Heterogeneous Treatment Effects under Network Interference: A Nonparametric Approach Based on Node Connectivity

Heejong Bong, Colin B. Fogarty, Elizaveta Levina and Ji Zhu
Department of Statistics, University of Michigan

November 15, 2024

Abstract

In network settings, interference between units makes causal inference more challenging as outcomes may depend on the treatments received by others in the network. Typical estimands in network settings focus on treatment effects aggregated across individuals in the population. We propose a framework for estimating node-wise counterfactual means, allowing for more granular insights into the impact of network structure on treatment effect heterogeneity. We develop a doubly robust and non-parametric estimation procedure, KECENI (Kernel Estimation of Causal Effect under Network Interference), which offers consistency and asymptotic normality under network dependence. The utility of this method is demonstrated through an application to microfinance data, revealing the impact of network characteristics on treatment effects.

Keywords: Causal inference; Augmented inverse propensity score; Network interference; Double robustness; Non-parametric estimation

1 Introduction

We focus on the challenge of estimating causal treatment effects in network data, where an individual's outcome may depend not only on their own treatment but also on the treatments received by others in their network. In network settings, the presence of interference between units invalidates traditional approaches to causal inference in both randomized experiments and observational studies. Interference necessitates both defining new causal estimands and developing new identification strategies while allowing the potential outcomes of a given unit to depend upon the interventions received by other units in the network. There is a large literature targeting average direct effects, meant to embody the average impact of an individual's own treatment assignment on the observed response; and/or average spillover effects, meant to represent the average impact on the observed responses of units of the treatments received by *other* individuals in the network. Still others consider the comparisons of counterfactual means, averaged across units in the network, under particular choices of treatment assignment. For representative examples, see van der Laan (2014), Liu et al. (2016), Aronow & Samii (2017), Forastiere et al. (2021), Ogburn et al. (2022), Leung (2022), Li & Wager (2022), Leung & Loupos (2022), Khatami et al. (2024).

While the estimands under consideration may vary from one work to the next, a common feature is that they are defined as averages of counterfactuals taken over the entire study population. While these aggregate treatment effects offer useful insights into the overall population, the population itself is composed of units that may vary considerably in terms of their positioning and connectivity in the network under consideration. It is natural to presume that the influence of neighbors' interventions may exhibit heterogeneity based on a unit's connectivity within the network. For example, our analysis of rural Indian village microfinance data in Section 6, using our proposed method, revealed that the spillover effects of neighbors' participation in savings self-help groups were larger for units with fewer neighbors. Aggregate causal estimands provide limited insight into this difference across different node types or network structures. As a result, these methods are unable to account for heterogeneity in treatment responses based on network characteristics.

In this work, we propose an approach to account for such heterogeneity by estimating node-wise counterfactual means, rather than focusing on population averages. This framework unifies the estimation of different types of aggregated treatment effects without the need for separate estimators for each. More importantly, it allows for flexibility in capturing heterogeneity in treatment effects based on network structure. For instance, we can estimate the average treatment effect for highly connected units (high-degree nodes) and compare it to less connected units, providing novel insights into how network characteristics influence treatment effects.

Our approach offers some additional advantages beyond our choice of estimand. First, it is non-parametric, relying on a smoothness assumption for counterfactual means over the space of local intervention-connectivity configurations. We also verify appealing theoretical properties including double robustness, consistency, and asymptotic normality under network dependence. A key contribution is the development of empirical process theory for augmented inverse propensity weighting (AIPW) estimators under interference, which accommodates the interdependence between units in a network.

In Section 2, we introduce the necessary notation, define our target estimand, and specify identifying assumptions. In Section 3, we present our proposed method, KECENI, and summarize the three-step estimation procedure in Algorithm 1. We prove both consistency and asymptotic normality of our estimators in Section 4. We compare KECENI's performance to existing methods using simulated data in Section 5 and demonstrate its ability to capture heterogeneous spillover effects through an application to Indian microfinance data in Section 6. The code for the methods developed in this paper is available in our Python package `KECENI`. The code vignettes used to generate the results in Sections 5 and 6 are accessible at `github.com/HeejongBong/KECENI`.

2 Node-wise Causal Effects under Network Interference

2.1 Stable Unit Treatment on Neighborhood Value Assumption

Suppose that each unit $i \in [n]$ receives a level of an intervention represented by a binary random variable $T_i \in \{0, 1\}$ and an observed outcome $Y_i \in \mathbb{R}$. Without further restrictions on the degree of interference, for each unit there are 2^n counterfactuals $(Y_i(t_1, \ldots, t_n) : t_1, \ldots, t_n \in \{0, 1\})$ corresponding to the effects of both unit i's own intervention and the interventions on other units; the observed outcome is $Y_i(T_1, \ldots, T_n)$. However, estimating all possible counterfactuals from a single observation of $\{(T_i, Y_i) : i \in [n]\}$ is infeasible without further modeling assumptions. This necessitates imposing restrictions on the nature of the interference present in the study. We assume that additional information is available regarding interference between units, represented by a network $\mathcal{G}$ with node set $[n]$ and edge set $\mathcal{E} \subseteq [n] \times [n]$. In this work, we only consider undirected and unweighted networks.

Estimation of counterfactuals is enabled by assuming that interference occurs only between adjacent units in the network $\mathcal{G}$.

Assumption 2.1 (Network Interference). Conditional on $\mathcal{G}$, for $i \in [n]$,

$$Y_i(t_1, \ldots, t_n) = Y_i(t'_1, \ldots, t'_n) \ \text{ if } \ t_j = t'_j, \ \ \forall j \in N_i,$$

where $N_i \equiv \{i\} \cup \{j \in [n] : (i, j) \in \mathcal{E}\}$.

This assumption is a slightly relaxed version of *stable unit treatment on neighborhood value assumption* (Forastiere et al. 2021, SUTNVA), removing the necessity of summary functions in defining effective treatment, while maintaining consistency in the interference structure. It follows from Assumption 2.1 that the counterfactual $Y_i(\cdot)$ can be written as a function of the intervention variables for neighboring units. For ease of notation, we denote the collection of neighborhood interventions as t_{N_i}, where the subscript specifies the set of collected units. Using this notation, $Y_i(t_1, \ldots, t_n)$ can be simplified to $Y_i(t_{N_i})$.

2.2 Causal Estimand: Definition, Motivation, Illustration

With the counterfactual outcomes defined, we now specify our target estimand as the *node-wise counterfactual mean*. Let i^* represent the target node in network $\mathcal{G}$. The estimand is the expected value of the node's counterfactual outcome, $Y_{i^*}(t^*_{N_{i^*}})$, accounting for uncertainty in both the counterfactual and the neighborhood covariates. Formally, we define it as:

$$\theta_{i^*}(t^*_{N_{i^*}}) \equiv \mathbb{E}[Y_{i^*}(t^*_{N_{i^*}})|\mathcal{G}] = \int \mathbb{E}[Y_{i^*}(t^*_{N_{i^*}})|X_{N_{i^*}} = x_{N_{i^*}}, \mathcal{G}]\, dP_{X_{N_{i^*}}|\mathcal{G}}(x_{N_{i^*}}),$$

where $t^*_{N_{i^*}} \in \{0,1\}^{|N_{i^*}|}$ represents a hypothetical intervention assignment to the neighborhood of unit i^*. This estimand was also referred to as the *unit potential outcome expectation* by Tchetgen Tchetgen et al. (2021). We refer to this as a *node-wise* estimand to distinguish it from *unit treatment effects*, which typically refer to conditional expectations of counterfactual outcomes given specific covariates.

Node-wise average treatment effects are differences between counterfactual means under two intervention scenarios: $\theta_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^{**}_{N_{i^*}})$, where $t^*_{N_{i^*}}$ and $t^{**}_{N_{i^*}}$ are user-specified neighborhood interventions. As we highlight below, these node-wise effects may be of interest in their own right, but can further serve as building blocks for estimands both familiar and new.

Example (Node-wise Direct Effect). For node i^*, the *direct effect* when its neighbors are assigned treatment allocation $t^*_{N_{i^*}}$ is $\theta_{i^*}(t_{i^*} = 1, t_{N_{i^*}\setminus\{i^*\}} = t^*_{N_{i^*}\setminus\{i^*\}}) - \theta_{i^*}(t_{i^*} = 0, t_{N_{i^*}\setminus\{i^*\}} = t^*_{N_{i^*}\setminus\{i^*\}})$.

Example (Node-wise Indirect Effect). For node i^*, the *indirect effect* of assigning treatments $t^*_{N_i^*}$ versus $t^{**}_{N_i^*}$ to the neighbors of unit i when $t_i^* = 0$ is $\theta_{i^*}(t_{i^*} = 0, t_{N_{i^*}\setminus\{i^*\}} = t^*_{N_{i^*}\setminus\{i^*\}}) - \theta_{i^*}(t_{i^*} = 0, t_{N_{i^*}\setminus\{i^*\}} = t^{**}_{N_{i^*}\setminus\{i^*\}})$.

Example (Overall Average Effect). When comparing two hypothetical interventions t^* and t^{**}, the overall *average treatment effect* across the nodes in the network is $n^{-1}\sum_{i=1}^n \left\{\theta_i(t^*_{N_i}) - \theta_i(t^{**}_{N_i})\right\}$.

Example (Average Effects: Network-Driven Subpopulations). Treatment effects may vary for node types with varying network characteristics, such as the connectivity of a given node: highly connected individuals may be affected by an intervention differently than poorly connected individuals, particularly if spillover effects are anticipated. By averaging our estimand subpopulations defined by node-characteristics, we can investigate such heterogeneity. For instance, by averaging the treatment effects over nodes with the same degree, we can express the average treatment effect for degree-k nodes as $\frac{1}{|\mathcal{V}_k|}\sum_{i\in\mathcal{V}_k}\left\{\theta_i(t^*_{N_i}) - \theta_i(t^{**}_{N_i})\right\}$, where $\mathcal{V}_k$ denotes the set of nodes with degree k. Comparing these averages across different values of k allows us to detect how treatment effects vary with network connectivity, revealing patterns of heterogeneity driven by the local network structure.

2.3 Strong Ignorability Condition for Identifiability

In observational studies assuming no interference between units, a common expedient for identifying average treatment effects is the assumption of *strong ignorability*. Strong ignorability requres (i) $(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i | X_i$; and (ii) $0 < \mathbb{P}[T_i = 1|X_i] < 1$. Variants of this assumption are widely used in the network interference literature. For example, van der Laan (2014), Tchetgen Tchetgen et al. (2021), Ogburn et al. (2022), and Leung (2022) adopt the most restrictive version, assuming $Y_{[n]}(\cdot) \perp\!\!\!\perp T_{[n]} | X_{[n]}$, where $Y_{[n]}(\cdot)$, $T_{[n]}$, and $X_{[n]}$ represent the collections of counterfactuals, interventions, and covariates for all units. In contrast, Forastiere et al. (2021) proposed a weaker identifiability condition using an exposure mapping $g_i : t_{N_i} \to g_i(t_{N_i})$ for each unit i, where $Y_i(t_{N_i}) = Y_i(t'_{N_i})$ if $g_i(t_{N_i}) = g_i(t'_{N_i})$. Their condition implies only that the interventions are independent of each unit's potential outcomes: $Y_i(t_{N_i}) \perp\!\!\!\perp g_i(T_{N_i}) | X_i$. Khatami et al. (2024) used a slightly weaker version with the condition on X_i replaced by that on the covariates of the neighboring units, X_{N_i}. Our approach provides a similar ignorability assumption without relying on an explicit exposure mapping, resulting in an assumption that is slightly stronger than those of Forastiere et al. (2021) and Khatami et al. (2024) but allows for more flexibility in defining counterfactuals, as formalized in Assumption 2.1.

Assumption 2.2 (Strong Ignorability under Network Interference). For any node i in the observed network $\mathcal{G}$,

$$(Y_i(t_{N_i}) : t_{N_i} \in \{0,1\}^{|N_i|}) \perp\!\!\!\perp T_{N_i} | (X_{N_i}, \mathcal{G}),$$

and for any $t_{N_i} \in \{0,1\}^{|N_i|}$,

$$\mathbb{P}[T_{N_i} = t_{N_i} | X_{N_i}, \mathcal{G}] > 0, \text{ a.s.}$$

This assumption is sufficient for the identifiability of $\theta_{i^*}(t^*_{N_{i^*}})$ under network interference.

Proposition 2.3 (Identifiability on the Observed Network). *Under Assumptions 2.1 and 2.2, $\theta_{i^*}(t^*_{N_{i^*}})$ is identifiable for arbitrary user-specified neighborhood intervention vectors, $t^*_{N_{i^*}}$.*

3 KECENI: Kernel Estimate of Causal Effect under Network Interference

Although our strong ignorability condition, Assumption 2.2, allows the target estimand $\theta_{i^*}(t^*_{N_{i^*}})$ to be expressed as a functional of the observed data distribution, nonparametric estimation from a single realization of $(Y_{[n]}, T_{[n]}, X_{[n]})$ remains challenging. A key difficulty in the presence of network interference is a lack of overlap with respect to the possible treatment assignments. In short, it is rare to find units i with the exact same local network configuration and neighborhood intervention T_{N_i} as the target unit i^* and intervention $t^*_{N_{i^*}}$ when estimating $\theta(t^*_{N_{i^*}})$.

Interestingly, a parallel issue arises in the conventional observational setting assuming no interference when the treatment under consideration is continuous and the estimand is the dose-response function. As in our setting, when targeting the dose-response function one cannot generally find observed units with the same dose as that received in the counterfactual setting of interest. Kennedy et al. (2017) tackled this challenge of non-overlap by proposing a framework for estimating causal effects of continuous interventions in dose-response functions. Their key insight was to borrow strength from units with similar intervention values, defined through a dissimilarity metric, allowing the estimation of causal effects even when exact matches are unavailable. We extend this idea to the network interference setting by introducing the Kernel Estimation of Causal Effects under Network Interference (KECENI).

As in the continuous treatment setting, KECENI addresses the non-overlap issue by borrowing strength from units with similar local network configurations and interventions. Our approach aligns with Auerbach & Tabord-Meehan (2021), who introduced a dissimilarity metric for comparing local configurations and proposed a k-nearest-neighbors estimator for $\theta_{i^*}(t^*_{N_{i^*}})$ by aggregating outcomes from units with the smallest dissimilarity. Their method focuses on settings without confounding between the intervention and outcome, making it effective when the treatment is not influenced by covariates. We address the more general case where confounding may exist by incorporating covariate information into the local configuration framework.

3.1 Definition of the KECENI Estimator

Inspired by Kennedy et al. (2017), we define *pseudo-outcomes* which leverage covariate information in a doubly robust manner in the setting of network interference. Specifically, we assume $\mathbb{E}[Y_i \mid T_{N_i}, X_{N_i}, \mathcal{G}] = \mathbb{E}[Y_i \mid T_{N_i}, X_{N_i}, \mathcal{G}_{N_i}]$ and $\mathbb{P}[T_{N_i} \mid X_{N_i}, \mathcal{G}] = \mathbb{P}[T_{N_i} \mid X_{N_i}, \mathcal{G}_{N_i}]$ and define

$$\begin{aligned}\xi_i(\overline{\mu}, \overline{\pi}) \equiv\ & \frac{Y_i - \overline{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})} \int \overline{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) \\ & + \int \overline{\mu}(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) dP_{X_{N_i}|\mathcal{G}}(x_{N_i}),\end{aligned}$$

where $\overline{\mu}$ and $\overline{\pi}$ are fixed approximations of the outcome regression function $\mu(t_{N_i}, x_{N_i}, g_{N_i}) = \mathbb{E}[Y_i \mid T_{N_i} = t_{N_i}, X_{N_i} = x_{N_i}, \mathcal{G}_{N_i} = g_{N_i}]$ and propensity score $\pi(t_{N_i}|x_{N_i}, g_{N_i}) = \mathbb{P}[T_{N_i} = t_{N_i} \mid X_{N_i} = x_{N_i}, \mathcal{G}_{N_i} = g_{N_i}]$, and $\mathcal{G}_{N_i}$ denotes the subnetwork induced by the node set N_i. These pseudeo-outcomes inherit the double robustness proved in from Kennedy et al. (2017): $\xi_i(\overline{\mu}, \overline{\pi})$ is unbiased for $\theta_i(t_{N_i})$ if either the outcome regression approximation or the proposensity score approximation are correct. The proof is in Appendix B.3.

Proposition 3.1. *If either $\overline{\mu} = \mu$ or $\overline{\pi} = \pi$, for any node i and neighborhood intervention t_{N_i},*

$$\mathbb{E}[\xi_i(\overline{\mu}, \overline{\pi}) | T_{N_i} = t_{N_i}, \mathcal{G}] = \theta_i(t_{N_i}).$$

We estimate $\theta_{i^*}(t^*_{N_{i^*}})$, the counterfactual mean of the target node i^* under assignment $t^*_{N_{i^*}}$, using kernel smoothing of the observed pseudo-outcomes ξ_i. This smoothing requires a user-specified dissimilarity metric Δ_i, which measures the difference between the connectivity characteristics of node i in network $\mathcal{G}$ and its neighborhood intervention assignment T_{N_i}, and those of node i^* under the intervention $t^*_{N_{i^*}}$. The smoothing is based on a kernel function $\kappa_\lambda(x)$, where $\lambda > 0$ is a bandwidth hyper-parameter that controls the degree of smoothing and is calibrated by cross-validation (see Appendix A.2 for details). This way, KECENI effectively addresses the non-overlap issue by borrowing information from units with close but not identical configurations to i^*, while adjusting for confounders through a doubly robust framework. The estimation procedure is summarized in Algorithm 1.

Algorithm 1 KECENI Estimation

1. **Estimate nuisance functions:** The estimates of these nuisance functions, denoted by $\widehat{\mu}$, $\widehat{\pi}$, and $\widehat{P}_{X_{[n]}|\mathcal{G}}$, are obtained through appropriate machine learning or non-parametric methods.

2. **Construct pseudo-outcomes:** We obtain the empirical pseudo-outcomes $\widehat{\xi}_i(\widehat{\mu}, \widehat{\pi})$, defined by substituting the estimated distribution $\widehat{P}_{X_{N_i}|\mathcal{G}}$ in place of $P_{X_{N_i}|\mathcal{G}}$ in the original definition of $\xi_i(\widehat{\mu}, \widehat{\pi})$.

3. **Apply kernel smoothing:** The estimator for $\theta_{i^*}(t^*_{N_{i^*}})$ is given by

$$\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) \equiv \widehat{D}^{-1} \sum_{i=1}^{n} \widehat{\xi}_i(\widehat{\mu}, \widehat{\pi}) \cdot \kappa_\lambda(\Delta_i), \tag{1}$$

   where $\widehat{D} \equiv \sum_{i=1}^{n} \kappa_\lambda(\Delta_i)$ normalizes the kernel weights to ensure they sum to one.

Remark 3.2. In some cases, it may be simpler to work with propensity scores conditional on covariates and the connectivity of two-hop neighborhoods, denoted by $N_i^{(2)} \equiv \cup_{j \in N_i} N_j$.

For instance, this applies when the intervention assignment is independent across units, conditional on the covariates and connectivity of adjacent units. In such cases, the propensity score conditional on two-hop neighborhoods is given by:

$$\pi(T_{N_i}|X_{N_i^{(2)}}, \mathcal{G}_{N_i^{(2)}}) = \prod_{j \in N_i} \pi^\circ(T_j|X_{N_j}, \mathcal{G}_{N_j}),$$

where π° represents the unit-level propensity score. The pseudo-outcome can then be defined using covariates and connectivity from two-hop neighborhoods. The theoretical properties we establish, including double robustness, hold regardless of how we condition on the network structure.

4 Consistency and Asymptotic Normality

In this section, we provide theoretical guarantees regarding the asymptotic properties and accuracy of the KECENI estimates under a set of regularity conditions.

4.1 Regularity Conditions

First, we adopt the *neighborhood dependence* assumption, originally introduced by Stein (1972). This assumption restricts the dependencies within a network by assuming that nonadjacent units are independent of each other. Under this assumption, we can analyze each unit's outcome in isolation from non-neighboring units, as if it were a separate experiment.

Assumption 4.1 (Neighborhood Dependence). Conditional on $\mathcal{G}$,

$$X_i \perp\!\!\!\perp X_{N_i^{\complement}}, \text{ and } (T_i, Y_i(\cdot)) \perp\!\!\!\perp (X_{N_i^{\complement}}, T_{N_i^{\complement}}, Y_{N_i^{\complement}}(\cdot))|X_{N_i}, \ \forall i,$$

where $N_i \equiv \{i\} \cup \{j \in [n] : (i, j) \in \mathcal{E}\}$ and $N_i^{\complement} = [n] \setminus N_i$.

Under the neighborhood dependence assumption, we develop a novel empirical process theory that is essential to justify using the same data for both nuisance parameter estimation and causal estimand estimation. To allow this, we impose a restriction on the size of the parameter space, which in our case, corresponds to the functional spaces of $\widehat{\mu}$ and $\widehat{\pi}$, quantified by *entropy numbers* (van der Vaart & Wellner 1996, p.83). Similar assumptions have been employed by Kennedy et al. (2017) and van der Laan (2014), arguing that this restriction is relatively mild. Additionally, this assumption can be further relaxed by employing sample-splitting techniques, as discussed in van der Laan et al. (2011, Chapter 27). We note that the factor $2^{|N_i|}$ in Eq. (2) is used to adjust increasing number of ways to allocate intervention to the unit i's neighborhood.

Assumption 4.2 (Finite Entropy Condition). The estimators $\widehat{\mu}$, $\widehat{\pi}$ and their limits $\overline{\mu}$, $\overline{\pi}$ are contained in uniformly bounded functional classes with finite uniform entropy integrals. In particular, the functional space $\mathcal{F}_\mu$ of $\widehat{\mu}$ and $\overline{\mu}$ is equipped with metric ℓ_∞-norm, and $\mathcal{F}_\pi$ of $\widehat{\pi}$ and $\overline{\pi}$ is equipped with metric

$$\|\pi\|_* \equiv \max_{i\in[n]} 2^{|N_i|} \cdot \|\pi(\cdot|\cdot, \mathcal{G}_{N_i})\|_\infty. \tag{2}$$

In addition, we require the following uniform boundedness assumptions to control the uncertainty of each pseudo-outcome ξ_i. The first is the bounded outcome assumption, which was also implicitly used in Kennedy et al. (2017) and van der Laan (2014). The second is a uniform version of the positivity assumption, commonly imposed in the causal inference literature, particularly in the context of inverse probability weighting and its extensions.

Assumption 4.3 (Uniform Boundedness). The outcome random variables Y_i are uniformly bounded. Additionally, both the estimator $\widehat{\pi}$ and its limit $\overline{\pi}$ have uniformly bounded inverses. Specifically, there exists a constant $M > 0$ such that

$$\max_{i\in[n]} \left\| \frac{1}{2^{|N_i|}\widehat{\pi}(\cdot \mid \cdot, \mathcal{G}_{N_i})} \right\|_\infty < M \quad \text{and} \quad \max_{i\in[n]} \left\| \frac{1}{2^{|N_i|}\overline{\pi}(\cdot \mid \cdot, \mathcal{G}_{N_i})} \right\|_\infty < M.$$

In Appendix B.4, we develop the novel empirical process theory we use to prove the consistency and asymptotic normality of KECENI estimates. This theory is broadly applicable to any empirical process defined as a weighted average of neighborhood-dependent functionals.

We assume that the dissimilarity metric Δ_i, defined as a function of $(i, i^*, T_{N_i}, t^*_{N_{i^*}}, \mathcal{G})$, appropriately captures the closeness between $\theta_i(T_{N_i})$ and $\theta_{i^*}(t^*_{N_{i^*}})$. This assumption, often referred to as the *smoothness condition*, is fundamental in non-parametric estimation. For instance, Auerbach & Tabord-Meehan (2021) introduced the concept of ϵ-*isomorphism*, which ensures that counterfactual means or quantiles vary smoothly with small changes in local configurations. This smoothness condition is essential for ensuring the effectiveness of their k-nearest neighbor estimator.

Assumption 4.4 (Lipschitz Continuity of Node-Wise Counterfactual Means)*.* There exists a constant $L > 0$ such that

$$|\theta_i(T_{N_i}) - \theta_{i^*}(t^*_{N_{i^*}})| \leq L \cdot \Delta_i, \tag{3}$$

for any units i.

Some may argue that imposing smoothness on the marginal expectation $\theta_i(t_{N_i}) = \mathbb{E}[Y_i(t_{N_i})|\mathcal{G}]$ is less practical than applying it to the conditional expectation $\mathbb{E}[Y_i(t_{N_i})|X_{N_i}, \mathcal{G}]$, as the latter accounts for covariates' influence on outcomes. In Appendix A.1, we demonstrate how marginal smoothness can be derived from conditional smoothness using appropriately chosen dissimilarity metrics.

Finally, we introduce the following condition to ensure the convergence of the covariate distribution estimator, $\widehat{P}_{X_{[n]}|\mathcal{G}}$ to the true distribution.

Assumption 4.5 (Uniform Convergence and Equicontinuity of the Covariate Distribution Estimator)*.* For each node i, let $\mathcal{F}_i$ be a functional space on $\mathbb{R}^{|N_i| \times p}$ with the ℓ_∞-norm. We assume there exist a sequence q_n converging to 0 and a totally bounded subset $\mathcal{F}$ of the direct sum space $\oplus_{i=1}^{\infty} \mathcal{F}_i$ (where each element has finitely many non-zero components) with the norm $\|(f_1, \ldots, f_n)\| = \sum_{i=1}^n \|f_i\|_\infty$ such that:

- $\widehat{P}_{X_{[n]}|\mathcal{G}}$ satisfies uniform convergence on $\mathcal{F}$:

$$\sup_{(f_1,\ldots,f_n)\in\mathcal{F}} \left| \sum_{i=1}^n \int f_i(x_{N_i})(dP_{X_{N_i}|\mathcal{G}} - d\widehat{P}_{X_{N_i}|\mathcal{G}})(x_{N_i}) \right| = O_p(q_n),$$

- $\widehat{P}_{X_{[n]}|\mathcal{G}}$ satisfies equicontinuity on $\mathcal{F}$:

$$\sup_{\|(f_1,\ldots,f_n)-(g_1,\ldots,g_n)\|<\delta_n} \left| \sum_{i=1}^n \int (f_i - g_i)(x_{N_i})(dP_{X_{N_i}|\mathcal{G}} - d\widehat{P}_{X_{N_i}|\mathcal{G}})(x_{N_i}) \right| = o_p(q_n),$$

  as $\delta_n \to 0$, and

- $\mathcal{F}$ contains all nuisance parameter estimates and their limits:

$$(\overline{\pi}(t_{N_i}|\cdot, \mathcal{G}_{N_i}) : i \in [n]) \in \mathcal{F}, \text{ and } (\overline{\mu}(t_{N_i}, \cdot, \mathcal{G}_{N_i}) : i \in [n]) \in \mathcal{F},$$

$$(\widehat{\pi}(t_{N_i}|\cdot, \mathcal{G}_{N_i}) : i \in [n]) \in \mathcal{F}, \text{ and } (\widehat{\mu}(t_{N_i}, \cdot, \mathcal{G}_{N_i}) : i \in [n]) \in \mathcal{F},$$

  for all $t_{[n]} \in \{0, 1\}^n$, almost surely.

In the absence of interference, the covariate distribution is often estimated by the empirical distribution of i.i.d. covariates X_i, reducing this condition to the standard Glivenko-Cantelli and Donsker properties of the two nuisance parameter spaces. The finite entropy condition of these spaces (Assumption 4.2) automatically guarantees these properties with $q_n = 1/\sqrt{n}$. This guarantee extends to the network interference setting when covariate vectors are i.i.d. and $\widehat{P}_{X_{[n]}|\mathcal{G}}$ is given by the product measure of the empirical distribution. We denote this *empirical product measure* as

$$\widehat{P}^{\otimes}_{X_{[n]}|\mathcal{G}} \equiv \otimes_{i=1}^n \widehat{P}_{X_i|\mathcal{G}}, \tag{4}$$

where $\widehat{P}_{X_i|\mathcal{G}}$ is the empirical measure of the observed covariates $\{X_i : i = 1, \ldots, n\}$.

$$Y_i \longleftarrow T_{N_i} \longleftarrow X_{N_i^{(2)}} \;\text{———}\; X_{N_j^{(2)}} \longrightarrow T_{N_j} \longrightarrow Y_j$$

Figure 1: Dependence of (Y_i, T_{N_i}, X_{N_i}) and (Y_j, T_{N_j}, X_{N_j}) between nodes i and j within 5 hops. The directed connections indicate structural relationships, while the undirected connection indicates correlation.

4.2 Consistency of the KECENI Estimator

Let $D \equiv \mathbb{E}[\sum_{i=1}^{n} \kappa_\lambda(\Delta_i) | \mathcal{G}]$ and $K_{\max} \equiv \max_{i \in [n]} K_i$, where K_i is the size of the node's 5-hop neighborhood $N_i^{(5)}$. The nodes j in $N_i^{(5)}$ satisfy $(Y_i, T_{N_i}, X_{N_i}) \not\perp\!\!\!\perp (Y_j, T_{N_j}, X_{N_j})$ as depicted in Fig. 1. K_i indicates the number of nodes with such dependency with node i.

Let $D \equiv \mathbb{E}\left[\sum_{i=1}^{n} \kappa_\lambda(\Delta_i) \mid \mathcal{G}\right]$ and define $K_{\max} \equiv \max_{i \in [n]} K_i$, where K_i represents the size of the node's 5-hop neighborhood, $N_i^{(5)}$. Nodes j in $N_i^{(5)}$ are those for which the variables (Y_i, T_{N_i}, X_{N_i}) and (Y_j, T_{N_j}, X_{N_j}) exhibit dependency, as shown in Fig. 1. The value K_i therefore reflects the number of nodes that have a dependency relationship with node i. The following theorem establishes consistency of estimated node-wise counterfactual means using KECENI.

Theorem 4.6. *Suppose that the network and hyperparameter are fixed with $\mathcal{G} = \mathcal{G}_n$ and $\lambda = \lambda_n$ at each $n \in \mathbb{N}$, respectively, and that κ has support $[-1, 1]$. Note that $K_{\max} = K_{\max,n}$ is also a function of n. Let $\overline{\mu}$ and $\overline{\pi}$ be functions to which $\widehat{\mu}$ and $\widehat{\pi}$ converge in the sense that $\|\widehat{\mu} - \overline{\mu}\|_\infty = o_p(1)$ and $\|\widehat{\pi} - \overline{\pi}\|_* = o_p(1)$. Suppose that $K_{\max} = O(\sqrt{n})$ and λ decreases with n such that $D = \Omega(\sqrt{n})$. Then, under the aforementioned assumptions, we have*

$$|\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}})| = O_p\left\{\sqrt{\frac{K_{\max}}{D}} + \lambda + \widetilde{r}_n(t^*_{N_{i^*}})\widetilde{s}_n(t^*_{N_{i^*}}) + q_n\right\}, \tag{5}$$

where

$$\widetilde{r}_n(t^*_{N_{i^*}}) \equiv \left\{\sup_{(i, t_{N_i}) : \Delta_i(t_{N_i}) \le \lambda} r(t_{N_i})\right\}, \; \widetilde{s}_n(t^*_{N_{i^*}}) \equiv \left\{\sup_{(i, t_{N_i}) : \Delta_i(t_{N_i}) \le \lambda} s(t_{N_i})\right\},$$

$$r(t_{N_i}) \equiv \left\| 2^{|N_i|}(\pi - \widehat{\pi})(t_{N_i} | X_{N_i}, \mathcal{G}_{N_i}) \right\|_2, \text{ and } s(t_{N_i}) \equiv \|(\mu - \widehat{\mu})(t_{N_i}, X_{N_i}, \mathcal{G}_{N_i})\|_2,$$

and $\Delta_i(t_{N_i})$ *is the dissimilarity metric of node* i *when* t_{N_i} *replaced* T_{N_i}*.*

The convergence rate in Eq. (5) consists of four components. The first two terms, $\sqrt{\frac{K_{\max}}{D}}$ and λ, correspond to the variance and bias errors, respectively. The impact of λ on the variance error term is implicit. As λ increases, D also increases, which in turn reduces the variance error. This aligns with the typical bias-variance tradeoff in kernel smoothing, where increasing λ allows for borrowing information from more dissimilar units, thereby reducing variance at the cost of increased bias. The third term represents the product of the local convergence rates of the nuisance estimates $\widehat{\mu}$ and $\widehat{\pi}$ towards their true values, μ and π. This term reflects the *rate double robustness* (Jiang et al. 2022) of the KECENI estimator, where the overall convergence rate can still reach the oracle rate—achievable when the nuisance parameters are known—despite the individual convergence rates of the nuisance estimators being slower, as long as their product is sufficiently sharp. The fourth term captures the convergence rate of the covariate distribution estimator. Unfortunately, the estimation error for the covariate distribution does not benefit from double robustness: for consistent estimation, the covariate distribution model must be correctly specified.

4.3 Asymptotic Normality of the KECENI Estimator

We next establish asymptotic normality of the KECENI estimator. Because our estimator builds on kernel smoothing estimator, the smoothing bias is comparable to the variance term even asymptotically. For this reason, we establish the asymptotic normality centered on the bias mean. First, we define the smoothing bias b_n and variance σ_n^2 by

$$\begin{aligned} & b_n \equiv \widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}}), \quad \sigma_n^2 \equiv \mathrm{Var}\left[D^{-1} \textstyle\sum_{i=1}^n \psi_i \middle| \mathcal{G} \right], \\ & \psi_i = \kappa_\lambda(\Delta_i)\ (\xi_i(\overline{\mu}, \overline{\pi}) - \widetilde{\theta}_{i^*}(t^*_{N_{i^*}})) + \mathbb{E}[\kappa_\lambda(\Delta_i)\ (\widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_i(T_{N_i}))|\mathcal{G}], \end{aligned} \tag{6}$$

where $\widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) \equiv D^{-1}\mathbb{E}[\sum_{i=1}^n \kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})|\mathcal{G}]$. Recall that $D \equiv \mathbb{E}[\sum_{i=1}^n \kappa_\lambda(\Delta_i)|\mathcal{G}]$. The following theorem establishes asymptotic normality of our KECENI-based estimators after de-biasing.

Theorem 4.7. *Suppose that* $\max\{\widetilde{r}_n(t^*_{N_{i^*}})\widetilde{s}_n(t^*_{N_{i^*}}), q_n, \frac{n^{1/3}K_{\max}^{2/3}}{D}\} = o_p(\sigma_n)$. *Then, under the same set of assumptions as in Theorem 4.6, we have*

$$\frac{1}{\sigma_n}(\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}}) - b_n) \xrightarrow{d} N(0,1).$$

The additional assumption controls the behavior of the nuisance parameters such that their impact becomes negligible in the asymptotic regime, and that certain moment conditions required for the central limit theorem are satisfied. This allows us to establish the asymptotic normality of the KECENI estimator.

Under the conditions of Theorem 4.7, the uncertainty in $\widehat{\theta}_{i^*}(t^*_{N_{i^*}})$ can be assessed by the variance of $\sum_{i=1}^n \widehat{\psi}_i$ scaled with $\widehat{D}^{-2}$, where $\widehat{\psi}_i$ is the plug-in estimator:

$$\begin{aligned}\widehat{\psi}_i \equiv\ & \kappa_\lambda(\Delta_i)\ (\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \widehat{\theta}_{i^*}(t^*_{N_{i^*}})) + \textstyle\sum_{t_{N_i}\in\{0,1\}^{|N_i|}} \kappa_\lambda(\Delta_i(t_{N_i}))\widehat{\varpi}(t_{N_i}|\mathcal{G})\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) \\ & - \textstyle\sum_{t_{N_i}\in\{0,1\}^{|N_i|}} \kappa_\lambda(\Delta_i(t_{N_i}))\widehat{\varpi}(t_{N_i}|\mathcal{G}) \displaystyle\int \widehat{\mu}(t_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i}),\end{aligned}$$

where $\Delta_i(t_{N_i})$ is the dissimilarity metric of node i when t_{N_i} replaced T_{N_i}, and $\widehat{\varpi}(t_{N_i}|\mathcal{G}) \equiv \int \widehat{\pi}(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i})$. To account for the network dependencies among $\widehat{\psi}_i$ in the variance estimation, we propose using the network *heteroskedasticity and autocorrelation consistent* (HAC) estimator (Kojevnikov et al. 2021). Alternatively, for better representation of higher-order moments in finite samples, the *block-based bootstrap* method (Kojevnikov 2021) can be applied to the $\widehat{\psi}_i$'s.

5 Simulation Study

5.1 Node-wise Counterfactual Means

First, we evaluate our proposed method for estimating node-wise counterfactual means through simulations. Since no competing methods exist for this specific task, we focus on demonstrating the theoretical properties outlined in Section 4, including consistency and asymptotic normality.

Setting. The simulation consisted of 80 repeated runs, each using the same network $\mathcal{G}$ to represent relationships among $n = 1000$ units. We sampled $\mathcal{G}$ once from a latent variable network model (Rastelli et al. 2016). In these models, the adjacency probability of each node pair is determined by the positions of the two nodes in a latent space. A common approach is to define the adjacency probability as a function of the distance between the latent positions. In this simulation, we set the latent positions Z_i to be i.i.d. samples from the Uniform distribution on $[-1, 1]^2$ and modeled $\mathcal{G}$ as a simple undirected network where the adjacency of each node pair was independent with probability

$$\mathbb{P}[(i, j) \in \mathcal{E}] = \mathbf{1}\{i \neq j\}\rho \exp\left(-e^{\beta \cdot \|Z_i - Z_j\|_\infty}\right). \tag{7}$$

Here ρ and β are parameters that govern the network structure. Specifically, a larger ρ results in a more densely connected network, and a larger β increases the proportion of connections between nodes that are closer in the latent space, relative to those that are farther apart. We set $\rho = 2$ and $\beta = 10$. Appendix C.1 provides supplementary figures illustrating the basic properties of the network.

With the network fixed, we sampled covariate feature vectors $X_i \in \mathbb{R}^3$, treatment variables $T_i \in \{0, 1\}$, and outcome variables $Y_i \in \mathbb{R}$ for nodes $i \in \{1, \ldots, n\}$. The covariate features were sampled from a multivariate Gaussian distribution with mean zero and an identity covariance matrix, independently across the nodes. Given the covariates, treatments were randomly assigned as $T_i \sim \text{Bernoulli}\left(\text{expit}(\langle X_i, (0.5, 0.5, 0.5)\rangle)\right)$ independently across i, where expit is the logistic sigmoid function: $\text{expit}(x) \equiv \frac{e^x}{1+e^x}$. Finally, we defined the outcomes as

$$
\begin{aligned}
Y_i = {} & 2 \cdot (T_i - 0.5) + 2 \cdot \text{Avg}(T_{N_i \setminus \{i\}} - 0.5) \\
& + \langle X_i, (-1.55, -1.55, -1.55)\rangle + \left\langle \text{Avg}(X_{N_i \setminus \{i\}}), (-1.55, -1.55, -1.55)\right\rangle + \epsilon_i,
\end{aligned}
\tag{8}
$$

where for node i, N_i is the neighborhood in $\mathcal{G}$, $\text{Avg}(T_{N_i \setminus \{i\}} - 0.5) \equiv \mathbf{1}\{|N_i| > 1\}(\frac{1}{|N_i|-1} \sum_{j \in N_i \setminus \{i\}} T_j - 0.5)$, $\text{Avg}(X_{N_i \setminus \{i\}}) \equiv \mathbf{1}\{|N_i| > 1\}(\frac{1}{|N_i|-1} \sum_{j \in N_i \setminus \{i\}} X_j)$, and ϵ_i is an independent standard Gaussian random error. The treatment assignment and outcome were designed so that $\mathbb{E}[Y_i | T_i = 1] \approx \mathbb{E}[Y_i | T_i = 0]$. In order to recover the correct treatment effect, one should properly account for the confounding by the covariate vector.

We selected a specific unit i^* from the observed units $\{1, \ldots, n\}$. We applied KECENI to estimate the target unit's expected outcome $\theta_{i^*}(t^*_{N_{i^*}})$ under a given treatment allocation scenario $t^*_{N_{i^*}} \in \{0,1\}^{|N_{i^*}|}$. We considered two treatment scenarios for $t^*_{N_{i^*}}$, denoted by $t^{(0)}_{N_{i^*}}$ and $t^{(1)}_{N_{i^*}}$. In the former, no units were treated (i.e., $t^{(0)}_i = 0, \forall i \in N_{i^*}$), while in the latter, all units were treated (i.e., $t^{(1)}_i = 1, \forall i \in N_{i^*}$). The selected unit i^* had the smallest ℓ_∞ norm of the latent position and two adjacent units. According to Eq. (8), the expected outcomes for these scenarios were $\theta_{i^*}(t^{(0)}_{N_{i^*}}) = -2$ and $\theta_{i^*}(t^{(1)}_{N_{i^*}}) = 2$, respectively.

Estimation Detail. Applying KECENI, we used propensity scores conditional on covariates and the connectivity of two-hop neighborhoods as described in Remark 3.2. For the dissimilarity metric, we used $\Delta_i \equiv \left\| (T_i, \text{Avg}(T_{N_i \setminus \{i\}} - 0.5)) - (t^*_{i^*}, \text{Avg}(t^*_{N_{i^*} \setminus \{i^*\}} - 0.5)) \right\|_1$. We used logistic regression for the propensity score model and linear regression for the outcome regression model, incorportating two-hop neighborhood covariates in both models as we described in Remark 3.2. Specifically, we fit:

$$\mathbb{P}[T_i = 1 | X_{N_i} = x_{N_i}, \mathcal{G}_{N_i}] \sim \text{LogisticReg}(x_i, \text{Avg}(x_{N_i \setminus \{i\}})),$$

$$\mathbb{E}[Y_i | T_{N_i} = t_{N_i}, X_{N_i^{(2)}} = x_{N_i^{(2)}}, \mathcal{G}_{N_i^{(2)}}] \sim \text{LinearReg}(t_i, \text{Avg}(t_{N_i \setminus \{i\}} - 0.5), x_i, \text{Avg}(x_{N_i \setminus \{i\}})), \tag{9}$$

where $N_i^{(2)}$ is the two-hop neighborhood of node i in $\mathcal{G}$. Using these models, we obtained $\widehat{\pi}^{\circ}(x_{N_i})$ and $\widehat{\mu}(t_{N_i}, x_{N_i^{(2)}})$ as the respective estimates. For the propensity score, we defined $\widehat{\pi}(t_{N_i} | x_{N_i^{(2)}}, \mathcal{G}_{N_i^{(2)}}) \equiv \prod_{j \in N_i} \{(1 - \widehat{\pi}^{\circ}(x_{N_j}))\mathbf{1}(t_j = 0) + \widehat{\pi}^{\circ}(x_{N_j})\mathbf{1}(t_j = 1)\}$ and used $\widehat{\mu}(t_{N_i}, x_{N_i^{(2)}}, \mathcal{G}_{N_i^{(2)}})$ for the outcome regression. The covariate distribution was modeled as the product empirical measure $\widehat{P}^{\otimes}_{X_{[n]}|\mathcal{G}}$ in Eq. (4).

Results. Fig. 2 presents the estimation and inference results from applying KECENI to the 80 simulated datasets. Fig. 2a shows the distribution of $\widehat{\theta}_{i^*}(t^{(1)}_{N_{i^*}})$ in light blue and $\widehat{\theta}_{i^*}(t^{(0)}_{N_{i^*}})$ in light red, with the true values indicated by solid blue and red lines, respectively. Both distributions are close to the true values, but there is a slight bias towards each other. This is the smoothing bias b_n as specified in Eq. (6). The bias is emphasized as the target treatments represent extreme cases ("all treated" vs. "no treated"). Consequently, the histograms show smoothing biases toward the center, corresponding to the overall average outcome in the observed data. This leads to an estimated difference that tends to be smaller than the true difference, as seen in the gray histogram of the difference estimate in Fig. 2b, with the true difference indicated by the solid black line. In Appendix C.3, we show that the bias reduces as the sample size increases, demonstrating the consistency of Theorem 4.6. The normal Q-Q plots in Figs. 2c and 2d demonstrate that the empirical order statistics of the estimates align well with the normal quantiles, supporting the asymptotic normality presented in Theorem 4.7.

Supplementary simulation results on estimation bias when network interference is ignored is available in Appendix C.2.

5.2 Demonstration of Double Robustness

In this subsection, we demonstrate the double robustness of KECENI using a simulation study, primarily comparing it to the G-computation algorithm (Robins 1986). Given $\widehat{\mu}$ and $\widehat{P}_{X_{[n]}|\mathcal{G}}$ as defined in Section 3, we define the G-computation estimator as

$$\widehat{\theta}_{i^*}^{G}(t_{N_{i^*}}^*) \equiv \int \widehat{\mu}(t_{N_{i^*}}^*, x_{N_{i^*}^{(2)}}, \mathcal{G}_{N_{i^*}^{(2)}}) d\widehat{P}_{X_{N_{i^*}^{(2)}}|\mathcal{G}}(x_{N_{i^*}^{(2)}}).$$

Similar to the no-interference setting, this estimator is vulnerable to misspecification of the outcome regression model. In contrast, KECENI posesses a double robustness property: it provides correct treatment effect estimates even if one of the outcome regression model or the propensity score model is misspecified, as long as the other is correctly specified. The following simulation demonstrates this double robustness in comparison with the G-computation.

Setting. We used the same 80 datasets with $n = 1000$ units as those in the previous subsection. This time the objective was to estimate the direct treatment effect on a target unit. Specifically, we chose a target unit i^* with four adjacent units and set $t_{N_{i^*}}^{(0)}$ so that $t_{i^*}^{(0)} = 0$ and $|N_{i^*}|^{-1} \sum_{i \in N_{i^*}} T_i = 0.5$. The second treatment scenario $t_{N_{i^*}}^{(1)}$ had the same treatment assignment as $t_{N_{i^*}}^{(0)}$ except that $t_{i^*}^{(1)} = 1$. The expected outcomes for these scenarios were $\theta_{i^*}(t_{N_{i^*}}^{(0)}) = -1$ and $\theta_{i^*}(t_{N_{i^*}}^{(1)}) = 1$, respectively.

Estimation Details. We investigated the sensitivity of the G-computation and KECENI to model misspecification. We quantified the extent of misspecification in the propensity

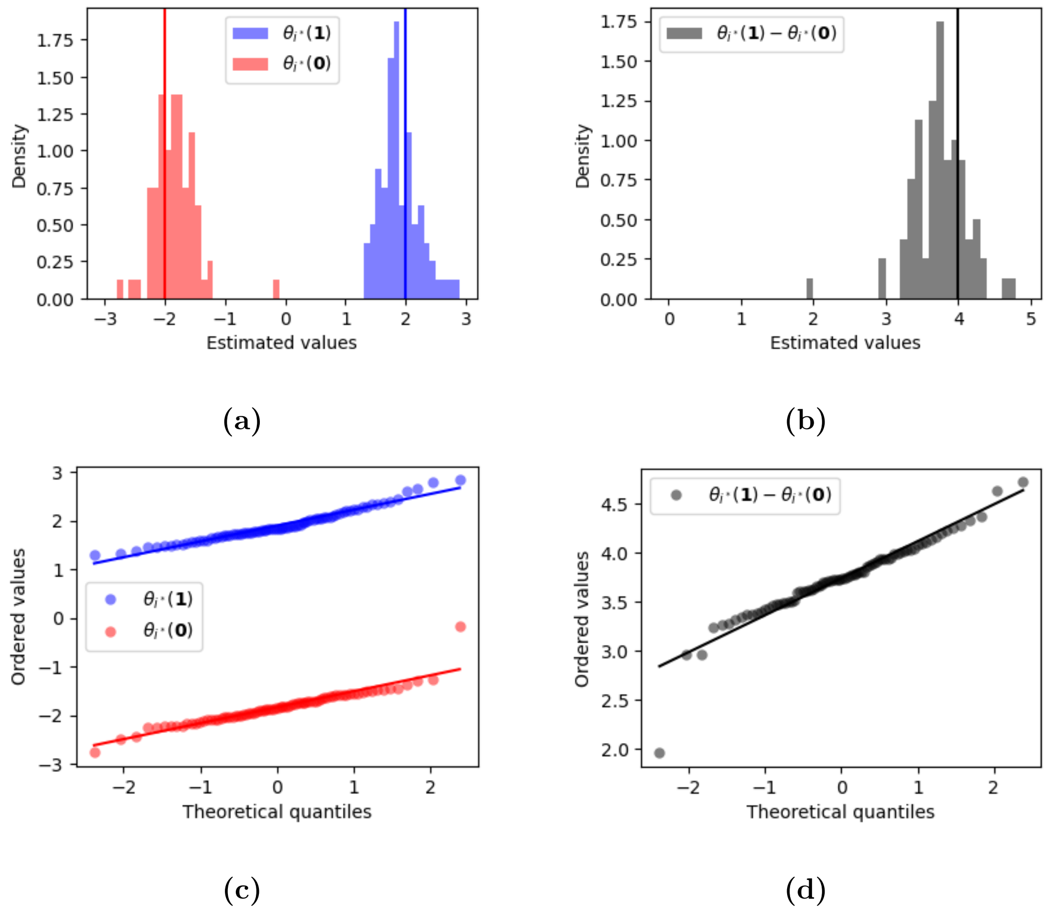
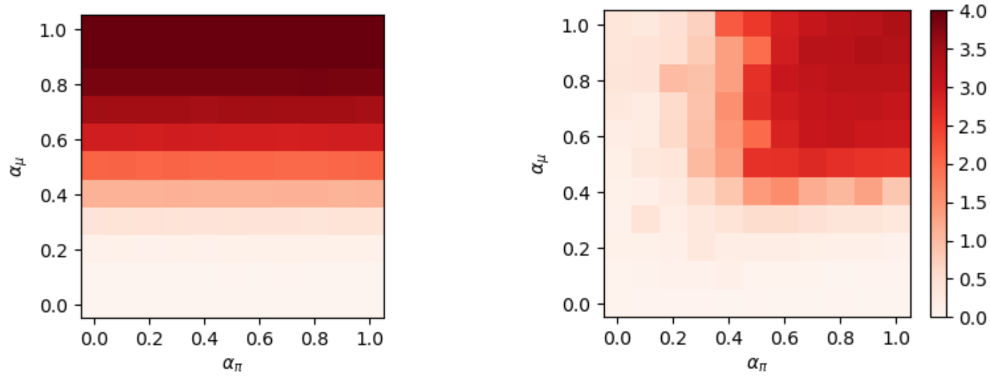


Figure 2: Histograms of (a) the KECENI estimates for the node-wise counterfactual means under two treatment scenarios and (b) the estimated treatment effects. (c,d) The normal Q-Q plots of the KECENI estimates and estimated treatment effects. The blue histogram and Q-Q plot correspond to the distribution of $\widehat{\theta}_{i*}(t^{(1)}_{N_{i*}})$ (all treated scenario), where the red ones are of $\widehat{\theta}_{i*}(t^{(0)}_{N_{i*}})$ (no treated scenario). The grey ones correspond to the estimated treatment effects, $\widehat{\theta}_{i*}(t^{(1)}_{N_{i*}}) - \widehat{\theta}_{i*}(t^{(0)}_{N_{i*}})$.



(a) **(b)**

Figure 3: Root-mean-square error of (a) G estimate and of (b) KECENI.

score and outcome regression models by α_π and α_μ, respectively, using the following models:

$$\mathbb{P}[T_i = 1 | X_{N_i} = x_{N_i}]$$

$$\sim \text{LogisticReg}((1-\alpha_\pi)x_i + \alpha_\pi x_i^2, \text{Avg}((1-\alpha_\pi)x_{N_i\setminus\{i\}} + \alpha_\pi x^2_{N_i\setminus\{i\}})),$$

$$\mathbb{E}[Y_i | T_{N_i} = t_{N_i}, X_{N_i^{(2)}} = x_{N_i^{(2)}}]$$

$$\sim \text{LinearReg}\begin{pmatrix} t_i, \text{Avg}(t_{N_i\setminus\{i\}} - 0.5), \\ (1-\alpha_\mu)x_i + \alpha_\mu x_i^2, \text{Avg}((1-\alpha_\mu)x_{N_i\setminus\{i\}} + \alpha_\mu x^2_{N_i\setminus\{i\}}) \end{pmatrix}.$$

These models are correctly specified when $\alpha_\pi = \alpha_\mu = 0$ and mostly misspecified when $\alpha_\pi = \alpha_\mu = 1$. We quantified the accuracy of G-computation and KECENI by RMSE as (α_π, α_μ) varied over $\{0, 0.1, \ldots, 0.9, 1\}^2$.

Results. Fig. 3 presents heatmaps of RMSEs across the 11×11 (α_π, α_μ)-tuples. Fig. 3a shows that the G-computation is sensitive to outcome regression misspecification but not to the propensity score model. In contrast, Fig. 3b demonstrates KECENI's double robustness: it remains accurate as long as either the propensity score or outcome regression model is close to the truth.

5.3 Average Treatment Effect

In this simulation study, we compare KECENI with existing methods for estimating average treatment effects under network interference. We included two methods with available code: AUTOGNET by Tchetgen Tchetgen et al. (2021) and TMLENET by Ogburn et al. (2022). Both methods use structural equation models: AUTOGNET employs an autoregressive parametric model for G-computation, while TMLENET uses summary-statistics-based structural equation models, assuming these statistics are known. We demonstrate the shortcomings of AUTOGNET and TMLENET when parametric modeling assumptions are violated and summary statistics are misspecified, repsectively. In contrast, we use the Wasserstein distance to model the nuisance functions μ and π via kernel regressions and to define our dissimilarity metrics Δ_i. This approach allows us to construct an estimator for the average treatment effect without relying on strict parametric forms or summary statistics.

Setting. The simulation consisted of 40 repeated runs, using the same network model $\mathcal{G}$ as that in the previous subsection with $n = 4000$ units. For the compatibility with existing methods, we sampled binary covariate feature vectors $X_i \in \{0, 1\}^2$, treatment variables $T_i \in \{0, 1\}$ and outcome variables $Y_i \in \{0, 1\}$. Each element of the covariate features was sampled from the Bernoulli distribution with probability 0.5 i.i.d. across the elements and nodes. These elements were denoted by $X_i^{(1)}$ and $X_i^{(2)}$, respectively. Treatments were randomly assigned as

$$T_i \sim \text{Bernoulli}\left(\text{expit}(5 \cdot \text{Avg}((X_{N_i \setminus \{i\}}^{(1)} - 0.5)(X_{N_i \setminus \{i\}}^{(2)} - 0.5)))\right),$$

independently across i, where expit is the logistic sigmoid function: $\text{expit}(x) \equiv \frac{e^x}{1+e^x}$. Outcomes $Y_i \in \{0, 1\}$ were given as

$$Y_i \sim \text{Bernoulli}\left(\text{expit}\left\{(T_i - 0.5) - 7\text{Avg}((X^{(1)}_{N_i \setminus \{i\}} - 0.5)(X^{(2)}_{N_i \setminus \{i\}} - 0.5))\right\}\right), \tag{10}$$

independently across i. The treatment assignment and outcome were designed so that the $\mathbb{E}[Y_i|T_i = 1] \approx \mathbb{E}[Y_i|T_i = 0]$. In order to recover the correct treatment effect, one should properly account for the confounding by the covariate vector.

We compared the performance of the three aforementioned methods in estimating the average counterfactual means over the population under two treatment assignments and their difference. We considered no-treated and all-treated scenarios: $t^{(0)}_i = 0, t^{(1)}_i = 1, \forall i$, respectively. The average counterfactual means for these scenarios were $n^{-1}\sum_{i=1}^n \theta_i(t^{(0)}_{N_i}) \approx 0.406$ and $n^{-1}\sum_{i=1}^n \theta_i(t^{(1)}_{N_i}) \approx 0.594$. The average treatment effect was approximately 0.188. For all three methods, we assume that the nature of the dependence structure yielding treatment assignments and potential outcomes are properly specified.

Estimation Details and Results. We consider the performance of AUTOGNET and TMLENET under misspecification of the parametric form of the structural equation model and the encoding of the summary statistics respectively. For AUTOGNET, we used covariates X_i directly withoutinteraction terms, leading to an incorrect structural equation model for the data $(X_i, T_i, Y_i : i = 1, \ldots, n)$ as per Tchetgen Tchetgen et al. (2021). For TMLENET, the treatment and covariate summary statistics were set to $(T_i, \text{Avg}(T_{N_i \setminus \{i\}} - 0.5))$ and $(X_i, \text{Avg}(X_{N_i \setminus \{i\}} - 0.5))$, respectively. Under these conditions, AUTOGNET and TMLENET produced significantly inaccurate estimates, failing to identify the treatment effect between the two scenarios (Figs. 4a and 4b).

In contrast, KECENI enables estimation under weaker assumptions, avoiding strict parametric forms or reliance on summary statistics in modeling the propensity score and outcome regression functions. We implement KECENI assuming that treatment is assigned independently across nodes given covariate vectors. Under this assumption, $\pi(t_{N_i}|x_{N_i^{(2)}}, \mathcal{G}_{N_i^{(2)}}) = \prod_{j \in N_i}\{(1-\pi^\circ(x_{N_j}))\mathbf{1}(t_j = 0) + \pi^\circ(x_{N_j})\mathbf{1}(t_j = 1)\}$, where $N_i^{(2)}$ is the 2-hop neighborhood of node i, and $\pi^\circ(x_{N_j})$ represents the probability of node j being treated ($T_j = 1$) given $X_{N_j} = x_{N_j}$. We further operationalize KECENI by assuming that the propensity score $\pi^\circ(x_{N_i})$ and outcome regression $\mu(t_{N_i}, x_{N_i^{(2)}}, \mathcal{G}_{N_i^{(2)}})$ change smoothly with the distribution of t_j and x_j over $j \in N_i \setminus \{i\}$. Specifically, we assume:

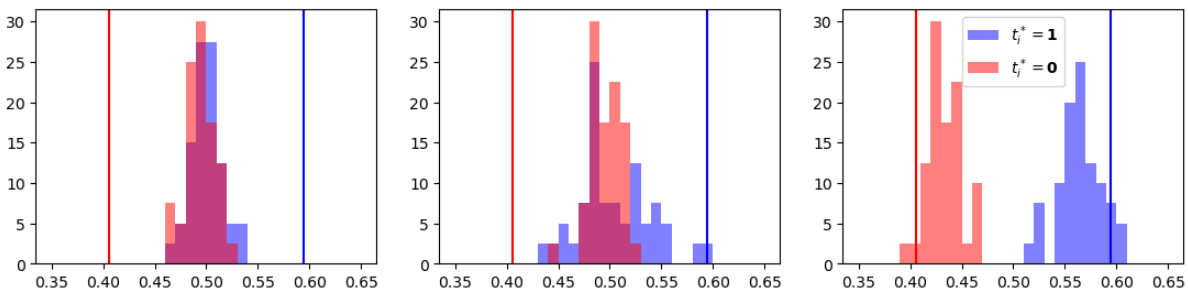
$$
\begin{aligned}
&|\mu(t_{N_i}, x_{N_i^{(2)}}, \mathcal{G}_{N_i^{(2)}}) - \mu(t'_{N_{i^*}}, x'_{N_{i^*}^{(2)}}, \mathcal{G}_{N_{i^*}^{(2)}})| \\
&\leq L_\mu \left(\|(t_i, x_i) - (t'_{i^*}, x'_{i^*})\|_1 + \text{Wasserstein}_{1,1}\left(\widehat{\mathbb{P}}\{(t_j, x_j)\}_{j \in N_i \setminus \{i\}}, \widehat{\mathbb{P}}\{(t'_j, x'_j)\}_{j \in N_{i^*} \setminus \{i^*\}}\right)\right); \\
&|\pi^\circ(x_{N_i}) - \pi^\circ(x'_{N_{i^*}})| \leq L_\pi \left(\|x_i - x'_{i^*}\|_1 + \text{Wasserstein}_{1,1}\left(\widehat{\mathbb{P}}\{x_j\}_{j \in N_i \setminus \{i\}}, \widehat{\mathbb{P}}\{x'_j\}_{j \in N_{i^*} \setminus \{i^*\}}\right)\right),
\end{aligned}
$$

for some $L_\mu, L_\pi > 0$ and any given t_{N_i}, $t'_{N_{i^*}}$, x_{N_i}, and $x'_{N_{i^*}}$, where $\widehat{\mathbb{P}}$ is the empirical probability measure on the respective set, and $\text{Wasserstein}_{1,1}(\cdot,\cdot)$ denotes the Wasserstein 1-distance with respect to the ℓ_1 metric. We then use kernel smoothing estimators on the Wasserstein distances to fit μ and π°, yielding the estimates $\widehat{\mu}$ and $\widehat{\pi}$. The Wasserstein distance was also used to define the dissimilarity metric Δ_i which relates the observed treatent assignments and network characteristics to those of the target node,

$$
\Delta_i \equiv |T_i - t^*_{i^*}| + \text{Wasserstein}_{1,1}\left(\widehat{\mathbb{P}}\{T_j\}_{j \in N_i \setminus \{i\}}, \widehat{\mathbb{P}}\{t^*_j\}_{j \in N_{i^*} \setminus \{i^*\}}\right).
$$

Finally, $\widehat{P}_{X_{[n]}|\mathcal{G}}$ is taken to be the empirical product measure in Eq. (4). Fig. 4c indicates that despite the presence of smoothing bias, our non-parametric KECENI implementation was able to successfully identify the non-zero treatment effect much more successfully than AUTOGNET and TMLENET.

.



(a) **(b)** **(c)**

Figure 4: Histogram of average counterfactual mean estimates under unknown data generating model and summary statistics by (a) AUTOGNET, (b) TMLENET and (c) KECENI.

6 Application to Social Networks and Microfinance in Indian Villages

In this section, we analyze the impact of participation in savings self-help groups (SHGs) on financial behavior, using microfinance data from rural Indian villages provided by Banerjee et al. (2013). The dataset was collected from a survey conducted in 75 villages in Karnataka, India. This survey included a village-level questionnaire on leadership, a full household census, and a detailed follow-up individual demographic survey based on stratified sampling by religion and geographic sublocation. In addition, socio-economic interactions between the subsampled individuals were captured in a 12-dimensional multilayer network, where each layer corresponded to a respective type of relationship.

In 37 of the surveyed villages, additional financial information was collected, including savings, SHG participation, and outstanding loans. Focused on these villages, Khatami et al. (2024) examined the causal effect of SHG participation on financial risk tolerance, measured by the presence of outstanding loans. We extend their analysis by investigating not only the average direct and spillover effects across the population, but also how these effects vary with both the intervention and network characteristics of outcome units, providing insights into the heterogeneity of causal effects across the social network.

We defined the interference network $\mathcal{G}$ with the same node set as the observed multilayer network (i.e., the subsampled villagers), and an edge set formed by the union of seven network layers. These layers correspond to the relationships: 'those who visit the respondent's home', 'those whose homes the respondent visits', 'kin in the village', 'nonrelatives with whom the respondent socializes', 'those to whom the respondent would lend money', 'those from whom the respondent would borrow material goods', and 'those with whom the respondent prays'. The network interference assumption (Assumption 2.1) was posited based on this composite network. For each node i, we define the direct effect of the intervention (participation in self-help groups, SHG) and the spillover effect of the neighborhood's interventions on the outcome (presence of outstanding loans) in the same way as Khatami et al. (2024), i.e.,

$$\mathrm{DE}_{i^*} \equiv \mathbb{E}[\theta_{i^*}(t_{i^*} = 1, t_{N_{i^*}} = T_{N_{i^*}})] - \mathbb{E}[\theta_{i^*}(t_{i^*} = 0, t_{N_{i^*}} = T_{N_{i^*}})],$$
$$\mathrm{SpE}_{i^*} \equiv \mathbb{E}[\theta_{i^*}(t_{i^*} = T_{i^*}, t_{N_{i^*}} = \mathbf{1})] - \mathbb{E}[\theta_{i^*}(t_{i^*} = T_{i^*}, t_{N_{i^*}} = \mathbf{0})].$$

To address potential confounding, we controlled for six demographic features $X_i \in \mathbb{R}^6$ of individuals (age, education, whether native to the village, employment status, whether they work outside the village, and saving status), after standardizing them.

We applied non-parametric KECENI using summary statistics to estimate DE_{i^*} and SpE_{i^*}, setting the nuisance parameter models as

$$\mathbb{P}[T_i = 1 | X_{N_i} = x_{N_i}] \sim \mathrm{KernelReg}(x_i, \mathrm{Avg}(x_{N_i \setminus \{i\}})),$$
$$\mathbb{E}[Y_i | T_{N_i} = t_{N_i}, X_{N_i^{(2)}} = x_{N_i^{(2)}}] \sim \mathrm{KernelReg}(t_i, \mathrm{Avg}(t_{N_i \setminus \{i\}} - 0.5), x_i, \mathrm{Avg}(x_{N_i \setminus \{i\}})),$$

where interventions T_i were assumed i.i.d. given the summary statistics. For the dissimilarity metric, we used $\Delta_i \equiv \left\| (T_i, \mathrm{Avg}(T_{N_i \setminus \{i\}} - 0.5), \log|N_i|) - (t^*_{i^*}, \mathrm{Avg}(t^*_{N_{i^*} \setminus \{i^*\}} - 0.5), \log|N_{i^*}|) \right\|_1$ and the empirical product measure in Eq. (4) for the covariate distribution estimate.

By aggregating the estimates $\widehat{\mathrm{DE}}_{i^*}$ and $\widehat{\mathrm{SpE}}_{i^*}$, we obtained estimates for the average direct effect (ADE) and average spillover effect (ASpE). The ADE estimate was 0.160 with a 95%-confidence interval of $(0.146, 0.170)$, calculated using block-based bootstrap. Our point estimate was smaller than the 0.315 reported by Khatami et al. (2024) and the other six baseline methods listed in their Appendix 12.2. However, our confidence interval demonstrated the significance of the ADE, in contrast to the wider interval $(-1.570, 2.200)$ reported by Khatami et al. (2024). For the ASpE, our estimate was 0.011, with a 95%-confidence interval of $(-0.001, 0.022)$. This result of an insignificant ASpE aligns with the estimate of 0.050 from Khatami et al. (2024) and other baseline estimates.

Although the ASpE results suggest no evidence of a spillover effect, a more detailed analysis using our flexible node-wise counterfactual framework reveals otherwise. Specifically, we investigate the spillover effect by intervening on the outcome node i^* at a fixed value t^*, rather than at the observed intervention T_{i^*}. We define, with a slight abuse of notation:

$$\mathrm{SpE}_{i^*}(t^*) \equiv \theta_{i^*}(t_{i^*} = t^*, t_{N_{i^*}\setminus\{i^*\}} = \mathbf{1}) - \theta_{i^*}(t_{i^*} = t^*, t_{N_{i^*}\setminus\{i^*\}} = \mathbf{0}),$$

for $t^* = 0, 1$, and compare the estimates $\widehat{\mathrm{SpE}}_{i^*}(0)$ and $\widehat{\mathrm{SpE}}_{i^*}(1)$. The average estimates of $\widehat{\mathrm{SpE}}_{i^*}(0)$ and $\widehat{\mathrm{SpE}}_{i^*}(1)$ across nodes were 0.001 and 0.058, with 95%-confidence intervals of $(-0.009, 0.012)$ and $(0.040, 0.075)$, respectively. These results indicate a significant spillover effect when the respondent were *intervened* to participate in SHG. The ASpE approach likely failed to detect this effect due to the relatively small observed proportion of SHG participants (17.1%).

We further explore the heterogeneity of the spillover effect based on node connectivity. Fig. 5 displays the distribution of the estimates $\widehat{\mathrm{SpE}}_{i^*}(0)$ and $\widehat{\mathrm{SpE}}_{i^*}(1)$, stratified by the degree of node i^* into five groups: (i) isolated nodes, (ii) nodes with degree 1 – 4, (iii) 5 – 8, (iv) 9 – 12, and (v) 13 or higher. When node i^* was not intervened ($t^*_{i^*} = 0$),

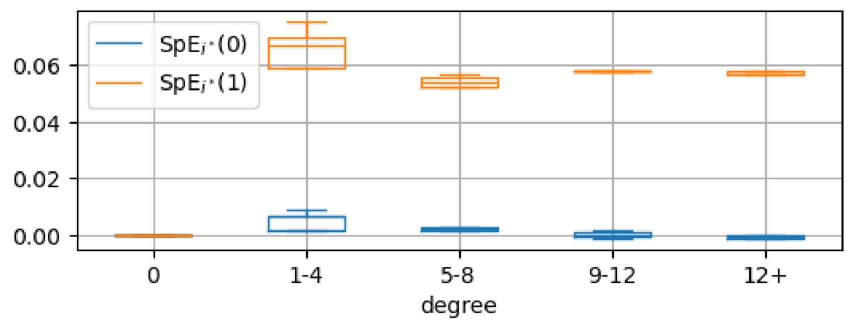


Figure 5: Boxplots of the node-wise spillover effect estimates stratified by the degree of the node in five groups. The blue boxplots represent the spillover effects when the node i^* was not intervened ($t^*_{i^*} = 0$), while the orange boxplots correspond to the intervened case ($t^*_{i^*} = 1$).

the spillover effect remained approximately constant and close to zero across all degrees. However, when the node was intervened ($t^*_{i^*} = 1$), we observed substantial heterogeneity in the spillover effect across node degrees. Excluding the trivial case of isolated nodes with zero spillover effect, we found that the spillover effect was generally larger for nodes with fewer connections (i.e., lower degree). This suggests that less connected individuals tend to experience greater influence from their neighbors' participation in self-help groups compared to highly connected individuals. In other words, individuals who are less integrated into the network appear to be more susceptible to the financial behaviors of their peers, potentially due to their increased reliance on fewer social connections. However, these findings are based solely on point estimates without proper uncertainty assessments or confidence intervals. They should not be interpreted as indicating any statistical significance.

These findings highlight the flexibility of our approach based on node-wise causal estimands and importance of accounting for node connectivity when assessing causal effects under network interference, as interventions may have different impacts depending on the local network structure.

7 Discussion

In this paper, we propose a novel approach to causal inference under network interference. By focusing on node-wise counterfactual means, we provide a flexible framework that accounts for heterogeneity in causal effects driven by local network characteristics, such as node degree. This approach enables us to examine how treatment effects vary across subgroups defined by their network connectivity. This advantage is well illustrated in our analysis of rural Indian microfinance data, where average spillover effects appeared insignificant, but substantial heterogeneity was observed at the individual node level. Specifically, we find that nodes with fewer connections experience stronger spillover effects, suggesting that less-connected individuals are more influenced by their neighbors' participation in financial programs.

We emphasize that in defining the target estimand $\theta_{i^*}(t^*_{N_{i^*}})$, the network $\mathcal{G}$ serves as a conditioning variable rather than an argument of the counterfactual. This distinction highlights that the dependency of $\theta_{i^*}(t^*_{N_{i^*}})$ on the network is associative, not causal. Consequently, our method does not imply any causal interpretation of the heterogeneity in causal effects observed in the real data analysis when intervening on the network structure. To make such claims, we would need to disentangle contagion effects from homophily (Shalizi & Thomas 2011). Achieving this would require strong assumptions, such as the availability of pre-network covariates unaffected by social factors and capable of controlling for confounding between the network and covariates due to homophily. While such assumptions may hold in specific contexts with sufficient domain knowledge, our work focuses on more general settings of network interference and does not pursue this direction.

Some regularity conditions in our theoretical analysis could be relaxed. For example, while we assume neighborhood dependence, this can be restrictive as it requires independence between units not directly connected in the network. Alternative dependence structures have been proposed in the literature. Bhattacharya et al. (2020) and Tchetgen Tchetgen et al. (2021) introduce a conditional independence framework, similar to the global Markov property in graphical models for causal inference. Another approach is the ψ-*dependence* framework of Kojevnikov et al. (2021), which assumes that dependence weakens as the shortest path distance increases. This generalizes several dependence structures, including strong mixing, neighborhood dependence, and functional dependence on independent innovations (Kojevnikov et al. 2021, Section 2.4). Extending our results to these weaker forms of network dependence is a direction for future research.

Additionally, our method assumes that the network structure $\mathcal{G}$ accurately captures the interference relationships among units and that we observe this network without error. In practice, real-world networks often exist latently, representing underlying social interactions, and are obtained through noisy measurements. This introduces a layer of complexity not accounted for in our current framework. There is a growing body of research addressing noisy observations of interference networks, such as Li et al. (2021), Li & Wager (2022), and Leung (2022). Extending our results to accommodate noisy network settings presents an interesting direction for future work.

SUPPLEMENTARY MATERIAL

Supplement manuscript: A PDF manuscript providing supplemental descriptions of proofs and extensive details regarding arguments conveyed in the main text. (.pdf file)

References

Aronow, P. M. & Samii, C. (2017), 'Estimating average causal effects under general interference, with application to a social network experiment', *Annals of Applied Statistics* **11**(4), 1912–1947.

Auerbach, E. & Tabord-Meehan, M. (2021), 'The local approach to causal inference under network interference', *arXiv preprint arXiv:2105.03810* .

Banerjee, A., Chandrasekhar, A. G., Duflo, E. & Jackson, M. O. (2013), 'The diffusion of microfinance', *Science* **341**(6144), 1236498.

Bang, H. & Robins, J. M. (2005), 'Doubly robust estimation in missing data and causal inference models', *Biometrics* **61**(4), 962–973.

Bhattacharya, R., Malinsky, D. & Shpitser, I. (2020), Causal inference under interference and network uncertainty, *in* 'Uncertainty in Artificial Intelligence', PMLR, pp. 1028–1038.

Forastiere, L., Airoldi, E. M. & Mealli, F. (2021), 'Identification and estimation of treatment and interference effects in observational studies on networks', *Journal of the American Statistical Association* **116**(534), 901–918.

Horvitz, D. G. & Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**(260), 663–685.

Jiang, K., Mukherjee, R., Sen, S. & Sur, P. (2022), 'A new central limit theorem for the augmented ipw estimator: Variance inflation, cross-fit covariance and beyond', *arXiv preprint arXiv:2205.10198* .

Kennedy, E. H., Ma, Z., McHugh, M. D. & Small, D. S. (2017), 'Non-parametric methods for doubly robust estimation of continuous treatment effects', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79**(4), 1229–1245.

Khatami, S. B., Parikh, H., Chen, H., Roy, S. & Salimi, B. (2024), 'Graph neural network based double machine learning estimator of network causal effects', *arXiv preprint arXiv:2403.11332* .

Kojevnikov, D. (2021), 'The bootstrap for network dependent processes', *arXiv preprint arXiv:2101.12312* .

Kojevnikov, D., Marmer, V. & Song, K. (2021), 'Limit theorems for network dependent random variables', *Journal of Econometrics* **222**(2), 882–908.

Leung, M. P. (2022), 'Causal inference under approximate neighborhood interference', *Econometrica* **90**(1), 267–293.

Leung, M. P. & Loupos, P. (2022), 'Graph neural networks for causal inference under network confounding', *arXiv preprint arXiv:2211.07823* .

Li, S. & Wager, S. (2022), 'Random graph asymptotics for treatment effect estimation under network interference', *Annals of Statistics* **50**(4), 2334–2358.

Li, W., Sussman, D. L. & Kolaczyk, E. D. (2021), 'Causal inference under network interference with noise', *arXiv preprint arXiv:2105.04518* .

Liu, L., Hudgens, M. G. & Becker-Dreps, S. (2016), 'On inverse probability-weighted estimators in the presence of interference', *Biometrika* **103**(4), 829–842.

Ogburn, E. L., Sofrygin, O., Diaz, I. & Van der Laan, M. J. (2022), 'Causal inference for social network data', *Journal of the American Statistical Association* pp. 1–15.

Rastelli, R., Friel, N. & Raftery, A. E. (2016), 'Properties of latent variable network models', *Network Science* **4**(4), 407–432.

Robins, J. (1986), 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect', *Mathematical Modelling* **7**(9-12), 1393–1512.

Ross, N. (2011), 'Fundamentals of Stein's method', *Probability Surveys* **8**, 210–293.

Sävje, F., Aronow, P. & Hudgens, M. (2021), 'Average treatment effects in the presence of unknown interference', *Annals of Statistics* **49**(2), 673.

Shalizi, C. R. & Thomas, A. C. (2011), 'Homophily and contagion are generically confounded in observational social network studies', *Sociological Methods & Research* **40**(2), 211–239.

Stein, C. (1972), A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *in* 'Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory', Vol. 6, University of California Press, pp. 583–603.

Tchetgen Tchetgen, E. J., Fulcher, I. R. & Shpitser, I. (2021), 'Auto-g-computation of causal effects on a network', *Journal of the American Statistical Association* **116**(534), 833–844.

van der Laan, M. J. (2014), 'Causal inference for a population of causally connected units', *Journal of Causal Inference* **2**(1), 13–74.

van der Laan, M. J., Rose, S. et al. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, Vol. 4, Springer.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, Springer, New York, NY.

Vershynin, R. (2018), *High-dimensional probability: An introduction with applications in data science*, Vol. 47, Cambridge university press.

Villani, C. et al. (2009), *Optimal transport: old and new*, Vol. 338, Springer.

Supplement to "Heterogeneous Treatment Effects under Network Interference: A Nonparametric Approach Based on Node Connectivity"

Heejong Bong, Colin B. Fogarty, Elizaveta Levina and Ji Zhu

Department of Statistics, University of Michigan

A Details about KECENI

A.1 Dissimilarity Metric and Pseudo-outcome

A key challenge and contribution of our methodological development is defining an appropriate dissimilarity metric. To be useful for the kernel smoothing step in Algorithm 1, this metric must capture the closeness between $\theta_i(T_{N_i})$ and $\theta_{i^*}(t^*_{N_{i^*}})$ under a practical set of assumptions. This requirement is commonly referred to as a *smoothness condition*, a fundamental aspect of non-parametric estimation. For example, Auerbach & Tabord-Meehan (2021) introduced the concept of ϵ-*isomorphism* which captures equivalence among local configurations with minor discrepancies, assuming that counterfactual quantities like $\mathbb{E}[Y_i(t_{N_i})|\mathcal{G}]$ vary smoothly with the degree of disagreement. This assumption was critical in enabling their non-parametric k-nearest neighbor estimator to function effectively. However, in the presence of confounding effects from covariates, it is insufficient to rely solely on smoothness in terms of interventions and connectivity. It becomes natural, and indeed necessary, to extend this smoothness assumption to the conditional expectation $\mu(t_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) \equiv \mathbb{E}[Y_i(t_{N_i})|X_{N_i} = x_{N_i}, \mathcal{G}_{N_i}]$, accounting for the covariates' influence. This shift acknowledges that confounding introduces additional variability in outcomes, requiring smoothness not just in the intervention and network structure but also in how these interact with covariate distributions.

To formalize this extended smoothness assumption, let δ define a dissimilarity metric among local intervention-covariate-connectivity configurations $(t_{N_i}, x_{N_i}, \mathcal{G}_{N_i})$.

Assumption A.1 (Lipschitz continuity of true conditional expectation). There exists a constant $L > 0$ such that

$$|\mu(t_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) - \mu(t'_{N_{i'}}, x'_{N_{i'}}, \mathcal{G}_{N_{i'}})| \leq L \cdot \delta\{(t_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t'_{N_{i'}}, x'_{N_{i'}}, \mathcal{G}_{N_{i'}})\}, \tag{11}$$

for any units i, i', local interventions $t_{N_i} \in \{0,1\}^{|N_i|}$, $t'_{N_{i'}} \in \{0,1\}^{|N_{i'}|}$, and local covariates $x_{N_i} \in \mathbb{R}^{|N_i| \times p}$, $x'_{N_{i'}} \in \mathbb{R}^{|N_{i'}| \times p}$.

Given this assumption, the goal is to establish a dissimilarity metric Δ_i such that

$$|\theta_i(T_{N_i}) - \theta_{i^*}(t^*_{N_{i^*}})| \leq L \cdot \Delta_i. \tag{12}$$

While a large Δ_i may trivially satisfy this inequality, it would offer little insight when aggregating pseudo-outcomes. Therefore, we seek the smallest Δ_i that guarantees Eq. (12) under Eq. (11). The solution is provided by Kantorovich duality (Villani et al. 2009, Theorem 5.10), which implies that the Wasserstein distance between $P_{X_{N_i}|\mathcal{G}}$ and $P_{X_{N_{i^*}}|\mathcal{G}}$, using the dissimilarity metric $\delta\{(T_{N_i}, \cdot, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, \cdot, \mathcal{G}_{N_{i^*}})\}$, achieves this minimal Δ_i. In the following theorem, we demonstrate that the Wasserstein distance not only satisfies Eq. (12) but also serves as the optimal metric, in a *minimax* sense, under the sole assumption of Eq. (11).

Theorem A.2 (Smoothness of $\theta_i(T_{N_i})$ with respect to Δ_i)**.** *Define*

$$\Delta_i \equiv \min_{\omega \in \Omega} \int \delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} d\omega(x_{N_i}, x'_{N_{i^*}}),$$

where Ω is the set of probability measures on the product space $\mathbb{R}^{|N_i| \times p} \times \mathbb{R}^{|N_{i^*}| \times p}$, with marginals $P_{X_{N_i}|\mathcal{G}}$ on $\mathbb{R}^{|N_i| \times p}$ and $P_{X_{N_{i^*}}|\mathcal{G}}$ on $\mathbb{R}^{|N_{i^*}| \times p}$. Then, if Assumption A.1 holds, *Eq.* (12) *holds. Conversely, there exist random variables* $Y_i(T_{N_i})$ *and* $Y_{i^*}(t^*_{N_{i^*}})$ *satisfying Assumption A.1 and* $|\theta_i(T_{N_i}) - \theta_{i^*}(t^*_{N_{i^*}})| = L \cdot \Delta_i$.

A potentially more effective estimate can be obtained by relaxing the marginal constraint on $\mathbb{R}^{|N_i|\times p}$. With this relaxation, the optimal ω exhibits a projective property: for each $x'_{N_{i^*}} \in \mathbb{R}^{|N_{i^*}|\times p}$, the conditional probability measure $\omega(\cdot, x'_{N_{i^*}})$ becomes a Dirac measure on x_{N_i} with the smallest $\delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\}$ (in almost sure sense). Following this property, we denote the marginal of the optimal ω on $\mathbb{R}^{|N_i|\times p}$ by $P_{i^*\to i}$. Under this marginal, $P_{i^*\to i}$, the counterfactual mean

$$\theta_{i^*\to i}(T_{N_i}) \equiv \int \mathbb{E}[Y_i(T_{N_i})|X_{N_i} = x_{N_i}, \mathcal{G}_{N_i}]dP_{i^*\to i}(x_{N_i}).$$

achieves the smallest error in approximating $\theta_{i^*}(t^*_{N_{i^*}})$ given the smoothness condition in Eq. (11). As a result, we can improve the estimation accuracy by modifying the pseudo-outcome to be an unbiased estimator of $\theta_{i^*\to i}(T_{N_i})$, rather than of $\theta_i(T_{N_i})$. This is done by defining the modified pseudo-outcome as

$$\begin{aligned}\xi_i(\widehat{\mu}, \widehat{\pi}) \equiv\ & \frac{Y_i - \widehat{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i})}{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\nu_{i^*\to i}(X_{N_i}) \int \widehat{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i})dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) \\ & + \int \widehat{\mu}(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i})\nu_{i^*\to i}(x_{N_i})dP_{X_{N_i}|\mathcal{G}}(x_{N_i}).\end{aligned}$$

where $\nu_{i^*\to i}$ is the Radon-Nikodym derivative of $P_{i^*\to i}$ with respect to $P_{X_{N_i}|\mathcal{G}}$. This modified pseudo-outcome is a doubly robust unbiased estimator of $\theta_{i^*\to i}(T_{N_i})$, as detailed in the following theorem, whose proof is provided in Appendix B.3.

Theorem A.3 (Double robustness of pseudo-outcomes)**.** *If either* $\overline{\mu} = \mu$ *or* $\overline{\pi} = \pi$, *for any target intervention* $t^*_{N_i}$,

$$\mathbb{E}[\xi_i(\overline{\mu}, \overline{\pi})|T_{N_i} = t^*_{N_i}] = \theta_{i^*\to i}(t^*_{N_i}).$$

In practice, obtaining the optimal ω can be computationally expensive, and $P_{i^*\to i}$ is often not absolutely continuous with $P_{X_{[n]}|\mathcal{G}}$, especially when using discrete $\widehat{P}_{X_{[n]}|\mathcal{G}}$. Below, we introduce two approximations, $\widetilde{\omega}$, for the optimal ω: a kernel smoother-based approach and a nearest-neighbor approach. We define $\widetilde{\Delta}_i$ and $\widetilde{\nu}_{i^*\to i}$ as follows:

$$\widetilde{\Delta}_i \equiv \int \delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} d\widetilde{\omega}(x_{N_i}, x'_{N_{i^*}}),$$

$$\widetilde{\nu}_{i^*\to i}(x_{N_i}) \equiv \int \frac{d\widetilde{\omega}(x_{N_i}, x'_{N_{i^*}})}{dP_{X_{N_i}|\mathcal{G}}(x_{N_i})}.$$

Note that the integration defining $\widetilde{\nu}_{i^*\to i}(x_{N_i})$ is taken over $x'_{N_{i^*}}$. Given fitted $\widehat{P}_{X_{[n]}|\mathcal{G}}$, we use Monte Carlo methods to approximations $\widehat{\Delta}_i$, $\widehat{\nu}_{i^*\to i}$, and accordingly

$$\begin{aligned}\widehat{\xi}_i(\widehat{\mu}, \widehat{\pi}) \equiv\ & \frac{Y_i - \widehat{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i})}{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})} \widehat{\nu}_{i^*\to i}(X_{N_i}) \int \widehat{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i}) \\ & + \int \widehat{\mu}(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) \widehat{\nu}_{i^*\to i}(x_{N_i}) d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i}).\end{aligned}$$

Kernel Smoother-Based Approach In this approach, we approximate the optimal ω using a kernel function $\mathbf{k}_h : [0,\infty) \to [0,\infty)$ with a bandwidth parameter $h > 0$. The Dirac conditional probability measure $\omega(\cdot, x'_{N_{i^*}})$ is then smoothed across the support of $P_{X_{N_i}|\mathcal{G}}$. Specifically, for each x_{N_i} and $x'_{N_{i^*}}$, we define

$$d\widetilde{\omega}(x_{N_i}, x'_{N_{i^*}}) \equiv \frac{\mathbf{k}_h \circ \delta\{(t_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\}}{\int \mathbf{k}_h \circ \delta\{(t_{N_i}, x''_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} dP_{X_{N_i}|\mathcal{G}}(x''_{N_i})} dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) dP_{X_{N_{i^*}}|\mathcal{G}}(x'_{N_{i^*}}).$$

This ensures that $\widetilde{\omega}$ has an absolutely continuous marginal on $\mathbb{R}^{|N_i|\times p}$ with respect to $P_{X_{N_i}|\mathcal{G}}$, while preserving some degree of the projective property.

Nearest Neighbor-Based Approach In this approach, we approximate the optimal ω such that the conditional measure $\omega(\cdot, x'_{N_{i^*}})$ is distributed across nearby points up to a proximity threshold, rather than concentrated solely at the nearest point. Specifically, let $\alpha \in [0,1]$ be a tuning parameter representing the threshold. For each x_{N_i} and $x'_{N_{i^*}}$, we define

$$d\widetilde{\omega}(x_{N_i}, x'_{N_{i^*}}) \equiv \frac{1}{\alpha} \mathbf{1}\left\{\delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} \le q_{i,\alpha}(x'_{N_{i^*}})\right\} dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) dP_{X_{N_{i^*}}|\mathcal{G}}(x'_{N_{i^*}}),$$

where $q_{i,\alpha}(x'_{N_{i^*}})$ is the α quantile of the distribution of $\delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\}$, with respect to $P_{X_{N_i}|\mathcal{G}}$:

$$q_{i,\alpha}(x'_{N_{i^*}}) \equiv \sup\left\{q : P_{X_{N_i}|\mathcal{G}}\left[\delta\{(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} \le q\right] \le \alpha\right\}.$$

A.2 Calibration of Hyper-parameters by Leave-neighborhood-out Cross-validation

KECENI relies on the bandwidth hyperparameter λ and tuning parameters h or α for approximating the dissimilarity metrics Δ_i and derivatives $\nu_{i^*\to i}$. The choice of these parameters can significantly influence the estimates and the resulting causal inferences. To reduce potential subjectivity in the results, we implement a data-driven calibration of the hyperparameters using a leave-neighborhood-out cross-validation approach.

Given a discrete set of candidate hyperparameter values, we first fit the nuisance functions $\widehat{\mu}$, $\widehat{\pi}$, and $d\widehat{P}_{X_{[n]}|\mathcal{G}}$ using the entire dataset. With these fitted nuisance functions, we compute the unmodified pseudo-outcome $\widehat{\xi}_i$ by setting $\nu_{i^*\to i}(X_{N_i}) = 1$. This serves as an unbiased estimator of $\theta_i(T_{N_i})$, independent of the hyperparameters. Next, we compute the leave-neighborhood-out estimator $\widehat{\theta}_i^{(-N_i^{(2)})}(T_{N_i})$ for $\theta_i(T_{N_i})$ for each candidate hyperparameter set, by fitting KECENI on a dataset where the two-hop neighborhood of unit i is left out.

Finally, we evaluate each hyperparameter set by calculating the mean squared error: $\frac{1}{n}\sum_{i=1}^n(\widehat{\xi}_i - \widehat{\theta}_i^{(-N_i^{(2)})}(T_{N_i}))^2$. The set with the lowest error is then selected.

A.3 Details about Inference

The KECENI estimate $\widehat{\theta}_{i^*}(t^*_{N_{i^*}})$ can be expressed as the ratio of two sums:

$$\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) = \frac{\sum_{i=1}^n \kappa_\lambda(\Delta_i)\widehat{\xi}_i(\widehat{\mu}, \widehat{\pi})}{\sum_{i=1}^n \kappa_\lambda(\Delta_i)}.$$

Under regularity conditions (as detailed in Section 4), both the numerator and denominator converge to normal distributions with appropriate scaling.

We approximate the variance term in $\widehat{\theta}_{i^*}(t^*_{N_{i^*}})$ by a linear approximation with respect to the errors in the denomiator and numerator (see Appendix B.6 for more details about the variance term):

$$\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) - \widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) \approx \sum_{i=1}^{n} \frac{1}{D} \left[\begin{aligned} &\kappa_\lambda(\Delta_i)\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \mathbb{E}[\kappa_\lambda(\Delta_i)\theta_{i^*\to i}(T_{N_i})|\mathcal{G}] \\ &- \widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) \left\{ \kappa_\lambda(\Delta_i) - \mathbb{E}[\kappa_\lambda(\Delta_i)|\mathcal{G}] \right\} \end{aligned} \right],$$

where $\widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) \equiv D^{-1}\mathbb{E}[\sum_{i=1}^n \kappa_\lambda(\Delta_i)\theta_{i^*\to i}(T_{N_i})|\mathcal{G}]$ and $D \equiv \mathbb{E}[\sum_{i=1}^n \kappa_\lambda(\Delta_i)|\mathcal{G}]$. According to the delta method, the variance of $\widehat{\theta}_{i^*}(t^*_{N_{i^*}})$ converges to the variance of the right hand side. We denote each summand by φ_i and approximate by

$$\widehat{\varphi}_i \equiv \frac{1}{\widehat{D}} \left[\kappa_\lambda(\Delta_i)\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \widehat{\mathbb{E}}[\kappa_\lambda(\Delta_i)\theta_{i^*\to i}(T_{N_i})|\mathcal{G}] - \widehat{\theta}_{i^*}(t^*_{N_{i^*}}) \left\{ \kappa_\lambda(\Delta_i) - \widehat{\mathbb{E}}[\kappa_\lambda(\Delta_i)|\mathcal{G}] \right\} \right],$$

where $\widehat{\mathbb{E}}$ represents the expectation over the fitted marginal propensity

$$\widehat{\varpi}(t_{N_i}|\mathcal{G}) \equiv \int \widehat{\pi}(t_{N_i}|x_{N_i},\mathcal{G}_{N_i}) d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i}).$$

Alternatively, we approximate the variance term in $\widehat{\theta}_{i^*}(t^*_{N_{i^*}})$ by a linear approximation with respect to the errors in $\widehat{\xi}_i(\widehat{\mu},\widehat{\pi})$ and Δ_i:

$$\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) - \widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) \approx \sum_{i=1}^{n} \left[\begin{aligned} &\frac{\kappa_\lambda(\Delta_i)\{\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \mathbb{E}[\widehat{\xi}_i(\widehat{\mu},\widehat{\pi})|\mathcal{G}]\}}{\sum_{j=1}^n \kappa_\lambda(\Delta_j)} \\ &+ \left\{ \widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \widehat{\theta}_{i^*}(t^*_{N_{i^*}}) \right\} \frac{\kappa_\lambda(\Delta_i) - \mathbb{E}[\kappa_\lambda(\Delta_i)|\mathcal{G}]}{\sum_{j=1}^n \kappa_\lambda(\Delta_j)} \end{aligned} \right],$$

where $\widetilde{\theta}_{i^*}(t^*_{N_{i^*}}) \equiv D^{-1}\mathbb{E}[\sum_{i=1}^n \kappa_\lambda(\Delta_i)\theta_{i^*\to i}(T_{N_i})|\mathcal{G}]$ and $D \equiv \mathbb{E}[\sum_{i=1}^n \kappa_\lambda(\Delta_i)|\mathcal{G}]$. We denote each summand by φ_i and approximate by

$$\begin{aligned} \widehat{\varphi}_i &\equiv \left[\frac{\kappa_\lambda(\Delta_i)\{\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \widehat{\theta}_{i^*}(t^*_{N_{i^*}})\}}{\sum_{j=1}^n \kappa_\lambda(\Delta_j)} + \left\{ \widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \widehat{\theta}_{i^*}(t^*_{N_{i^*}}) \right\} \frac{\kappa_\lambda(\Delta_i) - \widehat{\mathbb{E}}[\kappa_\lambda(\Delta_i)|\mathcal{G}]}{\sum_{j=1}^n \kappa_\lambda(\Delta_j)} \right] \\ &= \frac{1}{\widehat{D}} \left[\left\{ \widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \widehat{\theta}_{i^*}(t^*_{N_{i^*}}) \right\} \{2 \cdot \kappa_\lambda(\Delta_i) - \widehat{\mathbb{E}}[\kappa_\lambda(\Delta_i)|\mathcal{G}]\} \right]. \end{aligned}$$

It is important to note that φ_i terms can be dependent on each other, even when $\widehat{\mu}$, $\widehat{\pi}$, and $\widehat{P}_{X_{[n]}|\mathcal{G}}$ are replaced by μ, π, and $P_{X_{[n]}|\mathcal{G}}$. This dependency arises from the network structure $\mathcal{G}$ in two ways: (1) $\xi_i(\mu, \pi)$ depends on Y_i, T_{N_i}, and X_{N_i}, meaning that if units i and j share a neighbor k, they rely on the same values of T_k and X_k; (2) the feature set (Y_i, T_i, X_i) is generally not independent across units, with dependencies structured according to the network $\mathcal{G}$. The specific dependency assumptions are outlined in Assumption 4.1. Assuming independence, as in the sample variance estimate $\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{\varphi}_i - \frac{1}{n}\sum_{j=1}^{n}\widehat{\varphi}_j)^2$, may significantly underestimate variability, leading to incorrect confidence interval coverage.

To address this, we propose using variability estimators that account for correlations within the network. When $\widehat{\theta}_{i^*}(t^*_{N_{i^*}})$ satisfies asymptotic normality, we recommend estimating the variance with the network *heteroskedasticity and autocorrelation consistent* (HAC) estimator (Kojevnikov et al. 2021). Alternatively, if normality is not guaranteed, we can apply the block-based bootstrap (Kojevnikov 2021) to assess the variability of $\widehat{\theta}_{i^*}(t^*_{N_{i^*}})$.

B Proofs

B.1 Proof of Identifiability (Theorem 2.3)

By G-computation, we have

$$\begin{aligned}
\theta_{i^*}(t^*_{N_{i^*}}) &= \mathbb{E}[Y_{i^*}(t^*_{N_{i^*}})|\mathcal{G}] \\
&= \int \mathbb{E}[Y_{i^*}(t^*_{N_{i^*}})|X_{N_{i^*}} = x_{N_{i^*}}, \mathcal{G}]dP_{X_{N_{i^*}}|\mathcal{G}}(x_{N_{i^*}}) \\
&\overset{\text{(i)}}{=} \int \mathbb{E}[Y_{i^*}(t^*_{N_{i^*}})|T_{N_{i^*}} = t^*_{N_{i^*}}, X_{N_{i^*}} = x_{N_{i^*}}, \mathcal{G}]dP_{X_{N_{i^*}}|\mathcal{G}}(x_{N_{i^*}}) \\
&\overset{\text{(ii)}}{=} \int \mathbb{E}[Y_{i^*}|T_{N_{i^*}} = t^*_{N_{i^*}}, X_{N_{i^*}} = x_{N_{i^*}}, \mathcal{G}]dP_{X_{N_{i^*}}|\mathcal{G}}(x_{N_{i^*}}),
\end{aligned}$$

where step (i) follows from the ignorability condition (Assumption 2.2), and step (ii) from the consistency condition (i.e., $Y_i = Y_i(T_{N_i})$). The positivity condition in Assumption 2.2

ensures that the conditional expectation is well-defined and identifiable, as the conditioning event has positive probability almost surely with respect to $dP_{X_{N_{i^*}}|\mathcal{G}}(x_{N_{i^*}})$. $\square$

B.2 Proof of Minimax Smoothness (Theorem A.2)

Let ω be the optimal joint probability measure on the product space $\mathbb{R}^{|N_i|\times p} \times \mathbb{R}^{|N_{i^*}|\times p}$, with marginals $P_{X_{N_i}|\mathcal{G}}$ on $\mathbb{R}^{|N_i|\times p}$ and $P_{X_{N_{i^*}}|\mathcal{G}}$ on $\mathbb{R}^{|N_{i^*}|\times p}$. Let $\omega_{X_{N_i}|X_{N_{i^*}}}$ be a version of conditional measure of X_{N_i} given $X_{N_{i^*}}$. Then

$$
\begin{aligned}
&|\theta_{i^*\to j}(T_{N_i}) - \theta_{i^*}(t^*_{N_{i^*}})| \\
&= \left| \int \mu(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) dP_{i^*\to i}(x_{N_i}) - \int \mu(t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}}) dP_{X_{N_{i^*}}|\mathcal{G}}(x'_{N_{i^*}}) \right| \\
&\leq \int\int |\mu(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) - \mu(t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})| \; d\omega_{X_{N_i}|X_{N_{i^*}}}(x_{N_i}|x'_{N_{i^*}}) dP_{X_{N_{i^*}}|\mathcal{G}}(x'_{N_{i^*}}) \\
&\overset{\text{(i)}}{\leq} L \int\int \delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} \; d\omega_{X_{N_i}|X_{N_{i^*}}}(x_{N_i}|x'_{N_{i^*}}) dP_{X_{N_{i^*}}|\mathcal{G}}(x'_{N_{i^*}}) \\
&= L \int \delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} \; d\omega(x_{N_i}, x'_{N_{i^*}}) = L\Delta_i,
\end{aligned}
$$

which is our first claim, where Assumption A.1 applies to (i). Now, consider the union of the two covariate space $\mathcal{X}_\cup \equiv \mathbb{R}^{|N_i|\times p} \cup \mathbb{R}^{|N_{i^*}|\times p}$, and let $\Omega_\cup$ be the set of probability measures on $\mathcal{X}_\cup \times \mathcal{X}_\cup$, where $\omega(x, x') \in \Omega_\cup$ if and only if it has marginals $\mathbf{1}\{x \in \mathbb{R}^{|N_i|\times p}\} P_{X_{N_i}|\mathcal{G}}(x)$ and $\mathbf{1}\{x' \in \mathbb{R}^{|N_{i^*}|\times p}\} P_{X_{N_{i^*}}|\mathcal{G}}(x')$. Then

$$
\begin{aligned}
L\Delta_i &\equiv \min_{\omega\in\Omega} \int L\delta\{(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x'_{N_{i^*}}, \mathcal{G}_{N_{i^*}})\} d\omega(x_{N_i}, x'_{N_{i^*}}) \\
&= \min_{\omega\in\Omega_\cup} \int L\delta\{(T_{N_i}, x, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x', \mathcal{G}_{N_{i^*}})\} d\omega(x, x').
\end{aligned}
$$

By the Kantorovich-Rubinstein formula (Villani et al. 2009, Particular Case 5.16),

$$
\begin{aligned}
L\Delta_i &= \min_{\omega\in\Omega_\cup} \int L\delta\{(T_{N_i}, x, \mathcal{G}_{N_i}), (t^*_{N_{i^*}}, x', \mathcal{G}_{N_{i^*}})\} d\omega(x, x') \\
&= \sup_{\mu\in\mathcal{M}} \left\{ \int \mu(x) d(\mathbf{1}\{x \in \mathbb{R}^{|N_i|\times p}\} P_{X_{N_i}|\mathcal{G}}(x)) - \int \mu(x') d(\mathbf{1}\{x' \in \mathbb{R}^{|N_{i^*}|\times p}\} P_{X_{N_{i^*}}|\mathcal{G}}(x')) \right\}
\end{aligned}
$$

where $\mathcal{M}$ is the set of L-Lipschitz functions in $L_1(\mathbb{R}^{|N_i|\times p} \cup \mathbb{R}^{|N_{i^*}|\times p})$. For the optimal μ, let $\mu_i(x) \equiv \mu(x)\mathbf{1}\{x \in \mathbb{R}^{|N_i|\times p}\}$ and $\mu_{i^*}(x) \equiv \mu(x')\mathbf{1}\{x' \in \mathbb{R}^{|N_{i^*}|\times p}\}$. Then any $Y_i(T_{N_i})$ and $Y_{i^*}(t^*_{N_{i^*}})$ with conditional expectation $\mathbb{E}[Y_i(t_{N_i})|X_{N_i} = x_{N_i}, \mathcal{G}] \mid_{t_{N_i}=T_{N_i}} = \mu_i(x_{N_i})$ and $\mathbb{E}[Y_{i^*}(t^*_{N_{i^*}})|X_{N_{i^*}} = x'_{N_{i^*}}, \mathcal{G}] = \mu_{i^*}(x'_{N_{i^*}})$ satisfy the second claim. $\square$

B.3 Proof of Double-robustness (Theorems 3.1 and A.3)

Below, we prove the version of Theorem A.3. For the version of Proposition 3.1, take $\nu_{i^*\to i} \equiv 1$ and then we obtain $\mathbb{E}[\xi_i(\overline{\mu}, \overline{\pi})|T_{N_i} = t^*_{N_i}, \mathcal{G}] = \theta_i(t^*_{N_i})$. Let

$$\varpi(t_{N_i}|\mathcal{G}) \equiv \int \pi(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) \text{ and } \overline{\varpi}(t_{N_i}|\mathcal{G}) \equiv \int \overline{\pi}(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) dP_{X_{N_i}|\mathcal{G}}(x_{N_i}).$$

Then,

$$\begin{aligned}
&\mathbb{E}[\xi_j(\overline{\mu}, \overline{\pi})|T_{N_i} = t^*_{N_i}, \mathcal{G}] \\
&\overset{\text{(i)}}{=} \int \mathbb{E}[\xi_j(\overline{\mu}, \overline{\pi})|T_{N_i} = t^*_{N_i}, X_{N_i} = x_{N_i}, \mathcal{G}] \ \frac{\pi(t^*_{N_i}|x_{N_i}, \mathcal{G}_{N_i})}{\varpi(t^*_{N_i}|\mathcal{G})} \ dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) \\
&\overset{\text{(ii)}}{=} \int \{\mu(t^*_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) - \overline{\mu}(t^*_{N_i}, x_{N_i}, \mathcal{G}_{N_i})\} \frac{\pi(t^*_{N_i}|x_{N_i}, \mathcal{G}_{N_i})/\varpi(t_{N_i}|\mathcal{G})}{\overline{\pi}(t^*_{N_i}|x_{N_i}, \mathcal{G}_{N_i})/\overline{\varpi}(t_{N_i}|\mathcal{G})} \ \nu_{i^*\to i}(x_{N_i}) \ dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) \\
&\quad + \int \overline{\mu}(t^*_{N_i}, x_{N_i}, \mathcal{G}_{N_i}) \nu_{i^*\to i}(x_{N_i}) dP_{X_{N_i}|\mathcal{G}}(x_{N_i}) \\
&\overset{\text{(iii)}}{=} \int \{\mu(t_{N_i}, x_{N_i}, \mathcal{G}_j) - \overline{\mu}(t_{N_i}, x_{N_i}, \mathcal{G}_j)\} \left\{\frac{\pi(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i})/\varpi(t_{N_i}|\mathcal{G})}{\overline{\pi}(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i})/\overline{\varpi}(t_{N_i}|\mathcal{G})} - 1\right\} \nu_{i^*\to i}(x_{N_i}) \ dP_{X_{N_i}}(x_{N_i}) \\
&\quad + \int \mu(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i}) \nu_{i^*\to i}(x_{N_i}) dP_{X_{N_i}}(x_{N_i}), \\
&\overset{\text{(iv)}}{=} \int \{\mu(t_{N_i}, x_{N_i}, \mathcal{G}_j) - \overline{\mu}(t_{N_i}, x_{N_i}, \mathcal{G}_j)\} \left\{\frac{\pi(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i})/\varpi(t_{N_i}|\mathcal{G})}{\overline{\pi}(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i})/\overline{\varpi}(t_{N_i}|\mathcal{G})} - 1\right\} \nu_{i^*\to i}(x_{N_i}) \ dP_{X_{N_i}}(x_{N_i}) \\
&\quad + \theta_{i^*\to i}(t^*_{N_i}),
\end{aligned}$$

where (i) follows the definition of conditional expectation, (ii) and (iii) is due to rearranging, and (iv) follows the definition of $\theta_{i^*\to i}(t^*_{N_i})$. So the last line shows that

$$\mathbb{E}[\xi_i(\overline{\mu}, \overline{\pi})|T_{N_i} = t^*_{N_i}, \mathcal{G}] = \theta_{i^*\to i}(t^*_{N_i}),$$

as long as either $\overline{\mu} = \mu$ or $\overline{\pi} = \pi$. $\square$

B.4 Empirical Process Theory of Weighted Average Process under Neighborhood Dependence

Assume that $P_{X_{[n]}|\mathcal{G}}$ is known. The KECENI estimator in Eq. (1) can then be viewed as a kernel-smoothed value of $\xi_i(\widehat{\mu}, \widehat{\pi})$. This formulation implies that our estimator behaves as a kernel-weighted empirical process $Z_n(\widehat{\mu}, \widehat{\pi})$ indexed by the nuisance parameters:

$$Z_n(\widehat{\mu}, \widehat{\pi}) = \frac{1}{\widehat{D}} \sum_{i=1}^{n} \kappa_\lambda(\Delta_i) \xi_i(\widehat{\mu}, \widehat{\pi}), \tag{13}$$

where $\widehat{D} \equiv \sum_{i=1}^{n} \kappa_\lambda(\Delta_i)$, $\mathbb{E}_n$ indicates the expectation over the empirical distribution which puts probability mass 1 on the observed data (Y, T, X), and $\mathbb{E}_{Y_{[n]}, X_{[n]}|T_{[n]}}$ indicates the conditional expectation of Y and X on observed T.

Empirical process theory provides the necessary tools to handle the nontrivial propagation of error from $\widehat{\mu}$ and $\widehat{\pi}$ into the empirical process. This is crucial when the same data is used for both the estimation of nuisance parameters and the kernel-weighted averaging. The key concept here is establishing *asymptotic equicontinuity* of the empirical process, which is defined as: for a proper scale sequence σ_n,

$$\sup_{\max\{\|\mu_1-\mu_2\|, \|\pi_1-\pi_2\|\}<\delta_n} \frac{1}{\sigma_n} (\mathbb{E}_n - \mathbb{E}_{(Y_{[n]}, X_{[n]})|T_{[n]}}) [Z_n(\mu_1, \pi_1) - Z_n(\mu_2, \pi_2)] \xrightarrow{p} 0, \tag{14}$$

conditional on $T_{[n]}$ and $\mathcal{G}$ in almost sure sense, under a proper metric $\|\cdot\|$ and for $\delta_n \to 0$. The condition on $T_{[n]}$ is imposed to fix the kernel weights, which is necessary to establish asymptotic equicontinuity.

While the asymptotic equicontinuity result of van der Laan (2014, Appendix (A3)) applies to neighborhood dependence, it is not efficient for kernel-weighted empirical processes, particularly in cases where many observations contribute very little compared to others. This situation often arises during the kernel smoothing step in KECENI, leading to non-tight error bounds when applying the result from van der Laan (2014). In this section, we extend their result by providing a more general form of asymptotic equicontinuity with improved rates for weighted empirical processes. Notably, our result reduces to that of van der Laan (2014) in the special case where all weights are equal, and the weighted averages revert to unweighted averages.

Equicontinuity of weighted average process We consider a weighted average process indexed by $\theta \in \mathcal{F}$, defined as

$$Z_n(\theta) = \frac{1}{\sqrt{W_n K_n}} \sum_{i=1}^{n} w_i f_i(\theta), \tag{15}$$

where $W_n \equiv \sum_{i=1}^n w_i$; w_i are fixed weights; given fixed θ, $f_i(\theta)$ are dependent to the observed data O and exhibit neighborhood dependence with respect to network $\mathcal{G}_n$; and K_n is the maximum degree of $\mathcal{G}_n$. We prove the stochastic equicontinuity of this process as follows.

Theorem B.1. *Assume that (1) for each i and p, there exists a universal constant C so that $\|(\mathbb{E}_n - \mathbb{E})[f_i(\theta_1) - f_i(\theta_2)]\|_p \le C d(\theta_1, \theta_2)$; (2) $\mathcal{F}$ is totally bounded with respect to this metric d; and (3) for some $\eta > 0$, $\int_0^\eta \psi_2^{-1}(N(\epsilon, \mathcal{F}, d))d\epsilon < \infty$, where $\psi_2(x) = \exp(x^2) - 1$, $\|\cdot\|_{\psi_2}$ is the respective Orlicz norm, and $N(\epsilon, \mathcal{F}, d)$ is the covering number of $\mathcal{F}$ by d-balls of size ϵ. Then the process Z_n is stochastically equicontinuous, i.e., for any sequence $\delta_n \to 0$ and $\eta > 0$,*

$$\mathbb{P}\left[\sup_{d(\theta_1,\theta_2)<\delta_n} \frac{1}{\sqrt{W_n K_n}}(\mathbb{E}_n - \mathbb{E})[Z_n(\theta_1) - Z_n(\theta_2)] > \eta\right] \to 0.$$

Proof. We use the theoretical argument in the proof of Theorem 2.2.4 and Corollary 2.2.5 in van der Vaart & Wellner (1996). van der Laan (2014) rewrote these results as follows (see Appendix (A3) therein): for a sequence $(\zeta_n(\theta) : n \in \mathbb{N}, \theta \in \mathcal{F})$ of random processes, if (1) $\|\zeta_n(\theta_1) - \zeta_n(\theta_2)\|_{\psi_2} \le C \cdot d(\theta_1, \theta_2)$ for some universal constant C and metric $d(\cdot,\cdot)$, (2) $\mathcal{F}$ is totally bounded with respect to this metric d, and (3) for some $\eta > 0$, $\int_0^\eta \psi_2^{-1}(N(\epsilon, \mathcal{F}, d))d\epsilon < \infty$, then the process ζ_n is stochastically equicontinuous.

Plugging in $\zeta_i(\theta) \equiv \frac{1}{\sqrt{W_n K_n}}(\mathbb{E}_n - \mathbb{E})[\sum_{i=1}^n w_i f_i(\theta)]$, it suffices to show that

$$\left\| \frac{1}{\sqrt{W_n K_n}}(\mathbb{E}_n - \mathbb{E}) \left[\textstyle\sum_{i=1}^n w_i(f_i(\theta_1) - f_i(\theta_2))\right] \right\|_{\psi_2} \leq C d(\theta_1, \theta_2), \tag{16}$$

for some universal constant C, where the Orlicz norm is defined conditional on T. This needs an Orlicz norm bound for a weighted average of neighborhood dependent random variables. We modify Lemma 6 of van der Laan (2014) to provide the following lemma:

Lemma B.2. *Assume that for each i and p, there exists a universal constant C so that $\|\zeta_i(\theta_1) - \zeta_i(\theta_2)\|_p \leq C d(\theta_1, \theta_2)$, $\mathbb{E}[\zeta_i(\theta_1) - \zeta_i(\theta_2)] = 0$, and ζ_i is neighborhood independent with respect to network $\mathcal{G}$, which assoiates each node i with neighborhood $N_i \subset \{1, \ldots, n\}$ and has maximum degree K. Then, there exists another universal constant C such that for any $w_i \in [0, 1]$,*

$$\|\textstyle\sum_{i=1}^n w_i(\zeta_i(\theta_1) - \zeta_i(\theta_2))\|_{\psi_2} \leq C\sqrt{W_n K_n}\ d(\theta_1, \theta_2).$$

This lemma directs to the desired result. □

Proof of Lemma B.2 Below we claim that there exists a universal constant C such that for any integer p, $\|\sum_i w_i(\zeta_i(\theta_1) - \zeta_i(\theta_2))\|_p \leq C\sqrt{p\, W_n K_n}\ d(\theta_1, \theta_2)$. Then by Proposition 2.5.2 of Vershynin (2018), $\|\sum_i w_i(\zeta_i(\theta_1) - \zeta_i(\theta_2))\|_{\psi_2} \leq C\sqrt{W_n K_n}\ d(\theta_1, \theta_2)$ for another universal constant C.

Following the proof of Lemma 5 in van der Laan (2014), for an integer p and $\vec{i} \equiv (i_1, \ldots, i_p) \in \{1, \ldots, n\}^p$, let $R(\vec{i})$ be an indicator that has value 1 if there exists an element i_k so that $i_l \notin N_{i_k}$, $\forall l \neq k$ or 0 otherwise. Then

$$\begin{aligned}
&\mathbb{E}[|\textstyle\sum_i w_i(\zeta_i(\theta_1) - \zeta_i(\theta_2))|^p] = \sum_{i_1,\ldots,i_p} \mathbb{E}[\prod_{k=1}^p w_{i_k} |\zeta_i(\theta_1) - \zeta_i(\theta_2)|] \\
&= \textstyle\sum_{i_1,\ldots,i_p} (1 - R(\vec{i}))\ w_{i_1} \cdots w_{i_p} \mathbb{E}[|\prod_k (\zeta_i(\theta_1) - \zeta_i(\theta_2))|] \\
&\leq \textstyle\sum_{i_1,\ldots,i_p} (1 - R(\vec{i}))\ w_{i_1} \cdots w_{i_p} (p^{-1} \sum_k \|\zeta_i(\theta_1) - \zeta_i(\theta_2)\|_p^p) \\
&\leq C \textstyle\sum_{i_1,\ldots,i_p} (1 - R(\vec{i}))\ w_{i_1} \cdots w_{i_p}\ d(\theta_1, \theta_2)^p
\end{aligned}$$

for some universal constant C. Now it suffices to show that $\sum_{i_1,\ldots,i_p}(1-R(\vec{i}))\ w_{i_1}\cdots w_{i_p} \leq C(pW_nK_n)^{p/2}$.

Let an indicator vector $B(\vec{i}) \in \{0,1\}^p$ to have $B_k(\vec{i}) = 1$ if $N_{i_k} \cap \cup_{l=1}^{k-1} N_{i_l} = \emptyset$ or 0 otherwise. According to the proof of Lemma 6 in van der Laan (2014), $B(\vec{i})$ has at most $p/2$ 1's. For a fixed $\vec{b} \equiv (b_1,\ldots,b_p)$,

$$\begin{aligned}
&\sum_{B(\vec{i})=b}(1-R(\vec{i}))\ w_{i_1}\cdots w_{i_p} \leq \sum_{B(\vec{i})=b} w_{i_1}\cdots w_{i_p} \\
&= \sum_{B(i_1,\ldots,i_{p-1})=(b_1,\ldots,b_{p-1})} w_{i_1}\cdots w_{i_{p-1}} \begin{pmatrix} \mathbf{1}\{b_p=1\}\sum_{i:N_i\cap(\cup_{k=1}^{p-1}N_{i_k})=\emptyset} w_i \\ +\,\mathbf{1}\{b_p=0\}\sum_{i:N_i\cap(\cup_{k=1}^{p-1}N_{i_k})\neq\emptyset} w_i \end{pmatrix} \\
&\leq \sum_{B(i_1,\ldots,i_{p-1})=(b_1,\ldots,b_{p-1})} w_{i_1}\cdots w_{i_{p-1}} \left(\mathbf{1}\{b_p=1\}\sum_i w_i + \mathbf{1}\{b_p=0\}\sup_{|I|=pK_n}\sum_{i\in I} w_i\right).
\end{aligned}$$

Since the p-th term in the summand of the last line is not dependent on $i_1,\ldots,i_{p-1}$, this boils down again to upper bound $\sum_{B(i_1,\ldots,i_{p-1})=(b_1,\ldots,b_{p-1})} w_{i_1}\cdots w_{i_{p-1}}$. Applying the above argument recursively, we obtain

$$\begin{aligned}
\sum_{B(\vec{i})=b}(1-R(\vec{i}))\ w_{i_1}\cdots w_{i_p} &\leq \prod_{k=1}^{p}\left(\mathbf{1}\{b_k=1\}\sum_i w_i + \mathbf{1}\{b_k=0\}\sup_{|I|=pK_n}\sum_{i\in I} w_i\right) \\
&= (\textstyle\sum_i w_i)^{\#\{k:b_k=1\}}(\sup_{|I|=pK_n}\sum_{i\in I} w_i)^{\#\{k:b_k=0\}} \\
&\leq (\textstyle\sum_i w_i)^{p/2}(\sup_{|I|=pK_n}\sum_{i\in I} w_i)^{p/2} \leq (\sum_i w_i)^{p/2}(pK_n)^{p/2},
\end{aligned}$$

where the last inequality follows that $w_i \in [0,1]$, $\forall i$. Because the final upper bound is not dependent on b, and there are at most 2^p different b's,

$$\sum_{i_1,\ldots,i_p}(1-R(\vec{i}))\ w_{i_1}\cdots w_{i_p} \leq 2^p(pK_nW_n)^{p/2},$$

which concludes this lemma's proof. $\square$

B.5 Proof of Consistency (Theorem 4.6)

In this proof, we provide a high-probability upper bound for the estimaion error, $|\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}})|$. First, we define an oracle estimator:

$$\overline{\theta}_{i^*}(t^*_{N_{i^*}}) \equiv \widehat{D}^{-1} \sum_{i=1}^{n} \kappa_\lambda(\Delta_i)\xi_i(\overline{\mu}, \overline{\pi}),$$

where $\widehat{D} \equiv \sum_{i=1}^{n} \kappa_\lambda(\Delta_i)$. Then we decompose the error into

$$\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}}) = \mathcal{R}_{n,1} + \mathcal{R}_{n,2} + \mathcal{R}_{n,3},$$

where

$$\mathcal{R}_{n,1} = \overline{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}}),$$
$$\mathcal{R}_{n,2} = \widehat{D}^{-1}(\mathbb{E}_n - \mathbb{E}) \left[\sum_{i=1}^{n} \kappa_\lambda(\Delta_i)(\widehat{\xi}_i(\widehat{\mu}, \widehat{\pi}) - \xi_i(\overline{\mu}, \overline{\pi})) \middle| \mathcal{G} \right],$$
$$\mathcal{R}_{n,3} = \widehat{D}^{-1}\mathbb{E} \left[\sum_{i=1}^{n} \kappa_\lambda(\Delta_i)(\widehat{\xi}_i(\widehat{\mu}, \widehat{\pi}) - \xi_i(\overline{\mu}, \overline{\pi})) \middle| \mathcal{G} \right],$$

$\mathbb{E}_n$ indicates the expectation over the empirical distribution which puts probability mass 1 on the observed data $(Y_{[n]}, T_{[n]}, X_{[n]})$, and $\mathbb{E}_{Y_{[n]}, X_{[n]}|T_{[n]}}$ indicates the conditional expectation of $Y_{[n]}$ and $X_{[n]}$ on observed $T_{[n]}$.

Bounding the decomposed errors, we show the following lemma.

Lemma B.3. *$|D - \widehat{D}| = O_p(\sqrt{K_{\max} D})$, and hence $D\widehat{D}^{-1} \xrightarrow{p} 1$ as long as $K_{\max} = o(D)$.*

Proof. We note that

$$\begin{aligned}
&\mathrm{Var}[\widehat{D}|\mathcal{G}] = \mathrm{Var}\left[\sum_{i=1}^{n} \kappa_\lambda(\Delta_i) \middle| \mathcal{G} \right] \\
&\le \frac{1}{2} \sum_{i=1}^{n} \sum_{j:j \in N_i^{(5)}} \mathrm{Cov}\left[\kappa_\lambda(\Delta_i), \kappa_\lambda(\Delta_j)|\mathcal{G}\right] \le \frac{1}{2} \sum_{i=1}^{n} \sum_{j:j \in N_i^{(5)}} \left(\mathrm{Var}\left[\kappa_\lambda(\Delta_i)|\mathcal{G}\right] + \mathrm{Var}[\kappa_\lambda(\Delta_j)|\mathcal{G}]\right) \\
&\le K_{\max} \sum_{i=1}^{n} \mathrm{Var}\left[\kappa_\lambda(\Delta_i)|\mathcal{G}\right] \le K_{\max} \sum_{i=1}^{n} \mathbb{E}\left[(\kappa_\lambda(\Delta_i))^2|\mathcal{G}\right] \le K_{\max} \sum_{i=1}^{n} \mathbb{E}\left[\kappa_\lambda(\Delta_i)|\mathcal{G}\right] = K_{\max} D.
\end{aligned}$$

By the Chebyshev inequality, we obtain the desired result. $\square$

The rest of the proof proceeds as follows:

- $\mathcal{R}_{n,1}$ is bounded using the Kernel smoother theory, based on Fan (1993). The variance may calculated under a network dependence structure.
- $\mathcal{R}_{n,2}$ is bounded using the newly developed empirical process theory under a network dependence structure.
- $\mathcal{R}_{n,3}$ is bounded using the double robustness of $\xi_j(\overline{\mu},\overline{\pi})$, $\|\widehat{\mu}-\overline{\mu}\|_\infty = o_p(1)$ and $\|\widehat{\pi}-\overline{\pi}\|_* = o_p(1)$ and the accuracy of the covariate distribution estimate $\widehat{P}_{X_{[n]}|\mathcal{G}}$.

B.5.1 Bounding $\mathcal{R}_{n,1}$

$$
\begin{aligned}
&|\overline{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}})| = |\widehat{D}^{-1}\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)\xi_i(\overline{\mu},\overline{\pi}) - \theta_{i^*}(t^*_{N_{i^*}})| \\
&\leq |\widehat{D}^{-1}\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))| + |\widehat{D}^{-1}\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)(\theta_i(T_{N_i}) - \theta_{i^*}(t^*_{N_{i^*}}))| \\
&\leq |\widehat{D}^{-1}\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)(\xi_j(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))| + L|\widehat{D}^{-1}\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)\cdot\Delta_i|,
\end{aligned}
$$

where the last equality follows the smoothness result in Assumption 4.4. Due to the uniform boundedness of ξ_j, for some $M_\xi > 0$

$$
\mathbb{E}[|\kappa_\lambda(\Delta_i)(\xi_j(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))|^2|T,\mathcal{G}] \leq 4M_\xi^2|\kappa_\lambda(\Delta_i)|^2,
$$

and as in the proof of Lemma B.3,

$$
\begin{aligned}
&\mathbb{E}[|\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))|^2|\mathcal{G}] \\
&\leq \sum_{i=1}^n \sum_{j:j\neq i} \mathbb{E}\left[\begin{array}{c} |\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))| \\ \times |\kappa_\lambda(\Delta_j)(\xi_j(\overline{\mu},\overline{\pi}) - \theta_j(T_{N_j}))| \end{array}\middle| \mathcal{G}\right] \\
&\leq \sum_{i=1}^n \sum_{j:j\neq i} \left(\begin{array}{c} \mathbb{E}\left[|\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))|^2 \middle| \mathcal{G}\right] \\ + \mathbb{E}\left[|\kappa_\lambda(\Delta_j)(\xi_j(\overline{\mu},\overline{\pi}) - \theta_j(T_{N_j}))|^2 \middle| \mathcal{G}\right] \end{array}\right) \\
&\leq K_{\max}\sum_{i=1}^n \mathbb{E}\left[|\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))|^2 \middle| \mathcal{G}\right] \leq 4K_{\max}M_\xi^2 \sum_{i=1}^n \mathbb{E}\left[|\kappa_\lambda(\Delta_i)|^2 \middle| \mathcal{G}\right] = O\left(K_{\max}D\right).
\end{aligned}
$$

Hence with Lemma B.3 the first term is upper bounded by

$$\begin{aligned}
&\mathbb{E}[|\widehat{D}^{-1}\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))|^2|\mathcal{G}] \\
&\leq O\left(\frac{1}{D^2}\sum_{i=1}^n\sum_{j:j\neq i}\mathbb{E}\left[\begin{array}{l}|\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))| \\ \\ \times|\kappa_\lambda(\Delta_j)(\xi_j(\overline{\mu},\overline{\pi}) - \theta_{i^*\to k}(T;\mathcal{G}))|\end{array}\middle|\mathcal{G}\right]\right) \\
&\leq O\left(\frac{1}{D^2}\sum_{i=1}^n\sum_{j:j\neq i}\left(\begin{array}{l}\mathbb{E}\left[|\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i}))|^2\middle|\,\mathcal{G}\right] \\ \\ +\,\mathbb{E}\left[|\kappa_\lambda(\Delta_j)(\xi_j(\overline{\mu},\overline{\pi}) - \theta_{i^*\to k}(T;\mathcal{G}))|^2\middle|\,\mathcal{G}\right]\end{array}\right)\right) \\
&\leq O\left(\frac{K_{\max}}{D}\right).
\end{aligned}$$

Similarly, due to the support of κ, the second term is upper bounded by

$$\widehat{D}^{-1}\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)\cdot\Delta_j(T_{N_i}) \leq \lambda.$$

In sum, by Markov's inequality,

$$|\mathcal{R}_{n,1}| = O_p\left(\sqrt{\frac{K_{\max}}{D}} + L\lambda\right).$$

B.5.2 Bounding $\mathcal{R}_{n,2}$

Recall that

$$\begin{aligned}
\overline{\varpi}(t_{N_i}|\mathcal{G}) &\equiv \int \overline{\pi}(t_{N_i}|x_{N_i},\mathcal{G}_{N_i})dP_{X_{N_i}|\mathcal{G}}(x_{N_i}), \text{ and} \\
\overline{m}(t_{N_i},\mathcal{G}) &\equiv \int \overline{\mu}(t_{N_i},x_{N_i},\mathcal{G}_{N_i})dP_{X_{N_i}|\mathcal{G}}(x_{N_i}).
\end{aligned}$$

In addition, we define

$$\begin{aligned}
\widehat{\varpi}(T_{N_i}|\mathcal{G}) &\equiv \int \widehat{\pi}(T_{N_i}|x_{N_i},\mathcal{G}_{N_i})d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i}|\mathcal{G}), \text{ and} \\
\widehat{m}(t_{N_i},\mathcal{G}) &\equiv \int \widehat{\mu}(T_{N_i},x_{N_i},\mathcal{G}_{N_i})d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i}|\mathcal{G}).
\end{aligned}$$

Then,

$$
\begin{aligned}
&\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \xi_i(\overline{\mu},\overline{\pi}) \\
&= \frac{Y_j - \widehat{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i})}{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\widehat{\varpi}(T_{N_i}|\mathcal{G}) + \widehat{m}(T_{N_i}, \mathcal{G}) \\
&\quad - \frac{Y_j - \overline{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\overline{\varpi}(T_{N_i}|\mathcal{G}) - \overline{m}(T_{N_i}, \mathcal{G}) \\
&= \frac{Y_j - \overline{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\left\{\widehat{\varpi}(T_{N_i}|\mathcal{G}) - \overline{\varpi}(T_{N_i}|\mathcal{G})\right\} \\
&\quad + (Y_j - \overline{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i}))\frac{\widehat{\varpi}(T_{N_i}|\mathcal{G})}{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\left\{\frac{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i}) - \overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\right\} \\
&\quad + \frac{\widehat{\varpi}(T_{N_i}|\mathcal{G})}{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\left\{\widehat{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i}) - \overline{\mu}(T_{N_i}, X_{N_i}, \mathcal{G}_{N_i})\right\} \\
&\quad + \left\{\widehat{m}(T_{N_i}, \mathcal{G}) - \overline{m}(T_{N_i}, \mathcal{G})\right\}.
\end{aligned}
$$

Due to the unifrom boundedness assumption (Assumption 4.3),

$$
\left|\frac{\widehat{\varpi}(T_{N_i}|\mathcal{G})}{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\right| \leq \left|\int \frac{\widehat{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i})}{\widehat{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})} d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i})\right| \leq M_\pi M.
$$

Similarly,

$$
\begin{aligned}
\left|\frac{\widehat{\varpi}(T_{N_i}|\mathcal{G}) - \overline{\varpi}(T_{N_i}|\mathcal{G})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\right| &\leq \left|\frac{\int \widehat{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i}) - \int \overline{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) dP_{X_{N_i}|\mathcal{G}}(x_{N_i})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\right| \\
&\leq \left|\frac{\int \{\widehat{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i}) - \overline{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i})\} d\widehat{P}_{X_{N_i}|\mathcal{G}}(x_{N_i})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\right| \\
&\quad + \left|\frac{\int \overline{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i})(d\widehat{P}_{X_{N_i}|\mathcal{G}} - dP_{X_{N_i}|\mathcal{G}})(x_{N_i})}{\overline{\pi}(T_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\right| \\
&\leq \|\widehat{\pi} - \overline{\pi}\|_* M + o_p(1),
\end{aligned}
$$

where the second term in the last inequality follows the uniform convergence of $\widehat{P}_{X_{[n]}|\mathcal{G}}$ over $\mathcal{F}_\pi$ (first part of Assumption 4.5). Similarly,

$$
|\widehat{m}(T_{N_i}, \mathcal{G}) - \overline{m}(T_{N_i}, \mathcal{G})| \leq \|\widehat{\mu} - \overline{\mu}\|_\infty + o_p(1).
$$

Therefore, $\|\widehat{\varpi} - \overline{\varpi}\|_* \to 0$ and $\|\widehat{m} - \overline{m}\|_\infty \to 0$ as $n \to \infty$.

Due to the uniform boundedness of $\overline{\mu}$, $\widehat{\mu}$, $\overline{\pi}$, $\widehat{\pi}$ and Y_i,

$$\|\xi_i(\widehat{\mu},\widehat{\pi}) - \xi_i(\overline{\mu},\overline{\pi})\|_\infty = O\left(\|\widehat{\pi} - \overline{\pi}\|_* + \|\widehat{\mu} - \overline{\mu}\|_\infty + \|\widehat{\varpi} - \overline{\varpi}\|_* + \|\widehat{m} - \overline{m}\|_\infty\right).$$

Furthermore,

$$\begin{aligned}
&\|\kappa_\lambda(\Delta_i)(\xi_i(\widehat{\mu},\widehat{\pi}) - \xi_i(\overline{\mu},\overline{\pi})) - \mathbb{E}[\kappa_\lambda(\Delta_i)(\xi_i(\widehat{\mu},\widehat{\pi}) - \xi_i(\overline{\mu},\overline{\pi}))|\mathcal{G}]\|_p \\
&\leq 2\|\xi_i(\widehat{\mu},\widehat{\pi}) - \xi_i(\overline{\mu},\overline{\pi})\|_\infty = O\left(\|\widehat{\pi} - \overline{\pi}\|_* + \|\widehat{\mu} - \overline{\mu}\|_\infty + \|\widehat{\varpi} - \overline{\varpi}\|_* + \|\widehat{m} - \overline{m}\|_\infty\right).
\end{aligned} \tag{17}$$

Now we prove the asymptotic negligibility of $\mathcal{R}_{n,2}$. Put $\theta = (\mu, \pi, m, \varpi) \in \mathcal{F}_\mu \times \mathcal{F}_\pi$, $d(\theta_1, \theta_2) \equiv \|\widehat{\pi} - \overline{\pi}\|_* + \|\widehat{\mu} - \overline{\mu}\|_\infty + \|\widehat{\varpi} - \overline{\varpi}\|_* + \|\widehat{m} - \overline{m}\|_\infty$, and

$$Z_n(\theta) = \frac{1}{\sqrt{\widehat{D}K_{\max}}}(\mathbb{E}_n - \mathbb{E}_{Y_{[n]},X_{[n]}|T_{[n]}})\left[\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)\xi_i(\mu,\pi)\right],$$

where $\widehat{D} \equiv \sum_{i=1}^n \kappa_\lambda(\Delta_i)$, $K_i \equiv |N_i^{(5)}|$ represent the 5-hop neighborhood degree of each unit i, and $K_{\max} \equiv \max_{i\in[n]} K_i$. Note that K_i is the number of other units j such that (Y_i, T_{N_i}, X_{N_i}) is dependent on (Y_j, T_{N_j}, X_{N_j}) through neighborhood interactions. To $Z_n(\theta)$, we apply the empirical process theory for weighted average process of neighborhood dependent observation, which we developed in Appendix B.4. Following the same argument as Eq. (17), $\|\zeta_j(\theta_1) - \zeta_j(\theta_2)\|_p \leq Cd(\theta_1, \theta_2)$, satisfying the condition (1) of Theorem B.1. The conditions (2) and (3) follow that the parameter spaces $\mathcal{F}_\pi$ and $\mathcal{F}_\mu$ are totally bounded with respect to $\|\cdot\|_*$ and $\|\cdot\|_\infty$, respectively, and satisfy $\int_0^\eta \psi_2^{-1}(N(\epsilon, \mathcal{F}_\pi, \|\cdot\|_*))d\epsilon < \infty$ and $\int_0^\eta \psi_2^{-1}(N(\epsilon, \mathcal{F}_\mu, \|\cdot\|_\infty))d\epsilon < \infty$ for some $\eta > 0$. Hence $\frac{1}{\sqrt{\widehat{D}K_{\max}}}(\mathbb{E}_n - \mathbb{E})[\sum_{i=1}^n \kappa_\lambda(\Delta_i)\xi_i(\cdot,\cdot)]$ satisfies stochastic equicontinuity conditional on T. That is, for any sequence $\delta_n \to 0$ and fixed $\eta > 0$,

$$\mathbb{P}\left[\sup \frac{1}{\sqrt{\widehat{D}K_{\max}}}(\mathbb{E}_n - \mathbb{E}_{Y_{[n]},X_{[n]}|T_{[n]}})\left[\textstyle\sum_{i=1}^n \kappa_\lambda(\Delta_i)(\xi_i(\mu_1,\pi_1) - \xi_i(\mu_2,\pi_2))\right] > \eta \,\middle|\, T_{[n]}, \mathcal{G}\right] \to 0$$

as $n \to \infty$, where the supremum is over the space of $(\mu_1, \mu_2, \pi_1, \pi_2)$ satisfying $\|\mu_1 - \mu_2\|_\infty < \delta_n$ and $\|\pi_1 - \pi_2\|_* < \delta_n$. Then by the dominated convergence theorem, the stochastic equicontinuity holds marginally (i.e., without the condition on $T_{[n]}$), and it follows Lemma B.3 that

$$\mathcal{R}_{n,2} = \widehat{D}^{-1}(\mathbb{E}_n - \mathbb{E}) \left[\textstyle\sum_{i=1}^{n} \kappa_\lambda(\Delta_i)(\xi_i(\mu_1, \pi_1) - \xi_i(\mu_2, \pi_2))|\mathcal{G}\right] = o_p\left(\sqrt{\frac{K_{\max}}{D}}\right).$$

B.5.3 Bounding $\mathcal{R}_{n,3}$

For any treatment t, network g and its node k, conditioning on $T_{N_i} = t_{N_i}$,

$$\begin{aligned}
&\mathbb{E}[\widehat{\xi}_i(\widehat{\mu}, \widehat{\pi}) - \xi_i(\overline{\mu}, \overline{\pi})|T_{N_i} = t_{N_i}, \mathcal{G}] \\
&= \mathbb{E}\left[\{\mu(t_{N_i}, X_{N_i}, \mathcal{G}_{N_i}) - \widehat{\mu}(t_{N_i}, X_{N_i}, \mathcal{G}_{N_i})\}\left\{\frac{\pi(t_{N_i}|X_{N_i}, \mathcal{G}_{N_i})/\varpi(t_{N_i}, \mathcal{G})}{\widehat{\pi}(t_{N_i}|X_{N_i}, \mathcal{G}_{N_i})/\widehat{\varpi}(t_{N_i}, \mathcal{G})}\right\}\right] \\
&\quad + \widehat{m}(t_{N_i}, \mathcal{G}) - \theta_i(t; \mathcal{G}) \\
&= \frac{\widehat{\varpi}(t_{N_i}|\mathcal{G})}{\varpi(t_{N_i}|\mathcal{G})}\mathbb{E}\left[(\mu - \widehat{\mu})(t_{N_i}, X_{N_i}, \mathcal{G}_{N_i})\frac{(\pi - \widehat{\pi})(t_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}{\widehat{\pi}(t_{N_i}|X_{N_i}, \mathcal{G}_{N_i})}\right] \\
&\quad + \frac{1}{\varpi(t_{N_i}|\mathcal{G})}\mathbb{E}\left[(\widehat{\pi} - \pi)(t_{N_i}|X_{N_i}, \mathcal{G}_{N_i})\right]\mathbb{E}\left[(\mu - \widehat{\mu})(t_{N_i}, X_{N_i}, \mathcal{G}_{N_i})\right] \\
&\quad + \frac{\mathbb{E}\left[(\mu - \widehat{\mu})(t_{N_i}, X_{N_i}, \mathcal{G}_{N_i})\right]}{\varpi(t_{N_i}|\mathcal{G})}\int \widehat{\pi}(t_{N_i}|x_{N_i}, \mathcal{G}_{N_i})(d\widehat{P}_{X_{N_i}|\mathcal{G}} - dP_{X_{N_i}|\mathcal{G}})(x_{N_i}) \\
&\quad + \int \widehat{\mu}(t_{N_i}, x_{N_i}, \mathcal{G}_{N_i})(d\widehat{P}_{X_{N_i}|\mathcal{G}} - dP_{X_{N_i}|\mathcal{G}})(x_{N_i}),
\end{aligned}$$

where the first equality is due to the double-robustness of $\xi_i(\overline{\mu}, \overline{\pi})$. Due to the uniform boundedness of π and $\widehat{\pi}$ and Hölder's inequality,

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{i=1}^{n} \kappa_\lambda(\Delta_i)(\widehat{\xi}_i(\widehat{\mu}, \widehat{\pi}) - \xi_i(\overline{\mu}, \overline{\pi}))\middle|\mathcal{G}\right]\right| \\
&= O\left(\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[\kappa_\lambda(\Delta_i) r(T_{N_i}) s(T_{N_i})\right] \\
&+ \left|\sum_{i=1}^{n} \mathbb{E}\left[\kappa_\lambda(\Delta_i)\int \widehat{\pi}(T_{N_i}|x_{N_i}, \mathcal{G}_{N_i})(d\widehat{P}_{X_{N_i}|\mathcal{G}} - dP_{X_{N_i}|\mathcal{G}})(x_{N_i})\right]\right| \\
&+ \left|\sum_{i=1}^{n} \mathbb{E}\left[\kappa_\lambda(\Delta_i)\int \widehat{\mu}(T_{N_i}, x_{N_i}, \mathcal{G}_{N_i})(d\widehat{P}_{X_{N_i}|\mathcal{G}} - dP_{X_{N_i}|\mathcal{G}})(x_{N_i})\right]\right|
\end{aligned}\right),
\end{aligned}$$

where $r(t_{N_i}) \equiv \left\|2^{|N_i|}(\pi - \widehat{\pi})(t_{N_i}|X_{N_i}, \mathcal{G}_{N_i})\right\|_2$ and $s(t_{N_i}) \equiv \|(\mu - \widehat{\mu})(t_{N_i}, X_{N_i}, \mathcal{G}_{N_i})\|_2$.

For the last two terms, by the equicontinuity of $\widehat{P}_{X_{[n]}|\mathcal{G}}$ (the second part of Assumption 4.5),

$$\left|\sum_{i=1}^{n} \int \frac{1}{D} \mathbb{E}\left[\kappa_\lambda(\Delta_i)(\widehat{\pi}-\overline{\pi})(T_{N_i}|x_{N_i},\mathcal{G}_{N_i})\right](d\widehat{P}_{X_{N_i}|\mathcal{G}} - dP_{X_{N_i}|\mathcal{G}})(x_{N_i})\right| = o_p(q_n),$$

and by the uniform convergence of $\widehat{P}_{X_{[n]}|\mathcal{G}}$ (the first part of Assumption 4.5),

$$\left|\sum_{i=1}^{n} \int \frac{1}{D} \mathbb{E}\left[\kappa_\lambda(\Delta_i)\overline{\pi}(T_{N_i}|x_{N_i},\mathcal{G}_{N_i})\right](d\widehat{P}_{X_{N_i}|\mathcal{G}} - dP_{X_{N_i}|\mathcal{G}})(x_{N_i})\right| = O_p(q_n).$$

Hence, the last two terms have convergence rate $O_p(q_n)$. For the first term, because κ_λ has support $[-\lambda, \lambda]$,

$$\frac{1}{D}\sum_{i=1}^{n} \mathbb{E}\left[\kappa_\lambda(\Delta_i) r(T_{N_i}) s(T_{N_i})\right] \leq \left\{\sup_{(i,t_{N_i}):\Delta_i(t_{N_i})\leq\lambda} r(t_{N_i})\right\}\left\{\sup_{(i,t_{N_i}):\Delta_i(t_{N_i})\leq\lambda} s(t_{N_i})\right\}$$

By this, we conclude that

$$|\mathcal{R}_{n,3}| = O_p\left(\widetilde{r}_n(t^*_{N_{i^*}})\widetilde{s}_n(t^*_{N_{i^*}}) + q_n\right).$$

B.6 Proof of Asymptotic Normality (Theorem 4.7)

Recall that the estimation error is decomposed into

$$\widehat{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}}) = \mathcal{R}_{n,1} + \mathcal{R}_{n,2} + \mathcal{R}_{n,3},$$

where

$$\begin{aligned}
\mathcal{R}_{n,1} &= \overline{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}}),\\
\mathcal{R}_{n,2} &= \widehat{D}^{-1}(\mathbb{E}_n - \mathbb{E})\left[\sum_{i=1}^{n}\kappa_\lambda(\Delta_i)(\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \xi_i(\overline{\mu},\overline{\pi}))\right],\\
\mathcal{R}_{n,3} &= \widehat{D}^{-1}\mathbb{E}\left[\sum_{i=1}^{n}\kappa_\lambda(\Delta_i)(\widehat{\xi}_i(\widehat{\mu},\widehat{\pi}) - \xi_i(\overline{\mu},\overline{\pi}))\right],
\end{aligned}$$

and $\mathbb{E}_n$ indicates the expectation over the empirical distribution which puts probability mass 1 on the observed data (Y, T, X).

In the proof of $\widehat{\theta}_{i^*}(t^*;\mathcal{G}^*)$'s consistency (Appendix B.5), we showed that $\mathcal{R}_{n,2}$ and $\mathcal{R}_{n,3}$ are $o_p(\sqrt{\frac{K_{\max}}{D}})$ and $O_p(r_n(t^*_{N_{i^*}})s_n(t^*_{N_{i^*}}) + q_n)$, respectively. Due to the condition on σ_n, $\frac{\mathcal{R}_{n,2}+\mathcal{R}_{n,3}}{\sigma_n} \xrightarrow{p} 0$. Now, it is sufficient to show that $\frac{1}{\sigma_n}(\mathcal{R}_{n,1} - b_n)$ converges to $N(0,1)$ in distribution.

B.6.1 Weak convergence of $\frac{1}{\sigma_n}(\mathcal{R}_{n,1} - b_n)$

$$
\begin{aligned}
\mathcal{R}_{n,1} - b_n &= \overline{\theta}_{i^*}(t^*_{N_{i^*}}) - \theta_{i^*}(t^*_{N_{i^*}}) - b_n \\
&= \frac{D-\widehat{D}}{D^2}\mathbb{E}\left[\sum_{i=1}^n \kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})\middle|\mathcal{G}\right] \\
&\quad + \frac{1}{D}\left(\sum_{i=1}^n (\kappa_\lambda(\Delta_i)\xi_i(\overline{\mu},\overline{\pi}) - \mathbb{E}[\kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})|\mathcal{G}])\right) \\
&\quad + \frac{(D-\widehat{D})^2}{D^2\widehat{D}}\mathbb{E}\left[\sum_{i=1}^n \kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})\middle|\mathcal{G}\right] \\
&\quad + \frac{D-\widehat{D}}{D\widehat{D}}\left(\sum_{i=1}^n (\kappa_\lambda(\Delta_i)\xi_i(\overline{\mu},\overline{\pi}) - \mathbb{E}[\kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})|\mathcal{G}])\right).
\end{aligned}
$$

As shown in the consistency proof, $\widehat{D}^{-1} = O_p(1)$, $\frac{\widehat{D}-D}{D} = O_p\left(\sqrt{\frac{K_{\max}}{D}}\right)$, and

$$
\begin{aligned}
&\frac{1}{D}\sum_{i=1}^n (\kappa_\lambda(\Delta_i)\xi_i(\overline{\mu},\overline{\pi}) - \mathbb{E}[\kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})|\mathcal{G}]) \\
&= O_p\left(\sqrt{\frac{K_{\max}}{D}}\right),
\end{aligned}
$$

due to the double-robust property of ξ_i as shown in Theorem A.3. (i.e., $\mathbb{E}[\xi_i(\overline{\mu},\overline{\pi}) - \theta_i(T_{N_i})|T,\mathcal{G}] = 0$, almost surely.) Thus under the condition of $\frac{n^{1/3}K_{\max}^{4/3}}{D} = o_p(\sigma_n)$ and $K_{\max} = O(\sqrt{n})$,

$$
\begin{aligned}
&\frac{(D-\widehat{D})^2}{D^2\widehat{D}}\mathbb{E}\left[\sum_{i=1}^n \kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})\middle|\mathcal{G}\right] \\
&+ \frac{D-\widehat{D}}{D\widehat{D}}\left(\sum_{i=1}^n (\kappa_\lambda(\Delta_i)\xi_i(\overline{\mu},\overline{\pi}) - \mathbb{E}[\kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})|\mathcal{G}])\right) \\
&= O_p\left(\frac{K_{\max}}{D}\right) = o_p(\sigma_n).
\end{aligned}
$$

For the first two terms,

$$\begin{aligned}
&\frac{D-\widehat{D}}{D^2}\mathbb{E}\left[\sum_{i=1}^{n}\kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})\middle|\mathcal{G}\right] \\
&= \frac{1}{D}\sum_{i=1}^{n}\left(\mathbb{E}[\kappa_\lambda(\Delta_i)|\mathcal{G}]-\kappa_\lambda(\Delta_i)\right)\widetilde{\theta}_{i^*}(t^*_{N_{i^*}}),
\end{aligned}$$

and hence

$$\begin{aligned}
&\frac{D-\widehat{D}}{D^2}\mathbb{E}\left[\sum_{i=1}^{n}\kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})\middle|\mathcal{G}\right] \\
&+\frac{1}{D}\left(\sum_{i=1}^{n}\left(\kappa_\lambda(\Delta_i)\xi_i(\overline{\mu},\overline{\pi})-\mathbb{E}[\kappa_\lambda(\Delta_i)\ \theta_i(T_{N_i})|\mathcal{G}]\right)\right) \\
&= \frac{1}{D}\sum_{i=1}^{n}\begin{pmatrix}\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi})-\widetilde{\theta}_{i^*}(t^*_{N_{i^*}})) \\ +\,\mathbb{E}[\kappa_\lambda(\Delta_i)\ (\widetilde{\theta}_{i^*}(t^*_{N_{i^*}})-\theta_i(T_{N_i}))|\mathcal{G}]\end{pmatrix}.
\end{aligned}$$

Let W_j be the summands; i.e.,

$$W_j \equiv \frac{1}{D}\begin{pmatrix}\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi})-\widetilde{\theta}_{i^*}(t^*_{N_{i^*}})) \\ +\,\mathbb{E}[\kappa_\lambda(\Delta_i)\ (\widetilde{\theta}_{i^*}(t^*_{N_{i^*}})-\theta_i(T_{N_i}))|\mathcal{G}].\end{pmatrix}$$

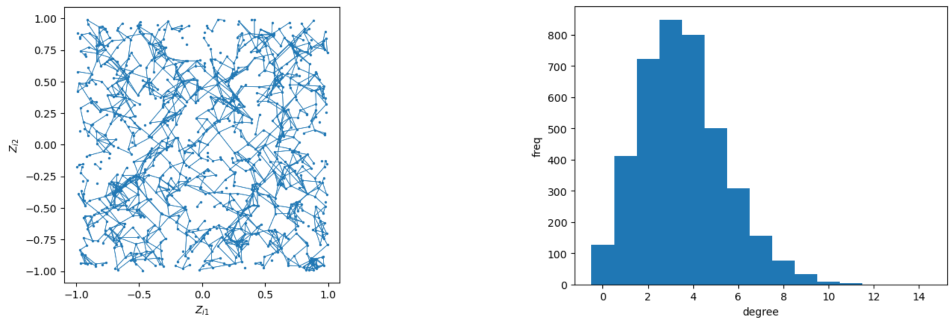
It is easily observed that $\mathbb{E}[W_j|\mathcal{G}]=0$, and due to the uniform boundedness of ξ_i, for some $M_\xi>0$

$$\begin{aligned}
\mathbb{E}[|W_j|^3|\mathcal{G}] &= \frac{1}{D^3}\mathbb{E}[|\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi})-\theta_i(T_{N_i}))|^3|\mathcal{G}] \\
&\leq \frac{M_\xi^3}{D^3}<\infty, \text{ and} \\
\mathbb{E}[W_j^4|\mathcal{G}] &= \frac{1}{D^4}\mathbb{E}[|\kappa_\lambda(\Delta_i)(\xi_i(\overline{\mu},\overline{\pi})-\theta_i(T_{N_i}))|^4|\mathcal{G}] \\
&\leq \frac{M_\xi^4}{D^4}<\infty.
\end{aligned}$$

More importantly, under the condition of $\frac{n^{1/3}K_{\max}^{2/3}}{D}=o_p(\sigma_n)$ and $K_{\max}=O(\sqrt{n})$,

$$\frac{K_{\max}^2}{\sigma_n^3}\sum_{i=1}^{n}\mathbb{E}[|W_j|^3|\mathcal{G}]+\sqrt{\frac{28}{\pi}}\frac{K_{\max}^{3/2}}{\sigma_n^2}\sqrt{\sum_{i=1}^{n}\mathbb{E}[W_j^4|\mathcal{G}]}=o_p(1).$$

That is, the right hand side in the Eq. (3.8) of Ross (2011) vanishes in probability. By Theorem 3.6 of Ross (2011), $\sum_j W_j$ satisfies the desired asymptotic normality.



(a) (b)

Figure C.1: **(a) The latent positions of the 1000 units, their adjacency in network $\mathcal{G}$ and (b) the histogram of their degrees.**

C Supplementary figures for simulation study

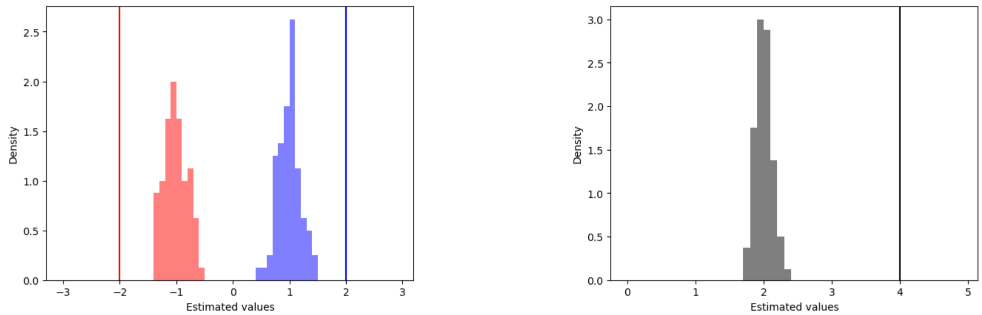
C.1 Characteristics of simulated network

Fig. C.1a shows the sampled latent positions and the corresponding adjacency structure among the units, while Fig. C.1b presents the histogram of the degree distribution for those units.

C.2 Bias if network intervention is ignored.

In this section, we highlight the potential bias of classical causal methods assuming no interference. These methods include augmented inverse propensity weighting (AIPW, Bang & Robins 2005), which imposes the stable unit treatment value assumption (SUTVA). Under SUTVA, each unit is assumed to be independent and identically distributed, making the individual treatment effect equivalent to the average treatment effect.

Fig. C.2 shows the distribution of the AIPW estimates from the same 80 datasets. The mean of the treatment effect estimates was 2.001, indicating a significant bias. This mean



(a) (b)

Figure C.2: **Histograms of (a) the AIPW estimates for the counterfactual means under two treatment scenarios and (b) the estimated treatment effects.**

value is close to the direct treatment effect, defined as

$$\mathbb{E}[Y_{i^*}(t_{i^*} = 1, t_{N_{i^*} \setminus i^*} = T_{N_{i^*} \setminus i^*})] - \mathbb{E}[Y_{i^*}(t_{i^*} = 0, t_{N_{i^*} \setminus i^*} = T_{N_{i^*} \setminus i^*})],$$

with a slight abuse of the notations. This result aligns with Sävje et al. (2021), who found that commonly used estimators for treatment effects without interference are consistent for the direct effect under certain experimental designs. Their Proposition 2 established this for the inverse propensity weighting estimator (Horvitz & Thompson 1952) under randomized experimental designs and degree restrictions. Our simulation study suggests a possible extension to doubly robust estimators, such as AIPW. Identifying specific conditions under which this argument holds for doubly robust estimators is a potential area for future research.

C.3 Dependency on Sample Size

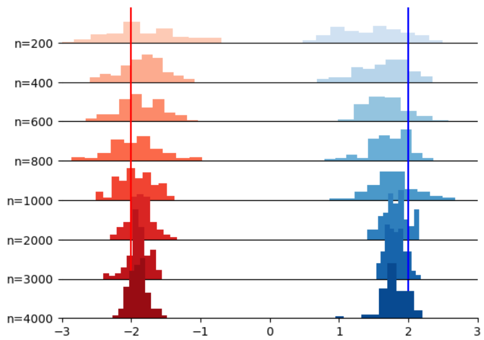
We further investigated the accuracy of KECENI in relation to the sample size n, representing the number of observed units or nodes in the network. We applied KECENI to simulation datasets with seven different sample sizes: $n \in \{250, 500, 750, 1000, 2000, 3000, 4000\}$. To isolate the effect of sample size, we generated a network $\mathcal{G}$ for each n while maintaining consistent network sparsity and connectivity of the target node. We first sampled $Z_i \in \mathbb{R}^2$ for $i \in \{1, \dots, 4000\}$ from a uniform distribution on $[-2, 2]^2$ and then generated a preliminary network $\mathcal{G}^{\text{pre}}$ from the latent variable network model with adjacency probability as given in Eq. (7). For each n, we defined $\mathcal{G}$ as a subgraph of $\mathcal{G}^{\text{pre}}$, selecting nodes with the smallest n values of $\|Z_i\|_\infty$. We then proceeded with the same feature generation and estimation procedures as in the previous simulation study.

Fig. C.3 summarizes the results of 80 simulation runs for each sample size n. Figs. C.3a and C.3b show the distributions of estimated counterfactual means and treatment effects for different values of n. As n increases, both the bias and variance of the KECENI estimates decrease, leading to more accurate treatment effect estimates. To estimate the convergence rate, we plotted the logarithms of the root-mean-square errors (RMSE) against $\log n$ in Fig. C.3c, which showed an excellent linear fit with a slope of -0.30. This rate, though slower than the parametric rate of $n^{-1/2}$, is expected due to the nonparametric nature of the kernel smoothing estimation. This result demonstrates the consistency of KECENI under the given simulation setting.

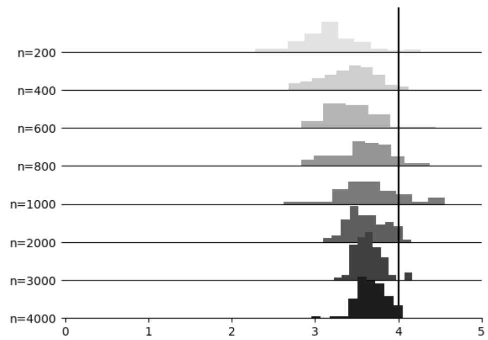
For average treatment effects, this bias has been theoretically derived and studied through simulations by Forastiere et al. (2021). Our simulation study extends the literature by demonstrating a similar bias in the estimation of counterfactual means for specific node types.

C.4 Histogram Results Demonstrating Double Robustness.

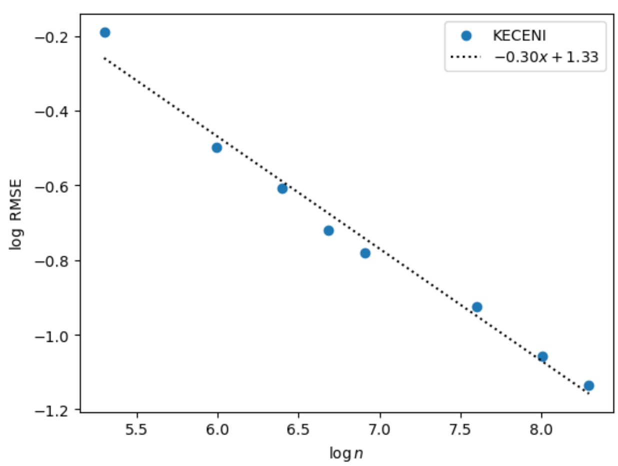
We estimated the treatment effect with KECENI using the kernel smoother approach based on the same δ and $\widehat{P}_{X_{[n]}|\mathcal{G}}$ as the previous example. For the correctly specified propensity



(a)



(b)



(c)

Figure C.3: **(a,b) Histogram of the estimated counterfactual means and the treatment effect by KECENI in different sample sizes. (c) Plot of their root mean squared errors against sample sizes in log scale.**

score and outcome regression models, we used the models in Eq. (9). For the misspecified models, we used quadratic models instead. That is, we fit

$$\mathbb{P}[T_i = 1 | X_{N_i} = x_{N_i}] \sim \text{LogisticRegression}(x_i^2, \text{Avg}(x_{N_i \setminus i}^2)),$$

$$\mathbb{E}[Y_i | T_{\mathcal{G}i} = t_{\mathcal{G}i}, X_{\mathcal{G}i} = x_{\mathcal{G}i}] \sim \text{LinearRegression}(t_i, \text{Avg}(t_{N_i \setminus i} - 0.5), x_i^2, \text{Avg}(x_{N_i \setminus i}^2))$$

to obtain propensity score estimate $\breve{\pi}(t_{\mathcal{G}i}, x_{\mathcal{G}i})$ and outcome estimate $\breve{\mu}(t_{\mathcal{G}i}, x_{\mathcal{G}i})$, where x_i^2 and $x_{N_i \setminus i}^2$ indicate the elementwise squares. We report the results of the G-computation and KECENI under four nuisance parameter sets: (i) both correctly specified, $(\widehat{\mu}, \widehat{\pi})$; (ii) outcome regression misspecified, $(\breve{\mu}, \widehat{\pi})$; (iii) propensity score misspecified, $(\widehat{\mu}, \breve{\pi})$ and (iv) both misspecified, $(\breve{\mu}, \breve{\pi})$.

Fig. C.4 and Fig. C.5 show the distributions of the G-computation and KECENI estimates, respectively, under four different nuisance parameter sets. Both methods performed well when both the propensity score and outcome regression were correctly specified (Figs. C.4a and C.5a). However, the G-computation failed to detect the treatment effect when the outcome regression was misspecified (Figs. C.4b and C.4d). In contrast, KECENI remained robust against outcome regression misspecification if the propensity score was correctly modeled (Fig. C.5b), and similarly robust to propensity score misspecification if the outcome regression was correct (Fig. C.5c). KECENI only failed when both models were incorrect (Fig. C.5d). These results demonstrate the double robustness of KECENI in estimating individual treatment effects.

C.5 Under known parametric model or summary statistics.

Let $W_i \equiv (X_i^{(1)} - 0.5)(X_i^{(2)} - 0.5)$ be the interaction terms in X_i for node i. We applied AUTOGNET, TMLENET, and KECENI using W_i as covariates. The data $(W_i, T_i, Y_i : i = 1, \ldots, n)$ satisfied the autoregressive model of Tchetgen Tchetgen et al. (2021) given the

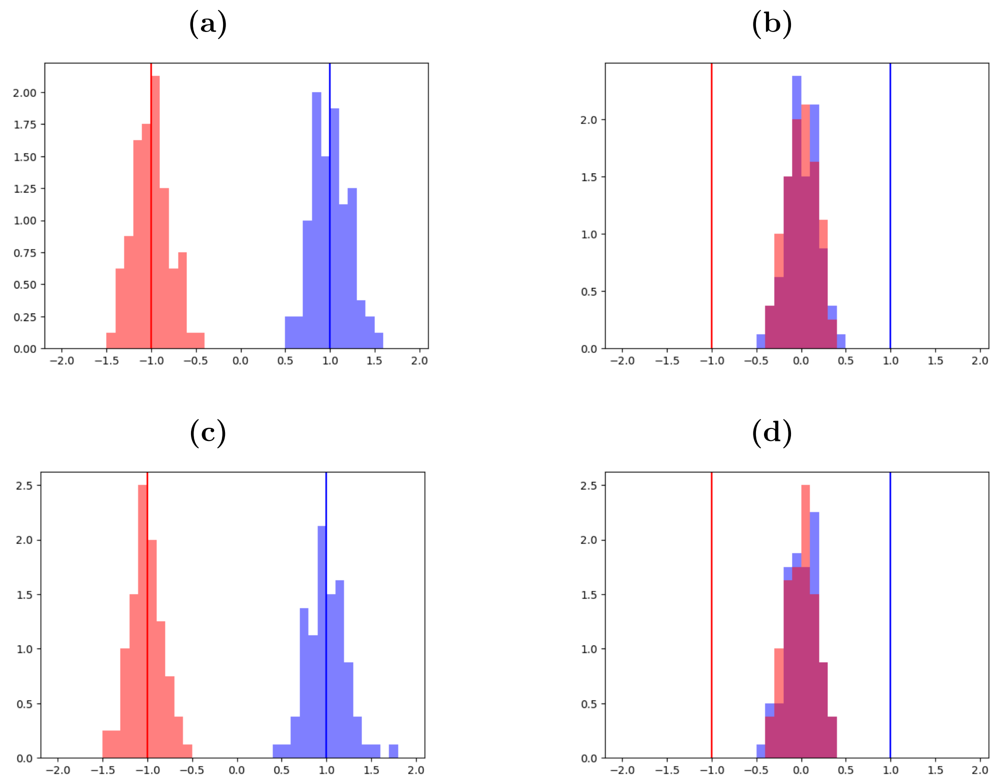


Figure C.4: **Histogram of estimated counterfactual mean by the G estimation** when (a) the both models were correct, (b) the outcome regression model was misspecified, (c) the propensity score model was misspecified and (d) the both models were misspecified.

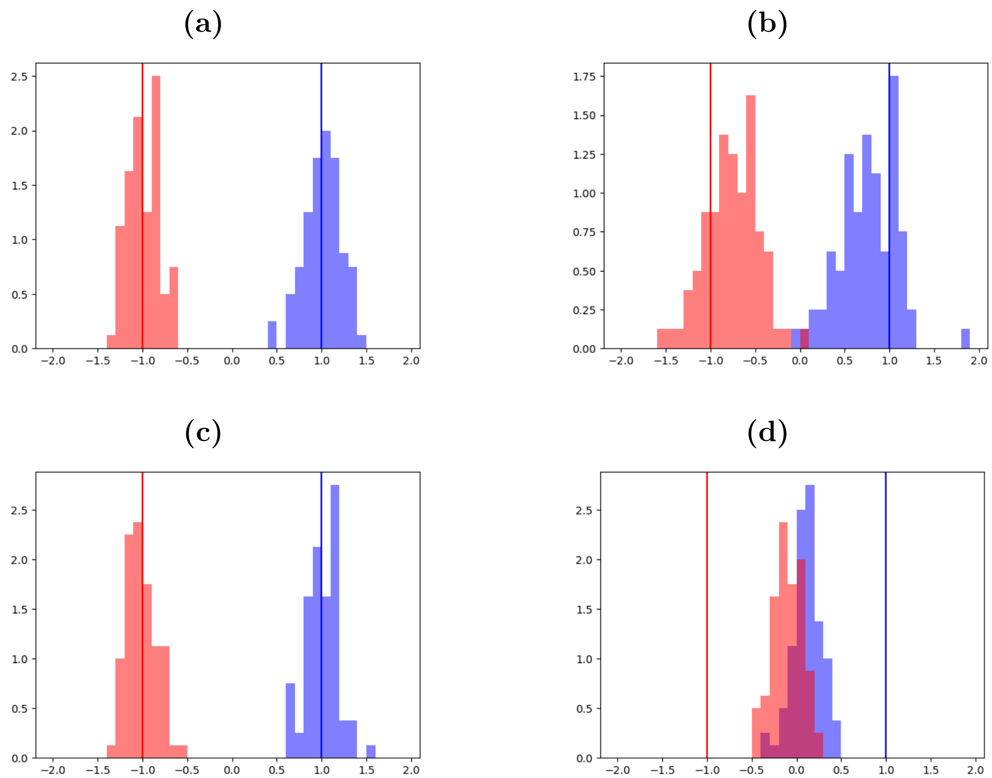


Figure C.5: **Histogram of estimated counterfactual mean by KECENI** when (a) the both models were correct, (b) the outcome regression model was misspecified, (c) the propensity score model was misspecified and (d) the both models were misspecified.

observed network $\mathcal{G}$. For TMLENET, we assumed known treatment and covariate summary statistics as $(T_i, \text{Avg}(T_{N_i\setminus i} - 0.5)$ and $(W_i, \text{Avg}(W_{N_i\setminus i}))$, respectively. For KECENI, we used a kernel smoother approach with the distance metric

$$\delta\{(t_{\mathcal{G}i}, w_{\mathcal{G}i}), (t^*_{\mathcal{G}i^*}, w_{\mathcal{G}i^*})\} \equiv \left\|(t_i, \text{Avg}(t_{N_i\setminus i} - 0.5), w_i, \text{Avg}(w_{N_i\setminus i})) - (t^*_{i^*}, \alpha_{i^*} \textstyle\sum_{j\in N_{i^*}} t^*_j, w_{i^*}, \text{Avg}(w_{N_{i^*}\setminus i^*}))\right\|_1$$

and the covariate distribution estimate $\widehat{P}_{X_{[n]}|\mathcal{G}} \equiv \prod_{i=1}^n \widehat{P}_{X_i|\mathcal{G}}$, where $\widehat{P}_{X_i|\mathcal{G}}$ is the empirical distribution of observed covariates. We considered two approaches for estimating propensity scores and outcome regressions. In the first, we assumed known parametric forms:

$$\mathbb{P}[T_i = 1|W_{N_i} = w_{N_i}] \sim \text{LogisticRegression}(w_i, \text{Avg}(w_{N_i\setminus i})),$$

$$\mathbb{E}[Y_i|T_{\mathcal{G}i} = t_{\mathcal{G}i}, W_{\mathcal{G}i} = w_{\mathcal{G}i}] \sim \text{LinearRegression}(t_i, \text{Avg}(t_{N_i\setminus i} - 0.5), w_i, \text{Avg}(w_{N_i\setminus i})),$$

yielding parametric estimates. We call this approach as the parametric KECENI and compare with AUTOGNET as the both methods assume known parametric models regarding the treatment and outcome processes. In the second approach, we assumed the forms were unknown and used non-parametric kernel smoothing, cross-validating the bandwidth. We treated treatments as i.i.d. given the summary statistics and assumed W_i vectors were i.i.d. across nodes. This approach is called as the non-parametric KECENI using known summary statistics and compared to TMLENET.

Fig. C.6 presents the estimation results of the four methods. When the structural equation models were correctly specified, AUTOGNET and TMLENET accurately estimated the average counterfactual means and treatment effects (Figs. C.6a and C.6b). The corresponding KECENI models also produced accurate estimates (Figs. C.6c and C.6d). The root-mean-square errors (RMSE) for the methods are shown in the top rows of Table C.1, where KECENI outperformed or matched the performance of its competitors.

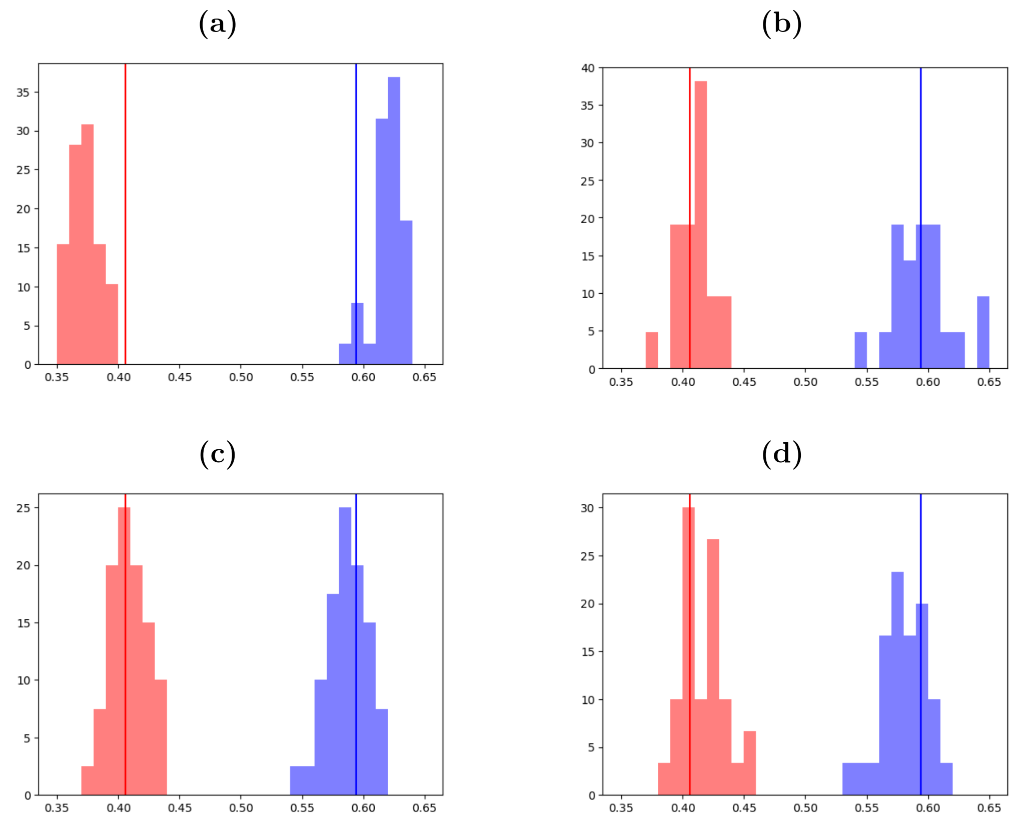


Figure C.6: **Histogram of counterfactual mean estimates under known parametric model or summary statistics by (a) AUTOGNET, (b) TMLENET, (c) parametric KECENI, and (d) non-parametric KECENI using known summary statistics.**

Method	**RMSE**
AUTOGNET, correct structural equation model	0.065
TMLENET, known summary statistics	0.035
KECENI, parametric	0.022
KECENI, non-parametric, known summary statistics	0.032
AUTOGNET, incorrect structural equation model	0.183
TMLENET, incorrect summary statistics	0.184
KECENI, fully non-parametric	0.063

Table C.1: Root-mean-square errors (RMSE) of the competing methods: AUTOGNET, TMLENET, KECENI.