Nonparametric Estimation of Treatment Effects with Endogenous Peers*

Haoran Pan†

October 6, 2025

Abstract

When individual outcomes depend on peer outcomes, treating an individual in a network affects all connected individuals. This causes the absence of a control group and threatens the validity of causal inference. Existing methods assume linear functional forms and exogenous networks, or exclude the dependence on peer outcomes. By introducing a nonparametric peer effect model, I prove that the treatment effect is identified by comparing individuals with the same neighbors but different treatment status, which does not rely on the above assumptions. Estimation is performed using a combination of a kernel estimator, which relaxes the same neighbor condition in finite samples, and the method of sieves. The consistency of the proposed estimator is then established. Application of this method to an anti-violence campaign suggests that the effect of the campaign on individual attitude is increasing in the average neighbor attitude.

*I am indebted to Hiroaki Kaido for his invaluable guidance and encouragement throughout this project. I am grateful to Iván Fernández-Val, Marc Rysman, Jean-Jacques Forneron for their suggestions and support. For helpful comments and discussions, I thank Dmitry Arkhangelsky, Krishna Dasaratha, Jihye Jeon, Shakeeb Khan, Zhongjun Qu, Pedro Sant'Anna, Stephen Wager, Linchuan Xu, José Zubizarreta, Liang Zhong, and participants of the BU Econometrics Seminar.

†Ph.D. Candidate in Economics, Boston University. Email: hpan51@bu.edu

1 Introduction

I study the identification and estimation of the treatment effect under network interference. The interference operates through endogenous peer effects (Manski (1993)), allowing individual outcomes to depend on peer outcomes. Treating an individual thereby affects every other individual in a connected network, leading to the absence of valid control group. For example, Cai et al. (2015) provide information on a weather insurance product to randomly selected farmers in villages, who spread the information to their peers. However, one may expect the communication process to continue and the treatment thus affects all acquaintances of the treated farmers. Comparing the outcome of treated and untreated farmers is subject to bias due to the informational spillover.

Existing methods related to this problem rely on strong assumptions. One approach is to assume linearity and construct instrumental variables (IV) under the assumption of exogenous networks (Bramoullé et al. (2009)). However, people with a stronger desire for information may form links selectively, which causes the network to be endogenous. Banerjee et al. (2024) show that the network structure can be directly affected by treatment assignment, which also points towards the endogeneity of networks. Another approach is to assume that the spillover only propagates in short distances and depends only on the treatment assignment. However, under endogenous peer effects, the spillover propagates to distant agents and depends on the shocks of other agents. More detailed discussions are in Section 2.

To address this problem, I construct a nonparametric peer effect model and define the treatment effect as the difference in expected outcome conditional on the average outcome of peers. Under the additively separable error assumption, I provide a novel identification argument of the proposed treatment effect by comparing the individuals with the same neighbors and different treatment status. This argument allows for endogenous networks and the dependence of the treatment assignment on networks, thereby allowing for some endogeneity of the treatment. A consistent nonparametric estimator of the treatment effect is then constructed. In more detail, the paper proceeds in three steps.

In the first step, I model the individual outcome as a nonparametric function of the average neighbor outcome, treatment status, and observed characteristics. The spillover is modeled as the average neighbor outcome, a low-dimensional statistic, which is similar to the literature on peer effects.[1] The flexible functional form allows for rich interaction between the treatment and spillover, and individual heterogeneity in terms of observed characteristics. I then define the treatment effect of interest as the difference in expected outcome, fixing the

[1] For example, the average is adopted in Calvó-Armengol et al. (2009), maximum appears in Tao and Lee (2014), minimum is modeled in Bietenbeck (2020), a CES-type aggregator is studied by Boucher et al. (2024), and the quantile is analyzed by Houndetoungan (2025).

level of average neighbor outcomes and observed characteristics. This relates to the optimal treatment assignment and reveals the relationship between treatment and spillover. For example, when the treatment effect decays with the average neighbor outcome, substitution is likely present.

In the second step, I prove that the proposed treatment effect is identified by comparing individuals with the same neighbors but different treatment statuses. The endogeneity of the average neighbor outcome intertwines with the nonparametric functional form, posing identification challenges. This is further exacerbated by the concern for endogenous networks, which limits our ability to find IVs. I solve this problem by comparing individuals whose shocks are correlated with the average neighbor outcome in the same manner. More specifically, under the additively separable error assumption, I show that the average neighbor outcome depends on the shocks of these individuals through a symmetric statistic. The equality in conditional mean of the shocks is then established, leading to identification. Since I am comparing individuals endogenous in the same way, the method can accommodate the endogeneity of networks, individuals characteristics, and the treatment assignment as a function of the endogenous networks and characteristics.

In the third step, this paper provides a nonparametric estimator of the treatment effect and establishes its consistency. The identification argument compares individuals with the same neighbors, which may lead to few observations in finite samples. However, standard multivariate kernels provide insufficient smoothing because the dimension of the set of neighbors grows at the same rate as the sample size. I tackle this problem by smoothing with respect to a low-dimensional variable defined as the ratio of the number of different links to the degree. Next, I relate this ratio to the degree of endogeneity, thereby quantifying the order of the bias from smoothing. A kernel estimator is then adopted to relax the same neighbor condition in finite samples, and the method of sieve is used to flexibly model the treatment effect as a function of the spillover and individual characteristics. The consistency of this estimator in L^2-norm is established.

The method is then applied to the data from Paluck et al. (2016) who studies the effect of anti-violence campaigns in schools. An index of individual attitude against violence is constructed as the outcome variable. The index is constructed such that students with more optimistic views of the degree of violence in their schools and more positive attitude towards anti-violence acts get higher score. The result suggests that the treatment effect is increasing in average neighbor attitude, which can be viewed as complementarity. Students may be further encouraged by their friends having more positive attitudes.

The paper is organized as follows: Section 2 reviews the related literature. Section 3 describes the model, establishing the existence and uniqueness of the reduced form and the structure of the reduced form. Section 4 studies the identification of the causal effect of interest. Section 5 discusses the estimator for the causal effect and its consistency. Section 6 provides simulation evidence and Section 7 applies the method to empirical data. Section 8 concludes. The figures and tables are collected in Appendix A. The technical lemmas and proofs of the results are in Appendix B.

2 Related Literature

This paper is related to the strand of literature studying the identification of peer effects. Using a linear-in-means model, Manski (1993) argues that the reflection problem hinders identification of group effects while the coefficients of individual characteristics directly affecting outcomes are identified. This paper shows that the identification of the effect of individual characteristics (treatment status) still holds under more flexible functional forms. My identification argument is similar to Graham and Hahn (2005), who show that the endogenous peer effect acts like group fixed effect when all nodes are connected within a block. I extend this idea by finding nodes with the same neighbors. These paper assume specific structure where people are connected to each other within a group. In a more general network setup, Bramoullé et al. (2009) show that the endogenous peer effect can be identified using exogenous characteristics of neighbors that are two (or more) steps away. This relies on the assumption of exogenous networks, which is relaxed in this paper. Nonetheless, the paper identifies the treatment effect instead of the endogenous peer effect. Griffith (2024) solves the endogeneity of the network by characterizing the endogeneity as an omitted variable problem. He adopts parametric model of network formation and identifies the latent variables affecting both network formation and outcome. In contrast, I adopt a more flexible functional form and do not explicitly model the network formation process. As mentioned above, I can only identify the treatment effect whereas Griffith (2024) identifies the peer effect as well. Other articles have considered identification through variance restrictions in the linear-in-means model (Graham (2008), Rose (2017)), and using panel data structure (Manresa (2013), Miraldo et al. (2021)). Rose and Yu (2022) consider misspecified peer groups. My approach imposes mean restrictions and uses cross-sectional data. In terms of modeling individual heterogeneity, Carrell et al. (2013), Masten (2018) and Griffith (2024) construct linear models with random coefficients to model heterogeneity. In contrast, the heterogeneity in this paper stems from the interaction between treatment, spillover, and observed characteristics under flexible functional form.

The interference structure studied in this paper bears resemblance to the literature on equilibrium treatment effect. Munro et al. (2025) and Munro (2025) study treatment effect in centralized markets. In Munro et al. (2025), agents affect each other through the market clearing price whereas the spillover is captured by average neighbor outcome in this paper.

They use mean-field approximation under the assumption of many agents and generates price variations through augmented experiments. I do not directly restrict the density of the network. However, when the network is dense, variations in the average neighbor outcome can be limited and the treatment effect is only identified on a few values of the average neighbor outcome. The setup in this paper is also related to Menzel (2025) who studies marginal effects conditional on the treatment status of neighbors that are two or more steps away. In contrast, this paper studies unconditional treatment effect that depends on average neighbor outcomes, which is assumed to be a sufficient statistic for the treatment status of other connected individuals.

In the statistics literature, Hudgens and Halloran (2008) construct average direct and indirect effects in two-stage, hierarchical randomized experiments, and the inference results are provided by Tchetgen and VanderWeele (2012). Under greater generality, Aronow and Samii (2017) captures the interference through an arbitrary known function of the treatment vector, which is referred to as the exposure map. Manski (2013) studies this problem under sampling-based uncertainty and call it effective treatments. This leads to natural definitions of direct (treatment) and indirect (spillover) effects. Leung (2020) applies this idea to network problems and assumes that the exposure map depends on the share of treated neighbors. Although the direct effect is easy to define, the indirect effect often differs across contexts and is sensitive to assumptions. Hu et al. (2022) provides a form of the indirect effect that requires less assumptions and is sensible under different kinds of interference pattern. Li and Wager (2022) build on this and provide estimation and inference results using graphon. Wang et al. (2025) study a similar problem in spatial contexts. Some recent studies relax the assumption of correct specification of the exposure map (Sävje et al. (2021), Leung (2022), Sävje (2024)). However, specifying a correct exposure map is difficult in the nonparametric peer effect model in this paper. The dependence on peer outcome generates high-dimensional exposure maps even in the linear model. Under the flexible functional form assumption, the entries of the exposure map become complex functions of treatment and individual shocks. Specifying an approximately correct exposure map as in Leung (2022) is difficult due to the same reason. As for the general direct and indirect effects, they are stated as functions of the treatment assignment. In contrast, the problem I study features spillover that depends on outcome instead of just the treatment assignment. This paper contributes to this literature by considering the case where the exposure map depends on endogenous variables other than the treatment assignment, and provides identification and estimation results.

In terms of proof strategy, this paper relates to the work by Sasaki (2025) on the consistency of GMM and M estimators for finite-dimensional parameters in the context of network ψ-dependence (Kojevnikov et al. (2021)). However, this paper focuses on infinite-dimensional parameters and provides a consistency result for a special case of M estimator for such parameter.

Finally, this paper relates to the strand of literature on information provision. Recent studies have examined information provision in diverse domains including insurance (Cai et al. (2015), Chemin (2018)), gun violence (Wood and Papachristos (2019)), corporate tax visits (Boning et al. (2020)), new technology (Beaman et al. (2021)), property rights (Aberra and Chemin (2025)), biased belief (Wagner et al. (2025)). I contribute to the literature by providing a new methodology to study treatment effect under information provision. In addition, I provide evidence of the nonlinearity of the treatment effect of anti-violence campaign, which is increasing in the average neighbor outcome.

3 Setup

For any matrix B, B_{ij} denote the (i,j)-th entry of B. For a random variable W, let $supp(W)$ denote its support. Bold-faced letters are used to denote vectors. For example, $\mathbf{Y}$ denote the vector $(Y_1, \cdots, Y_n)'$. Functions with vector-valued outputs are also denoted in bold face.

Assume that the researcher observes n agents represented as nodes in a network with adjacency matrix A. Let $\tilde{A}$ denote the row-normalized adjacency matrix. Let $Y_i \in supp(Y) \subseteq \mathbb{R}$ be the outcome of node i, $T_i \in \{0,1\}$ be the treatment status of node i. Also let $X_i \in supp(X) \subseteq \mathbb{R}^{d_X}$ be the characteristics of node i and $v_i \in supp(v) \subseteq \mathbb{R}^{d_v}$ be the unobserved shock received by node i. It is assumed that the researcher observes $\{Y_i, T_i, X_i\}_{i=1}^n$ and the adjacency matrix A. The focus of this paper is on experimental contexts and the treatment T is assumed to be randomly assigned.

Example 3.1 *Cai et al. (2015) provide information on a weather insurance product to randomly-chosen farmers in rural Chinese villages. The information provision takes the form of information sessions where staffs convey important details about the product including price and coverage. A farmer i is treated ($T_i = 1$) if he/she attends the information session. One outcome variable (Y) of interest is the knowledge of farmers regarding the insurance product. This is measured by the share of correctly answered questions regarding the details of the product by the farmer. Examples of individual characteristics (X) include education level, age, income, past experience of drought. The network adjacency matrix A is measured by asking the farmers to list their friends. A farmer i is connected to another farmer j ($A_{ij} = 1$) if i lists j as his/her friend.*

To model the dependence on peer outcome, I consider the following nonlinear peer effect model:

$$\forall i: \quad Y_i = g(D_i, T_i, X_i, v_i)$$
$$D_i \coloneqq \frac{1}{n_i}\sum_{j=1}^{n} A_{ij}Y_j \quad n_i \coloneqq \sum_{j=1}^{n} A_{ij} \tag{1}$$

The knowledge of node i (Y_i) is affected by the treatment status of i (T_i), the average knowledge of the neighbors (D_i), and the characteristics of i (X_i). However, D_i is affected by the knowledge of the neighbors' neighbors. Therefore, the knowledge of node i affected by every other node that can reach i. This captures knowledge transmission and its dependence on the network structure. The model is a continuous analogue of the contagion model where the transmission of actions occurs if the share of such action among neighbors exceeds a certain threshold (Morris (2000), Centola and Macy (2007), Centola (2010)).

Example 3.2 *Let Y_i be a binary variable. $Y_i = 1$ represents taking a certain action. The contagion model can be written as $Y_i = 1\{D_i \geq \frac{\alpha}{n_i}\}$ where $n_i = \sum_j A_{ij}$ is the degree of node i and $D_i = \frac{1}{n_i}\sum_j A_{ij}Y_j$ is the share of neighbors taking the action. α is contagion threshold for the number of sources. $\alpha = 1$ is the simple contagion model and $\alpha \geq 2$ captures complex contagion. We could also allow for the threshold α to differ across individuals. For instance, $\alpha_i = p(X_i, T_i, v_i)$ allows the threshold to depend on the characteristics, treatment status, and unobserved shocks of that individual. A treatment that subsidizes the action would reduce the threshold.*

The contagion model is typically adopted in the literature to capture the spread of actions in a network. A treatment that encourages some people to take the action could lead to a spread of adoption. However, the outcomes are likely continuous when studying information provision. As an example, Cai et al. (2015) measures the knowledge of farmers about an insurance product by computing the share of correctly answered questions about the product. The contagion model relates to the model in Equation 1 if we take $g = g_1 \circ g_2$ where $g_1(y) = 1\{y \geq 0\}$ and $g_2(d, t, x, v) = d - p(t, x, v)$ for some threshold function $p()$.

Example 3.3 *Consider $g(D_i, T_i, X_i, v_i) = \tilde{g}(\beta_1 D_i + \beta_2 T_i + X_i'\beta_3) + v_i$. Assume that $\tilde{g}$ is a strictly concave function and $\beta_1 > 0, \beta_2 > 0$. Being treated ($T_i = 1$) improves one's knowledge. Having more knowledgeable neighbors (higher D_i) diminishes this effect. The specification adopted in Cai et al. (2015) assumes that $\tilde{g}$ is a linear function which implies constant treatment effect.*

Equation (1) naturally leads to the following causal objects:

$$\tau_T(d, x) \coloneqq E_v[g(d, 1, x, v) - g(d, 0, x, v)] \tag{2}$$
$$\tau_D(d, d'; t, x) \coloneqq E_v[g(d, t, x, v) - g(d', t, x, v)] \tag{3}$$

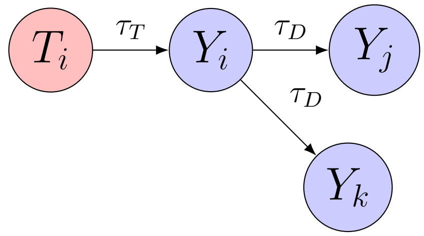


Figure 1: Causal Effects

The first term $\tau_T(d, x)$ is the difference in counterfactual outcomes under different treatment status, conditional on the observed characteristics and the average knowledge of the neighbors. The causal effect τ_T can be viewed as evaluating the immediate impact of treatment at a particular level of spillover, sharing similar intuition as the Average Partial Causal Effect in Bugni et al. (2025). The causal effect in Bugni et al. (2025) is 'immediate' as it does not account for adjustment of actions induced by the treatment. The effect τ_T is 'immediate' as it does not account for the subsequent informational spillover between nodes. A similar nonlinear treatment effect is considered in the shift-share design by Garzon and Possebom (2025) where the nonlinearity is with respect to the continuous treatment intensity. The identification argument in this paper can also accommodate continuous treatment (T_i) but the major focus is on the nonlinearity with respect to the spillover. The second term $\tau_D(d, d'; t, x)$ is the spillover effect. It measures the difference in expected outcome under different levels of average neighbor outcome. To better understand the causal effects, consider assigning treatment to an individual i in network. The treatment leads to an immediate increase in i's knowledge measured by τ_T. Through communication, i spreads the knowledge to the neighbors, which measured by τ_D. This is depicted in Figure 1.

The causal effects are useful for three reasons. First, the value of $\tau_T(d, x)$ at different values of d provides information on substitutability / complementarity between the information obtained from neighbors and the treatment. If the $\tau_T(d, x)$ is decreasing in d, substitution between the informational treatment and spillover is likely present. Second, the treatment effect and the spillover effect together enables the measurement of the impact of a counterfactual treatment assignment to another network, which is made precise in Example 3.4. Third, under some particular network structure, $\tau_T(d, x)$ along determines the optimal treatment assignment, which is illustrated in Proposition 3.1.

Example 3.3. (Continued) *The causal effects take the following form:*

$$\begin{aligned} \tau_T(d, x) &= \tilde{g}(\beta_1 d + \beta_2 + x'\beta_3) - \tilde{g}(\beta_1 d + x'\beta_3) \\ \tau_D(d, d'; t, x) &= \tilde{g}(\beta_1 d + \beta_2 t + x'\beta_3) - \tilde{g}(\beta_1 d' + \beta_2 t + x'\beta_3) \end{aligned} \tag{4}$$

When $\tilde{g}$ is a concave function and $\beta_1, \beta > 0$, $\tau_T(d, x)$ is a decreasing function for any given x. In this case, treatment and spillover are substitutes for knowledge acquisition. A sparse assignment may be optimal when treating a fixed number of nodes. If instead $\tilde{g}$ is convex, then the assignment should be clustered to utilize the complementarity.

Example 3.4 *Consider a network of n' nodes and the row-normalized adjacency matrix $\bar{A}$. The vector of initial knowledge level is $\mathbf{Y}$. Consider assigning a vector of treatment $\mathbf{t}$. Let $\mathbf{Y}^*$ be the equilibrium knowledge after the treatment $\mathbf{t}$. This example decomposes $\mathbf{Y}^* - \mathbf{Y}$ as a sum of the causal effects τ_T, τ_D.*

For simplicity, ignore the covariates $\mathbf{x}$. Also, assume that $g(d, t, v) = \bar{g}(d, t) + v$. This assumption is the key to identification which is later imposed in Assumption 4.1. Let $\odot$ denote the point-wise multiplication. Define the following objects:

$$
\begin{aligned}
&\mathbf{Y}_{(0)} \coloneqq \mathbf{g}(\bar{A}\mathbf{Y}, \mathbf{0}, \mathbf{v}) = \mathbf{Y} \\
&\mathbf{Y}_{(1)} \coloneqq \mathbf{g}(\bar{A}\mathbf{Y}, \mathbf{t}, \mathbf{v}) \\
&\mathbf{Y}_{(s)} \coloneqq \mathbf{g}(\bar{A}\mathbf{Y}_{(s-1)}, \mathbf{t}, \mathbf{v}) \quad s \geq 2 \\
&\triangle_{(1)} \coloneqq Y_{(1)} - Y_{(0)} = \mathbf{g}(\bar{A}\mathbf{Y}, \mathbf{t}, \mathbf{v}) - \mathbf{g}(\bar{A}\mathbf{Y}, \mathbf{0}, \mathbf{v}) = \bar{\mathbf{g}}(\bar{A}\mathbf{Y}, \mathbf{t}) - \bar{\mathbf{g}}(\bar{A}\mathbf{Y}, \mathbf{0}) = \tau_T(\bar{A}\mathbf{Y}) \odot \mathbf{t} \\
&\triangle_{(s)} \coloneqq \mathbf{Y}_{(s)} - \mathbf{Y}_{(s-1)} = \bar{\mathbf{g}}(\bar{A}\mathbf{Y}_{(s-1)}, \mathbf{t}) - \bar{\mathbf{g}}(\bar{A}\mathbf{Y}_{(s-2)}, \mathbf{t}) = \tau_D\left(\bar{A}\mathbf{Y}_{(s-1)}, \bar{A}\mathbf{Y}_{(s-2)}; \mathbf{t}\right) \quad s \geq 2
\end{aligned} \tag{5}
$$

As will be shown in the proof of Proposition 3.2, the contraction mapping theorem ensured by Assumption 3.2 implies that:

$$
\mathbf{Y}^* - \mathbf{Y} = \sum_{s=1}^{\infty} \triangle_{(s)} = \tau_T(\bar{A}\mathbf{Y}) \odot \mathbf{t} + \sum_{s=2}^{\infty} \tau_D\left(\bar{A}\mathbf{Y}_{(s-1)}, \bar{A}\mathbf{Y}_{(s-2)}; \mathbf{t}\right) \tag{6}
$$

The term $\triangle_{(1)}$ is the immediate change in knowledge induced by the treatment assignment while $\triangle_{(s)}$ for $s \geq 2$ are the changes in knowledge due to the spillover effect. The above process can be viewed as an infinite-step adjustment to the new equilibrium where s represents the step. Initially, the treatment assignment induces an immediate impact $\triangle_{(1)}$. In step 2 and onward, the knowledge level in the network keeps adjusting through the spillover effects.

More can be said regarding the optimal assignment under specific network structure and sign restriction of the treatment effect (Assumption 3.3, 3.4 below). Formally, consider the problem of assigning treatment to $m < n'$ individuals in the network with the goal of maximizing the average outcome $\sum_{i=1}^{n'} Y_i$. This can be viewed as a result of budget-constrained maximization problem where treating each node is equally costly.

Proposition 3.1 *Continue with the setup in Example 3.4. Assume that Assumption 3.2, 3.3, 3.4 hold. Further assume that the network $\bar{A}$ is fully connected: $\bar{A}_{ij} = \frac{1}{n'}$ for any i, j.*

For two treatment $\mathbf{t}_1, \mathbf{t}_2$, denote the result equilibrium knowledge $\mathbf{Y}_1^*, \mathbf{Y}_2^*$. Then the following holds: $\mathbf{1}'\mathbf{Y}_1^* > \mathbf{1}'\mathbf{Y}_2^*$ if and only if $\tau_T(\bar{A}\mathbf{Y})'\mathbf{t}_1 > \tau_T(\bar{A}\mathbf{Y})'\mathbf{t}_2$.

This result suggests that the determination of optimal treatment assignment in a network boils down to the comparison of τ_T when all agents are connected to each other within a block. This highlights the importance of τ_T.

3.1 Existence and Uniqueness of the Reduced Form

The data $(\mathbf{X}, \mathbf{T}, \mathbf{v}, A)$ is assumed to be drawn from an underlying distribution. Restrictions will be placed on this underlying distribution when it comes to the identification part. The goal of this current section is to establish the existence and uniqueness of the reduced form equation. To proceed, I make the following assumptions:

Assumption 3.1 *The true data generating process for the knowledge vector $\mathbf{Y}$ is:*

$$\begin{aligned} \forall i: \quad & Y_i = g(D_i, T_i, X_i, v_i) \\ & D_i := \frac{1}{n_i}\sum_j A_{ij}Y_j \quad n_i := \sum_j A_{ij} \end{aligned} \tag{7}$$

where $g : \mathcal{G} \to supp(Y)$ is some measurable function and $\mathcal{G}$ is a polish space such that $supp(Y) \times \{0,1\} \times supp(X) \times supp(v) \subseteq \mathcal{G}$.

Assumption 3.2 *There exists a constant $\kappa \in (0,1)$ such that $\frac{\partial}{\partial D}g(D_i, T_i, X_i, v_i) \leq \kappa < 1$ for all realization of D_i, T_i, X_i, v_i*

This assumption restricts the strength of knowledge spillover and prevents explosive behavior. In words, the assumption requires that if the average knowledge of node i's neighbors increases by a unit, the resulting increase in i's knowledge is less than one unit. If the increase is more than one, treating every node could lead to unbounded knowledge. This assumption restricts the influence of distant nodes and is important to establish the ψ-dependence condition for the consistency of the estimator. The following two examples illustrate this assumption under the linear-in-means model and the model considered in Example 3.3.

Example 3.5 *Assume that $Y_i = g(D_i, T_i, X_i, v_i) = \beta_1 D_i + \beta_2 T_i + X_i\beta_3 + v_i$ holds for all i. The equation can thus be written in vector form: $\mathbf{Y} = \beta_1\mathbf{D} + \beta_2\mathbf{T} + \mathbf{X}\beta_3 + \mathbf{v}$ and $\mathbf{D} = \tilde{A}\mathbf{Y}$. The assumption that $|\frac{\partial}{\partial D}g(D_i, T_i, X_i, v_i)| = |\beta_1| < 1$ implies that the matrix $I - \beta\tilde{A}$ is diagonal dominant, hence invertible. The unique reduced form is $\mathbf{Y} = (I - \beta\tilde{A})^{-1}(\beta_2\mathbf{T} + \mathbf{X}\beta_3 + \mathbf{v})$. When $|\beta_1| > 1$, the system becomes explosive.*

Example 3.3. (Continued) *The restriction that* $\frac{\partial}{\partial D} g(D_i, T_i, X_i, v_i) \leq \kappa < 1$ *amounts to* $|\beta_1 \frac{d}{dy} \tilde{g}(y)| \leq \kappa < 1$ *for all values of* y*.*

Assumption 3.1 states the data generating process for $\mathbf{Y}$ as a solution to a system of equations and Assumption 3.2 implies that such solution exists and is unique, which is formalized by point 1 of Proposition 3.2. The following assumptions add additional sign restrictions on the structural model.

Assumption 3.3 $g(D_i, 1, X_i, v_i) - g(D_i, 0, X_i, v_i) \geq 0$ *for all realizations of* D_i, X_i, v_i

Assumption 3.4 *For any value* $d > d'$, $g(d, T_i, X_i, v_i) - g(d', T_i, X_i, v_i)$ *takes the same sign as* $g(D_i, 1, X_i, v_i) - g(D_i, 0, X_i, v_i)$ *for all realizations of* T_i, X_i, v_i

Assumption 3.3 is a monotonicity assumption on the treatment effect in the knowledge equation allowing for heterogeneous effects. Assumption 3.4 restricts the sign of the effect of the average knowledge of the neighbors to be the same across different values, and the same as the treatment effect. For example, if being treated increases knowledge, having more knowledgeable peers should also increase knowledge. As a side note, the sign can be negative in Assumption 3.3, 3.4, provided that they are the same. This assumption is reasonable in the informational treatment context since communication between agents and the information provision are unlikely to deprecate understanding.

Example 3.3. (Continued) *Assumption 3.3 can be satisfied if* $\tilde{g}(.)$ *is monotonic. Assumption 3.4 can be satisfied if* $\tilde{g}(.)$ *is monotonic and* β_1, β_2 *take the same sign.*

The following proposition establishes the existence of a unique reduced form, which justifies Assumption 3.1 and is important for the subsequent analysis of the identification.

Proposition 3.2 *Let* $g : supp(Y) \times \{0, 1\} \times supp(X) \times supp(v) \to supp(Y)$ *be some unknown measurable function as in Equation (7). Assume Assumption 3.2 holds. The following statements are true:*

1. *The simultaneous equation system stated in Equation (7) admits a unique solution:* $\mathbf{Y} = \mathbf{r}(\mathbf{T}, \mathbf{X}, \mathbf{v})$ *where* $\mathbf{r} : \{0, 1\}^n \times supp(X)^n \times supp(v)^n \to supp(Y)^n$ *is some measurable function.*

2. *Let* r_i *be the i-th entry of* $\mathbf{r}$ *in the point above. If Assumption 3.3, 3.4 also hold,* $Y_i = r_i(\mathbf{T}, \mathbf{X}, \mathbf{v})$ *is non-decreasing in* T_j *for any* j *and strictly increasing for some* j*.*

Furthermore, the above conclusions still hold if D_i *is replaced by* $\sum_i w_i Y_i$ *where* w_i *is such that* $w_i \in [0, 1]$ *and* $\sum_i w_i = 1$*.*

The above proposition suggests that the equilibrium knowledge $\mathbf{Y}$ can be expressed as a function of $\{\mathbf{T}, \mathbf{X}, \mathbf{v}\}$. The identification results in the subsequent section use this property to separate restrictions on v_i, v_j from the restrictions on $\{T_k, X_k, v_k\}_{k \neq i,j}$.

Remark 3.1 *From the above proposition, it follows that the following system of knowledge equations also admits a unique reduced form:*

$$\begin{aligned}
\forall i: \quad Y_i &= g(D_i^*, T_i, X_i, v_i) \\
D_i^* &= \frac{1}{\sum_j A_{ij} w_i(X_j, T_j, v_j)} \sum_j A_{ij} w_i(X_j, T_j, v_j) Y_j \\
w_i(X_j, T_j, v_j) &\geq 0
\end{aligned}$$

where $w_i(X_j, v_j)$ represents the weight placed by i on neighbor j, which depends on both observed and unobserved characteristics. This system of equations allows each agent i to place different weights on the knowledge of different neighbors (Griffith (2024)).

3.1.1 Reduced Form and Exposure Map

It is helpful to consider the structure of the reduced form equation $\mathbf{r}$. Start with the linear-in-means model as in Example 3.5 where $Y_i = g(D_i, T_i, X_i, v_i) = \beta_1 D_i + \beta_2 T_i + X_i \beta_3 + v_i$ for all i. Assume that $|\beta_1| < 1$. The reduced form in matrix notation is thus

$$\mathbf{Y} = \sum_{t=0}^{\infty} \beta_1^t \tilde{A}^t (\beta_2 \mathbf{T} + \mathbf{X}\beta_3 + \mathbf{v}) \tag{8}$$

where $\tilde{A}^t$ is the t-th power of the row-normalized adjacency matrix $\tilde{A}$. The term $\tilde{A}^t_{ij} > 0$ if j can be reached from i with t steps (a walk of length t exists between i, j). Another property of $\tilde{A}^t$ is such that the row-sum equals one. This implies that knowledge depends on the average of the treatment status, characteristics, and shocks of the neighbors at different distances. More formally, we can express the outcome of node i as:

$$Y_i(\mathbf{T}, \mathbf{X}, \mathbf{v}) = Y_i(\{(\tilde{A}^t)'_i(\mathbf{T}, \mathbf{X}, \mathbf{v})\}_{t=0}^{\infty})$$

where $(\tilde{A}^t)_i$ is the i-th row of the matrix $\tilde{A}^t$. This implies that, holding fixed $\mathbf{X}, \mathbf{v}$, for any two vectors of treatment assignment $\mathbf{T}, \tilde{\mathbf{T}}$:

$$Y_i(\mathbf{T}) = Y_i(\tilde{\mathbf{T}}) \quad \text{iff} \quad (\tilde{A}^t)'_i \mathbf{T} = (\tilde{A}^t)'_i \tilde{\mathbf{T}} \ \forall t = 0, \cdots, \infty$$

The vector $\{(\tilde{A}^t)'_i \mathbf{T}\}_{t=0}^{\infty}$ is called the exposure map, which summarizes the spillover as a function of the vector $\mathbf{T}$. In the present context, it is a high-dimensional statistic that consists of average treatment status at different distances. Since $|\beta_1| < 1$, the influence of distant nodes decays at a geometric rate. This model thus falls within the framework of approximate neighborhood interference by Leung (2022). Essentially, we can compare individuals with the same value of $\{(\tilde{A}^t)_i' \mathbf{T}\}_{t=0}^M$ for some value M that increases with the sample size.

Next consider the nonlinear model and expand the terms:[2]

$$\begin{aligned} Y_i = g(D_i, T_i, X_i, v_i) &= g\left(\sum_j \tilde{A}_{ij} g(D_j, T_j, X_j, v_j), T_i, X_i, v_i\right) \\ &= g\left(\sum_j \tilde{A}_{ij} g\left(\sum_k \tilde{A}_{jk} g(D_k, T_k, X_k, v_k), T_j, X_j, v_j\right), T_i, X_i, v_i\right) \end{aligned}$$

Define $N_i(L) \coloneqq \{j : \tilde{A}_{ij}^L > 0\}$ as the set of neighbors that can be reached within L steps from node i. Let $\mathbf{D}_i(L) \coloneqq (D_j : j \in N_i(L))$ be the vector of average knowledge of the nodes that can be reached in L steps from i. Denote the following set of variables:

$$\begin{aligned} \sigma_{j,L} &\coloneqq g(D_j, T_j, X_j, v_j) \ \forall j \in N_i(L) \\ \sigma_{j,l} &\coloneqq g\left(\sum_k \tilde{A}_{jk} \sigma_{k,l+1}, T_j, X_j, v_j\right) \ \forall j \in N_i(L-l), 1 \le l \le L-1 \\ \sigma_{i,0} &\coloneqq Y_i = g\left(\sum_j \tilde{A}_{ij} \sigma_{j,1}, T_i, X_i, v_i\right) \end{aligned} \tag{9}$$

The above process represents expanding the knowledge equation L times starting from node i. The term $\sigma_{j,l}$ represents the knowledge of a node j that can be reached within l steps from node i, and the term $\sigma_{i,0}$ is the knowledge of node i. The above representation suggests that for arbitrary L, Y_i depends on $\mathbf{D}_i(L)$ and $\{T_j, X_j, v_j\}$ for all $j \in \bigcup_{l \le L} N_i(l)$.

Unlike the linear case, the dependence of Y_i on $\mathbf{T}$ does not boil down to low-dimensional statistics such as the share of treated neighbors at different distances. This is because g is unknown and not additively separable in T. Specifying an approximately correct exposure map would require knowledge of the unknown function g. The true exposure map is now a high-dimensional vector consisting of unknown statistics. Imposing assumptions on the structure of the exposure map can be viewed as implicitly imposing restrictions on the function g.

Another difficulty that arises in the nonlinear model is the non-separability of $\mathbf{T}, \mathbf{v}$. The treatment effect depends on the potential outcome of the other individuals. In the linear

[2] The argument in Proposition 3.2 shows that such iteration eventually converges to the reduced form.

model described by Equation (8), $\mathbf{v}, \mathbf{T}$ are additively separable and the resulting exposure map $\{(\tilde{A}^t)_i'\mathbf{T}\}_{t=0}^{\infty}$ is exogenous under the sampling-based uncertainty view. This breaks down in the nonlinear context since g is not additively separable in D_i, T_i.[3] This means that the true exposure map also depends $\mathbf{v}$, leading to endogeneity. Comparison of the average outcome across different levels of exposure suffers from the endogeneity problem.

4 Identification

4.1 Identification of the Treatment Effect

The key challenge to identification is the correlation between D_i and v_i. Consider the specification $Y_i = \tilde{g}(D_i, T_i, X_i) + v_i$. The naive difference $E[Y_i|D_i = d, T_i = 1, X_i = x] - E[Y_j|D_j = d, T_j = 0, X_j = x]$ does not identify $\tau_T(d, x)$ because the conditioning events reflect $E[v_i|D_i = d, T_i = 1, X_i = x] \neq E[v_j|D_i = d, T_j = 0, X_i = x]$. To see this, consider the case depicted in Figure 2 with two pairs of links $(i, k), (j, l)$ and i being the only treated individual (colored in red). Assume that the outcome follows a linear-in-means model:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 T_i + X_i\beta_3 + v_i \tag{10}$$

Assume that the treatment and spillover effects are both positive ($\beta_1 > 0$ and $\beta_2 > 0$). The naive difference estimator is comparing nodes i, j with $D_i = D_j, X_i = X_j, T_i \neq T_j$. However, D_i is an increasing function of T_i, v_i, which implies that $v_i < v_j$. In words, an individual is more knowledgeable if he/she is treated or has higher level of shocks. This then increases the average neighbor knowledge D_i through spillover. However, if i, j have the same average neighbor knowledge but different treatment status, the shocks of i must be lower. This is the problem of endogenous peer effects.

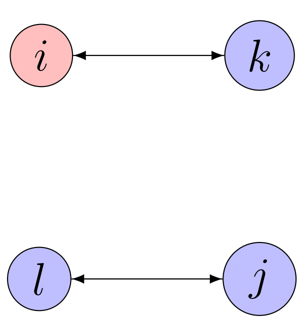


Figure 2: Endogeneity

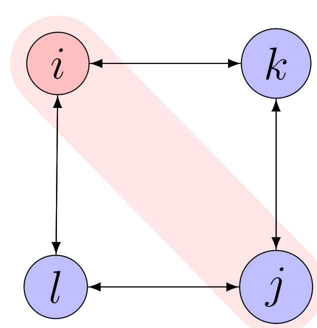


Figure 3: Identification

[3] Even if $g(D_i, T_i, X_i, v_i) = \tilde{g}(D_i, T_i, X_i) + v_i$, the terms $\mathbf{v}, \mathbf{T}$ are still non-separable due to the simultaneity problem induced by D_i, which is an equilibrium object.

If the model is fully parametric as in Equation (10), and the network is exogenous, instrumental variable (IV) approach using the share of treated neighbors as the instrument would suffice. Under flexible functional form, one may still use nonparametric IV (Newey and Powell (2003)) or generalized IV (Chesher and Rosen (2017)) to relax the parametric structure. However, when the network is endogenous, it is in general hard to find suitable instruments. Endogeneity of the network is a valid concern because individuals with better understanding may form links in different ways than their less knowledgeable peers.

This paper takes another route to address this problem. Under the additively separable structure, the essential condition needed is $E[v_i - v_j | D_i = D_j, T_i = 1, T_j = 0, X_i = X_j, \mathcal{E}] = 0$ for some conditioning event $\mathcal{E}$.[4] Suppose that the event $\mathcal{E}$ is such that $E[v_i - v_j | D_i = D_j, T_i = 1, T_j = 0, X_i = X_j, \mathcal{E}] = E[v_i - v_j | s(v_i, v_j), T_i = 1, T_j = 0, X_i = X_j, \mathcal{E}]$ where $s(v_i, v_j)$ is a symmetric function. If in addition that v_i, v_j are i.i.d. conditional on $\{T_i = 1, T_j = 0, X_i = X_j, \mathcal{E}\}$, it is true that $E[v_i - v_j | s(v_i, v_j), T_i = 1, T_j = 0, X_i = X_j, \mathcal{E}] = 0$, which is proven in Lemma B.3. The overall argument is that conditional on the event $\mathcal{E}$, the endogenous variables D_i, D_j depend on v_i, v_j only through a symmetric function $s(v_i, v_j) = s(v_j, v_i)$. Since v_i, v_j are conditionally i.i.d., their expectation remains the same after further conditioning on a symmetric variable $s(v_i, v_j)$.

The above analysis immediately highlights the relation between the method in this paper and the control function approach (see Wooldridge (2015) for a review). The above argument can be understood as follows: conditional on $\mathcal{E}$, the symmetric quantity $s(v_i, v_j)$ is a control function for D_i, D_j. However, this paper does not estimate this quantity, in contrast to the control function approach (i.e. Newey et al. (1999)). This is because the control function is a symmetric quantity, which suffices to establish the required equality in conditional expectation. This is also a weaker result compared to the control function approach, which typically establishes conditional exogeneity.

It remains to find the event $\mathcal{E}$. Recall that $Y_k = \tilde{g}(D_k, T_k, X_k) + v_k$ depends on v_i through D_k, and the dependence of D_k on v_i happens through the quantity $A_{ki}v_i$.[5] This implies that D_i, D_j depends on v_i, v_j through the vector $\{A_{ki}v_i + A_{kj}v_j\}_{k \neq i,j}$. If $A_{ki} = A_{kj}$ for all k, it follows that D_k depends only on the quantity $v_i + v_j$, which is a symmetric function of v_i, v_j. One candidate for the event $\mathcal{E}$ is thus $\{A_{ki} = A_{kj} \text{ for all } k\}$. This ensures that D_i, D_j depends on v_i, v_j through a symmetric function $s(v_i, v_j) = v_i + v_j$. The idea is illustrated in Figure 3. The two individuals i, j are such that $A_{ki} = A_{kj}, A_{li} = A_{lj}$. The above argument ensures that $E[v_i - v_j | D_i = D_j, T_i = 1, T_j = 0, X_i = X_j] = 0$ and thus provides the identification result. This intuition is formalized by Proposition 4.1 below under the following assumptions:

[4] To see this, $E[Y_i - Y_j | D_i = D_j = d, T_i = 1, T_j = 0, X_i = X_j = x, \mathcal{E}] = \tilde{g}(d, 1, x) - \tilde{g}(d, 0, x) + E[v_i - v_j | D_i = D_j = d, T_i = 1, T_j = 0, X_i = X_j = x, \mathcal{E}]$.

[5] To see this, $D_k = \sum_q \tilde{A}_{kq} Y_q = \sum_q \tilde{A}_{kq}[g(D_q, T_q, X_q) + v_q]$.

Assumption 4.1 *For all i:*

1. $Y_i = g(D_i, T_i, X_i, v_i) = \bar{g}(D_i, T_i, X_i) + v_i$
2. $v_i \perp (\mathbf{T}, \mathbf{X}_{-i}, \mathbf{v}_{-i})$ *conditional on* $X_i, \{A_{ki}\}_{k=1}^n$
3. $v_i \perp \{A_{ql}\}_{l \neq i}$ *conditional on* $X_i, \{A_{ki}\}_{k=1}^n$
4. v_i, v_j *are identically distributed conditional on* $X_i = X_j, A_{ki} = A_{kj}$ *for all* k

Assumption 4.1 has three components. The first assumption on additive separability restricts the degree of unobserved heterogeneity, excluding random coefficients on treatment T_i. This paper focuses on the heterogeneity of the treatment effect across different levels of average neighbor outcomes, rather than the heterogeneity of unobserved characteristics.

The second part is a conditional exogeneity assumption. The conditional exogeneity of $\mathbf{T}$ allows the treatment assignment to depend on individual characteristics and network structure but not unobserved shocks. The conditional exogeneity of $\{A_{ql}\}_{l \neq i}$, which is the adjacency matrix A without column i, is a bit subtle. Under undirected network, this assumption says that v_i is uncorrelated with the link structure of other individuals, conditional on i's link structure. This allows for A_{ql} to depend on for example, A_{qi}, A_{li}, which represents a taste for transitivity. However, under undirected networks, this assumption implies that v_i only affects A_{ki} but not A_{ik}. In words, v_i only affects if others link to i but not whether i link to others. The following example illustrates this possibility.

Example 4.1 *An individual k obtains utility $U_{ki} = \varphi(v_i, X_i, X_k) - c$ from linking with i. The part $\varphi(v_i, X_i, X_k)$ represents the benefit of linking with i and is characterized by the characteristics of i (both observed and unobserved). The term c is a cost of forming links. The link $A_[ki]$ is formed according to a threshold-crossing rule $A_{ki} = \mathbb{1}\{U_{ki} \geq \epsilon_{ki}\}$ where ϵ_{ki} is a random shock. As a result, v_i only affects A_{ki} but not A_{ik}.*

The third part requires that the shocks be i.i.d. conditional on $X_i = X_j, A_{ki} = A_{kj}$ for all k. This allows the underlying network formation process to be driven by some unobserved heterogeneity correlated with v_i, provided that such heterogeneity is the same across nodes sharing the same neighbors. This shares similarity with the model in Auerbach (2022). However, notice that there is no restriction imposed on the level of the first moment of v_i (i.e. $E[v_i] = 0$). This is because the identification argument relies on taking differences $(Y_i - Y_j)$ and is unaffected by the levels of $E[v_i]$ provided that it is common across individuals.

Proposition 4.1 *Let Assumption 3.2, 4.1 hold. Then the treatment effect $\tau_T(d,x)$, for arbitrary (d,x) in the support, is identified by the following equation:*

$$\begin{aligned}\tau_T(d,x) &= \bar{g}(d,1,x) - \bar{g}(d,0,x) \\ &= E[Y_i - Y_j | A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, D_j = D_i = d, X_i = X_j = x]\end{aligned} \tag{11}$$

The identification argument takes the form of a simple differencing argument. In undirected networks, it is the difference in expected outcome between individuals of the same neighbors and characteristics but different treatment status. One implication is that the result can be extended to observational studies if the assumption of selection on observables holds. Essentially, the required condition is that among the individuals with the same neighbors and characteristics, the treatment is as if randomly assigned.

The result shares similar intuition as Zeleneev (2020) and Auerbach (2022). Zeleneev (2020) controls for unobserved heterogeneity by controlling for the residuals. In the model studied by Auerbach (2022), individuals with the same link structure have the same unobserved characteristics. Controlling for link structure solves the endogeneity problem created by the unobserved characteristics. The conditioning event of $A_{ki} = A_{kj}$ for all k also connects to the identification results in Graham and Hahn (2005). They show that group average outcome acts like a group fixed effect, which disappears when differencing within group. In their setup, every individual is connected to every other individual in the group. This implies $A_{ki} = A_{kj}$ for all k if i, j belong to the same group. This identification idea is also broadly related to the papers that control for unobserved heterogeneity using specific network structures (Graham (2017), Gao (2020), Gao et al. (2023)).

The conditioning event $A_{ki} = A_{kj}$ for all k if i, j may be justified by certain network formation models. The most notable example is a special case of the stochastic block model where $P(A_{ij} = 1) = p > 0$ if i, j belongs to the same group and 0 otherwise. Assuming that the group size is bounded, the conditioning event is observed more frequently as the number of blocks tends to infinity, which is inherently the 'many network asymptotics'. One caveat is that this conditioning event has different implications depending on whether self-links are allowed (i.e. $A_{ii} = 1$). If self-links are ruled out, the event implies that i, j cannot be linked, as in Figure 3.

The identification argument can be applied to both directed and undirected networks. In undirected networks, $A_{ki} = A_{kj}$ implies that $A_{ik} = A_{jk}$, which leads to $D_i = D_j$. Thus, the above argument cannot identify the spillover effect $\tilde{g}(d,t,x) - \tilde{g}(d',t,x)$ in undirected networks due to the lack of variation in D. In contrast, the argument can be applied directly to identify the spillover effect in directed networks because $A_{ki} = A_{kj}$ does not imply that $A_{ik} = A_{jk}$, leading to variation in $D_i - D_j$. This is shown in Corollary 4.1.

Corollary 4.1 *Let Assumption 3.2, 4.1 hold. Then the spillover effect $\tau_D(d,d';t,x)$ under* ***directed*** *networks, for arbitrary (d,d',t,x) in the support, is identified by the following equation:*

$$\begin{aligned} &\tau_D(d,d';t,x) = \bar{g}(d,t,x) - \bar{g}(d',t,x) \\ &= E[Y_i - Y_j | A_{ki} = A_{kj} \ \forall k, T_i = 1, T_j = 0, D_i = d, D_j = d', X_i = X_j = x] \end{aligned} \tag{12}$$

5 Estimation in Undirected Networks

This section presents the estimation procedure and the consistency of the estimator in undirected networks. Proposition 4.1 shows that $\{\tau_T(d,x)\}_{(d,x)}$ is identified through a set of moment restrictions.

$$E[Y_i - Y_j - \tau_T(d,x) | A_{ki} = A_{kj} \ \forall k, T_i = 1, T_j = 0, D_j = D_i = d, X_i = X_j = x] = 0 \quad \forall (d,x) \tag{13}$$

Although this may be estimated by applying kernel-based methods, the number of observations satisfying $D_i = D_j = d$ can be small. For the observed characteristics X, random sampling guarantees that there will be samples with X close enough. In contrast, D is an equilibrium quantity with complicated dependence on $\{\mathbf{T}, \mathbf{X}, \mathbf{v}\}$ and the network structure A. Random sampling may not be able to guarantee enough samples with D being close. This relates to the problem of 'thin sets' in Khan and Tamer (2010) and can lead to slow convergence.

Since Equation (13) conditions on the realization of D_i, X_i, the result also holds by interacting with a measurable function $m(D_i, X_i)$. For all values of (d,x):

$$E[(Y_i - Y_j - \tau_T(d,x)) m(d,x) | A_{ki} = A_{kj} \ \forall k, T_i = 1, T_j = 0, D_j = D_i = d, X_i = X_j = x] = 0 \tag{14}$$

Since the above holds for all values of d, x, integrating out (d,x) yields:

$$E[(Y_i - Y_j - \tau_T(D_i, X_i)) m(D_i, X_i) | A_{ki} = A_{kj} \ \forall k, T_i = 1, T_j = 0, D_i = D_j, X_i = X_j] = 0 \tag{15}$$

This leads to the following equivalent characterization of τ_T:

Corollary 5.1 *Under the assumptions of Proposition 4.1, the following holds:*

$$\tau_T(D_i, X_i) = \arg\min_{q \in \mathcal{Q}} E\left[(Y_i - Y_j - q(D_i, X_i))^2 \middle| A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, X_i = X_j\right] \tag{16}$$

where $\mathcal{Q}$ is the space of square-integrable functions.

To see the connection, the objective function in Equation (16) can be expanded as:

$$\begin{aligned}
&E\left[(Y_i - Y_j - q(D_i, X_i))^2 \middle| A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, X_i = X_j\right] \\
&= E\left[(\tau_T(D_i, X_i) + v_i - v_j - q(D_i, X_i))^2 \middle| A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, X_i = X_j\right] \\
&= E\left[(\tau_T(D_i, X_i) - q(D_i, X_i))^2 + (v_i - v_j)^2 \middle| A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, X_i = X_j\right] \\
&\quad + 2E\left[(\tau_T(D_i, X_i) - q(D_i, X_i))(v_i - v_j) \middle| A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, X_i = X_j\right] \\
&= E\left[(\tau_T(D_i, X_i) - q(D_i, X_i))^2 + (v_i - v_j)^2 \middle| A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, X_i = X_j\right]
\end{aligned} \tag{17}$$

where the cross-term vanishes due to Equation (15). Also, $(v_i - v_j)^2$ is independent of the choice of q. If the space of function $\mathcal{Q}$ is approximated by a linear sieve space (i.e. $q(D_i, X + i) = \sum_{r=1}^{R} \gamma_r b_r(D_i, X_i)$ for some basis functions $\{b_r\}_{r=1}^{R}$), the coefficients γ_r can be estimated directly through least squares, easing the computation.

This leads us to define the population objective function as

$$L(q) := E\left[(\tau_T(D_i, X_i) - q(D_i, X_i))^2 + (v_i - v_j)^2 \middle| A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, X_i = X_j\right] \tag{18}$$

The relevant part for the minimization problem is $(\tau_T(D_i, X_i) - q(D_i, X_i))^2$. The term $(v_i - v_j)^2$ acts as a level-shifter that is independent of the choice of $q(.)$. As a result, this term can be omitted.

Let b denote the bandwidth and $K_1\left(\frac{s_{ij}}{b}\right)$ be a kernel applied to the difference between the i-th column and the j-th column of the adjacency matrix A. The term s_{ij} defines a notion of distance between the i-th column and the j-th column of the adjacency matrix A. This quantifies the deviation from the condition $\{A_{ki} = A_{kj}\ \forall k\}$. Detailed discussions of the choice of s_{ij} and its implications are given in Subsection 5.1. Let K_2 be a kernel applied to the variable X. Define the weight ω_{ij} as follows:

$$\omega_{ij} := \frac{K_1\left(\frac{s_{ij}}{b}\right) K_2\left(\frac{X_i - X_j}{b}\right) \mathbb{1}\{T_j \neq T_i\}}{\sum_{j \neq i} K_1\left(\frac{s_{ij}}{b}\right) K_2\left(\frac{X_i - X_j}{b}\right) \mathbb{1}\{T_j \neq T_i\}} \tag{19}$$

Denote $\mathcal{T} := \{i : T_i = 1, \exists j \text{ s.t. } \omega_{ij} > 0\}$. The sample objective function can be defined as follows:

$$L_n(q; b) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \sum_{j \neq i} [(T_i - T_j)(Y_i - Y_j) - |T_i - T_j| q(D_i, X_i)]^2 \omega_{ij} \tag{20}$$

The sample objective function L_n can be viewed as estimating the conditional expectation $E[[(T_i - T_j)(Y_i - Y_j) - |T_i - T_j| q(D_i, X_i)]^2 | A_j = A_i, X_j = X_i]$ using the Nadaraya-Watson estimator. The exception is that we are only comparing nodes with different treatment statuses to arrive at the treatment effect. The estimator of the causal effect is defined as

$$\hat{\tau}_T(D_i, X_i) := \arg\min_{q \in \mathcal{Q}_k} L_n(q; b) \tag{21}$$

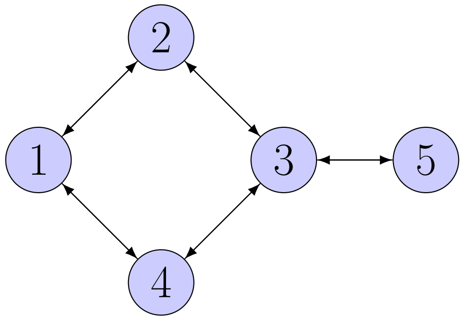
where $\mathcal{Q}_k$ is some sieve space. The goal of the rest of this section is to show that $\|\hat{\tau}_T(D_i, X_i) - \tau_T(D_i, X_i)\|_2 \xrightarrow{p} 0$ as $n \to \infty$.

5.1 Kernel

The requirement that $A_{ki} = A_{kj}$ for all k places heavy restriction on the data. To deal with this, this subsection considers a kernel on the ℓ^2-norm of the difference in columns of $\tilde{A}$. Let $\iota(i)$ be a vector with 1 at the i-th position and 0 elsewhere. Define $\iota(j)$ in the same way. Let s_{ij} be a function that depends on the difference between the i-th and the j-th column of A:

$$s_{ij} := s(A, i, j) := \tilde{s}(\|A(\iota(i) - \iota(j))\|_2) \tag{22}$$

The vector $\tilde{A}(\iota(i) - \iota(j))$ is the difference between the i-th and the j-th column of the adjacency matrix A. It has non-zero entries only in places where $A_{ki} \neq A_{kj}$, which happens when node k is connected to only one of i, j but not both. It is immediate that $A_{ki} = A_{kj}$ for all k if and only if $s_{ij} = 0$ if $\tilde{s}(a) \neq 0$ for any $a \neq 0$. The identification argument in Proposition 4.1 can be regarded as conditioning on $s_{ij} = 0$, which guarantees that D_i, D_j depends on v_i, v_j only through the symmetric function $v_i + v_j$. This then implies the key identification result $E[v_i - v_j | T_i, T_j, X_i = X_j, D_i = D_j, s_{ij} = 0] = 0$. However, when $s_{ij} \neq 0$, these results no longer hold and there is a bias from smoothing. Define the main version of



$$s = \begin{pmatrix} 0 & 2 & 0.5 & 2 & 3 \\ & 0 & 2.5 & 0 & 1 \\ & & 0 & 2.5 & 4 \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}$$

Figure 4: Illustration of s_{ij}

s_{ij} as follows:

$$s_{ij} \coloneqq \tilde{s}(\|A(\iota(i) - \iota(j))\|_2) \coloneqq \frac{1}{\min\{n_i, n_j\}} \|A(\iota(i) - \iota(j))\|_2 \tag{23}$$

where $n_i \coloneqq \sum_j A_{ij}$ is the degree of node i. When A_{ij} takes value in $\{0, 1\}$, the term s_{ij} can be interpreted as the ratio of the number of different links between i and j to the minimum degree of i, j. One immediate observation is that $s_{ij} \geq 2$ if i and j share no link in common. For concreteness, consider the example given in Figure 4. I only state the upper-triangular part of the matrix s_{ij} since it is symmetric. In addition, the diagonal elements s_{ii} equal zero by construction.

The following lemma quantifies the bias as a function of s_{ij}.

Lemma 5.1 *Let Assumption 3.2, 4.1 hold. Further assume that* $\frac{\max_k n_k}{\min_k n_k} \leq C$ *for some constant* C *and that* $\sup_{A_i, X_i} E[v_i^2 | A_i, X_i] < \infty$. *Then the following holds:*

$$|E[\ell(D_i(\mathbf{T}, \mathbf{X}, \mathbf{v}, A))(v_i - v_j) | X_i, X_j, T_i, T_j, s_{ij}]| \leq \tilde{C} s_{ij} \tag{24}$$

where $\tilde{C}$ *is a constant independent of* $\mathbf{T}, \mathbf{X}, \mathbf{v}, A$ *and* s_{ij} *is defined as in Equation (23).*

If assumption $\frac{\max_k n_k}{\min_k n_k} \leq C$ *for some constant* C *fails, the above conclusion still holds with* s_{ij} *replaced by the following quantity:*

$$s_{ij} \coloneqq \tilde{s}(\|A(\iota(i) - \iota(j))\|_2) \coloneqq \frac{1}{\min_{k: n_k > 0} n_k} \|A(\iota(i) - \iota(j))\|_2 \tag{25}$$

The appearance of s_{ij} may seem unnatural at first glance since none of D_i, v_i, v_j explicitly depends on s_{ij}. However, D_i implicitly depends on s_{ij} through its dependence on the network structure. The term $E[l(D_i)(v_i - v_j)]$ can be thought of as a measure of endogeneity. Recall that the characterization in Equation (16) requires the cross-term to vanish. Lemma 5.1 can be viewed as quantifying the magnitude of the cross-term when $s_{ij} \neq 0$. The assumption of $\frac{\max_k n_k}{\min_k n_k} \leq C$ states that the number of links for each node is of the same order of magnitude.

This is imposed by the literature on spatial autoregression (for example, Assumption 3 in Lee (2002)) to restrict correlation across spatial units. This can be rationalized by a variant of the link formation in Bickel and Chen (2009): $A_{ij} = \mathbb{1}\{\rho_n h(\psi_i, \psi_j) \geq \epsilon_{ij}\}$ where ψ_i, ψ_j are i.i.d. individual characteristics, ϵ_{ij} are i.i.d. dyad-level shocks, and ρ_n is a deterministic sequence that controls the sparsity of the network. For example, $\rho_n = C$ leads to dense networks where each node has degree of order $O(n)$. If we impose $\underline{h} \leq \inf_{a,b} h(a,b) \leq \sup_{a,b} h(a,b) \leq \bar{h}$, the condition $\frac{\max_k n_k}{\min_k n_k} \leq C$ is satisfied by $C = 2\frac{\bar{h}}{\underline{h}}$ for large n.

To illustrate the intuition of the lemma, consider a unit change in v_i (or v_j), holding v_i+v_j constant. When $s_{ij} = 0$, such a change does not affect D_i because D_i only depends on v_i+v_j. When $s_{ij} > 0$, there are units linked to only one of i, j but not both, and their outcomes are affected by this change, which then propagates in the network through spillover. This leads to the correlation between D_i and $v_i - v_j$, even after conditioning on $v_i + v_j$. Lemma 5.1 shows that this effect depends on two quantities: (1) the number of nodes that are linked to only one of i, j, (2) the magnitude of immediate change in the outcome of these nodes caused by a unit change in v_i (or v_j), holding $v_i + v_j$ constant. The first quantity is precisely $\|A(\iota(i) - \iota(j))\|_2$. For the second quantity, consider a node k that is linked to i but not j. The immediate effect of a unit change in v_i on the outcome of $Y_k = g(D_k, T_k, X_k) + v_k$ can be written as $\frac{\partial}{\partial D_k} g(D_k, T_k, X_k) \frac{\partial D_k}{\partial v_i}$. The first quantity $\frac{\partial}{\partial D_k} g(D_k, T_k, X_k)$ is bounded in absolute value by κ by Assumption 3.2, and the second quantity $\frac{\partial D_k}{\partial v_i}$ equals $\frac{1}{n_k}$ by definition. The effect thus depends on the degree of k. If k has many neighbors, a change in the outcome of one of its neighbors does not affect D_k by much and Y_k will thus stay approximately the same. However, the assumption $\frac{\max_k n_k}{\min_k n_k} \leq C$ implies $\frac{1}{n_k}\|A(\iota(i) - \iota(j))\|_2 \leq \frac{C}{n_i}\|A(\iota(i) - \iota(j))\|_2 \coloneqq s_{ij}$. Therefore, the overall effect will be bounded by constant multiples of s_{ij}.

The following corollary proves a similar result based on the assumption of a bounded matrix norm.

Corollary 5.2 *Let Assumption 3.2, 4.1 hold. In addition, assume that the operator norm of the adjacency matrix is bounded:* $\|A\| \frac{\kappa}{\min_k n_k} \leq \tilde{\kappa} < 1$ *for some constant* $\tilde{\kappa}$*, where* κ *is the bound imposed in Assumption 3.2. Finally, assume that* $\sup_{A_i, X_i} E[v_i^2 | A_i, X_i] < \infty$ *Then the following holds:*

$$|E[\ell(D_i(\mathbf{T}, \mathbf{X}, \mathbf{v}, A))(v_i - v_j) | X_i, X_j, T_i, T_j, s_{ij}]| \leq C_1 s_{ij} + C_2 \sqrt{s_{ij}} \tag{26}$$

where C_1, C_2 *are constants independent of* $\mathbf{T}, \mathbf{X}, \mathbf{v}, A$ *and* s_{ij} *is defined as in Equation (23).*

The difference from Lemma 5.1 is that we are now imposing assumptions on the norm of the adjacency matrix. The assumption can be satisfied if the norm of the adjacency matrix is bounded $\|A\| = O(1)$ and the minimal degree $\min_k n_k$ diverges. For the bound on $\|A\|$, it suffices to bound the largest eigenvalue when A is symmetric (i.e. undirected network). This is similar to Assumption A2 in de Paula et al. (2024). The authors assume that the maximum eigenvalue norm of $\rho_0 A$ is strictly less than one so that $(I - \rho_0 A)^{-1}$ is well-defined. Assumption 4.1 in Menzel (2025) also shares similar flavor. The core of this type of assumption is to guarantee that the propagation of shocks is not explosive. The bound on $\|A\|$ can be equivalently viewed as restricting the degree of concentration in the network, which is illustrated in the following two examples:

Example 5.1 (Star) *Consider the case of a star network. With $n = 4$ nodes, the adjacency matrix can be written as:*

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

where node 1 is the central node. Figure 5 includes the picture of the star network. Consider a vector v with entries $v_i = \frac{1}{\sqrt{n}}$ for all i. $\|Av\|_2^2 = \frac{(n-1)^2}{n} + \frac{n-1}{n} = n - 1$. This shows that $\|A\| \geq \sqrt{n-1}$.

Example 5.2 (Ring) *Consider the case of a ring, where $A_{ij} = 1$ if and only if $j = i + 1$ or if $i = 1, j = n$. For $n = 4$, the adjacency matrix can be written as:*

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

Figure 6 includes the picture of the ring network. For any vector v with $\|v\|_2 = 1$, it is immediate that $\|Av\|_2^2 = \sum_{i=1}^{n-2}(v_i + v_{i+2})^2 + (v_2 + v_n)^2 + (v_1 + v_{n-1})^2 \leq 4\sum_{i=1}^{n} v_i^2 = 4$. Thus, $\|A\|$ is bounded.

5.2 Consistency

With the kernels defined, it remains to establish the consistency of the proposed estimator in Equation (21), which relies on the following sets of assumptions.

Assumption 5.1 $\sup_{d,t,x} |\tilde{g}(d,t,x)| \leq \bar{y}$ *for some constant $\bar{y}$ and* $\sup_{A_i, X_i} E[|v_i|^4 | A_i, X_i] < \infty$

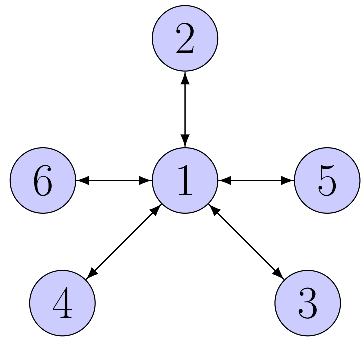


Figure 5: Star

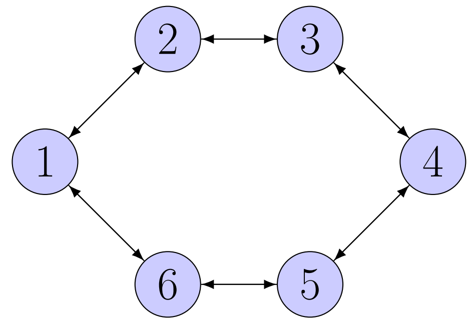


Figure 6: Ring

The boundedness assumption is reasonable in the context of information provision. The outcome variables tend to be measures related to knowledge and attitudes, which are usually bounded.

Assumption 5.2 *For any values of d, t, $\tilde{g}(d, t, x)$ is a Lipschitz function with respect to x with Lipschitz constant $Lip(x)$.*

The Lipschitz condition imposes smoothness restrictions on the function $g(d, t, x)$, which ensures that $|\tilde{g}(d, t, x) - \tilde{g}(d, t, x')| = o(1)$ as we restrict $\|x - x'\|_2 = o(1)$ through the kernel.

Assumption 5.3

1. *The kernels are bounded: $\|K_1\|_\infty < \infty$, $\|K_2\|_\infty < \infty$*
2. *The choice of bandwidth is such that $b \to 0$*

Assumption 5.4 *The conditional probability of receiving treatment is strictly bounded from below and above: $P(T_i = 1|X_i = x) \in [\underline{\pi}, \bar{\pi}]$ for all values of x.*

For any two nodes i, j, let $\ell(i, j)$ be the distance of the shortest path between i, j (i.e., the smallest integer k such that $A^k_{ij} > 0$ and $A^{l'}_{ij} = 0$ for all $k' < k$). Define the following quantities as in Kojevnikov et al. (2021):

$$N^\partial_n(i; s) := \{j \in 1, \cdots, n : \ell(j, i) = s\} \tag{27}$$

$$\delta^\partial_n(s; k) := \frac{1}{n} \sum_i |N^\partial_n(i; s)|^k \tag{28}$$

The first quantity $N^\partial_n(i; s)$ is the collection of nodes that are s-step away from i. The subscript n allows such set to vary with the sample size n. The second quantity $\delta^\partial_n(s; k)$ is the average of the k-th power of the number of neighbors that are s-step away. When $k = 1$, this becomes the average number of neighbors that are s-step away which is a measure of concentration. If the nodes are within short distances from each other (high $N_n^\partial(i;s)$ for small s), the network is concentrated. When the network is overly concentrated, changes in the outcome of a few nodes could have out-sized influence on the average outcome of the whole network. The next assumption places a restriction on the level of concentration in the network.

Assumption 5.5 *$\frac{1}{n}\sum_{s\geq 1}\delta_n^\partial(s;1)\kappa^s \xrightarrow{a.s.} 0$ where κ is the bound on the derivative $|\frac{\partial}{\partial d}g(d,t,x)|$ in Assumption 3.2.*

Assumption 3.2 imposes a bound κ on the magnitude of the spillover effect, so the interference decays at a geometric rate. This ensures that distant nodes have diminishing influence. However, Assumption 3.2 does not place restrictions on the network structure and distant nodes may not exist as in the star network. Assumption 5.5 fills this gap by requiring that the network is not overly clustered. It is adapted from Kojevnikov et al. (2021) and is a key condition for the law of large numbers.[6] In words, this assumption requires that the average number of neighbors at any distance be small relative to the number of nodes. The examples below illustrate this assumption in two different networks.

Example 5.1. (Continued) *Consider the case of a star network. For the central node i, $N_n^\partial(i;s) = n-1$ for $s=1$ and 0 for $s>1$. For the peripheral nodes, $N_n^\partial(i;s) = 1$ for $s=1$, $n-2$ for $s=2$, and 0 for $s>2$. It follows that $\frac{1}{n}\sum_{s\geq 1}\delta_n^\partial(s;1)\kappa^s = \frac{1}{n}[\frac{2n-2}{n}\kappa + \frac{(n-1)(n-2)}{n}\kappa^2] \xrightarrow{n\to\infty} \kappa^2 \neq 0$.*

Example 5.2. (Continued) *Consider the case of a ring, where $A_{ij}=1$ if and only if $j=i+1$ or if $i=1, j=n$. When n is an odd number, $|N_n^\partial(i;s)| = 2$ for all s. When n is an even number, $|N_n^\partial(i;s)| = 2$ for all $s<\frac{n}{2}$ and $|N_n^\partial(i;s)| = 1$ for $s=\frac{n}{2}$. Assume without loss that n is an odd number. We have $\frac{1}{n}\sum_{s\geq 1}\delta_n^\partial(s;1)\kappa^s = \frac{1}{n}\sum_{s=1}^{\frac{n-1}{2}} 2\kappa^s \leq \frac{1}{n}\frac{2}{1-\kappa} = O(\frac{1}{n})$.*

The assumption fails under the star network where any two peripheral nodes are 2-step away from each other, and any peripheral node is 1-step away from the central node. In contrast, the ring network is more spread out and satisfies the assumption.

There is another angle to interpret this assumption. Rewrite $\frac{1}{n}\sum_{s\geq 1}\delta_n^\partial(s;1)\kappa^s$ as follows:

$$\frac{1}{n}\sum_{s\geq 1}\delta_n^\partial(s;1)\kappa^s = \frac{1}{n}\sum_i\left(\frac{1}{n}\sum_{s\geq 1}|N_n^\partial(i,1)|\kappa^s\right) \tag{29}$$

The term $\frac{1}{n}\sum_{s\geq 1}|N_n^\partial(i,1)|\kappa^s$ can be regarded as the upper bound on the effect of a change in the outcome of node i on the average outcome in the entire network. The example below

[6] In Kojevnikov et al. (2021), this assumption is stated with κ^s replaced by $\theta_{n,s}$, which bounds the covariance between the outcome of nodes that are of distance s-away. In this paper, Lemma B.1 shows that $\theta_{n,s}$ behaves like κ^s.

illustrates this idea in the linear-in-means model. In this regard, $\frac{1}{n}\sum_{s\geq 1}\delta_n^{\partial}(s;1)\kappa^s$ is the average influence of a change in the outcome of a single node on the average outcome of the network. The law of large number requires that such effect shrinks to zero.

Example 5.3 *Consider $Y_i = \beta_1 D_i + \beta_2 T_i + X_i\beta_3 + v_i$ with $\beta_2 = \beta_3 = 0$. The system admits a unique reduced form when $|\beta_1| \leq \kappa < 1$: $\mathbf{Y} = (I - \beta_1\tilde{A})^{-1}\mathbf{v} = (I + \sum_{s=1}^{\infty}\beta_1^s\tilde{A}^s)\mathbf{v}$. Let $\iota(i)$ be a vector with the i-th entry equal to 1 and all other entries equal to 0. A unit change in v_i on $\frac{1}{n}\sum_{k=1}^{n} Y_k$ can be written as $\frac{1}{n}\mathbf{1}'(I-\beta_1\tilde{A})^{-1}\iota(i)$. Let $I_k, \tilde{A}_k$ be the k-th row of $I, \tilde{A}$ respectively.*

$$
\begin{aligned}
\left|\frac{1}{n}\mathbf{1}'(I-\beta_1\tilde{A})^{-1}\iota(i)\right| &= \left|\frac{1}{n}\sum_{k=1}^{n}\left(I_k + \sum_{s=1}^{\infty}\beta_1^s\tilde{A}_k^s\right)\iota(i)\right| \\
&= \left|\frac{1}{n}\sum_{l=1}^{n}\sum_{k\in N_n^{\partial}(i,l)}\left(\sum_{s=l}^{\infty}\beta_1^s\tilde{A}_k^s\right)\iota(i) + \frac{1}{n}\left(I_i + \sum_{s=1}^{\infty}\beta_1^s\tilde{A}_i^s\right)\iota(i)\right| \\
&\leq \left|\frac{1}{n}\frac{1}{1-\beta_1}\right| + \frac{1}{n}\sum_{l=1}^{n}\sum_{k\in N_n^{\partial}(i,l)}\left|\beta_1^l\frac{1}{1-\beta_1}\right| \qquad (|\tilde{A}_{ji}^s| \leq 1) \\
&= \left|\frac{1}{n}\frac{1}{1-\beta_1}\right| + \frac{1}{n}\sum_{l=1}^{n}|N_n^{\partial}(i,l)|\left|\beta_1^l\frac{1}{1-\beta_1}\right| \\
&\leq \frac{1}{n}\frac{1}{1-\kappa_1} + \frac{1}{n}\sum_{l=1}^{n}|N_n^{\partial}(i,l)|\kappa_1^l\frac{1}{1-\kappa_1} \\
&= O\left(\frac{1}{n}\sum_{l=1}^{n}|N_n^{\partial}(i,l)|\kappa_1^l\right)
\end{aligned}
$$

The following assumption imposes regularity assumptions on the sieve space

Assumption 5.6

1. *The sieve space $\mathcal{Q}_k$ are compact under the L^2-norm*
2. *$\mathcal{Q}_k \subseteq \mathcal{Q}_{k+1} \subseteq \mathcal{Q}$ for all k.*
3. *There exists a sequence $\pi_k\tau_T \in \mathcal{Q}_k$ such that $\|\pi_k\tau_T - \tau_T\|_2 \to 0$ as $k \to \infty$*

The zero covariance condition in Equation (15) is the theoretical underpinning for characterizing $\tau_T(D_i, X_i)$ as the unique minimizer of L^2 distance. However, this condition relies on $s_{ij} = 0$. The following two assumptions deals with the bias from smoothing by allowing for $s_{ij} \neq 0$. Assumption 5.7 approaches this problem from the 'many network asymptotics' where the network consists of blocks of bounded size and the number of blocks diverges with the sample size. Assumption 5.8 adopts the 'large network asymptotics' where observations do not belong to separate blocks. The consistency result requires that one of these two assumptions hold.

Assumption 5.7 *1. The event $\{A_i = A_j\}$ happens with positive probability that is bounded from below: $0 < \underline{p} \leq P(A_i = A_j)$*

2. *A_i has finite support*

3. *$K_2 = 0$ for any $\|X_i - X_j\|_2 \geq C'b$ for some constant C'*

4. *The conditional density of $v_i|A_i, X_i$ is near identical for close x: For any ϵ, there exists δ such that $|f_{v_i|A_i,X_i}(v|a,x) - f_{v_j|A_j,X_j}(v|a,x')| < \epsilon$ for any $\|x - x'\| < \delta$.*

5. *The conditional density $f_{v_i|A_i,X_i}(v|a,x)$ is bounded from below: $0 < \underline{f} \leq \inf_{v,a,x} f_{v_i|A_i,X_i}(v|a,x)$*

This assumption is imposed to deal with the endogeneity of D_i using the 'many network asymptotics'. In the empirical example, there are 28 treated schools. It is reasonable to assume that the probability of two students from the same school have the same set of friends as nonzero. For the second assumption, A_i takes a finite value within each school. For b small enough, this implies that $K_1(\frac{\|A_i - A_j\|}{b}) = K_1(0)\mathbb{1}\{A_i = A_j\}$ for any K_1 that is supported on a bounded interval. As a result, there is no smoothing with respect to $A_i - A_j$ asymptotically. These two assumptions can also be satisfied when people form groups as in Chemin (2018) and people are connected to all others in the same group. The event $A_i = A_j$ is thus equivalent to i, j belonging to the same group. For n large enough, $K_1 \neq 0$ only if i, j belong to the same group. The third assumption also requires that K_2 have bounded support. It can accommodate both discrete and continuous variables. The fourth and the fifth assumptions are technical assumptions and can be replaced with the following alternative assumption: For any ϵ, there exists δ such that $|f_{v_i|A_i,X_i}(v|a,x) - f_{v_j|A_j,X_j}(v|a,x')| < \epsilon f_{v_i|A_i,X_i}(v|a,x)$ for any $\|x - x'\| < \delta$. As shown in Proposition 4.1, $E[v_i - v_j|A_i = A_j, X_i = X_j, D_i] = 0$. The fourth and fifth assumptions ensure that this difference in conditional expectation is small when the condition $\{X_i = X_j\}$ is relaxed. Overall, Assumption 5.7 requires that $A_i = A_j$ holds strictly asymptotically (no smoothing with respect to $A_i - A_j$ asymptotically) but allows for $\|X_i - X_j\|_2$ to deviate from zero. To approach the finite sample bias, one may need to refer to Lemma 5.1 or Corollary 5.2.

Assumption 5.8

1. *For any Lipschitz function $l(.)$, the following holds: $|E[l(D_i)(v_i - v_j)|A_i, A_j, X_i, X_j]| \leq t_l(s_{ij})$ where t_l is some continuous function with $t_l(0) = 0$*

2. *K_1 has bounded support on $[0, C]$*

3. $|\mathcal{T}| \to \infty$ *as* $n \to \infty$

Under the 'single large network asymptotics', it may not be feasible to assume that the event $A_{ki} = A_{kj}\ \forall k$ happens with strictly positive probability. This assumption deals with the bias from relaxing this constraint. For point 1 to hold, we could apply Lemma 5.1 or Corollary 5.2. If all nodes have degrees of similar order of magnitude, Lemma 5.1 can be applied. If assumptions on the norm of the adjacency matrix A can be imposed, Corollary 5.2 can be adopted.

The second and third points may seem non-standard at first glance. In textbook nonparametric analysis, symmetric kernel is usually adopted to eliminate the bias from smoothing. However, the support of A_i (or $\tilde{A}_i$) is a subset of the space of sequences which is of infinite dimension. This is also referred to as functional data. Unlike finite-dimensional problems, applying product kernels may lead to significant under-smoothing, or even no sample being used. The literature studying the Nadaraya-Watson estimator under functional data thus constructs kernels on the difference in norms (Ferraty et al. (2010), Hong and Linton (2016)).

The final assumption requires that the effective sample size tends to infinity. It can be related to the assumption of $nb^d \to \infty$ when smoothing with respect to a d-dimensional variable. This assumption imposes an upper bound on the speed at which b tends to zero and ensures that the effective sample size tends to infinity. It also restricts the sequence of networks. This is illustrated in the following three examples. For the first example, assume that the network is undirected and all network links are i.i.d. Bernoulli random variables with probability p. For large n, each individual has degree near np and the number of different links for two arbitrary individuals is near $2np(1-p)$. It is expect that $s_{ij} \approx \frac{2(1-p)}{p}$. Unless $p \to 1$, it is unlikely that $|\mathcal{T}| \to \infty$. For the second example, assume that there are G groups. Each individual joins G_i groups where G_i is a random variable and $A_{ij} = 1$ if i, j share at least one group in common. This resembles the informal groups in Chemin (2018). Under this setup, we have positive probability of observing two individuals with the same neighbors and $|\mathcal{T}| \to \infty$ holds trivially. The third example is a network formation process based on homophily. Each individual draws $\xi_i \in [0, 1]$ and $A_{ij} = \mathbb{1}\{|\xi_i - \xi_j| \leq \epsilon\}$. For n large enough, pairs with $\xi_i \approx \xi_j$ are observed, which leads to $s_{ij} \approx 0$, are observed. In general, the required condition is more likely to hold when the network is generated by some underlying low-dimensional variables.

The following theorem establishes the consistency of the proposed estimator.

Theorem 1 *Assume that Assumption 3.2, 4.1, 5.1 - 5.6 hold. Further assume that Assumption 5.7 or Assumption 5.8 holds. Then*

$$\|\tau_T(D_i, X_i) - \hat{\tau}_T(D_i, X_i)\|_2 \xrightarrow{p} 0$$

where $\hat{\tau}_n(D_i, X_i)$ is defined in Equation (21).

6 Simulation

This section presents simulation evidence on the performance of the proposed estimator. The outcome is generated according to the following equation:

$$Y_i = 0.5 + 0.3T_i + 0.4D_i + 0.2T_iD_i + v_i \tag{30}$$

which leads to the following causal effect:

$$\tau_T(D_i, X_i) = 0.3 + 0.2D_i \tag{31}$$

The error term v_i follows a standard normal distribution and the treatment assignment T_i are i.i.d. Bernoulli random variable with $P(T_i = 1) = 0.3$. This section studies the performance of the estimator under both the 'many network' asymptotics and the 'single large network' asymptotics. For the 'many network' asymptotics, the network is generated as a block-diagonal matrix where each block represents a school as in the empirical application. Each block consists of $n = 500$ nodes. Within each block, the links are generated from one of the following data-generating processes (DGP):

1. **Network DGP 1**: There are 12 groups and each individual joins a random number G_i of groups. The number of groups G_i follows the following distribution: $\min\{12, 1+\tilde{G}_i\}$ where $\tilde{G}_i$ is a Poisson random variable with parameter $\lambda = 3$.

2. **Network DGP 2**: Each individual obtains an i.i.d. draw from the uniform distribution on $[0, 1]$, denoted as ξ_i. The undirected network is generated from a variant of the model in Auerbach (2022): $A_{ij} = \mathbb{1}\{\rho_n\sqrt{|\xi_i - \xi_j|} - \nu(\xi_i + \xi_j) \geq (1 - 2\nu)\eta_{ij}$. The parameter ρ_n is set at $60\frac{\log n}{n}$ and ν is set at $\nu = 0.3$. The dyad-level shocks η_{ij} are i.i.d. Uniform $[0, 1]$ variables.

3. **Network DGP 3**: Each individual randomly joins one of the 50 groups. Let $C_i \in \{1, 2, \cdots, 50\}$ denote the group that i joins. The network is formed by $A_{ij} = \mathbb{1}\{C_i = C_j\}$, i.e. all individuals within a group are linked to each other.

The summary statistics on degree, Y, D, and link differences (s_{ij}) for Network DGP 1, 2, 3 are contained in Table 1, 2, 3 respectively. Network DGP 1 can be considered as the baseline DGP while DGP 2 and 3 add additional challenges. As shown in Table 2, the minimum link difference is bounded away from zero, which challenges the identification argument. This adds bias to the estimator due to endogeneity. DGP 3 guarantees observations with the same neighbors but the number of such observations is limited. In addition, under DGP 3, two nodes have either $s_{ij} = 0$ (same link structure) or $s_{ij} > 1$. This implies that any choice of bandwidth with $0 < b \leq 1$ yields the same result. This can be viewed as adding challenges through a higher variance. Since the identification argument in Proposition 4.1 hinges on $s_{ij} = 0$, one would expect that the bias of the estimator in Network DGP 3 is less than that under Network DGP 2. However, recall that the endogeneity in estimation stems from the covariance of a function of D_i and $v_i - v_j$, which is illustrated in Equation (17). When D_i exhibits limited variation, this covariance term may also be small due to the Cauchy-Schwarz inequality. This is analogous to the consistency result in Lee (2002). As the network under DGP 2 is denser compared to DGP 3, this bias need not be much larger than in DGP 3.

For the 'single large network' asymptotics, the network is generated according to the following process:

1. **Network DGP 4**: This is a variant of the Network DGP 1. There are 10 groups and each individual joins a random number G_i of groups. The number of group G_i follows the following distribution: $\min\{10, 1+\tilde{G}_i\}$ where $\tilde{G}_i$ follows a Poisson distribution with parameter $\lambda = 1$.

The summary statistics are contained in Table 4. As the number of nodes expands, there will be more nodes with similar share of neighbors. This can be seen from Table 4 where the minimum Link difference decreases with the sample size.

For each simulated dataset, the estimation problem is:

$$\max_{q \in \mathcal{Q}} L_n(q; b) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \sum_{j \neq i, C_{ij}=1} [(T_i - T_j)(Y_i - Y_j) - |T_i - T_j| q(D_i)]^2 \omega_{ij} \tag{32}$$

The treatment effect τ_T is approximated by a linear combination of basis functions $q(D_i) = \sum_{r=1}^{R} \gamma_r B_r(D_i)$ where $\{\gamma_r\}_{r=1}^{R}$ is the set of coefficients to be estimated using weighted least squares. The estimated treatment effect $\hat{\tau}_T$ is equal to

$$\hat{\tau}_T(D_i) = \sum_{r=1}^{R} \hat{\gamma}_r B_r(D_i) \tag{33}$$

where the coefficients $\hat{\gamma}$ are estimated as follows:

$$\begin{aligned}
\hat{\gamma} &= (\mathbf{q}'\Omega\mathbf{q})^{-1}\mathbf{q}'\Omega\mathbf{y} \\
\mathbf{q} &\coloneqq (|T_i - T_j| q(D_i))_{i,j} \\
\mathbf{y} &\coloneqq ((T_i - T_j)(Y_i - Y_j))_{i,j}
\end{aligned}$$

where Ω is the diagonal matrix with diagonal entries ω_{ij}. The basis functions are Bernstein polynomials. The probability density function of the truncated normal distribution on $[0, 1]$ is adopted as the kernel K_1:

$$K_1\left(\frac{x}{b}\right) = \frac{1}{b}\frac{\phi\left(\frac{x}{b}\right)}{\Phi\left(\frac{1}{b}\right) - \Phi\left(0\right)}$$

where ϕ, Φ are the probability density function and the cumulative density function of the standard normal distribution, respectively.

The L^2 loss is used to evaluate the performance of the estimator. It is calculated as follows:

$$\hat{\ell}([d_1, d_M]) := \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\tau}_T(d_m) - \tau_T(d_m)\right)^2 \tag{34}$$

where $\{d_m\}_{m=1}^{M}$ is a set of grid on the interval $[d_1, d_M]$. The baseline interval $[d_1, d_M]$ is chosen to be the minimum and maximum of D. For example, $[d_1, d_M] = [0.1, 0.85]$ with grid size 0.001 for Network DGP 2. For Network DGP 3, the original interval is $[-0.2, 2]$ with grid size 0.002. Since the basis functions may behave poorly near the boundary due to limited number of observations, the L^2 loss is also computed on truncated intervals. For example, the L^2 loss is computed on the truncated interval $[0.15, 0.8]$ for Network DGP 2.

For Network DGP 1-3, I vary (1) the number of schools, (2)the bandwidth, and (3) the degree of the basis function. One exception is that the bandwidth is fixed for Network DGP 2. This is because any bandwidth $0 < b \leq 1$ yields the same result as mentioned above. For Network DGP 4, the number of nodes is set to different values instead of the number of schools. The number of simulation repetitions is set to 2000, and the L^2-loss results under Network DGP 1-4 are contained in Table 9-12 respectively.

From both tables, it is clear that the L^2 loss decreases with the sample size, which coincides with the consistency result in Theorem 1. This can be seen by comparing the results under different number of schools for Network DGP 1-3 or the results under different n for Network DGP 4. Higher degrees and lower bandwidth increases the L^2 loss but the effect is much less pronounced when we look at truncated intervals. In addition, the L^2 loss under Network DGP 2 and 3 are comparable except when the degree equals 8. This confirms the intuition that the endogeneity problem may be less of a concern when D_i exhibits limited variation.

#School	Degree			Y			D			Link Difference		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max
5	376	142	498	0.91	-2.64	4.51	0.70	0.23	1.00	0.58	0	2.50
10	376	140	499	0.91	-2.82	4.69	0.70	0.22	1.03	0.58	0	2.57
15	376	138	499	0.91	-2.94	4.79	0.70	0.22	1.04	0.58	0	2.61
20	376	137	499	0.91	-3.02	4.86	0.70	0.21	1.05	0.58	0	2.63

Table 1: Summary Statistic under Network DGP 1

#School	Degree			Y			D			Link Difference		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max
5	181	85	432	0.74	-2.81	4.29	0.33	0.13	0.80	1.52	0.18	4.96
10	181	83	435	0.74	-2.98	4.48	0.33	0.12	0.83	1.52	0.17	5.13
15	181	82	436	0.74	-3.09	4.60	0.33	0.12	0.84	1.52	0.17	5.22
20	181	81	437	0.74	-3.16	4.67	0.33	0.11	0.85	1.52	0.16	5.28

Table 2: Summary Statistic under Network DGP 2

#School	Degree			Y			D			Link Difference		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max
5	11	3	20	0.84	-2.86	4.65	0.54	-0.23	1.84	2.41	0	9.18
10	11	2	21	0.84	-3.05	4.84	0.54	-0.29	1.97	2.41	0	11.11
15	11	2	21	0.84	-3.15	4.97	0.54	-0.32	2.06	2.41	0	12.44
20	11	2	22	0.84	-3.24	5.05	0.54	-0.34	2.11	2.41	0	13.09

Table 3: Summary Statistic under Network DGP 3

n	Degree			Y			D			Link Difference		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max
500	176	85	405	0.74	-2.33	3.83	0.34	0.14	0.76	1.69	0.007	4.33
1000	353	179	836	0.74	-2.55	4.04	0.32	0.15	0.75	1.68	0.003	4.19
1500	530	275	1271	0.74	-2.67	4.15	0.32	0.15	0.75	1.68	0.002	4.14
2500	884	468	2154	0.73	-2.81	4.29	0.31	0.16	0.74	1.67	0.001	4.09

Table 4: Summary Statistic under Network DGP 4

7 Empirical Application

The empirical application is based on the network experiment conducted by Paluck et al. (2016). The authors study the impact of an anti-conflict intervention on social norms among adolescents in schools. There are 56 public middle schools that participated in the study. Half of these schools are randomly selected to receive an anti-conflict intervention. Within each treated school, a subset of students is designated as the seed group based on their covariates. Half of the students in the seed group are selected to participate in the intervention by block randomization. The treated students participate in bi-monthly meetings with trained research assistants. During training sessions, research assistants help students identify common conflict behaviors in their schools and encourage them to oppose such conflicts in public. The authors perform two waves of surveys. The first wave of survey is conducted before the intervention and the second wave occurs after the treatment. In each survey, students are asked to answer questions related to social norms and their own attitudes. I work with the 28 treated schools with 10,056 students in total.

Within each school, the network is measured by asking students to list up to ten students at their school whom they chose to spend time with in the past few weeks. The resulted network is directed. In the empirical application, I work with the undirected network which assumes that i, j is linked if $A_{ij} = 1$ or $A_{ji} = 1$.

Past studies adopt the indicator variable for wearing an orange wristband as the outcome variable (Paluck et al. (2016), Aronow and Samii (2017), Leung (2020)). The wristband is disseminated as a reward to those students engaging in conflict-mitigating behaviors. Since the current paper focus on continuous outcome variables, an index for anti-conflict attitude is adopted as the outcome variable. The index is constructed based on all the 33 variables in the section 'Respondent Attitudes' contained in the Wave II survey. These are binary questions that measure individual attitude towards conflicts in the school. As an example, the variable 'CSCAW2' contains the binary response towards the question 'If we want, students can change the amount of conflict at our school'. However, higher value of the binary variable does not necessarily correspond to a more positive attitude. For instance, the variable 'CILW2' contains the binary response to the question 'I have had a lot of conflict with other students at this school' and a higher value indicates a more negative attitude.[7] To this end, I redefine the binary variables such that a higher value represents a more positive attitude. In the case of the variable 'CILW2', this is done by working with the variable '1 - CILW2' instead of 'CILW2'. After this transformation, the index is created as an average across the binary responses. The summary statistics are listed in Table 5. Samples with responses outside $\{0, 1\}$ are excluded, leaving 5,802 individuals in the sample.

Some variables in the 'Respondent Attitudes' section do not directly reflect the attitude of the respondent. For example, the variable 'CBIW2' collects the response to the question 'Boys at this school are involved in a lot of conflict'. As a robustness check, I include only the set of variables that directly reflects the respondent attitude. The outcome variable is then constructed based on the following eight survey questions in the second wave:

1. If we want, students can change the amount of conflict at our school
2. I'd like to help change the amount of conflict at our school with a group of other students

[7]There are 17 such variables.

3. I think teachers and the bullying (harassment, intimidation & bullying: HIB) rules of this school help solve student conflicts

4. I can help change the way students at this school act around each other

5. I feel like I belong at this school

6. I have had a lot of conflict with other students at this school

7. Sometimes you have to be mean to others as a way to survive at this school

8. I've stayed home from school because of problems with other students

All questions are binary and the answers are either 0 (no) or 1 (yes). Answering 1 (yes) reflects a positive attitude for questions 1-5 and a negative attitude for 6-8. As argued above, the roles of 1 and 0 are reversed for question 6-8 and construct the index Y as the average answer for question 1-8. The summary statistics for this alternative construction of the index are listed in Table 8 in Appendix A. Samples with responses outside $\{0,1\}$ are excluded, leaving 10,056 individuals in the sample.

Variable	Mean	Standard Deviation	Min	Max	Sample Size
Y_i	0.679	0.137	0.182	1	4,756
D_i	0.681	0.071	0.242	0.939	4,694
T_i	0.064	0.245	0	1	4,756
n_i (Degree)	6.135	2.998	1	21	4,694

Notes: This table contains the summary statistics for the index constructed based on the questions in the section 'Respondent Attitudes' contained in the Wave II survey. The variable Y_i is constructed based on all questions in the section. Each sample is an individual. Samples are excluded if (1) answer does not fall in $\{0,1\}$ for the binary questions, (2) contain missing values for any variables listed in the table.

Table 5: Summary Statistics

Mean	Standard Deviation	Min	Max	First Quartile	Third Quartile	Sample Size
3.062	1.532	0	21	2.200	3.333	51,757

Notes: This table contains the summary statistics for the maximum share of different links s_{ij} as defined in Equation 23. Each observations is a pair of individuals in the same school.

Table 6: Summary Statistics of s_{ij}

Start with the baseline estimation without covariates. Let C_{ij} be an indicator variable that equals one if individuals i, j are in the same school. Only within-school comparisons are made since the survey collects the links within schools. As stated in the simulation section, the estimation problem is:

$$\max_{q\in\mathcal{Q}} L_n(q;b) = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\sum_{j\neq i, C_{ij}=1}[(T_i - T_j)(Y_i - Y_j) - |T_i - T_j|q(D_i)]^2\omega_{ij} \tag{35}$$

The treatment effect τ_T is approximated by a linear combination of basis functions $q(D_i) = \sum_{r=1}^{R}\gamma_r B_r(D_i)$ where $\{\gamma_r\}_{r=1}^{R}$ is the set of coefficients. The estimated treatment effect $\hat{\tau}_T$ is equal to

$$\hat{\tau}_T(D_i) = \sum_{r=1}^{R}\hat{\gamma}_r B_r(D_i) \tag{36}$$

The coefficients $\hat{\gamma}$ are estimated through weighted least squares and the basis functions are Bernstein polynomials. The probability density function of the truncated normal distribution on $[0,1]$ is adopted as the kernel K_1:

$$K_1\left(\frac{x}{b}\right) = \frac{1}{b}\frac{\phi\left(\frac{x}{b}\right)}{\Phi\left(\frac{1}{b}\right) - \Phi\left(0\right)}$$

where ϕ, Φ are the probability density function and the cumulative density function of the standard normal distribution, respectively.

I consider three values for the bandwidth $b \in \{0.2, 0.5, 1\}$ and three values for the number of basis functions $R \in \{3, 5, 8\}$. To avoid problems at the boundary, I truncate the range of the plot on both sides, and the plot for $b = 1, R = 3$ is shown in Figure 7. The blue lines are point-wise 95% confidence intervals obtained from bootstrapping the schools. The full results are shown in Figure 9. The results for the alternative definition of the outcome variable are included in Figure 10.

Due to the limited sample size, I also consider a partial linear model to incorporate control variables:

$$Y_i = \bar{g}(D_i, T_i, X_i) + v_i = \breve{g}(D_i, T_i) + X_i'\beta + v_i \tag{37}$$

The control variables X include gender, indicators for (1) white, (2) mother went to college, (3) live with both parents, (4) have older siblings, (5) hang out with boys and girls at school. The summary statistics of the control variables are included in Table 7. This is estimated by approximating $\breve{g}$ with smoothing splines (Section 5.4 in Hastie et al. (2009)). The estimation problem is as follows:

$$\begin{aligned}
&\min_{\gamma,\beta} \quad \frac{1}{|\tilde{\mathcal{T}}|}\sum_{i\in\tilde{\mathcal{T}}}\sum_{j}\left(Y_i - Y_j - \sum_{r=1}^{R}\gamma_r B_r(D_i) - (X_i - X_j)\beta\right)^2 \tilde{\omega}_{ij} + \lambda_1\left(\sum_r \gamma_r^2\right) + \lambda_2\left(\sum_k \beta_k^2\right) \\
&\quad \tilde{\omega}_{ij} := \frac{K_1(\frac{s_{ij}}{b})\mathbb{1}\{T_i \neq T_j\}}{\sum_j K_1(\frac{s_{ij}}{b})\mathbb{1}\{T_i \neq T_j\}} \\
&\quad \tilde{\mathcal{T}} := \left\{ i \in \{1,\cdots,n\} \,\middle|\, \sum_j K_1(\frac{s_{ij}}{b})\mathbb{1}\{T_i \neq T_j\} > 0 \right\}
\end{aligned} \tag{38}$$

The basis function B_r are natural cubic splines with knots at each unique value of the data point D_i, and are defined as in Equation (5.4), (5.5) in Hastie et al. (2009). The number of basis functions R is equal to the number of distinct values of D_i observed in the sample. The penalty parameters λ_1, λ_2 are chosen by five-fold (leave-one-out) cross validation, with the following criteria function:

$$CV(\lambda) := \frac{1}{|\tilde{\mathcal{T}}|}\sum_{i\in\tilde{\mathcal{T}}}\sum_{j}\left(Y_i - Y_j - \sum_{r=1}^{R}\check{\gamma}_r(\lambda; i,j)B_r(D_i) - (X_i - X_j)\check{\beta}(\lambda; i,j)\right)^2 \tilde{\omega}_{ij} \tag{39}$$

where $\check{\gamma}_r(\lambda; i,j)$ and $\check{\beta}(\lambda; i,j)$ are estimated using observations from clusters different from i, j. The result is shown in Figure 8.

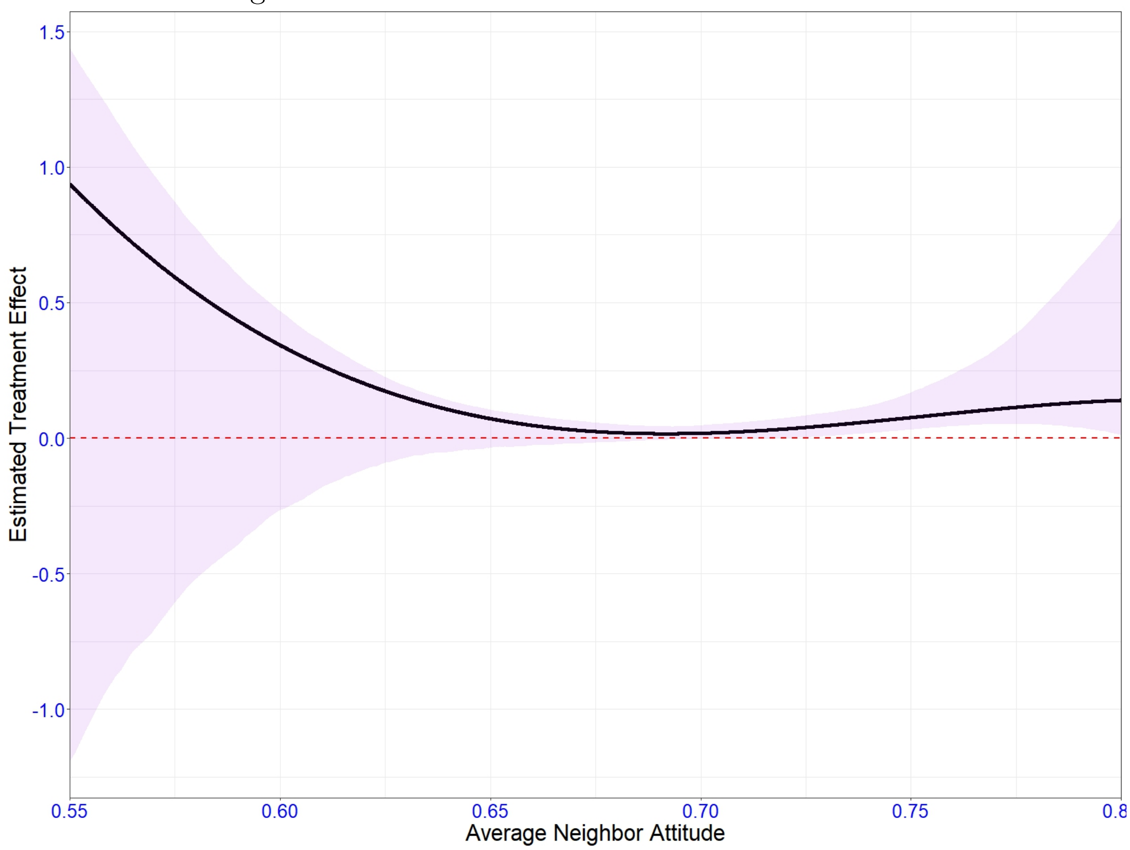
Variable	Mean	Standard Deviation	Sample Size
Male	0.554	0.497	4,756
White	0.641	0.480	4,756
Mother Went to College	0.720	0.449	4,756
Live with Both Parents	0.736	0.441	4,756
Have Older Siblings	0.627	0.484	4,756
Hang out with Boys and Girls	0.712	0.453	4,756

Notes: This table contains the summary statistics for the binary control variables. The variable Y_i is constructed based on all questions in the section. Each sample is an individual. Samples are excluded if (1) answer does not fall in $\{0,1\}$ for the binary questions, (2) contain missing values for any variables listed in the Table 5.

Table 7: Summary Statistics for Control Variables

Overall, the results suggest considerable non-linearity in the treatment effect. The treatment effect is higher for students whose friends have more positive attitudes, suggesting the presence of complementarity. Students benefit from treatment directly, but this can be enhanced by discussing the information with their peers. When the peers have more positive attitude, the benefit from discussing with peers also increases. If the policy maker intends to carry this treatment to another network (school), he/she may target the individuals with more optimistic friends. This again highlights the advantage of the method proposed in this paper which explicitly shows the relationship between treatment and spillover.

Figure 7: Estimated Treatment Effect under $b = 1$



Notes: This plot shows the estimated $\tau_T(D_i)$ at different values of D_i in [0.55, 0.8] under the bandwidth choice $b = 1$. The outcome variable is the index constructed using all the questions in the section 'Respondent Attitudes' contained in the Wave II survey. The x-axis is D_i and the y-axis is τ_T. The blue lines are point-wise 95% confidence intervals obtained from bootstrapping the schools. The basis function is Berstein polynomials of degree 3.

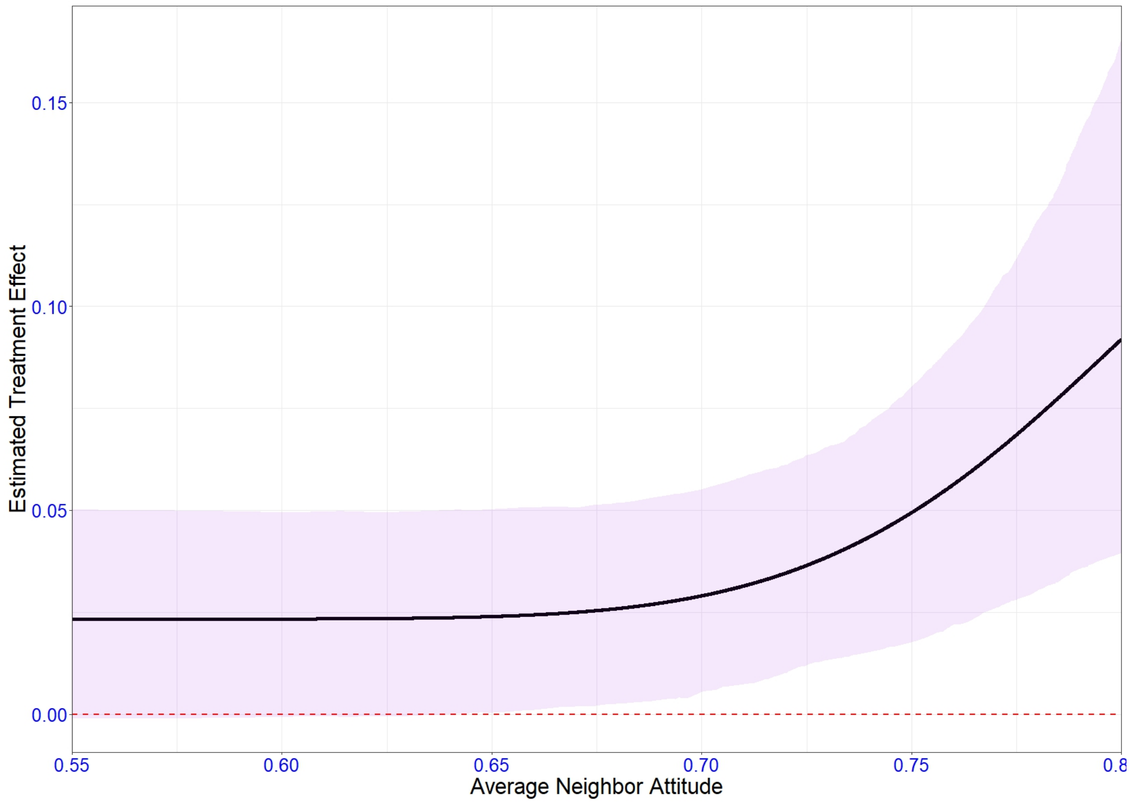
8 Conclusion

This paper studies the treatment effect under the presence of endogenous peer influence in networks. A nonlinear peer effect model is constructed, based on which the causal effects are defined. Identification of the treatment effect is obtained by comparing nodes with different treatment status but the same link structure. The identification argument can be extended to observational studies under the assumption of selection on observables. Although the identification argument places considerable restrictions on the data, I develop a kernel on the maximum share of different links which enables smoothing in finite samples. Consistency of the estimator is established, and the estimator is applied to the empirical example, illustrating the presence of nonlinearity of the anti-conflict intervention. The method in this paper also applies to other contexts. For example, the importance of endogenous peer effects is highlighted in other contexts including adolescent smoking (Nakajima (2007)), academic performance (Calvó-Armengol et al. (2009)). The treatment effect of policy interventions likely depends on the average neighbor outcomes. For example, consider the information sessions on the adverse impact of smoking. The effect of these sessions may be attenuated when peers are intense smokers: they may simply discourage the treated individual.

The proposed method has three advantages. First, it does not rely on parametric assumptions and exogeneity of the network. The latter is a typical assumption for constructing instrumental variables. Second, the method does not restrict the distance of spillover. Third, the method works under some types of cross-cluster interference. For example, people in a village choose to join a subset of groups. Chemin (2018) faces this challenge and states the result as a lower bound since the control groups are affected by such cross-cluster interference.

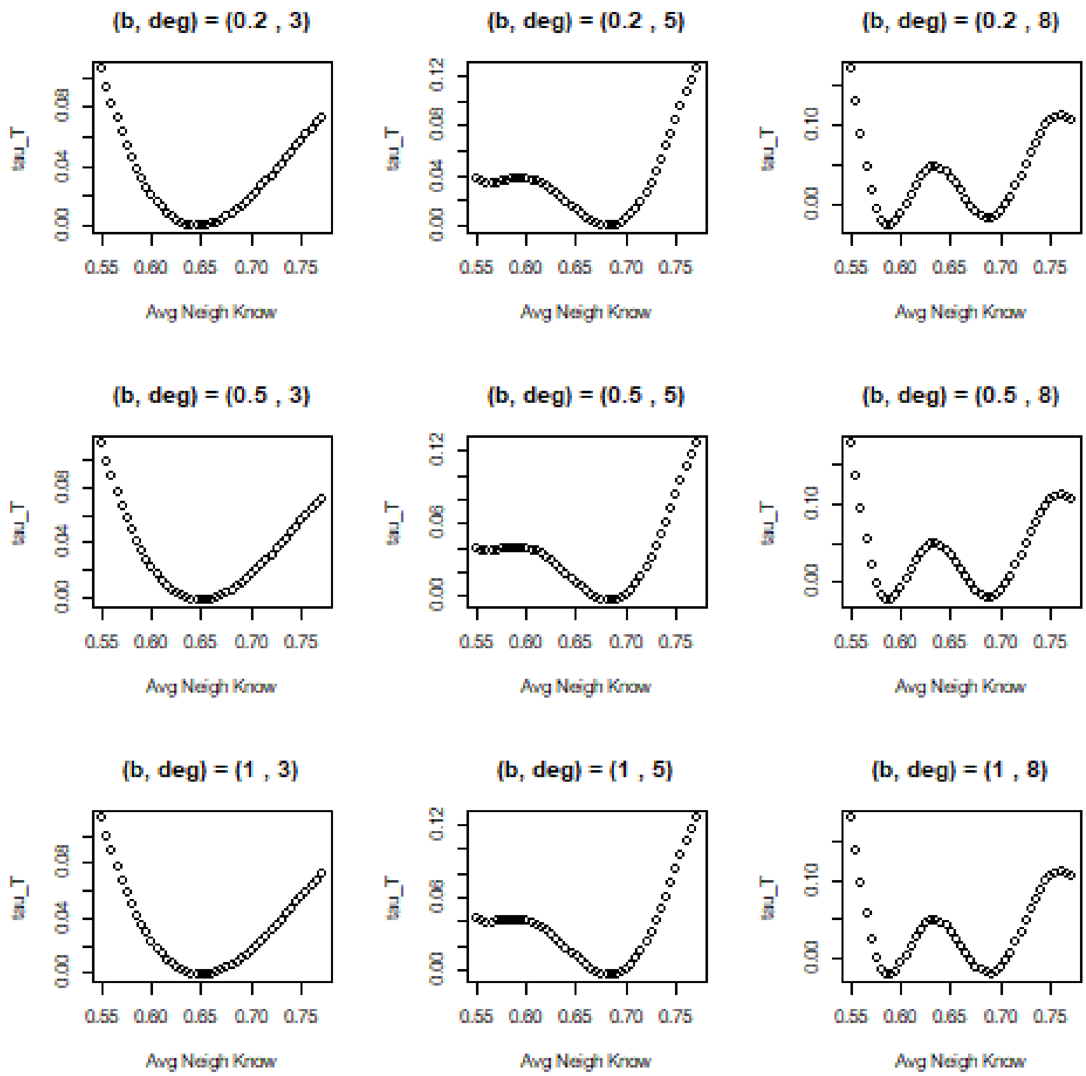
The paper also faces two major limitations. First, the functional form assumption of additively separable errors is hard to accommodate discrete outcomes. Second, the identification argument of the same set of neighbors places strong restriction on the data. It may fail in scenarios where networks are formed with limited level of dependence. This is because the number of ways to form links (2^n) is larger than the sample size (n). The kernel proposed in this paper relaxes this restriction in finite samples but one still needs to take a stance on how the network is formed.

Figure 8: Estimated Treatment Effect under Partial Linear Model



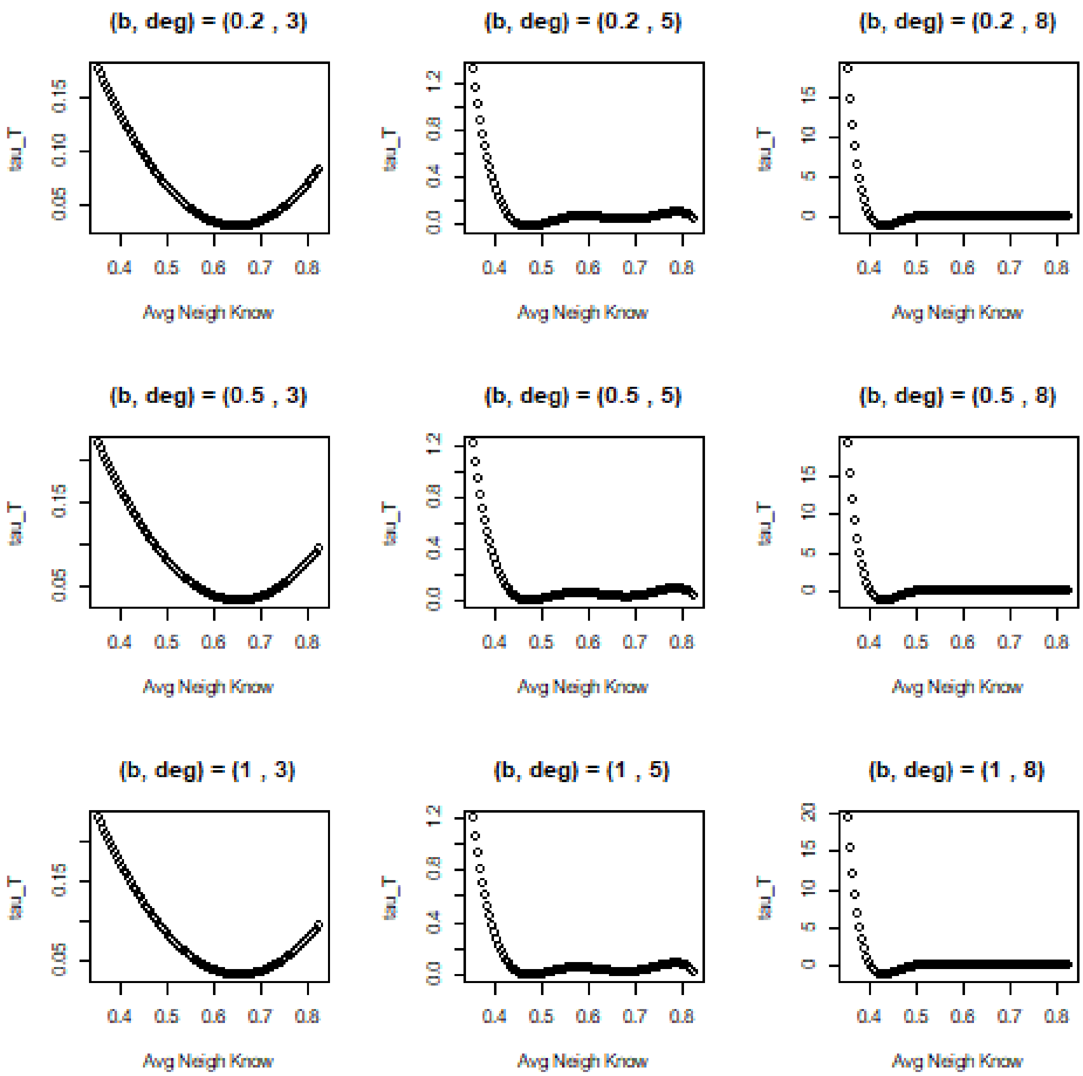
Notes: This plot shows $\tau_T(D_i)$ at different values of D_i in [0.55, 0.8] under the bandwidth choice $b = 1$. The outcome variable is the index constructed using all the questions in the section 'Respondent Attitudes' contained in the Wave II survey. The specification for the outcome variable is a partial linear model as in Equation (37) and the control variables include gender, indicators for (1) white, (2) mother went to college, (3) live with both parents, (4) have older siblings, (5) hang out with boys and girls at school. The basis functions are natural cubic splines with knots being all unique values of D_i in the sample. The coefficients are estimated by generalized ridge regression where the penalty term is chosen from five-fold (leave-one-out) cross-validation. The x-axis is D_i and the y-axis is $\hat{\tau}_T$. The purple area are point-wise 95% confidence intervals obtained from bootstrapping the schools.

Figure 9: Estimated Treatment Effect under Different Bandwidth and Degree



Notes: This plot shows $\tau_T(D_i)$ at different values of D_i in [0.55, 0.77]. The outcome variable is the index constructed using all the questions in the section 'Respondent Attitudes' contained in the Wave II survey. The range plotted corresponds to the range in the data truncated by 0.1 both to the left and to the right. It attempts to deal with the problem at the boundary. The x-axis is D_i and the y-axis is $\hat{\tau}_T$.

Figure 10: Estimated Treatment Effect Based on Partial Index



Notes: This plot shows $\tau_T(D_i)$ at different values of D_i in [0.36, 0.83]. The outcome variable is the partial index constructed using only the eight questions directly reflecting individual attitude towards conflict. This range corresponds to the range in the data truncated by 0.1 both to the left and to the right. It attempts to deal with the problem at the boundary. The x-axis is D_i and the y-axis is $\hat{\tau}_T$. The purple area are point-wise 95% confidence intervals obtained from bootstrapping the schools.

References

Aberra, Adam and Matthieu Chemin, "Know your rights! A field experiment on legal knowledge, property rights, and investment in Kenya," *Journal of Law, economics, & organization*, 2025, *41* (1), 348–378.

Aronow, Peter M. and Cyrus Samii, "Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment," *The annals of applied statistics*, 2017, *11* (4), 1912–1947.

Auerbach, Eric, "Identification and Estimation of a Partially Linear Regression Model Using Network Data," *Econometrica*, 2022, *90* (1), 347–365.

Banerjee, Abhijit, Emily Breza, Arun G Chandrasekhar, Esther Duflo, Matthew O Jackson, and Cynthia Kinnan, "Changes in Social Network Structure in Response to Exposure to Formal Credit Markets," *The Review of economic studies*, 2024, *91* (3), 1331–1372.

Beaman, Lori, Ariel Ben Yishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak, "Can Network Theory-Based Targeting Increase Technology Adoption?," *The American economic review*, 2021, *111* (6), 1918–1943.

Bickel, Peter J. and Aiyou Chen, "A nonparametric view of network models and Newman–Girvan and other modularities," *Proceedings of the National Academy of Sciences*, 2009, *106* (50), 21068–21073.

Bietenbeck, Jan, "The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR," *Journal of the European Economic Association*, 2020, *18* (1), 392–426.

Boning, William C., John Guyton, Ronald Hodge, and Joel Slemrod, "Heard it Through the Grapevine: Direct and Network Effects of a Tax Enforcement Field Experiment," *Journal of public economics*, 2020, *190*, 104261–.

Boucher, Vincent, Michelle Rendall, Philip Ushchev, and Yves Zenou, "Toward a General Theory of Peer Effects," *Econometrica*, 2024, *92* (2), 543–565.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, "Identification of Peer Effects through Social Networks," *Journal of econometrics*, 2009, *150* (1), 41–55.

Bugni, Federico A., Ivan A. Canay, and Steve McBride, "Decomposition and Interpretation of Treatment Effects in Settings with Delayed Outcomes," 2025.

Cai, Jing, Alain De Janvry, and Elisabeth Sadoulet, "Social Networks and the Decision to Insure," *American economic journal. Applied economics*, 2015, *7* (2), 81–108.

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou, "Peer effects and social networks in education," *The Review of economic studies*, 2009, *76* (4), 1239–1267.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West, "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, 2013, *81* (3), 855–882.

Centola, Damon, "The Spread of Behavior in an Online Social Network Experiment," *Science (American Association for the Advancement of Science)*, 2010, *329* (5996), 1194–1197.

_ **and Michael Macy**, "Complex Contagions and the Weakness of Long Ties," *The American journal of sociology*, 2007, *113* (3), 702–734.

Chemin, Matthieu, "Informal Groups and Health Insurance Take-up Evidence from a Field Experiment," *World development*, 2018, *101*, 54–72.

Chen, Xiaohong, "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in "Handbook of Econometrics," Vol. 6, Elsevier B.V, 2007, pp. 5549–5632.

Chesher, Andrew and Adam M. Rosen, "Generalized Instrumental Variable Models," *Econometrica*, 2017, *85* (3), 959–989.

de Paula, Áureo, Imran Rasul, and Pedro C L Souza, "Identifying Network Ties from Panel Data: Theory and an Application to Tax Competition," *The Review of Economic Studies*, 09 2024, *92* (4), 2691–2729.

Ferraty, Frédéric, Ali Laksaci, Amel Tadj, and Philippe Vieu, "Rate of Uniform Consistency for Nonparametric Estimates with Functional Variables," *Journal of Statistical Planning and Inference*, 2010, *140* (2), 335–352.

Gao, Wayne Yuan, "Nonparametric Identification in Index Models of Link Formation," *Journal of econometrics*, 2020, *215* (2), 399–413.

_ **, Ming Li, and Sheng Xu**, "Logical Differencing in Dyadic Network Formation Models with Nontransferable Utilities," *Journal of econometrics*, 2023, *235* (1), 302–324.

Garzon, Luigi and Vitor Possebom, "Nonlinear Treatment Effects in Shift-Share Designs," 2025.

Graham, Bryan S., "Identifying social interactions through conditional variance restrictions," *Econometrica*, 2008, *76* (3), 643–660.

_ , "An Econometric Model of Network Formation With Degree Heterogeneity," *Econometrica*, 2017, *85* (4), 1033–1063.

— **and Jinyong Hahn**, "Identification and Estimation of the Linear-in-means Model of Social Interactions," *Economics letters*, 2005, *88* (1), 1–6.

Griffith, Alan, "Random Assignment with Nonrandom Peers: A Structural Approach to Counterfactual Treatment Assessment," *The review of economics and statistics*, 2024, *106* (3), 859–871.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction* Springer series in statistics, 2nd ed. ed., New York: Springer, 2009.

Hong, Seok Young and Oliver Linton, "Asymptotic Properties of a Nadaraya-Watson Type Estimator for Regression Functions of Infinite Order," 2016.

Houndetoungan, Aristide, "Quantile Peer Effect Models," 2025.

Hu, Yuchen, Shuangning Li, and Stefan Wager, "Average direct and indirect causal effects under interference," *Biometrika*, 2022, *109* (4), 1165–1172.

Hudgens, Michael G and M. Elizabeth Halloran, "Toward Causal Inference With Interference," *Journal of the American Statistical Association*, 2008, *103* (482), 832–842.

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 2010, *78* (6), 2021–2042.

Kojevnikov, Denis, Vadim Marmer, and Kyungchul Song, "Limit Theorems for Network Dependent Random Variables," *Journal of econometrics*, 2021, *222* (2), 882–908.

Lee, Lung-Fei, "Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models," *Econometric theory*, 2002, *18* (2), 252–277.

Leung, Michael P., "Treatment and Spillover Effects Under Network Interference," *The review of economics and statistics*, 2020, *102* (2), 368–380.

— , "Causal Inference Under Approximate Neighborhood Interference," *Econometrica*, 2022, *90* (1), 267–293.

Li, Shuangning and Stefan Wager, "Random Graph Asymptotics for Treatment Effect Estimation under Network Interference," 2022.

Manresa, Elena, "Estimating the structure of social interactions using panel data," *Unpublished Manuscript. CEMFI, Madrid*, 2013, *1.*

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of economic studies*, 1993, *60* (3), 531–542.

— , "Identification of Treatment Response with Social Interactions," *The econometrics journal*, 2013, *16* (1), S1–S23.

Masten, MATTHEW A., "Random Coefficients on Endogenous Variables in Simultaneous Equations Models," *The Review of economic studies*, 2018, *85* (2 (303)), 1193–1250.

Menzel, Konrad, "Fixed-Population Causal Inference for Models of Equilibrium," 2025.

Miraldo, Marisa, Carol Propper, and Christiern Rose, "Identification of Peer Effects using Panel Data," 2021.

Morris, Stephen, "Contagion," *The Review of economic studies*, 2000, *67* (1), 57–78.

Munro, Evan, "Causal Inference under Interference through Designed Markets," 2025.

— , Xu Kuang, and Stefan Wager, "Treatment Effects in Market Equilibrium," 2025.

Nakajima, Ryo, "Measuring Peer Effects on Youth Smoking Behaviour," *The Review of economic studies*, 2007, *74* (3), 897–935.

Newey, Whitney K., "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 1991, *59* (4), 1161–1167.

— and James L. Powell, "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 2003, *71* (5), 1565–1578.

— , — , and Francis Vella, "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 1999, *67* (3), 565–603.

Paluck, Elizabeth Levy, Hana Shepherd, and Peter M. Aronow, "Changing climates of conflict: A social network experiment in 56 schools," *Proceedings of the National Academy of Sciences - PNAS*, 2016, *113* (3), 566–571.

Rose, Christiern and Lizi Yu, "Identification of Peer Effects with Miss-specified Peer Groups: Missing Data and Group Uncertainty," 2022.

Rose, Christiern D., "Identification of peer effects through social networks using variance restrictions," *The econometrics journal*, 2017, *20* (3), S47–S60.

Sasaki, Yuya, "GMM and M Estimation under Network Dependence," 2025.

Sävje, F, "Causal inference with misspecified exposure mappings: separating definitions and assumptions," *Biometrika*, 2024, *111* (1), 1–15.

Sävje, Fredrik, Peter M. Aronow, and Michael G. Hudgens, "Average Treatment Effects in the Presence of Unknown Interference," *The Annals of statistics*, 2021, *49* (2), 673–701.

Tao, Ji and Lung fei Lee, "A Social Interaction Model with an Extreme Order Statistic," *The econometrics journal*, 2014, *17* (3), 197–240.

Tchetgen, Eric J Tchetgen and Tyler J VanderWeele, "On Causal Inference in the Presence of Interference," *Statistical methods in medical research*, 2012, *21* (1), 55–75.

Wagner, Zachary, Corrina Moucheraud, Manisha Shah, Alexandra Wollum, Willa Friedman, and William H Dow, "Reducing Bias Among Health Care Providers: Experimental Evidence from Tanzania, Burkina Faso and Pakistan," *The Economic Journal*, 06 2025, *135* (670), 1891–1922.

Wang, Ye, Cyrus Samii, Haoge Chang, and PM Aronow, "Design-based Inference for Spatial Experiments under Unknown Interference," *The annals of applied statistics*, 2025, *19* (1).

Wood, George and Andrew V. Papachristos, "Reducing Gunshot Victimization in High-risk Social Networks through Direct and Spillover Effects," *Nature human behaviour*, 2019, *3* (11), 1164–1170.

Wooldridge, Jeffrey M., "Control Function Methods in Applied Econometrics," *The Journal of human resources*, 2015, *50* (2), 420–445.

Zeleneev, Andrei, "Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity," *Department of Economics, Princeton University*, 2020.

A Tables and Figures

Variable	Mean	Standard Deviation	Min	Max	Sample Size
Y_i	0.633	0.218	0	1	8,163
D_i	0.633	0.096	0.125	0.929	8,152
T_i	0.068	0.252	0	1	8,163
n_i (Degree)	9.894	3.876	1	32	8,152

Notes: This table contains the summary statistics for the index constructed based on the questions in the section 'Respondent Attitudes' contained in the Wave II survey. The variable Y_i is constructed based on the eight questions that more directly reflects individual attitude. Each sample is an individual. Samples are excluded if (1) answer does not fall in $\{0, 1\}$ for the binary questions, (2) contain missing values for any variables listed in the table.

Table 8: Summary Statistics under Alternative Construction of Y_i

			L^2 Loss on Interval				
Degree	Bandwidth	#School	$[0.2, 1.05]$	$[0.25, 1]$	$[0.3, 0.95]$	$[0.35, 0.9]$	$[0.4, 0.85]$
3	0.2	5	0.022	0.015	0.012	0.011	0.010
		10	0.014	0.009	0.007	0.006	0.005
		15	0.010	0.007	0.005	0.004	0.004
		20	0.009	0.005	0.004	0.003	0.003
	0.5	5	0.017	0.012	0.010	0.009	0.009
		10	0.011	0.007	0.006	0.005	0.005
		15	0.009	0.006	0.004	0.004	0.003
		20	0.007	0.005	0.004	0.003	0.003
	1	5	0.016	0.012	0.010	0.009	0.009
		10	0.011	0.007	0.006	0.005	0.005
		15	0.008	0.006	0.004	0.004	0.004
		20	0.007	0.005	0.004	0.003	0.003
5	0.2	5	0.030	0.018	0.015	0.013	0.012
		10	0.021	0.010	0.008	0.007	0.006
		15	0.017	0.008	0.006	0.005	0.004
		20	0.014	0.006	0.004	0.004	0.003
	0.5	5	0.024	0.014	0.012	0.011	0.011
		10	0.016	0.009	0.007	0.006	0.006
		15	0.013	0.007	0.005	0.004	0.004
		20	0.011	0.005	0.004	0.003	0.003
	1	5	0.023	0.014	0.012	0.011	0.010
		10	0.016	0.008	0.007	0.006	0.006
		15	0.012	0.006	0.005	0.004	0.004
		20	0.011	0.006	0.004	0.004	0.003
8	0.2	5	0.039	0.021	0.017	0.015	0.015
		10	0.028	0.013	0.009	0.008	0.007
		15	0.023	0.009	0.006	0.006	0.005
		20	0.021	0.008	0.005	0.004	0.004
	0.5	5	0.032	0.017	0.014	0.013	0.013
		10	0.023	0.010	0.008	0.007	0.006
		15	0.019	0.008	0.006	0.005	0.005
		20	0.018	0.007	0.005	0.004	0.004
	1	5	0.031	0.017	0.014	0.013	0.012
		10	0.022	0.010	0.008	0.007	0.006
		15	0.019	0.008	0.006	0.005	0.005
		20	0.017	0.007	0.005	0.004	0.004

Table 9: L^2 Loss under Network DGP1

Notes: This table shows the L^2 loss under Network DGP 1. The results include different configurations of the number of basis function (degree), bandwidth, number of schools. The L^2 loss is calculated according to Equation (34) on different choices of $[d_1, d_M]$ reported in Column 4-8. The number of individuals (nodes) in each school (block) is set at 500. The number of simulation repetition is set at 2000.

			L^2 Loss on Interval				
Degree	Bandwidth	#School	$[0.1, 0.85]$	$[0.15, 0.8]$	$[0.2, 0.75]$	$[0.25, 0.7]$	$[0.3, 0.65]$
3	0.2	5	0.021	0.014	0.011	0.010	0.009
		10	0.013	0.008	0.006	0.005	0.005
		15	0.010	0.006	0.004	0.004	0.003
		20	0.008	0.004	0.003	0.003	0.002
	0.5	5	0.016	0.011	0.009	0.008	0.008
		10	0.009	0.006	0.005	0.004	0.004
		15	0.007	0.004	0.003	0.003	0.003
		20	0.006	0.003	0.002	0.002	0.002
	1	5	0.014	0.010	0.009	0.008	0.008
		10	0.009	0.006	0.004	0.004	0.004
		15	0.007	0.004	0.003	0.003	0.003
		20	0.006	0.003	0.002	0.002	0.002
5	0.2	5	0.030	0.016	0.014	0.012	0.012
		10	0.020	0.009	0.007	0.006	0.006
		15	0.016	0.007	0.005	0.004	0.004
		20	0.014	0.005	0.004	0.003	0.003
	0.5	5	0.022	0.013	0.011	0.010	0.010
		10	0.015	0.007	0.006	0.005	0.005
		15	0.012	0.005	0.004	0.003	0.003
		20	0.010	0.004	0.003	0.003	0.002
	1	5	0.021	0.012	0.010	0.010	0.010
		10	0.014	0.007	0.005	0.005	0.005
		15	0.011	0.005	0.004	0.003	0.003
		20	0.010	0.004	0.003	0.002	0.002
8	0.2	5	0.038	0.020	0.016	0.015	0.014
		10	0.028	0.011	0.008	0.007	0.007
		15	0.023	0.008	0.006	0.005	0.005
		20	0.021	0.007	0.004	0.004	0.004
	0.5	5	0.031	0.015	0.013	0.012	0.012
		10	0.022	0.009	0.007	0.006	0.006
		15	0.018	0.006	0.004	0.004	0.004
		20	0.016	0.005	0.003	0.003	0.003
	1	5	0.029	0.015	0.012	0.011	0.011
		10	0.021	0.008	0.006	0.006	0.006
		15	0.017	0.006	0.004	0.004	0.004
		20	0.016	0.005	0.003	0.003	0.003

Table 10: L^2 Loss under Network DGP2

Notes: This table shows the L^2 loss under Network DGP 2. The results include different configurations of the number of basis function (degree), bandwidth, number of schools. The L^2 loss is calculated according to Equation (34) on different choices of $[d_1, d_M]$ reported in Column 4-8. The number of individuals (nodes) in each school (block) is set at 500. The number of simulation repetition is set at 2000.

			L^2 Loss on Interval				
Degree	Bandwidth	#School	$[-0.2, 2]$	$[-0.15, 1.95]$	$[-0.05, 1.85]$	$[0.1, 1.7]$	$[0.3, 1.5]$
3	(0,1]	5	0.028	0.023	0.017	0.012	0.010
		10	0.021	0.017	0.012	0.008	0.006
		15	0.017	0.014	0.009	0.006	0.004
		20	0.015	0.012	0.008	0.005	0.004
5	(0,1]	5	0.055	0.046	0.038	0.028	0.016
		10	0.036	0.028	0.021	0.014	0.009
		15	0.030	0.023	0.016	0.011	0.007
		20	0.027	0.021	0.014	0.010	0.005
8	(0,1]	5	11.145	11.378	9.425	4.052	0.633
		10	0.875	0.879	0.688	0.269	0.048
		15	0.674	0.631	0.476	0.183	0.031
		20	0.378	0.370	0.272	0.088	0.013

Table 11: L^2 Loss under Network DGP3

Notes: This table shows the L^2 loss under Network DGP 3. The results include different configurations of the number of basis function (degree), number of schools. All bandwidth satisfying $0 < b \leq 1$ yields the same L^2 loss as argued in the text. The L^2 loss is calculated according to Equation (34) on different choices of $[d_1, d_M]$ reported in Column 4-8. The number of individuals (nodes) in each school (block) is set at 500. The number of simulation repetition is set at 2000.

			L^2 Loss on Interval				
Degree	Bandwidth	n	$[0.1, 0.85]$	$[0.15, 0.8]$	$[0.2, 0.75]$	$[0.25, 0.7]$	$[0.3, 0.65]$
3	0.2	500	0.073	0.049	0.039	0.035	0.031
		1000	0.047	0.030	0.023	0.019	0.017
		1500	0.039	0.023	0.016	0.013	0.011
		2500	0.029	0.017	0.011	0.008	0.007
	0.5	500	0.062	0.042	0.034	0.030	0.027
		1000	0.042	0.027	0.021	0.017	0.015
		1500	0.035	0.021	0.015	0.012	0.011
		2500	0.026	0.015	0.010	0.008	0.007
	1	500	0.062	0.042	0.034	0.030	0.026
		1000	0.041	0.027	0.020	0.017	0.015
		1500	0.035	0.021	0.015	0.012	0.011
		2500	0.026	0.015	0.010	0.008	0.006
5	0.2	500	0.106	0.077	0.064	0.050	0.042
		1000	0.070	0.048	0.037	0.028	0.022
		1500	0.056	0.036	0.028	0.020	0.015
		2500	0.044	0.025	0.018	0.013	0.010
	0.5	500	0.091	0.066	0.054	0.043	0.036
		1000	0.064	0.043	0.034	0.025	0.020
		1500	0.053	0.033	0.025	0.018	0.014
		2500	0.042	0.024	0.017	0.012	0.009
	1	500	0.090	0.066	0.054	0.042	0.035
		1000	0.063	0.043	0.034	0.025	0.020
		1500	0.053	0.033	0.025	0.018	0.014
		2500	0.042	0.024	0.017	0.012	0.009
8	0.2	500	1.225	0.889	0.325	0.116	0.070
		1000	1.178	0.926	0.386	0.117	0.044
		1500	0.287	0.201	0.087	0.042	0.027
		2500	0.135	0.089	0.047	0.027	0.019
	0.5	500	0.986	0.712	0.260	0.095	0.059
		1000	1.103	0.868	0.362	0.109	0.041
		1500	0.285	0.198	0.083	0.039	0.026
		2500	0.130	0.085	0.044	0.026	0.018
	1	500	0.941	0.679	0.249	0.093	0.059
		1000	1.016	0.796	0.333	0.103	0.040
		1500	0.290	0.203	0.085	0.039	0.026
		2500	0.130	0.085	0.044	0.026	0.018

Table 12: L^2 Loss under Network DGP4

Notes: This table shows the L^2 loss under Network DGP 4. The results include different configurations of the number of basis function (degree), bandwidth, number of nodes (n). The L^2 loss is calculated according to Equation (34) on different choices of $[d_1, d_M]$ reported in Column 4-8. The number of simulation repetition is set at 2000.

B Proofs

B.1 ψ-dependence

The proof of consistency uses the definition of ψ-dependence and the law of large numbers for ψ-dependent variables from Kojevnikov et al. (2021). For any two nodes i, j, let $\ell(i, j)$ be the distance of the shortest path between i, j (i.e., the smallest integer k such that $A^k_{ij} > 0$ and $A^{l'}_{ij} = 0$ for all $k' < k$). For any two sets $A, B \subseteq \mathbb{N}_n$ where $\mathbb{N}_n$ is the collection of nodes, let $\ell(A, B) \coloneqq \{\min_{i,j} \ell(i, j), i \in A, j \in B\}$. Denote $Y_B \coloneqq (Y_i : i \in B)$ for any set $B \subseteq \mathbb{N}_n$. Let $\mathcal{L}_a$ be the set of bounded Lipschitz function from $\mathbb{R}^a \to \mathbb{R}$.

Definition 1 *A triangular array $\{Y_i\}_{i=1}^n$ is called ψ-dependent, if for each n there exists a sequence $\{\theta_{n,s}\}_{s\geq 0}$, $\theta_{n,0} = 1$ and a collection of non-random functional $(\psi_{a,b})_{a,b\in\mathbb{N}} : \mathcal{L}_a \times \mathcal{L}_b \to [0, \infty)$, such that for all $A, B \in \mathcal{P}_n(a, b, s)$ with $s > 0$ and all $f \in \mathcal{L}_a$ and $g \in \mathcal{L}_b$:*

$$|Cov(f(Y_A), g(Y_B))| \leq \psi_{a,b}(f, g)\theta_{n,s} \tag{40}$$

where

$$\mathcal{P}_n(a, b, s) \coloneqq \{(A, B) : A, B \subseteq \mathbb{N}_n, |A| = a, |B| = b, \ell(A, B) \geq s\} \tag{41}$$

Assumption B.1 *There exists a finite integer $S \geq 1$ such that $(v_i, X_i, T_i) \perp (v_j, X_j, T_j)$ for any i, j with $\ell(i, j) \geq S$*

Lemma B.1 *Assume that Assumption 3.2, 5.1, B.1 hold. Then $\{Y_i\}_{i=1}^n$ is ψ-dependent with*

$$\theta_{n,s} = \begin{cases} 1 & s \leq 2S + 1 \\ \tilde{\theta}_{n,s} & s > 2S + 1 \end{cases}$$
$$\psi_{a,b}(f, g) = 4[aLip(f)\|g\|_\infty + bLip(g)\|f\|_\infty + \|f\|_\infty\|g\|_\infty]$$
$$\tilde{\theta}_{n,s} \coloneqq \max\left\{\max_{i\in A} E[|Y_i - Y_i(\bar{y}; \lfloor s\rfloor)|], \max_{j\in B} E[|Y_j - Y_j(\bar{y}; \lfloor s\rfloor)|]\right\}$$

and

$$\tilde{\theta}_{n,s} \leq \frac{\kappa^s}{1-\kappa} E[|Y_i|]$$

Proof. Recall that in Equation (9), we have $Y_i = Y(\{D_j\}_{j\in N_i(L)}, \{T_k, X_k, v_k\}_{k\in N_i(l),1\leq l\leq L-1})$ where we expand the structural equation L times. Let

$$Y_i(a; L) \coloneqq Y(\{a\}_{j\in N_i(L)}, \{T_k, X_k, v_k\}_{k\in N_i(l),1\leq l\leq L-1}) \tag{42}$$

where we replace the value of D_j by a for all $j \in N_i(L)$. By definition of S, $Y_i(a, L) \perp Y_j(a, L)$ if $\ell(i, j) > 2S$. In addition, Assumption

For any $s \leq 2S + 1$, $|Cov(f(Y_A), g(Y_B))| \leq 4\|f\|_\infty \|g\|_\infty$ by boundedness of Y and f, g. For $s > 2S + 1$,

$$\begin{aligned}
&|Cov(f(Y_A), g(Y_B))| \\
&= |Cov(f(Y_A) - f(Y_A(\bar{y}; \lfloor s \rfloor)) + f(Y_A(\bar{y}; \lfloor s \rfloor)), g(Y_B))| \\
&\leq |Cov(f(Y_A) - f(Y_A(\bar{y}; \lfloor s \rfloor)), g(Y_B))| \\
&\quad + |Cov(f(Y_A(\bar{y}; \lfloor s \rfloor)), g(Y_B) - g(Y_B(\bar{y}; \lfloor s \rfloor)) + g(Y_B(\bar{y}; \lfloor s \rfloor)))| \\
&\leq |Cov(f(Y_A) - f(Y_A(\bar{y}; \lfloor s \rfloor)), g(Y_B))| + |Cov(f(Y_A(\bar{y}; \lfloor s \rfloor)), g(Y_B) - g(Y_B(\bar{y}; \lfloor s \rfloor)))| \\
&\hspace{20em} (Cov(f(Y_A(\bar{y}; \lfloor s \rfloor)), g(Y_B(\bar{y}; \lfloor s \rfloor))) = 0) \\
&\leq 2E[|f(Y_A) - f(Y_A(\bar{y}; \lfloor s \rfloor))|]\|g\|_\infty + 2E[|g(Y_B) - g(Y_B(\bar{y}; \lfloor s \rfloor))|]\|f\|_\infty \\
&\leq 2aLip(f) \max_{i \in A} E[|Y_i - Y_i(\bar{y}; \lfloor s \rfloor)|]\|g\|_\infty + 2bLip(g) \max_{j \in B} E[|Y_j - Y_j(\bar{y}; \lfloor s \rfloor)|]\|f\|_\infty \\
&\hspace{12em} (f, g \text{ are Lipschitz functions with bounded Lipschitz constants}) \\
&\leq 2aLip(f)\tilde{\theta}_{n,s}\|g\|_\infty + 2bLip(g)\tilde{\theta}_{n,s}\|f\|_\infty
\end{aligned}$$

which suggests that we can take $\psi_{a,b}(f, g) = 4[aLip(f)\|g\|_\infty + bLip(g)\|f\|_\infty + \|f\|_\infty \|g\|_\infty]$. In addition $\tilde{\theta}_{n,s} \leq \frac{\kappa^s}{1-\kappa} E[|Y_i|]$ by Assumption 3.2, B.1. ■

Corollary B.1 *Assume that Assumption 3.2, B.1 hold. Then $\{D_i\}_{i=1}^n$ is ψ-dependent with*

$$\begin{aligned}
\theta_{n,s} &= \begin{cases} 1 & s \leq 2S + 3 \\ \tilde{\theta}_{n,s} & s > 2S + 3 \end{cases} \\
\psi_{a,b}(f, g) &= 4[aLip(f)\|g\|_\infty + bLip(g)\|f\|_\infty + \|f\|_\infty \|g\|_\infty] \\
\tilde{\theta}_{n,s} &:= \max\left\{\max_{i \in A} E[|Y_i - Y_i(\bar{y}; \lfloor s \rfloor)|], \max_{j \in B} E[|Y_j - Y_j(\bar{y}; \lfloor s \rfloor)|]\right\}
\end{aligned}$$

and

$$\tilde{\theta}_{n,s} \leq \frac{\kappa^s}{1-\kappa} E[|Y_i|]$$

The proof is exactly the same as the one in Lemma B.1 and the change from $2S+1$ to $2S+3$ is due to the fact that D_i is the average of nodes that are one-step away from i.

B.2 Technical Lemma

The assumption of bounded derivative also ensures that the effect decays at a geometric rate, or faster, formalized by the following lemma.

Lemma B.2 *Assume Assumption 3.2, 3.1, 3.3, 3.4 hold. At initial treatment* $\mathbf{T}$ *with* $T_i = 0$*, let* $\tilde{\mathbf{Y}}$ *be the corresponding knowledge. For treatment* $\mathbf{T}^*$ *such that* $T_j^* = T_j \ \forall j \neq i$ *and* $T_i^* = 1$*, denote the resulting knowledge as* $\mathbf{Y}^*$*. Then the following holds:*

$$|Y_j^* - \tilde{Y}_j| \leq \frac{\kappa^\ell}{1-\kappa}\left(\max_{k:A_{ik}=1} \frac{1}{n_k}\right) |g(\tilde{D}_i, 1, X_i, v_i) - g(\tilde{D}_i, 0, X_i, v_i)| \leq \frac{\kappa^\ell}{1-\kappa}\left(\max_{k:A_{ik}=1} \frac{1}{n_k}\right) |Y_i^* - \tilde{Y}_i|$$

where $\tilde{D}_i = \frac{1}{\sum_j A_{ij}} \sum_j A_{ij}\tilde{Y}_j$ *and* ℓ *is the length of the shortest path connecting* i, j*.*

Proof. Define $\mathbf{Y}_{(0)}$ to be such that $Y_{k(0)} = \tilde{Y}_k$ if $k \neq i$ and $Y_{i(0)} = g(\tilde{D}_i, 1, X_i, v_i)$. Further define $\mathbf{Y}_{(n)} = \mathbf{g}(\tilde{A}\mathbf{Y}_{(n-1)}, \mathbf{X}, \mathbf{v})$. As shown in the proof of Proposition 3.2, $|Y_{j(n)} - Y_{j(n-1)}| \leq \kappa\tilde{A}_j|Y_{j(n-1)} - Y_{j(n-2)}|$. This implies that $|Y_{j(n')} - Y_{j(0)}| = 0$ for any $n' < \ell$. Also,

$$\begin{aligned} |Y_j^* - Y_{j(0)}| &\leq \frac{1}{1-\kappa}\|\mathbf{Y}_{(\ell)} - \mathbf{Y}_{(\ell-1)}\|_\infty \leq \frac{\kappa^\ell}{1-\kappa}\|\mathbf{Y}_{(1)} - \mathbf{Y}_{(0)}\|_\infty \\ &= \frac{\kappa^\ell}{1-\kappa}\left(\max_{k:A_{ik}=1} \frac{1}{n_k}\right) |g(\tilde{D}_i, 1, X_i, v_i) - g(\tilde{D}_i, 0, X_i, v_i)| \end{aligned}$$

■

Lemma B.3 *Let* $\{Y_i, W_i, B_i\}_{i=1}^n$ *be a set of random variables and let the bold-faced letter denote the entire vector. For instance,* $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$*. Consider the conditional expectation* $E[Y_i - Y_j|\mathcal{C}(\mathbf{W}, \mathbf{B}, \mathbf{Y})]$ *for some event* $\mathcal{C}$*. Assume that the following holds:*

$$E[Y_i - Y_j|\mathcal{C}(\mathbf{W}, \mathbf{B}, \mathbf{Y})] = E[Y_i - Y_j|\mathcal{C}'(\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(Y_i, Y_j))]$$

for some symmetric function $h(a, b) = h(b, a)$*. Also assume that* Y_i, Y_j *are i.i.d. conditional on* $\mathbf{B}$ *and* $\mathbf{Y} \perp \mathbf{W}$ *conditional on* $\mathbf{B}$*.*

Then the conditional expectation equals zero:

$$E[Y_i - Y_j|\mathcal{C}(\mathbf{W}, \mathbf{B}, \mathbf{Y})] = 0$$

Proof.

$$
\begin{aligned}
& E[Y_i - Y_j|\mathcal{C}(\mathbf{W}, \mathbf{B}, \mathbf{Y})] \\
&= E[Y_i - Y_j|\mathcal{C}'(\mathbf{T}, \mathbf{X}, \mathbf{V}_{-ij}, h(V_i, V_j))] \\
&= E\left[E[Y_i - Y_j|\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(V_i, V_j), \mathcal{C}'(\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(Y_i, Y_j))]\middle|\mathcal{C}'(\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(Y_i, Y_j))\right] \\
&= E\left[E[Y_i - Y_j|\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(Y_i, Y_j)]\middle|\mathcal{C}'(\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(Y_i, Y_j))\right] \\
&= E\left[E[Y_i - Y_j|\mathbf{X}, h(Y_i, Y_j)]\middle|\mathcal{C}'(\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(Y_i, Y_j))\right] \\
& \hspace{15em} ((Y_i, Y_j) \perp (\mathbf{W}, \mathbf{Y}_{-ij}) \text{ conditional on } \mathbf{B}) \\
&= E\left[0\middle|\mathcal{C}'(\mathbf{W}, \mathbf{B}, \mathbf{Y}_{-ij}, h(Y_i, Y_j))\right] \qquad \text{(Lemma B.4)} \\
&= 0
\end{aligned}
$$

where the second last equality follows from substituting $Y_i = V_1, Y_j = V_2$ and $\mathbf{X} = S$ in Lemma B.4. ■

Lemma B.4 *Assume the following holds for the variables $V_1, V_2 \in \mathbb{R}$, $S \in \mathbb{R}^k$ and the function $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$:*

1. *Conditional i.i.d.: V_1, V_2 are i.i.d. conditional on S*
2. *Symmetry: $h(a, b) = h(b, a)$*

Then, under the above assumptions, the following holds:

$$E[V_1 - V_2|h(V_1, V_2) = u, S = s] = 0 \quad \forall (u, s)$$

Proof. For any value t, define $h_1(V; t) := h(t, V)$ and $h_2(V; t) := h(V, t)$. By symmetry, we have $h_1(V; t) = h_2(V; t)$.

It suffices to show that V_1, V_2 have the same conditional distribution.

$$
\begin{aligned}
& f_{V_1|h(V_1,V_2),S}(v|u,s) \\
&= \frac{f_{V_1,h(V_1,V_2),S}(v,u,s)}{f_{h(V_1,V_2),S}(u,s)} = \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{V_1,h(V_1,V_2)|S}(v,u|s) \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h(V_1,V_2)|V_1,S}(u|v,s) f_{V_1|S}(v|s) \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h(v,V_2)|V_1,S}(u|v,s) f_{V_1|S}(v|s) \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h_1(V_2;v)|S}(u|s) f_{V_1|S}(v|s) \qquad (V_1, V_2 \text{ are independent conditional on } S) \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h_1(V_1;v)|S}(u|s) f_{V_2|S}(v|s) \\
&\qquad\qquad\qquad\qquad (V_1, V_2 \text{ are identically distributed conditional on } S) \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h_2(V_1;v)|S}(u|s) f_{V_2|S}(v|s) \qquad (h_1(V_1;v) = h_2(V_1;v) \text{ by symmetry}) \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h(V_1,V_2)|V_2,S}(u|v,x) f_{V_2|S}(v|s) \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{V_2,h(V_1,V_2)|S}(v,u|s) \\
&= f_{V_2|h(V_1,V_2),S}(v|u,s)
\end{aligned}
$$

We have shown that V_1, V_2 has the same density conditional on $(h(V_1, V_2), S)$, which implies the desired equality in first moment. ■

Lemma B.5 *Assume the following holds for the variables $V_1, V_2 \in \mathbb{R}$, $S \in \mathbb{R}^k$ and the function $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$:*

1. *Conditional independence:* V_1, V_2 *are independent conditional on* S
2. *Symmetry:* $h(a,b) = h(b,a)$
3. *Near identical distribution:* $\sup_v |f_{V_1|S}(v|s) - f_{V_2|S}(v|s)| \leq \epsilon$
4. *Bounded density:* $0 < \underline{f} \leq \inf f_{V_2|S}(v|s)$

Then, under the above assumptions, there exists some constant C such that the following holds:

$$
|E[V_1 - V_2|h(V_1,V_2) = u, S = s]| < \frac{\epsilon}{\underline{f}} |E[V_1|h(V_1,V_2) = u, S = s]| \quad \forall (u)
$$

Proof. For any value t, define $h_1(V;t) := h(t,V)$ and $h_2(V;t) := h(V,t)$. By symmetry, we have $h_1(V;t) = h_2(V;t)$.

It suffices to show that V_1, V_2 have the same conditional distribution.

$$
\begin{aligned}
&|f_{V_1|h(V_1,V_2),S}(v|u,s) - f_{V_2|h(V_1,V_2),S}(v|u,s)| \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h_1(V_2;v)|S}(u|s) |f_{V_1|S}(v|s) - f_{V_2|S}(v|s)| && \text{(Proof in Lemma B.4)} \\
&\leq \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h_1(V_2;v)|S}(u|s)\epsilon \\
&= \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{h_1(V_2;v)|S}(u|s) \frac{f_{V_2|S}(v|s)}{f_{V_2|S}(v|s)}\epsilon \\
&\leq \frac{f_S(s)}{f_{h(V_1,V_2),S}(u,s)} f_{V_2,h(V_1,V_2)|S}(v,u|s) \left(1 + \frac{\epsilon}{\underline{f}}\right) && \text{(bounded density)} \\
&= f_{V_2|h(V_1,V_2),S}(v|u,s) \left(1 + \frac{\epsilon}{\underline{f}}\right)
\end{aligned}
$$

The result follows from

$$
\begin{aligned}
|E[V_1 - V_2|h(V_1,V_2) = u, S = s]| &\leq |\int v|f_{V_1|h(V_1,V_2),S}(v|u,s) - f_{V_2|h(V_1,V_2),S}(v|u,s)|dv| \\
&\leq |\int v\frac{\epsilon}{\underline{f}} f_{V_2|h(V_1,V_2),S}(v|u,s)dv| \\
&= \frac{\epsilon}{\underline{f}} |E[V_1|h(V_1,V_2) = u, S = s]
\end{aligned}
$$

■

Lemma B.6 *Let Assumption 4.1, 5.1, 5.5 hold. Further assume that the kernel K_1 has bounded support. Then the following holds:*

$$
Var\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}(v_i - v_j)\omega_{ij}\right) = o(1) \tag{43}
$$

Proof. Define the following objects:

$$
\begin{aligned}
m_{ij} &:= (v_i - v_j)\omega_{ij} \\
M_i &:= \sum_{j\neq i} m_{ij}
\end{aligned} \tag{44}
$$

Firstly, $Var(M_i) \leq C_M$ for some constant C_M by the boundedness of $E[|v_i|^4|A_i, X_i]$ in Assumption 5.1, and that $\sum_j \omega_{ij} = 1$, $\omega_{ij} \geq 0$. For the covariance, first realize that

$Cov(M_i, M_k) = 0$ for any $k \in N_n^{\partial}(i; s)$ with $s \geq 5$.

$$Cov(M_i, M_k) = Cov\left(\sum_{j \neq i} m_{ij}, \sum_{l \neq k} m_{kl}\right)$$

By the conditional independence assumption on v_i in Assumption 4.1, $Cov(m_{ij}, m_{kl}) \neq 0$ only under the event $\{i = l\} \cup \{k = j\} \cup \{j = l\}$ (i.e. there is overlapping in the index). For any $k \in N_n^{\partial}(i; s)$ with $s \geq 5$, it must be that $\omega_{ik} = 0$ and $\omega_{ki} = 0$ for n large enough since they share no node in common and that K_1 is compactly supported. It remains to consider the case where $\{j = l\}$. For $\omega_{ij} \neq 0$, it must be that j be at most 2-step away from i. This is because i, j must share common links for $\omega_{ij} \neq 0$. Similarly, for $\omega_{kl} \neq 0$, l must be at most 2-step away from k. However, when $k \in N_n^{\partial}(i; s)$ with $s \geq 5$, there is no node that is within 2-step away from both i, k. Thus the covariance term equals zero.

It follows that

$$\begin{aligned}
\frac{1}{n^2}\sum_i \sum_{k \neq i} |Cov(M_i, M_k)| &= \frac{1}{n^2}\sum_i \sum_{s=1}^{\infty} \sum_{k \in N_n^{\partial}(i;s)} |Cov(M_i, M_k)| \\
&= \frac{1}{n^2}\sum_i \sum_{s=1}^{4} \sum_{k \in N_n^{\partial}(i;s)} |Cov(M_i, M_k)| \\
&\qquad (Cov(M_i, M_k) = 0 \text{ for any } k \in N_n^{\partial}(i; s) \text{ with } s \geq 5) \\
&\leq \frac{1}{n^2}\sum_i \sum_{s=1}^{4} \sum_{k \in N_n^{\partial}(i;s)} 2C_M \qquad \text{(Cauchy-Schwarz inequality)} \\
&\leq 2C_M \frac{1}{n}\sum_i \frac{1}{n}\sum_{s=1}^{4} |N_n^{\partial}(i; s)| \\
&= 2C_M \frac{1}{n}\sum_{s=1}^{4} \delta_n^{\partial}(s; 1) = o(1) \quad (\text{by Assumption 5.5 and } \kappa \in (0,1))
\end{aligned}$$

Therefore,

$$\begin{aligned}
Var\left(\frac{1}{n}\sum_{i=1}^{n} M_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n} Var(M_i) + \frac{1}{n^2}\sum_i \sum_{k \neq i} Cov(M_i, M_k) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n} C_M + \frac{1}{n^2}\sum_i \sum_{k \neq i} |Cov(M_i, M_k)| = o(1)
\end{aligned}$$

■

Lemma B.7 *Assume that Assumption 3.2, 4.1, 5.1 - 5.6 hold. Further assume that Assumption 5.7 or Assumption 5.8 holds. For any q, $L_n(q; b) \xrightarrow{p} L(q) + C$ for some constant*

C independent of q.

Proof. For the proof, I will write $\frac{1}{n}\sum_{i=1}^{n}$ instead of $\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}$ for ease of notation. The intuition remains the same since $|\mathcal{T}|$ also diverges.

Define the following quantities:

$$\begin{aligned} m_{1,ij} &:= [|T_i - T_j|\tau_T(D_i, X_i) - |T_i - T_j|q(D_i, X_i)] \\ m_{2,ij} &:= [g(D_i, T_j, X_j) - g(D_j, T_j, X_j)] \\ m_{3,ij} &:= [(T_i - T_j)(v_i - v_j)] \end{aligned}$$

Expand L_n:

$$\begin{aligned} &L_n(q; b) \\ &:= \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}[(T_i - T_j)(Y_i - Y_j) - |T_i - T_j|q(D_i, X_i)]^2\omega_{ij} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}[m_{1,ij}^2 + m_{2,ij}^2 + m_{3,ij}^2 + 2m_{1,ij}m_{2,ij} + 2m_{1,ij}m_{3,ij} + 2m_{2,ij}m_{3,ij}]\omega_{ij} \end{aligned}$$

Step 1: First show that the cross term $\frac{1}{n}\sum_i\sum_j m_{1,ij}m_{3,ij}\omega_{ij}$ vanishes. This is carried out in two steps. In Step 1.1, I show that $\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}[m_{1,ij}m_{3,ij}\omega_{ij} - E[m_{1,ij}m_{3,ij}\omega_{ij}]] = o_p(1)$ through L^2 convergence. In Step 1.2, I show that $E[m_{1,ij}m_{3,ij}\omega_{ij}] = o(1)$. It is useful to rewrite the summation as follows:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} m_{1,ij}m_{3,ij}\omega_{ij} \\ &= \frac{1}{n}\sum_{i}(\tau_T(D_i, X_i) - q(D_i, X_i))\left[v_i - \sum_{j:T_j=0} v_j\omega_{ij}\right] \qquad (\textstyle\sum_j \omega_{ij} = 1) \end{aligned}$$

Define the following objects:

$$\begin{aligned} \gamma_n(A_i, X_i) &:= \sum_{j} v_j\omega_{ij} \\ \gamma(A_i, X_i) &:= E[v_j|A_j = A_i, X_j = X_i, T_j = 0] \end{aligned}$$

Notice that $\gamma(A_i, X_i) = E[T_jv_j|A_j = A_i, X_j = X_i] = E[T_jv_j|A_j = A_i, X_j = X_i, D_j = D_i]$ so we could equivalently think of it as $\gamma(A_i, X_i, D_i)$.

Step 1.1: The following result holds:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}[m_{1,ij}m_{3,ij}\omega_{ij} - E[m_{1,ij}m_{3,ij}\omega_{ij}]] \\
&= \frac{1}{n}\sum_{i}(\tau_T(D_i, X_i) - q(D_i, X_i))[v_i - \gamma_{2,n}(A_i, X_i)] - E[T_i(\tau_T(D_i, X_i) - q(D_i, X_i))[v_i - \gamma_{2,n}(A_i, X_i)]]) \\
&= o_p(1)
\end{aligned}
$$

The convergence result follows from the following L^2 convergence argument: Assumption 5.1 implies that

$$
\begin{aligned}
Var\left(\frac{1}{n}\sum_{i}(\tau_T(D_i, X_i) - q(D_i, X_i))[(v_i - \gamma_{2,n}(A_i, X_i))]\right) &\leq 4\bar{y}Var\left(\frac{1}{n}\sum_{i}(v_i - \gamma_{2,n}(A_i, X_i))\right) \\
&= o(1) \qquad \text{(Lemma B.6)}
\end{aligned}
$$

This implies that $Var(\frac{1}{n}\sum_i(\tau_T(D_i, X_i) - q(D_i, X_i))[(v_i - \gamma_{2,n}(A_i, X_i))]) \to 0$.

Step 1.2: Now show that $E[m_{1,ij}m_{3,ij}\omega_{ij}] = o(1)$ under either Assumption 5.8 or Assumption 5.7.

Step 1.2.1: Consider first the case where Assumption 5.7 holds. Define $\tilde{\omega}_j := \omega_j \mathbb{1}\{A_i = A_j, \|X_i - X_j\|_2 \leq C'b\}$.

$$
\begin{aligned}
\omega_j - \tilde{\omega}_j &= \omega_j \mathbb{1}\{A_i \neq A_j, \|X_i - X_j\|_2 \leq C'b\} + \omega_j \mathbb{1}\{\|X_i - X_j\|_2 > C'b\} \\
&= \omega_j \mathbb{1}\{A_i \neq A_j, \|X_i - X_j\|_2 \leq C'b\} \qquad \text{(compact support of } K_2)
\end{aligned}
$$

However, since K_1 has compact support and A_i has finite support, $K_1(\frac{s_{ij}}{b}) = 0$ for any $A_i \neq A_j$ for n large enough. As a result, $\omega_j - \tilde{\omega}_j = 0$ for n large enough.

These imply that

$$
\begin{aligned}
E[v_i - \gamma_{2,n}(A_i, X_i)|A_i, D_i, X_i] &= E\left[\sum_j \tilde{\omega}_j v_i - \sum_j \tilde{\omega}_j v_j | A_i, D_i, X_i\right] \\
&\qquad \text{(by } \textstyle\sum_j \tilde{\omega}_j = 1 \text{ for large } n) \\
&= E\left[\sum_j \tilde{\omega}_j E[v_i - v_j | A_i = A_j, X_i, X_j, \|X_i - X_j\|_2 \leq C'b, D_i] | A_i, D_i, X_i\right] \\
&\qquad (\tilde{\omega}_j = 0 \text{ for } A_i \neq A_j \text{ or } \|X_i - X_j\|_2 > C'b) \\
&= E\left[\sum_j \tilde{\omega}_j E[v_i - v_j | A_i = A_j, X_i, X_j, \|X_i - X_j\|_2 \leq C'b, v_i + v_j] | A_i, D_i, X_i\right]
\end{aligned}
$$

(D_i depends only on $v_i + v_j$ when $A_i = A_j$ by the identification argument in Proposition 4.1)

Applying Lemma B.5, we have

$$
\begin{aligned}
& |E[v_i - \gamma_{2,n}(A_i, X_i)|A_i, D_i, X_i]| \\
& \leq E\left[\sum_j \tilde{\omega}_j |E[v_i - v_j|A_i = A_j, X_i, X_j, \|X_i - X_j\|_2 \leq C'b, v_i + v_j]||A_i, D_i, X_i\right] \\
& \leq E\left[\sum_j \tilde{\omega}_j |E[v_i|A_i, X_i, v_i + v_j]| \frac{\eta}{\underline{f}} |A_i, D_i, X_i\right] && \text{(Lemma B.5)} \\
& \leq E\left[\sum_j \tilde{\omega}_j E[|v_i||A_i, X_i, v_i + v_j] \frac{\eta}{\underline{f}} |A_i, D_i, X_i\right] = E\left[\sum_j \tilde{\omega}_j E[|v_i||A_i, X_i, D_i] \frac{\eta}{\underline{f}} |A_i, D_i, X_i\right] \\
& && \text{(conditional independence as argued in Proposition 4.1)} \\
& = E[|v_i||A_i, X_i, D_i] \frac{\eta}{\underline{f}} && \text{(by } \textstyle\sum_j \tilde{\omega}_j = 1)
\end{aligned}
$$

where $|f_{V_i|A_i,X_i}(v|a,x) - f_{V_i|A_i,X_i}(v|a,x')| < \eta$ for any $\|x - x'\|_2 \leq C'b$. To apply Lemma B.5, replace V_1, V_2 by v_i, v_j, $h(V_1, V_2)$ by $v_1 + v_2$, S by $(X_i, \|X_j - X_i\|_2)$.

These imply that

$$
\begin{aligned}
& |E[T_i(\tau_T(D_i, X_i) - q(D_i, X_i))[v_i - \gamma_{2,n}(A_i, X_i)]]| \\
& = |E[T_i(\tau_T(D_i, X_i) - q(D_i, X_i))[E[v_i - \gamma_{2,n}|A_i, D_i, X_i]]]| \\
& \leq E[|T_i(\tau_T(D_i, X_i) - q(D_i, X_i))||E[v_i - \gamma_{2,n}|A_i, D_i, X_i]|] \\
& \leq E[|T_i(\tau_T(D_i, X_i) - q(D_i, X_i))|E[|v_i||A_i, X_i, D_i]] \frac{\eta}{\underline{f}} \\
& = E[|T_i(\tau_T(D_i, X_i) - q(D_i, X_i))||v_i|] \frac{\eta}{\underline{f}} && \text{(Law of iterated expectation)} \\
& \leq \|\tau_T(D_i, X_i) - q(D_i, X_i)\|_\infty E[v_i^2] \frac{\eta}{\underline{f}} = o(1)
\end{aligned}
$$

The $o(1)$ result holds since η can be made arbitrarily small. Therefore, $E[m_{1,ij} m_{3,ij} \omega_{ij}] = o(1)$ as desired.

Step 1.2.2: Now consider the case where Assumption 5.8 holds. Then we have

$$
\begin{aligned}
& E[T_i(\tau_T(D_i,X_i)-q(D_i,X_i))[v_i-\gamma_{2,n}(A_i,X_i)]] \\
& \leq E\left[T_i(\tau_T(D_i,X_i)-q(D_i,X_i))\left[\sum_j \omega_{ij}\mathbb{1}\{s_{ij}\leq Cb\}v_i-\sum_j v_j\omega_{ij}\mathbb{1}\{s_{ij}\leq Cb\}\right]\right] \\
& \qquad\qquad\qquad\qquad \text{(support condition on } K_1 \text{ imposed by Assumption 5.8 and } \textstyle\sum_j \omega_j = 1) \\
& = E\left[\sum_j T_i\omega_{ij}\mathbb{1}\{s_{ij}\leq Cb\}E[(\tau_T(D_i,X_i)-q(D_i,X_i))(v_i-v_j)|A_i,A_j,T_i,T_j,X_i,X_j]\right] \\
& \leq E\left[\mathbb{1}\{s_{ij}\leq Cb\}t_l(s_{ij})\right] \qquad\qquad (\textstyle\sum_j \omega_{ij}=1) \\
& = o(1)
\end{aligned}
$$

The convergence result follows from point 1 of Assumption 5.8. In addition $\tau_T(D_i,X_i)-q(D_i,X_i)$ is a Lipschitz function with respect to D_i for any X_i.

Step 2: Next show that the terms involving $m_{2,ij}$ vanishes almost surely. By Assumption 3.2, 5.2, $m_{2,ij}\leq Lip(x)(X_i-X_j)+\kappa(D_i-D_j)$. In addition, $|D_i-D_j|\leq s_{ij}2\bar{y}+(\tilde{A}_i-\tilde{A}_j)'\mathbf{v}$. It follows that

$$
|m_{2,ij}|\leq |Lip(x)(X_i-X_j)|+2\kappa s_{ij}\bar{y}+\kappa|(\tilde{A}_i-\tilde{A}_j)'\mathbf{v}|
$$

Define $\bar{m}_{2,ij}\coloneqq |Lip(x)(X_i-X_j)|+2\kappa s_{ij}\bar{y}$. For an arbitrary ϵ, there exists a pair (δ_x,δ_s) such that $|\bar{m}_{2,ij}|<\epsilon$ if $|X_i-X_j|<\delta_x$ and $s_{ij}<\delta_s$. However, $b\to 0$ implies that $\omega_{ij}\to 0$ for any $|X_i-X_j|>\delta_x$ or $s_{ij}>\delta_s$. Since $b\to 0$ as $n\to\infty$, there exists n_ϵ such that $b<\min\{\delta_x,\delta_s\}$ for all $n>n_\epsilon$. This implies that for any $n>n_\epsilon$:

$$
\begin{aligned}
\left|\frac{1}{n}\sum_i\sum_{j\neq i}\bar{m}^2_{2,ij}\omega_{ij}\right| &\leq \frac{1}{n}\sum_i\sum_{j\neq i}|\bar{m}^2_{2,ij}\omega_{ij}| \\
&= \frac{1}{n}\sum_i\sum_{j\neq i}|\bar{m}^2_{2,ij}\omega_{ij}(\mathbb{1}\{|X_i-X_j|<\delta_x, s_{ij}<\delta_\epsilon\})| \\
&\leq \frac{1}{n}\sum_i\sum_{j\neq i}\epsilon^2\omega_{ij}=\epsilon^2
\end{aligned}
$$

Therefore, $P(\left|\frac{1}{n}\sum_i\sum_{j\neq i}\bar{m}^2_{2,ij}\omega_{ij}\right|>\epsilon^2)=0$ for all $n>n_\epsilon$. By the Borel-Cantelli Lemma, it follows that

$$
P\left(\left|\frac{1}{n}\sum_i\sum_{j\neq i}\bar{m}^2_{2,ij}\omega_{ij}\right|>\epsilon^2 \text{ infinitely often}\right)=0
$$

Since ϵ is arbitrary, we have

$$\frac{1}{n}\sum_i\sum_{j\neq i}\bar{m}_{2,ij}^2\omega_{ij}\xrightarrow{a.s.}0$$

I show convergence in L^2 norm for the term involving $|(\tilde{A}_i-\tilde{A}_j)'\mathbf{v}|$. Define

$$r_{ij} := [(\tilde{A}_i-\tilde{A}_j)'\mathbf{v}\omega_{ij}]^2 R_i := \sum_j r_{ij}$$

It follows that

$$\begin{aligned}
E\left[\left|\frac{1}{n}\sum_i\sum_{j\neq i}|(\tilde{A}_i-\tilde{A}_j)'\mathbf{v}|^2\omega_{ij}\right|^2\right] &= E\left[\left|\frac{1}{n}\sum_i R_i\right|^2\right]\\
&= \frac{1}{n^2}\sum_i E[R_i^2]+\frac{1}{n^2}\sum_i\sum_{j\neq i}E[R_iR_j]\\
&= o(1)
\end{aligned}$$

The last equality follows from two arguments. First, as in the proof of Lemma B.6

$$\left|\frac{1}{n^2}\sum_i\sum_{j\neq i}E[R_iR_j]\right| = o(1)$$

Second, $E[R_i^2]$ is bounded, which ensures $\frac{1}{n^2}\sum_i E[R_i^2]=o(1)$.

The convergence in probability result thus follows:

$$\begin{aligned}
\left|\frac{1}{n}\sum_i\sum_{j\neq i}m_{2,ij}^2\omega_{ij}\right| &\leq 2\left|\frac{1}{n}\sum_i\sum_{j\neq i}\bar{m}_{2,ij}^2\omega_{ij}\right|\\
&\quad + 2\left|\frac{1}{n}\sum_i\sum_{j\neq i}\kappa^2|(\tilde{A}_i-\tilde{A}_j)'\mathbf{v}|^2\omega_{ij}\right|\\
&\xrightarrow{p} 0
\end{aligned}$$

The same analysis applies to $\frac{1}{n}\sum_i\sum_{j\neq i}m_{2,ij}m_{1,ij}\omega_{ij}$ by bounding $m_{1,ij}$ using Assumption 5.1.

Step 3: I show that $\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}(m_{1,ij}^2\omega_{ij} - E[m_{1,ij}^2\omega_{ij}]) = o_p(1)$. However, notice that

$$E[\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} m_{1,ij}^2\omega_{ij}] = E[(\tau_T(D_i, X_i) - q(D_i, X_i))^2 | i \in \mathcal{T}]$$

by $\sum_j \omega_{ij} = 1$. The desired result hence follows.

To show the first convergence result, it suffices to show that $Var(\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} m_{1,ij}^2\omega_{ij}) = o(1)$. This follows from the proof of Theorem 3.1 in Kojevnikov et al. (2021). To apply the result in Kojevnikov et al. (2021), I verify that the assumptions hold. By Corollary B.1, D_i is ψ-dependent with $\theta_{n,s} = C\kappa^s$ where C is some constant. In addition $\psi_{a,b}(f,g) = 4[aLip(f)\|g\|_\infty + bLip(g)\|f\|_\infty + \|f\|_\infty\|g\|_\infty]$ satisfies Assumption 2.1 in Kojevnikov et al. (2021). Assumption 3.1 and 3.2 in Kojevnikov et al. (2021) are implied by Assumption 5.1, 5.5. Finally, τ, q are both Lipschitz.

Step 4: $\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} m_{3,ij}^2\omega_{ij} \xrightarrow{p} C$ where C is some constant independent of q. To show this, it suffices to show that its variance tends to zero. Define $M_i \coloneqq \sum_{j\neq i} m_{3,ij}^2\omega_{ij}$. Firstly, $Var(M_i) \leq C_M$ for some constant C_M by the boundedness of $E[|v_i|^4|A_i, X_i]$ and that $\sum_j \omega_{ij} = 1$, $\omega_{ij} \geq 0$. For the covariance, first realize that $Cov(M_i, M_k) = 0$ for any $k \in N_n^\partial(i;s)$ with $s \geq 5$. To see this, first recall that $m_{3,ij} = (T_i - T_j)(v_i - v_j)$.

$$Cov(M_i, M_k) = Cov\left(\sum_{j\neq i} m_{3,ij}^2\omega_{ij}, \sum_{l\neq k} m_{3,kl}^2\omega_{kl}\right)$$

By the conditional independence assumption on v_i, $Cov(m_{3,ij}^2, m_{3,kl}^2) \neq 0$ only under the event $\{i = l\} \cup \{k = j\} \cup \{j = l\}$ (i.e. there is overlapping in the index). For any $k \in N_n^\partial(i;s)$ with $s \geq 5$, it must be that $\omega_{ik} = 0$ and $\omega_{ki} = 0$ for n large enough since they share no node in common and that K_1 is compactly supported. It remains to consider the case where $\{j = l\}$. For $\omega_{ij} \neq 0$, it must be that j be at most 2-step away from i. This is because, i, j must share common links for $\omega_{ij} \neq 0$. Similarly, for $\omega_{kl} \neq 0$, l must be at most 2-step away from k. However, when $k \in N_n^\partial(i;s)$ with $s \geq 5$, there is no node that is within 2-step away from both i, k. Thus the covariance term equals zero.

It follows that

$$\begin{aligned}
\frac{1}{n^2}\sum_i\sum_{k\neq i}|Cov(M_i,M_k)| &= \frac{1}{n^2}\sum_i\sum_{s=1}^{\infty}\sum_{k\in N_n^{\partial}(i;s)}|Cov(M_i,M_k)| \\
&= \frac{1}{n^2}\sum_i\sum_{s=1}^{4}\sum_{k\in N_n^{\partial}(i;s)}|Cov(M_i,M_k)| \\
&\qquad\qquad (Cov(M_i,M_k)=0 \text{ for any } k\in N_n^{\partial}(i;s) \text{ with } s\geq 5) \\
&\leq \frac{1}{n^2}\sum_i\sum_{s=1}^{4}\sum_{k\in N_n^{\partial}(i;s)}2C_M \qquad \text{(Cauchy-Schwarz inequality)} \\
&\leq 2C_M\frac{1}{n}\sum_i\frac{1}{n}\sum_{s=1}^{4}|N_n^{\partial}(i;s)| \\
&= 2C_M\frac{1}{n}\sum_{s=1}^{4}\delta_n^{\partial}(s;1)=o(1) \quad \text{(by Assumption 5.5 and } \kappa\in(0,1))
\end{aligned}$$

Therefore,

$$\begin{aligned}
Var\left(\frac{1}{n}\sum_{i=1}^{n}M_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n}Var(M_i)+\frac{1}{n^2}\sum_i\sum_{k\neq i}Cov(M_i,M_k) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}C_M+\frac{1}{n^2}\sum_i\sum_{k\neq i}|Cov(M_i,M_k)|=o(1)
\end{aligned}$$

■

B.3 Proof of Results in the Paper

B.3.1 Proof of Proposition 3.1

Proof. Denote $\mathbf{Y}_{(s)}(\mathbf{t}), \triangle_{(s)}(\mathbf{t})$ as the resulting value of $\mathbf{Y}_{(s)}, \triangle_{(s)}$ defined in Example 3.4 as a function of the treatment assignment $\mathbf{t}$.

Only if direction: $\tau_T(\bar{A}\mathbf{Y})'\mathbf{t}_1 > \tau_T(\bar{A}\mathbf{Y})'\mathbf{t}_2$ implies that $\mathbf{1}'\triangle_{(1)}(\mathbf{t}_1) > \mathbf{1}'\triangle_{(1)}(\mathbf{t}_2)$. Since $\bar{A}$ is fully connected, this implies that $\bar{A}\triangle_{(1)}(\mathbf{t}_1) > \bar{A}\triangle_{(1)}(\mathbf{t}_2)$ (element-wise comparison). Since $\mathbf{Y}_{(1)}(\mathbf{t}_1) - \triangle_{(1)}(\mathbf{t}_1) = \mathbf{Y} = \mathbf{Y}_{(1)}(\mathbf{t}_2) - \triangle_{(1)}(\mathbf{t}_2)$, the above implies $\bar{A}\mathbf{Y}_{(1)}(\mathbf{t}_1) > \bar{A}\mathbf{Y}_{(1)}(\mathbf{t}_2)$. By Assumption 3.4, this implies $\mathbf{Y}_{(2)}(\mathbf{t}_1) > \mathbf{Y}_{(2)}(\mathbf{t}_2)$. Perform induction along this along and we have $\mathbf{Y}_{(s)}(\mathbf{t}_1) > \mathbf{Y}_{(s)}(\mathbf{t}_2)$ for all s. Since $\mathbf{Y}_1^* = \lim_s \mathbf{Y}_{(s)}(\mathbf{t}_1)$ and $\mathbf{Y}_2^* = \lim_s \mathbf{Y}_{(s)}(\mathbf{t}_2)$, it follows that $\mathbf{Y}_1^* \geq \mathbf{Y}_2^*$. However, using the same argument, we can also show that $\triangle_{(s)}(\mathbf{t}_1) > \triangle_{(s)}(\mathbf{t}_2)$ for all s, which implies $\mathbf{Y}_1^* > \mathbf{Y}_2^*$ by the infinite sum representation.

If direction: suppose $\tau_T(\bar{A}\mathbf{Y})'\mathbf{t}_1 \leq \tau_T(\bar{A}\mathbf{Y})'\mathbf{t}_2$, the above proof shows that $\mathbf{1}'\mathbf{Y}_1^* \leq \mathbf{1}'\mathbf{Y}_2^*$, a contradiction. ■

B.3.2 Proof of Proposition 3.2

Proof. Write the knowledge equation in matrix form: $\mathbf{Y} = \mathbf{g}(\mathbf{D}, \mathbf{T}, \mathbf{X}, \mathbf{v}) = \mathbf{g}(\tilde{A}\mathbf{Y}, \mathbf{T}, \mathbf{X}, \mathbf{v})$ where $\tilde{A}$ is the row-normalized version of A. The existence and uniqueness of the reduced form can be framed as the existence and uniqueness of the fixed point of $\mathbf{Y} = \mathbf{g}(\tilde{A}\mathbf{Y}, \mathbf{T}, \mathbf{X}, \mathbf{v})$. For some starting value $\mathbf{Y}_{(0)}$, define $\mathbf{Y}_{(n)} \coloneqq \mathbf{g}(\tilde{A}\mathbf{Y}_{(n-1)}, \mathbf{T}, \mathbf{X}, \mathbf{v})$.

By the mean-value theorem and boundedness of the derivative $\frac{\partial Y_i}{\partial D_i} \leq \kappa < 1$, we have that for any i:

$$
\begin{aligned}
|g(\tilde{A}\mathbf{Y}, T_i, X_i, v_i) - g(\tilde{A}\mathbf{Y}^*, T_i, X_i, v_i)| &= |\frac{\partial}{\partial D} g(\tilde{A}\tilde{\mathbf{Y}}, T_i, X_i, v_i)\tilde{A}_i(\mathbf{Y} - \mathbf{Y}^*)| \\
&\leq \frac{\kappa}{\sum_j A_{ij}} |\sum_j A_{ij}(Y_j - Y_j^*)| \\
&\leq \frac{\kappa}{\sum_j A_{ij}} \sum_j A_{ij}|(Y_j - Y_j^*)| \\
&\leq \frac{\kappa}{\sum_j A_{ij}} \sum_j A_{ij}\|(\mathbf{Y} - \mathbf{Y}^*)\|_\infty = \kappa\|(\mathbf{Y} - \mathbf{Y}^*)\|_\infty
\end{aligned}
$$

This implies

$$
\|\mathbf{g}(\tilde{A}\mathbf{Y}, \mathbf{T}, \mathbf{X}, \mathbf{v}) - \mathbf{g}(\tilde{A}\mathbf{Y}^*, \mathbf{T}, \mathbf{X}, \mathbf{v})\|_\infty \leq \kappa\|(\mathbf{Y} - \mathbf{Y}^*)\|_\infty
$$

is a contraction for any realization of $\mathbf{T}, \mathbf{X}, \mathbf{v}$ under the distance induced by the ℓ_∞ norm. By the Banach fixed point theorem, there is a unique fixed point.

Consider two treatment vectors $\mathbf{T}, \mathbf{T}^*$ such that $T_i = T_i^*$ for all $i \neq j$ and $T_j = 0, T_j^* = 1$. Denote the resulting knowledge as $\mathbf{Y} = r(\mathbf{T}, \mathbf{X}, \mathbf{v})$ and $\mathbf{Y}^* = r(\mathbf{T}^*, \mathbf{X}, \mathbf{v})$. For two vectors $\mathbf{a}, \mathbf{b}$, define $\mathbf{a} < \mathbf{b}$ as $a_i \leq b_i$ for all i with strict inequality for at least one i. By definition, $\mathbf{Y} = \mathbf{g}(\tilde{A}\mathbf{Y}, \mathbf{T}, \mathbf{X}, \mathbf{v})$. By Assumption 3.3, $\mathbf{Y} < \mathbf{g}(\tilde{A}\mathbf{Y}, \mathbf{T}^*, \mathbf{X}, \mathbf{v}) \coloneqq \mathbf{Y}_{(1)}$. By Assumption 3.4, $\mathbf{Y}_{(2)} = \mathbf{g}(\tilde{A}\mathbf{Y}_{(1)}, \mathbf{T}, \mathbf{X}, \mathbf{v}) > \mathbf{g}(\tilde{A}\mathbf{Y}, \mathbf{T}, \mathbf{X}, \mathbf{v}) > \mathbf{Y}$. By induction, we can show that $\mathbf{Y}_{(n)} > \mathbf{Y}$ for all n. As argued above, $\lim_{n\to\infty} \|\mathbf{Y}_{(n)} - \mathbf{Y}^*\|_\infty \to 0$, this implies that $\mathbf{Y}^* \geq \mathbf{Y}$. By Assumption 3.3, $\mathbf{Y} < \mathbf{g}(\tilde{A}\mathbf{Y}, \mathbf{T}^*, \mathbf{X}, \mathbf{v})$, which implies that $\mathbf{Y}^* > \mathbf{Y}$.

The above proof only uses the fact that $\sum_j \tilde{A}_{ij} = 1$ and $\tilde{A}_{ij} \geq 0$ and thus also holds for any row-normalized matrix B with non-negative entries. ■

B.3.3 Proof of Proposition 4.1

Proof. Let $\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v}) \coloneqq \{A_{ki} = A_{kj}\ \forall k, T_i = 1, T_j = 0, D_j = D_i = d, X_i = X_j = x\}$ denote the conditioning event.

Substituting the additive separability structure in Assumption 4.1:

$$\begin{aligned}&E[Y_i - Y_j|\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v})]\\&= \bar{g}(d,1,x) - \bar{g}(d,0,x) + E[v_i - v_j|\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v})]\end{aligned}$$

To complete the proof, I apply Lemma B.3 to show that $E[v_i - v_j|\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v})] = 0$. To this end, I show that the following holds:

$$\begin{aligned}\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v}) &= \mathcal{E}'(\mathbf{T}, \mathbf{X}, A, \mathbf{v}_{-ij}, h(v_i, v_j))\\h(v_i, v_j) &\coloneqq v_i + v_j\end{aligned}$$

which is equivalent to showing $D_i = D_j = d$ can be written as a restriction on $\mathbf{T}, \mathbf{X}, A, \mathbf{v}_{-ij}, h(v_i, v_j)$. This is because $D_i = D_j = d$ is the only restriction in $\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v})$ that involves v_i, v_j. After this, substituting $v_i = Y_i, v_j = Y_j$, $\mathbf{T} = \mathbf{W}, (\mathbf{X}, \tilde{A}_i, \tilde{A}_j) = \mathbf{B}, \mathbf{v} = \mathbf{Y}$, $h(v_i, v_j) = h(Y_i, Y_j)$ in Lemma B.3 yields the desired result.

For simplicity, consider first the case where $n_k = 2$ for all k such that $A_{ki} = A_{kj} = 1$ (i.e. all common neighbors of i, j have only two degrees). Then, conditional on other restrictions in $\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v})$, the event $D_i = D_j = d$ can be written as:

$$\begin{aligned}d = D_i = D_j &= \frac{1}{n_i} \sum_{k: A_{ki} A_{kj} = 1} \bar{g}(D_k, T_k, X_k, v_k) \quad (A_{ki} = A_{kj} \text{ for all } k \text{ conditional on } \mathcal{E}(d, x))\\&= \frac{1}{n_i} \sum_{k: A_{ki} A_{kj} = 1} \bar{g}(\tilde{A}_{ki} Y_i + \tilde{A}_{kj} Y_j, T_k, X_k, v_k) \qquad (\ n_k = 2)\\&= \frac{1}{n_i} \sum_{k: A_{ki} A_{kj} = 1} \bar{g}\left(\frac{1}{n_k}(v_i + v_j) + \frac{1}{n_k}\bar{g}(d,1,x) + \frac{1}{n_k}\bar{g}(d,0,x), T_k, X_k, v_k\right)\\&\qquad (\tilde{A}_{ki} = \tilde{A}_{kj} = \tfrac{1}{n_k} \text{ by assumption})\\&= \frac{1}{n_i} \sum_{k: A_{ki} A_{kj} = 1} \bar{g}\left(\frac{1}{n_k} h(v_i, v_j) + \frac{1}{n_k}\bar{g}(d,1,x) + \frac{1}{n_k}\bar{g}(d,0,x), T_k, X_k, v_k\right)\\&\qquad (h(v_i, v_j) \coloneqq v_i + v_j)\end{aligned}$$

This implies that $D_i = D_j = d$ can be written as a restriction on $\mathbf{T}, \mathbf{X}, A, \mathbf{v}_{-ij}, h(v_i, v_j)$, which is the desired result.

For the more general case, conditional on other restrictions in $\mathcal{E}(\mathbf{T}, \mathbf{X}, A, \mathbf{v})$, the event

$D_i = D_j = d$ is equivalent to

$$n_i d = \sum_{k:A_{ki}A_{kj}=1} \bar{g}\left(d'_k + \frac{1}{n_k}\sum_{q\neq i,j} A_{kq}Y_q, T_k, X_k, v_k\right)$$
$$d'_k := \frac{1}{n_k}(\bar{g}(d,1,x) + \bar{g}(d,0,x) + v_i + v_j)$$
$$= \frac{1}{n_k}(\bar{g}(d,1,x) + \bar{g}(d,0,x) + h(v_i, v_j))$$

As opposed to the $n_k = 2$ case, there is an additional term $\frac{1}{n_k}\sum_{q\neq i,j} A_{kq}Y_q$ which depends on the knowledge of nodes linked to neither i nor j. We want to show that it is a function of $h(v_i, v_j)$ and $\mathbf{T}, \mathbf{X}, \mathbf{v}_{-ij}$.

$$Y_k = \begin{cases} \bar{g}\left(d'_k + \frac{1}{n_k}\sum_{q\neq i,j} A_{kq}Y_q, T_k, X_k, v_k\right) & A_{ki} = A_{kj} = 1 \\ \bar{g}\left(\frac{1}{n_k}\sum_q A_{kq}Y_q, T_k, X_k, v_k\right) & A_{ki} = A_{kj} = 0 \end{cases}$$

Since $|\frac{\partial \bar{g}}{\partial D}| < \kappa < 1$, this system has a unique reduced form that depend only on d'_k and $\{X_k, T_k, v_k\}_{k\neq i,j}$ as in the proof of Proposition 3.2. This implies that $D_i = D_j = d$ can be written as a restriction on $(h(v_i + v_j), \mathbf{v}_{-ij}, \mathbf{T}, \mathbf{X})$. ■

B.3.4 Proof of Lemma 5.1

Proof. Define Λ as a diagonal matrix with entries $\Lambda_{ii} = \frac{\partial}{\partial d} g(D_i, T_i, X_i)$. I present the proof under the definition of

$$s_{ij} = \frac{1}{\min_{k:n_k>0} n_k}\|A(\iota(i) - \iota(j))\|_2$$

The result for the version of s_{ij} defined in Equation (23) follows immediately by the assumption of $\frac{\max_k n_k}{\min_k n_k} \leq C$.

In the subsequent proof, it is assumed that $s_{ij} > 0$. If $s_{ij} = 0$, we have $A_{ki} = A_{kj}$ for all k and Proposition 4.1 shows that $E[l(D_i)(v_i - v_j)] = 0$.

When $A_{ki} = A_{kj}$ for all k, the identification argument in Proposition 4.1 shows that D_i depends on v_i, v_j only through the quantity $v_i + v_j$. In other words, $\frac{\partial D_i}{\partial v_i} - \frac{\partial D_i}{\partial v_j} = 0$. To arrive at the desired result, I show in Step 1 that $|\frac{\partial D_i}{\partial v_i} - \frac{\partial D_i}{\partial v_j}| \leq \frac{s_{ij}}{1-\kappa}$. In Step 2, I show that $|E[l(D_i)(v_i - v_j)]| = O\left(\frac{s_{ij}}{1-\kappa}\right)$.

Step 1: The vector $\mathbf{Y}$ satisfies a system of nonlinear equations:

$$f_i = Y_i - g(\sum_j \tilde{A}_{ij}Y_j, T_i, X_i) - v_i = 0 \quad \forall i$$

Let J be the Jacobian matrix with ij-th entry $\frac{\partial f_i}{\partial Y_j}$. One can show that $J = I - \Lambda\tilde{A}$. J is invertible since it is diagonal-dominant by Assumption 3.2. Let $\mathcal{A}(i,j) \coloneqq \{k : A_{ki} \neq A_{kj}\}$. The l-th entry of the vector $\Lambda\tilde{A}(\iota(i) - \iota(j))$ satisfies the following inequality:

$$
\begin{aligned}
|[\Lambda\tilde{A}(\iota(i) - \iota(j))]_l| &= \left|\frac{1}{n_l}\Lambda_{ll}(A_{li} - A_{lj})\right| \\
&\leq \kappa\frac{1}{n_l}\mathbb{1}\{A_{li} \neq A_{lj}\} && (|\Lambda_{kk}| \leq \kappa \text{ by Assumption 3.2}) \\
&\leq \kappa\frac{1}{\min_k n_k}\mathbb{1}\{A_{li} \neq A_{lj}\} && (\text{definition of } s_{ij}) \\
&= \kappa s_{ij}\frac{1}{|\mathcal{A}(i,j)|}\mathbb{1}\{A_{li} \neq A_{lj}\}
\end{aligned}
$$

This implies that

$$
\begin{aligned}
|\tilde{A}_i'(\Lambda\tilde{A})^{s+1}(\iota(i) - \iota(j))| &= \tilde{A}_i'(\Lambda\tilde{A})^s[\Lambda\tilde{A}(\iota(i) - \iota(j))] \\
&\leq \kappa^s\tilde{A}_i'\tilde{A}^s|\Lambda\tilde{A}(\iota(i) - \iota(j))| \\
&\leq \kappa^s\tilde{A}_i'\mathbf{1}\sum_l [|\Lambda\tilde{A}(\iota(i) - \iota(j))|]_l \\
&\leq \kappa^s\sum_l \kappa s_{ij}\frac{1}{|\mathcal{A}(i,j)|}\mathbb{1}\{A_{li} \neq A_{lj}\} \\
&= \kappa^{s+1}s_{ij}
\end{aligned}
$$

The first inequality follows from two separate argument. First $|\tilde{A}v| \leq \tilde{A}|v|$ (element-wise comparison) for any vector v since $\tilde{A}$ has non-negative entries (the absolute value is taken with respect to each element in the vector). Second, $|\Lambda_{ii}| \leq \kappa$ by Assumption 3.2. The second inequality follows from the observation that $\tilde{A}^s$ is the s-th power of a Markov transition probability matrix $\tilde{A}$. As a result, the k-the entry of the vector $\tilde{A}v$ is bounded by $[\tilde{A}^s v]_k \leq \sum_j |v_j|$ for any k and any vector v. In vector notation, this implies $\tilde{A}^s v \leq \mathbf{1}\sum_j |v_j|$.

It follows that

$$
\begin{aligned}
\left|\frac{\partial D_i}{\partial v_i} - \frac{\partial D_i}{\partial v_j}\right| &= \left|\tilde{A}_i'\left(\frac{\partial \mathbf{Y}}{\partial v_i} - \frac{\partial \mathbf{Y}}{\partial v_j}\right)\right| \\
&= |\tilde{A}_i'J^{-1}(\iota(i) - \iota(j))| = |\tilde{A}_i'(I - \Lambda\tilde{A})^{-1}(\iota(i) - \iota(j))| \\
& \qquad\qquad\qquad\qquad\qquad\qquad (\text{Implicit Function Theorem}) \\
&= \left|\tilde{A}_i'\sum_{l=0}^{\infty}(\Lambda\tilde{A})^l(\iota(i) - \iota(j))\right| \\
&\leq \sum_{s=0}^{\infty}\kappa^s s_{ij} = \frac{s_{ij}}{1-\kappa}
\end{aligned}
$$

If $\frac{\max_k n_k}{\min_k n_k} \leq C$ for some constant C, the above bound holds by observing that $\frac{1}{\min_k n_k} \leq \frac{C}{\min\{n_i, n_j\}}$.

Step 2: One can write D_i as $D_i(v_i, v_i + v_j, \mathbf{v}_{-ij}, \mathbf{T}, \mathbf{X}, A)$. Since $v_i + v_j, \mathbf{v}_{-ij}, \mathbf{T}, \mathbf{X}, A$ will be conditioned upon, we write it as $D_i(v_i)$. Let $\bar{D} = D_i(\frac{v_i+v_j}{2})$. Denote $\mathcal{E}$ as the conditioning event $v_i + v_j = \bar{v}, \mathbf{X}, \mathbf{T}, A, \mathbf{v}_{-ij}$

$$\begin{aligned}
|E[l(D_i)(v_i - v_j)|\mathcal{E}]| &= |E[(l(D_i) - l(\bar{D}) + l(\bar{D}))(v_i - v_j)|\mathcal{E}]| \\
&= |E[(l(D_i) - l(\bar{D}))(v_i - v_j)|\mathcal{E}] + E[l(\bar{D})(v_i - v_j)|\mathcal{E}]| \\
&= |E[(l(D_i) - l(\bar{D}))(v_i - v_j)|\mathcal{E}]| \qquad \text{(Lemma B.3)} \\
&\leq E[|l(D_i) - l(\bar{D})||v_i - v_j||\mathcal{E}] \\
&\leq E\left[Lip(l)\left|\frac{\partial D_i}{\partial v_i} - \frac{\partial D_i}{\partial v_j}\right||v_i||v_i - v_j||\mathcal{E}\right] \\
&\leq Lip(l)\frac{s_{ij}}{1-\kappa}E[|v_i||v_i - v_j||\mathcal{E}]
\end{aligned}$$

The law of iterated expectation then yields the desired results:

$$\begin{aligned}
|E[l(D_i)(v_i - v_j)|A_i, A_j, X_i, X_j]| &\leq E[|E[l(D_i)(v_i - v_j)|A_i, A_j, X_i, X_j, \mathcal{E}]||A_i, A_j, X_i, X_j] \\
&= E[|E[l(D_i)(v_i - v_j)|\mathcal{E}]||A_i, A_j, X_i, X_j] \\
&\qquad \text{(The event } \mathcal{E} \text{ includes } A_i, A_j, X_i, X_j) \\
&\leq Lip(l)\frac{s_{ij}}{1-\kappa}E[E[|v_i||v_i - v_j||\mathcal{E}]|A_i, A_j, X_i, X_j] \\
&= Lip(l)\frac{s_{ij}}{1-\kappa}E[|v_i||v_i - v_j||A_i, A_j, X_i, X_j] \\
&\leq Lip(l)\frac{s_{ij}}{1-\kappa}E[v_i^2 + |v_i v_j||A_i, A_j, X_i, X_j] \\
&\qquad \text{(triangular inequality)} \\
&\leq Lip(l)\frac{s_{ij}}{1-\kappa}E[2v_i^2 + v_j^2|A_i, A_j, X_i, X_j] \\
&\qquad \text{(Cauchy Schwarz inequality)}
\end{aligned}$$

The desired result follows from the uniform boundedness of $E[v_i^2|A_i, X_i]$. ■

B.3.5 Proof of Corollary 5.2

Proof. Define $\tilde{\Lambda}$ as a diagonal matrix with entries $\tilde{\Lambda}_{ii} = \frac{1}{n_i}\frac{\partial}{\partial d}g(D_i, T_i, X_i)$. In the subsequent proof, it is assumed that $s_{ij} > 0$. If $s_{ij} = 0$, we have $A_{ki} = A_{kj}$ for all k and Proposition 4.1 shows that $E[l(D_i)(v_i - v_j)] = 0$.

To arrive at the desired result, I show in Step 1 that $|\frac{\partial D_i}{\partial v_i} - \frac{\partial D_i}{\partial v_j}| \leq \frac{s_{ij}}{1-\kappa}$. In Step 2, I show that $|E[l(D_i)(v_i - v_j)]| = O\left(\frac{s_{ij}}{1-\kappa}\right)$.

Step 1: The vector $\mathbf{Y}$ satisfies a system of nonlinear equations:

$$f_i = Y_i - g(\sum_j \tilde{A}_{ij} Y_j, T_i, X_i) - v_i = 0 \quad \forall i$$

Let J be the Jacobian matrix with ij-th entry $\frac{\partial f_i}{\partial Y_j}$. One can show that $J = I - \tilde{\Lambda} A$. J is invertible since it is diagonal-dominant by Assumption 3.2. Let $\mathcal{A}(i,j) := \{k : A_{ki} \neq A_{kj}\}$.

It follows that

$$\begin{aligned}
\left|\frac{\partial D_i}{\partial v_i} - \frac{\partial D_i}{\partial v_j}\right| &= \left|\tilde{A}_i' \left(\frac{\partial \mathbf{Y}}{\partial v_i} - \frac{\partial \mathbf{Y}}{\partial v_j}\right)\right| \\
&= |\tilde{A}_i' J^{-1}(\iota(i) - \iota(j))| = |\tilde{A}_i'(I - \tilde{\Lambda} A)^{-1}(\iota(i) - \iota(j))| \\
&\hspace{6cm} \text{(Implicit Function Theorem)} \\
&= \left|\tilde{A}_i' \sum_{l=0}^{\infty} (\tilde{\Lambda} A)^l (\iota(i) - \iota(j))\right| \\
&\leq |A_i'(\iota(i) - \iota(j))| + \sum_{l=1}^{\infty} \left|\tilde{A}_i'(\tilde{\Lambda} A)^l (\iota(i) - \iota(j))\right| \\
&\leq \frac{|A_{ii} - A_{ij}|}{n_i} + \sum_{l=1}^{\infty} \|\tilde{A}_i\|_2 \|(\tilde{\Lambda} A)^l (\iota(i) - \iota(j))\|_2 \quad \text{(Cauchy-Schwarz inequality)} \\
&\leq \frac{|A_{ii} - A_{ij}|}{n_i} + \sum_{l=1}^{\infty} \|\tilde{A}_i\|_2 \|(\tilde{\Lambda} A)^{l-1}\| \|A(\iota(i) - \iota(j))\|_2 \\
&\hspace{6cm} \text{(definition of matrix norm)} \\
&\leq \frac{|A_{ii} - A_{ij}|}{n_i} + \sum_{l=1}^{\infty} \frac{1}{\sqrt{n_i}} \|\tilde{\Lambda}\|^{l-1} \|A\|^{l-1} \|(\iota(i) - \iota(j))\|_2 \\
&\hspace{6cm} \text{(for two matrices } C, D\text{: } \|CD\| \leq \|C\|\|D\|) \\
&\leq \frac{|A_{ii} - A_{ij}|}{n_i} + \frac{1}{\sqrt{n_i}} \sum_{l=1}^{\infty} \left(\frac{\kappa}{\min_k n_k}\right)^{l-1} \|A\|^{l-1} \|A(\iota(i) - \iota(j))\|_2 \\
&\hspace{6cm} \text{(by } \|\tilde{\Lambda}\| \leq \tfrac{\kappa}{\min_k n_k}) \\
&\leq \frac{|A_{ii} - A_{ij}|}{n_i} + \frac{1}{\sqrt{n_i}} \sum_{l=1}^{\infty} \left(\frac{\kappa}{\min_k n_k}\right)^{l-1} \|A\|^{l-1} \sqrt{|\mathcal{A}(i,j)|} \\
&\hspace{6cm} \text{(by } \|A(\iota(i) - \iota(j))\|_2 \leq \sqrt{|\mathcal{A}(i,j)|}) \\
&\leq \frac{1}{n_i} + \sqrt{\frac{|\mathcal{A}(i,j)|}{n_i}} \frac{1}{1 - \|A\| \frac{\kappa}{\min_k n_k}} \\
&\leq s_{ij} + \sqrt{s_{ij}} \frac{1}{1 - \|A\| \frac{\kappa}{\min_k n_k}}
\end{aligned}$$

If $\frac{\max_k n_k}{\min_k n_k} \leq C$ for some constant C, the above bound holds by observing that $\frac{1}{\min_k n_k} \leq$

$\frac{C}{\min\{n_i,n_j\}}$.

Step 2: One can write D_i as $D_i(v_i, v_i+v_j, \mathbf{v}_{-ij}, \mathbf{T}, \mathbf{X}, A)$. Since $v_i+v_j, \mathbf{v}_{-ij}, \mathbf{T}, \mathbf{X}, A$ will be conditioned upon, we write it as $D_i(v_i)$. Let $\bar{D} = D_i(\frac{v_i+v_j}{2})$. Denote $\mathcal{E}$ as the conditioning event $v_i+v_j = \bar{v}, \mathbf{X}, \mathbf{T}, A, \mathbf{v}_{-ij}$

$$\begin{aligned}
|E[l(D_i)(v_i-v_j)|\mathcal{E}]| &= |E[(l(D_i)-l(\bar{D})+l(\bar{D}))(v_i-v_j)|\mathcal{E}]| \\
&= |E[(l(D_i)-l(\bar{D}))(v_i-v_j)|\mathcal{E}] + E[l(\bar{D})(v_i-v_j)|\mathcal{E}]| \\
&= |E[(l(D_i)-l(\bar{D}))(v_i-v_j)|\mathcal{E}]| \qquad \text{(Lemma B.3)} \\
&\leq E[|l(D_i)-l(\bar{D})||v_i-v_j||\mathcal{E}] \\
&\leq E\left[Lip(l)\left|\frac{\partial D_i}{\partial v_i} - \frac{\partial D_i}{\partial v_j}\right||v_i||v_i-v_j||\mathcal{E}\right] \\
&\leq Lip(l)\left(s_{ij} + \sqrt{s_{ij}}\frac{1}{1-\|A\|\frac{\kappa}{\min_k n_k}}\right)E[|v_i||v_i-v_j||\mathcal{E}]
\end{aligned}$$

The rest of the proof follows in exactly the same way as in Lemma 5.1. ■

B.3.6 Proof of Theorem 1

Proof. I apply Theorem 3.1 in Chen (2007) to establish consistency of the proposed estimator.

Condition 3.1: By the identification argument in Proposition 4.1,

$$q_0 := \tau_T(D_i, X_i) = \arg\min_{q\in\mathcal{Q}} L(q) \tag{45}$$

In addition, for any q such that $d(q, q_0) > \epsilon$, we have $L(q) > \epsilon$. Condition 3.1 is satisfied with $\delta(k) = 1$ and $g(\epsilon) = \epsilon$.

Condition 3.2 and 3.4 are implied by Assumption 5.6.

Condition 3.3 holds since $L(q)$ is continuous w.r.t in the L^2 norm. Condition 3.3 (ii) is implied by this continuity result since $\liminf_k \delta(k) > 0$.

Condition 3.5 (i): I apply Theorem 2.1 in Newey (1991) to establish the condition. Assumption 1 in Newey (1991) is implied by Assumption 5.6. Assumption 2 is the result of

Lemma B.7.

$$\begin{aligned}
\hat{L}_n(q;b) - \hat{L}_n(\tilde{q};b) &= \frac{2}{n}\sum_{i=1}^{n}\sum_{j\neq i}(T_i - T_j)(Y_i - Y_j)|T_i - T_j|[\tilde{q}(D_i, X_i) - q(D_i, X_i)]\omega_{ij} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}[q^2(D_i, X_i) - \tilde{q}^2(D_i, X_i)]\omega_{ij} \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\sum_{j\neq i}(T_i - T_j)(Y_i - Y_j)|T_i - T_j|\|q - \tilde{q}\|_\infty \omega_{ij} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\|q - \tilde{q}\|_\infty^2 \omega_{ij} \\
&= \frac{2}{n}\sum_{i=1}^{n}\sum_{j\neq i}[(T_i - T_j)(Y_i - Y_j) + 1]\omega_{ij}\|q - \tilde{q}\|_\infty(1 + \|q - \tilde{q}\|_\infty)
\end{aligned}$$

Let $B_n \coloneqq \frac{4}{n}\sum_{i=1}^{n}\sum_{j\neq i}[(T_i - T_j)(Y_i - Y_j) + 1]\omega_{ij}$. Assumption 3A in Newey (1991) requires that $B_n = O_p(1)$. To show this, notice that

$$\begin{aligned}
\left|\frac{4}{n}\sum_{i=1}^{n}\sum_{j\neq i}(T_i - T_j)(Y_i - Y_j)\omega_{ij}\right| &\leq 8\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}|(Y_i - Y_j)\omega_{ij}| \\
&= 8\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}|(g(D_i, T_i, X_i) - g(D_j, T_j, X_j) + v_i - v_j)\omega_{ij}| \\
&\leq 8\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}|(g(D_i, T_i, X_i) - g(D_j, T_j, X_j))\omega_{ij}| + |(v_i - v_j)\omega_{ij}| \\
&\leq 16|\bar{y}| + 8\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}|(v_i - v_j)\omega_{ij}| \\
&\qquad \text{(By boundedness in Assumption 5.1 and } \textstyle\sum_j \omega_{ij} = 1) \\
&= O_p(1) \qquad \text{(By Lemma B.6)}
\end{aligned}$$

In addition, by the norm inequality, $\|q - \tilde{q}\|_\infty \leq \|q - \tilde{q}\|_2$ since $\mathcal{Q}_k$ is a finite-dimensional space. These imply that Assumption 3A in Newey (1991) is satisfied. Equicontinuity of $L(q)$ holds by the same argument as above. Assumption 3A and Assumption 1, 2 in Newey (1991) implies the required conditions for Theorem 2.1 in the paper.

As pointed out in Chen (2007) (page 42), $\liminf_k \delta(k) > 0$ implies that Condition 3.5 (iii) is automatically satisfied and Condition 3.5 (ii) is implied by Condition 3.5 (i).

■