad X, A \\
& \mathrm{P}\,(\mathbf{D}'_{\mathbf{i}} \quad | \mathbf{a} \quad □, \mathbf{b} + □\ , \ \mathbf{V}'_{\mathbf{i}} \quad | \boldsymbol{\alpha} \quad □, \boldsymbol{\beta} + □ \quad X, A \\
& \qquad + \underbrace{\mathrm{P}\,(|\mathbf{D}_{\mathbf{i}} \quad \mathbf{D}'_{\mathbf{i}}| \quad □ \quad X, A + \mathrm{P}\,(|\mathbf{V}_{\mathbf{i}} \quad \mathbf{V}'_{\mathbf{i}}| \quad □ \quad X, A}_{\mathbf{R}_0} .
\end{aligned}
$$

By (C.3), the right-hand side equals

$$\mathrm{P}\,(\mathbf{D}'_{\mathbf{i}} \quad | \mathbf{a}, \mathbf{b}$$

Combining (C.6) and (C.7) and using the law of iterated expectations,

$$|p_t(i, \boldsymbol{X}, \boldsymbol{A} \quad p_t(i, \boldsymbol{X}_{\mathcal{N}(i, r_\lambda(s+1))}, \boldsymbol{A}_{\mathcal{N}(i, r_\lambda(s+1))} \; | \quad \lambda_n(s \; +1 \quad +2R_0.$$

roo o (C.2) Noting that $\mu_t(i, \boldsymbol{X}, \boldsymbol{A} \quad \mathrm{E}|Y_i 1_i(t \quad \boldsymbol{X}, \boldsymbol{A} \;\; p_t(i, \boldsymbol{X}, \boldsymbol{A}$, we first bound the numerator. For $B \quad \mathcal{N}(i, s$, define $Y_i' \quad g_{n(i,s)}(i, \boldsymbol{D}_B', \boldsymbol{X}_B, \boldsymbol{A}_B, \boldsymbol{\varepsilon}_B$. By Lemma C.2,

$$|\mathrm{E}|Y_i 1_i(t \quad \boldsymbol{X}, \boldsymbol{A} \quad \mathrm{E}|Y_i' 1_i(t \quad \boldsymbol{X}, \boldsymbol{A} \,| \quad \gamma_n(s \quad +\Lambda_n(i, s \; n(i, s \; \eta_n(s$$

where, using Assumption 4(b),

$$\begin{aligned}
|\mathbf{R}_1| \quad & \boldsymbol{\lambda}_{\mathbf{n}}(2\mathbf{s} \quad + 2\ \boldsymbol{\gamma}_{\mathbf{n}}(\mathbf{s} \quad + \Lambda_{\mathbf{n}}(\mathbf{i},\mathbf{s}\ \mathbf{n}(\mathbf{i},\mathbf{s}\ \boldsymbol{\eta}_{\mathbf{n}}(\mathbf{s} \quad + \mathbf{C}'(1 \quad + \mathbf{n}(\mathbf{i}, 1 \quad \boldsymbol{\eta}_{\mathbf{n}}(\mathbf{s} \quad , \\
|\mathbf{R}_2| \quad & \mathbf{C}\ \boldsymbol{\lambda}_{\mathbf{n}}(2\mathbf{s} \quad + (1 \quad + \mathbf{n}(\mathbf{i}, 1 \quad \boldsymbol{\eta}_{\mathbf{n}}(2\mathbf{s} \quad 1 \quad , \quad \text{and} \\
|\mathbf{R}_3| \quad & \mathbf{C}''(|\mathbf{R}_1| \quad + |\mathbf{R}_2|
\end{aligned}$$

for some universal $\mathbf{C}''$ 0. Substituting $\mathbf{s}$ 2 for $\mathbf{s}$ yields the result. ■

$$\mathbf{a}\ \mathrm{C} \qquad n \quad \mathbf{B_i} \quad \mathcal{N}(\mathbf{i},\mathbf{s} \quad \mathbf{D}'_{\mathbf{j}} \quad \mathbf{h}_{\mathbf{n}(\mathbf{j},\mathbf{s})}(\mathbf{j}, X_{\mathbf{B}_j}, A_{\mathbf{B}_j}, \nu_{\mathbf{B}_j} \quad D'_{\mathbf{B}_i} \quad (\mathbf{D}'_{\mathbf{j}}$$

for $|\mathbf{R}_1|$ $\mathbf{y_n}(\mathbf{s}$ $+\Lambda_{\mathbf{n}}(\mathbf{i},\mathbf{s}$ $\mathbf{n}(\mathbf{i},\mathbf{s}$ $\eta_{\mathbf{n}}(\mathbf{s}$. ■

a C n $\mathbf{Y}_{\mathbf{i}}', \mathbf{D}_{\mathbf{i}}'$ n n $\mathbf{1_i}(\mathbf{t}'$ 1 $\mathbf{D}_{\mathbf{i}}'$ $\mathbf{d},$ $\check{\mathbf{r}}_{\mathbf{j}=1}^{\mathbf{n}} \mathbf{A}_{\mathbf{ij}}\mathbf{D}_{\mathbf{j}}'$
$\mathbf{\Delta}$ n A u on , , , n , $\mathbf{C}$ 0 u o ny
$\mathbf{n}$ $\mathbb{N}$, $\mathbf{i}$ $\mathcal{N}_{\mathbf{n}}$, n $\mathbf{s} \succ 0$,

$$\mathrm{E}|\mathbf{Y_i}|\mathbf{1_i}(\mathbf{t} \quad \mathbf{1_i}(\mathbf{t}'| \;\; \boldsymbol{X}, \boldsymbol{A} \quad \mathbf{C}\,(1 \;+\mathbf{n}(\mathbf{i}, 1 \;\; \eta_{\mathbf{n}}(\mathbf{s} \; .$$

Proof. Recall the definition of $\mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\beta},$ prior to (C.3). Define $\mathbf{V_i}$ $\check{\mathbf{r}}_{\mathbf{j}=1}^{\mathbf{n}} \mathbf{A}_{\mathbf{ij}}\mathbf{D}_{\mathbf{j}}$, $\mathbf{V}_{\mathbf{i}}'$ $\check{\mathbf{r}}_{\mathbf{j}=1}^{\mathbf{n}} \mathbf{A}_{\mathbf{ij}}\mathbf{D}_{\mathbf{j}}'$, and $\mathcal{C}$ $|\mathbf{D_i}$ $\mathbf{D}_{\mathbf{i}}'|$ $, |\mathbf{V_i}$ $\mathbf{V}_{\mathbf{i}}'|$. Then

$$\begin{aligned} &\mathrm{E}|\mathbf{Y_i}|\mathbf{1_i}(\mathbf{t} \quad \mathbf{1_i}(\mathbf{t}'| \;\; \boldsymbol{X} \quad \boldsymbol{x}, \boldsymbol{A} \quad \boldsymbol{a} \\ &\qquad \mathrm{E}|\mathbf{Y_i}|\mathbf{1_i}(\mathbf{t} \quad \mathbf{1_i}(\mathbf{t}'| \;\; \mathcal{C}, \boldsymbol{X} \quad \boldsymbol{x}, \boldsymbol{A} \quad \boldsymbol{a} \;\; +\mathbf{C}\,\mathrm{P}(\mathcal{C}^{\mathbf{c}} \;\; \boldsymbol{X} \quad \boldsymbol{x}, \boldsymbol{A} \quad \boldsymbol{a} \end{aligned} \quad \text{(C.11)}$$

for some universal $\mathbf{C}$ 0 by Assumptions 1 and 4(a). By Assumption 3,

$$\mathbf{1_i}(\mathbf{t} \quad 1 \;\; \mathbf{D_i} \;\; |\mathbf{a}, \mathbf{b} \; , \;\; \mathbf{V_i} \;\; |\boldsymbol{\alpha}, \boldsymbol{\beta} \qquad \text{and} \quad \mathbf{1_i}(\mathbf{t}' \quad 1 \;\; \mathbf{D}_{\mathbf{i}}' \;\; |\mathbf{a}, \mathbf{b} \; , \;\; \mathbf{V}_{\mathbf{i}}' \;\; |\boldsymbol{\alpha}, \boldsymbol{\beta} \quad .$$

Under event $\mathcal{C}$,

$$\begin{aligned} 1 \;\; \mathbf{D_i} \;\; |\mathbf{a}, \mathbf{b} \; , \;\; \mathbf{V_i} \;\; |\boldsymbol{\alpha}, \boldsymbol{\beta} \; (\quad & 1 \;\; \mathbf{D}_{\mathbf{i}}' \; +(\mathbf{D_i} \quad \mathbf{D}_{\mathbf{i}}' \quad |\mathbf{a}, \mathbf{b} \; , \;\; \mathbf{V}_{\mathbf{i}}' \; +(\mathbf{V_i} \quad \mathbf{V}_{\mathbf{i}}' \quad |\boldsymbol{\alpha}, \boldsymbol{\beta} \\ & \qquad\qquad\qquad\qquad\qquad\qquad \mathrm{q} \; ; \; \mathrm{v} \;\; \mathbf{_i D} \\ & 1 \;\; \mathbf{D}_{\mathbf{i}}' \quad |\mathbf{a} \quad , \mathbf{b} \; + \; , \;\; \mathbf{V}_{\mathbf{i}}' \quad |\boldsymbol{\alpha} \quad \mathbf{s} \end{aligned}$$]JR01]JR Tf 0Td []JR Tf

Lipschitz functions on $\mathbb{R}^d$, Lip(f be the Lipschitz constant of f $\mathcal{L}_d$, and

$$\mathcal{P}_n(h,h';s \qquad (H,H' : H,H' \quad \mathcal{N}_n,\ H \quad h,\ H' \quad h', \square_A(H,H' \succ s\ .$$

D ɳ ƫ on C 1 A triangular array Z_i $\substack{n \\ i=1}$ is *on on* ψ *n n n* $\mathcal{F}_n$ if there exist C (0, and an $\mathcal{F}_n$-measurable sequence $\psi_n(s$ $_{s,n\in\mathbb{N}}$ with $\psi_n(0$ 1 for all n such that

$$|\mathrm{Cov}(f(Z_H\ , f'(Z_{H'}\ | \quad Chh'(\|f\|_\infty\ +\mathrm{Lip}(f \quad (\|f'\|_\infty\ +\mathrm{Lip}(f' \quad \psi_n(s \quad \text{a.s.} \tag{C.12}$$

for all n,h,h' $\mathbb{N}$; s 0; f $\mathcal{L}_h$; f' $\mathcal{L}_{h'}$; and $(H,H'$ $\mathcal{P}_n(h,h';s$. We call $\psi_n(s$ the *n n o n* of Z_i $\substack{n \\ i=1}$.

a C *n A u on , , , n , n n , o n*
t,t' $\mathcal{T}$, $\phi_{t,t'}(i$ $\substack{n \\ i=1}$ *on on* ψ *n n n* $(X,A$ *n on*
n n o n $\psi_n(s$ *n n* (15)

Proof. Let $\mathcal{F}_n$ be the σ-algebra generated by $(X,A$, $(h,h'$ $\mathbb{N}$ $\mathbb{N}$, $(f,f'$ $\mathcal{L}_h$ $\mathcal{L}_{h'}$, s 0, and $(H,H'$ $\mathcal{P}_n(h,h';s$. Define Z_i $\phi_{t,t'}(i$, Z_H $(Z_i$ $_{i\in H}$, ξ $f(Z_H$, ζ $f'(Z_{H'}$, and

$$D_i^{(s)} \quad h_{n(i,s)}(i, X_{\mathcal{N}(i,s)}, A_{\mathcal{N}(i,s)}, \nu_{\mathcal{N}(i,s)}\ .$$

For $D^{(s)}_{\mathcal{N}(i,s')}$ $(D_j^{(s)}$ $_{j\in\mathcal{N}(i,s')}$, let

$$\begin{aligned} 1_i^{(s)}(t &\quad 1\ f_{n(i,s/2)}(i, D^{(s/2)}_{\mathcal{N}(i,s/2)}, A_{\mathcal{N}(i,s/2)} \quad t\ , \\ Y_i^{(s)} &\quad g_{n(i,s/2)}(i, D^{(s/2)}_{\mathcal{N}(i,s/2)}, X_{\mathcal{N}(i,s/2)}, A_{\mathcal{N}(i,s/2)}, \varepsilon_{\mathcal{N}(i,s/2)}\ , \\ Z_i^{(s)} &\quad \frac{1_i^{(s)}(t\ (Y_i^{(s)} \quad \mu_t(i,X,A}{p_t(i,X,A} + \mu_t(i,X,A \\ &\quad \frac{1_i^{(s)}(t'\ (Y_i^{(s)} \quad \mu_{t'}(i,X,A}{p_{t'}(i,X,A} \quad \mu_{t'}(i,X,A \quad \tau_i(t,t'\ . \end{aligned}$$

Finally, let $\xi^{(s)}$ $f((Z_i^{(s)}$ $_{i\in H}$ and $\zeta^{(s)}$ $f'((Z_i^{(s)}$ $_{i\in H'}$.

By Assumption 6(a), $(\mathbf{Z}_{\mathbf{i}}^{(\mathbf{s}/2,\boldsymbol{\xi})} \quad {}_{\mathbf{i}\in\mathbf{H}} \qquad (\mathbf{Z}_{\mathbf{j}}^{(\mathbf{s}/2,\boldsymbol{\zeta})} \quad {}_{\mathbf{j}\in\mathbf{H}^1} \quad \mathcal{F}_{\mathbf{n}}$, so

$$\begin{aligned}
|\mathrm{Cov}(\boldsymbol{\xi},\boldsymbol{\zeta} \quad \mathcal{F}_{\mathbf{n}} \,| \quad & |\mathrm{Cov}(\boldsymbol{\xi} \quad \boldsymbol{\xi}^{(\mathbf{s}/2)},\boldsymbol{\zeta} \quad \mathcal{F}_{\mathbf{n}} \,| \;+|\mathrm{Cov}(\boldsymbol{\xi}^{(\mathbf{s}/2)},\boldsymbol{\zeta} \quad \boldsymbol{\zeta}^{(\mathbf{s}/2)} \quad \mathcal{F}_{\mathbf{n}} \,| \\
& 2\|\mathbf{f}'\|_{\infty}\mathbf{E}||\boldsymbol{\xi} \quad \boldsymbol{\xi}^{(\mathbf{s}/2)}| \quad \mathcal{F}_{\mathbf{n}} \quad +2\|\mathbf{f}\|_{\infty}\mathbf{E}||\boldsymbol{\zeta} \quad \boldsymbol{\zeta}^{(\mathbf{s}/2)}| \quad \mathcal{F}_{\mathbf{n}} \\
& 2 \;\; \mathbf{h}\|\mathbf{f}'\|_{\infty}\mathrm{Lip}(\mathbf{f} \quad +\mathbf{h}'\|\mathbf{f}\|_{\infty}\mathrm{Lip}(\mathbf{f}' \quad \max_{\mathbf{i}\in\mathcal{N}}
\end{aligned}$$

$\mathcal{L}_{\mathbf{h}} \quad \mathcal{L}_{\mathbf{h}^1}, \mathbf{s} \quad 0, (\mathbf{H}, \mathbf{H}' \quad \mathcal{P}_{\mathbf{n}}(\mathbf{h}, \mathbf{h}'; \mathbf{s} \ ,$

$$\mathbf{Y}_{\mathbf{i}}^{(\mathbf{s})} \quad \mathbf{g}_{\mathbf{n}(\mathbf{i},\mathbf{s})}(\mathbf{i}, \boldsymbol{D}_{\mathcal{N}(\mathbf{i},\mathbf{s})}, \boldsymbol{X}_{\mathcal{N}(\mathbf{i},\mathbf{s})}, \boldsymbol{A}_{\mathcal{N}(\mathbf{i},\mathbf{s})}, \varepsilon_{\mathcal{N}(\mathbf{i},\mathbf{s})} \ ,$$

$\boldsymbol{\xi} \quad \mathbf{f}((\mathbf{Y}_{\mathbf{i}} \ {}_{\mathbf{i}\in\mathbf{H}} \ , \boldsymbol{\zeta} \quad \mathbf{f}'((\mathbf{Y}_{\mathbf{i}} \ {}_{\mathbf{i}\in\mathbf{H}^1} \ , \boldsymbol{\xi}^{(\mathbf{s})} \quad \mathbf{f}((\mathbf{Y}_{\mathbf{i}}^{(\mathbf{s})} \ {}_{\mathbf{i}\in\mathbf{H}} \ ,$ and $\boldsymbol{\zeta}^{(\mathbf{s})} \quad \mathbf{f}'((\mathbf{Y}_{\mathbf{i}}^{(\mathbf{s})} \ {}_{\mathbf{i}\in\mathbf{H}^1} \ .$ By Assumption 6(a),

$$\begin{aligned}
|\mathrm{Cov}(\boldsymbol{\xi}, \boldsymbol{\zeta} \quad \mathcal{F}'_{\mathbf{n}} \ | \quad & |\mathrm{Cov}(\boldsymbol{\xi} \quad \boldsymbol{\xi}^{(\mathbf{s}/2)}, \boldsymbol{\zeta} \quad \mathcal{F}'_{\mathbf{n}} \ | \ + |\mathrm{Cov}(\boldsymbol{\xi}^{(\mathbf{s}/2)}, \boldsymbol{\zeta} \quad \boldsymbol{\zeta}^{(\mathbf{s}/2)} \quad \mathcal{F}'_{\mathbf{n}} \ | \\
& 2\|\mathbf{f}'\|_\infty \mathbf{E} \ |\boldsymbol{\xi} \quad \boldsymbol{\xi}^{(\mathbf{s}/2)}| \quad \mathcal{F}'_{\mathbf{n}} \quad + 2\|\mathbf{f}\|_\infty \mathbf{E} \ |\boldsymbol{\zeta} \quad \boldsymbol{\zeta}^{(\mathbf{s}/2)}| \quad \mathcal{F}'_{\mathbf{n}} \\
& 2 \ \mathbf{h}\|\mathbf{f}'\|_\infty \mathrm{Lip}(\mathbf{f} \quad + \mathbf{h}'\|\mathbf{f}\|_\infty \mathrm{Lip}(\mathbf{f}' \quad \max_{\mathbf{i}\in\mathcal{N}_n} \mathbf{E} \ |\mathbf{Y}_{\mathbf{i}} \quad \mathbf{Y}_{\mathbf{i}}^{(\mathbf{s}/2)}| \quad \mathcal{F}'_{\mathbf{n}} \\
& 2 \ \mathbf{h}\|\mathbf{f}'\|_\infty \mathrm{Lip}(\mathbf{f} \quad + \mathbf{h}'\|\mathbf{f}\|_\infty \mathrm{Lip}(\mathbf{f}' \quad \boldsymbol{\gamma}_{\mathbf{n}}(\mathbf{s} \ 2 \ ,
\end{aligned}$$

the last line using Assumption 2. Given $\boldsymbol{\psi}$-dependence, the claim follows from Corollary A.2 of Kojevnikov et al. (

$\hat{\mathsf{p}}_{\mathsf{t}}(\mathsf{i}, X, A$. For some universal constants C, C' 0, $\mathbf{E}|\mathsf{R}_{1\mathsf{t}}^2$ equals

$$\frac{1}{\mathsf{m}_{\mathsf{n}}} \underset{\mathsf{i}\in\mathcal{M}_n}{\ddot{\mathsf{y}}} \underset{\mathsf{j}\in\mathcal{M}_n}{\ddot{\mathsf{y}}} \mathbf{E} \;{}^{\textbf{”}}\mathbf{E}\,|(\mathsf{Y}_{\mathsf{i}} \quad \mu_{\mathsf{i}} \;(\mathsf{Y}_{\mathsf{j}} \quad \mu_{\mathsf{j}} \quad D, X, A \quad \mathbf{1}_{\mathsf{i}}(\mathsf{t}\; \mathbf{1}_{\mathsf{j}}(\mathsf{t}\; ($$

$\hat{\mathbf{p}}_{\mathbf{t}}(\mathbf{i}, \boldsymbol{X}, \boldsymbol{A}$, and

$$\Delta_{\mathbf{i}}(\mathbf{t} \quad (\hat{\boldsymbol{\mu}}_{\mathbf{t}}(\mathbf{i} \quad \boldsymbol{\mu}_{\mathbf{t}}(\mathbf{i} \quad {}^{\mathbf{p}_{\mathbf{t}}(\mathbf{i}} \quad 1_{\mathbf{i}}(\mathbf{t}$$

References

A on an E Ya av, “On the Bottleneck of Graph Neural Networks and its Practical Implications,” in “International Conference on Learning Representations” 2021.

Aronow an C a , “Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment,” *Ann o A* , 2017, (4), 1912–1947.

At D E s an G I ns, “Exact **p**-Values for Network Interference,” *ou n o A n A o on*, 2018, (521), 230–240.

– **G I ns J t r an E unro**, “Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations,” *ou n o ono* , 2021.

/BPC1 ID EIQ 0.28.601000cmBT/R20211.9552Tf 1001101.76489.96Tm (,)-390.203(G)0.743337(.)-380.153(5.6 0.2.4m)
b25(a)-2(n)875((L)-0.63102-0.73312,75(o)-2.37)1.38865(o)-2(u)2.92127(r)0.861963..0916(t)S(m)-3-2.nm

Bronst n , "Do We Need Deep Graph Neural Networks?," `https //towardsdatascience com/do-we-need-deep-graph-neural-networks -be 2d3ec cp` 2020. Accessed: 2022-07-02.

— **J Bruna . Co n an ov** , "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges," *X n X* , 2021.

C n Z ar C n an J Bruna, "On the Equivalence Between Graph Isomorphism Testing and Function Approximation with GNNs," in "Advances in Neural Information Processing Systems," Vol. 32 2019.

C rno u ov D C tv ov D r r E Du o C Hans n w an J o ns, "Double/Debiased Machine Learning for Treatment and Structural Parameters," *ono ou n* , 2018, , C1–C68.

Corso G Cava D B a n an ov , "Principal Neighbourhood Aggregation for Graph Nets," in "Advances in Neural Information Processing Systems," Vol. 33 2020, pp. 13260–13271.

D Groot , "Reaching a Consensus," *ou n o A n A o on*, 1974, (345), 118–121.

D ra a F C Gar a J no Donovan an A an B rra, "Identifying Causal Effects in Experiments with Spillovers and Non-compliance," *ou n o ono* , 2023, *5* (2), 1589–1624.

Dw v C Jos aur nt Y B n o an X Br sson, "Benchmarking Graph Neural Networks," *X n X* , 2022.

E n r C po n an B ann, "Treatment Effect Estimation from Observational Network Data Using Augmented Inverse Probability Weighting and Machine Learning," *X n X 5* , 2022.

Farr , "Robust Inference on Average Treatment Effects with Possibly more Covariates than Observations," *X n X* , 2018.

— **an an sra**, "Deep Neural Networks for Estimation and Inference," *ono* , 2021, (1), 181–213.

F an J nss n, “Fast Graph Representation Learning with PyTorch Geometric,” *X n X* , 2019.

Forast r E A rd an F a , “Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks,” *o u n o A n A o on*, 2021, (534), 901–918.

Gro , “The Logic of Graph Neural Networks,” in “2021 36th Annual ACM/IEEE Symposium on Logic in Computer Science” IEEE 2021, pp. 1–17.

H X an on , “Measuring Diffusion over a Large Network,” *o o no u o o n* , 2024.

Horn t n o an H t , “Multilayer Feedforward Networks are Universal Approximators,” *u o* , 1989,

, r an B a, “The Iteration Number of Colour Refinement,” in “47th International Colloquium on Automata, Languages, and Prog

[illegible] **ann** [illegible], “Causal Inference for Spatial Treatments,” *X* [illegible] *n*
X [illegible]

Zop [illegible], “1-WL Expressiveness Is (Almost) All You Need,” *X* *n* *X* [illegible] *5*, 2022.