Evaluating Policy Effects under Network Interference without Network Information: A Transfer Learning Approach

Tadao Hoshino[*]

October 17, 2025

Abstract

This paper develops a sensitivity analysis framework that transfers the average total treatment effect (ATTE) from source data with a fully observed network to target data whose network is completely unknown. The ATTE represents the average social impact of a policy that assigns the treatment to every individual in the dataset. We postulate a covariate-shift type assumption that both source and target datasets share the same conditional mean outcome. However, because the target network is unobserved, this assumption alone is not sufficient to pin down the ATTE for the target data. To address this issue, we consider a sensitivity analysis based on the uncertainty of the target network's degree distribution, where the extent of uncertainty is measured by the Wasserstein distance from a given reference degree distribution. We then construct bounds on the target ATTE using a linear programming-based estimator. The limiting distribution of the bound estimator is derived via the functional delta method, and we develop a wild bootstrap approach to approximate the distribution. As an empirical illustration, we revisit the social network experiment on farmers' weather insurance adoption in China by Cai et al. (2015).

---

[*]School of Political Science and Economics, Waseda University, 1-6-1 Nishi-waseda, Shinjuku-ku, Tokyo 169-8050, Japan. Email: thoshino@waseda.jp.

1 Introduction

Randomized controlled trials (RCTs) have long been the gold standard for estimating causal effects. However, it is rare that the group of individuals of interest for whom researchers or policymakers wish to know causal effects precisely coincides with the experimental sample. In many cases, the purpose of conducting an RCT is to determine in advance whether a treatment of concern yields positive effects so that it can then be introduced to a target population of real interest.

Nevertheless, the causal effects estimated from the experimental data cannot, in general, be directly applied to the non-experimental target data. To transfer estimation results from the source to the target data, we need to employ some data-adaptation techniques – *causal transfer learning*, transfer learning methods to infer causal effects in the target data, optimal treatment rules, and so forth. There is a rapidly growing body of literature developing transfer learning methods in this context (e.g., Stuart *et al.*, 2011; Hartman *et al.*, 2015; Buchanan *et al.*, 2018; Wu and Yang, 2023, among many others). For comprehensive surveys and tutorials, see, for example, Dahabreh *et al.* (2020) and Degtiar and Rose (2023).

Meanwhile, causal inference under network interference has gained increasing attention in the literature across economics, education, epidemiology, political science, and related areas. In these literature, performing an RCT has become one of major approaches for estimating treatment effects and *spillover effects* – the effects of others' treatments on one's own outcome (e.g., Bond *et al.*, 2012; Cai *et al.*, 2015; Paluck *et al.*, 2016; Carter *et al.*, 2021, among many others). While these studies have revealed both own and spillover effects in their experimental samples to some extent, policymakers ultimately may wish to extrapolate such findings to larger populations of their real concern. However, to the best of our knowledge, in contrast to the rich body of studies without network interactions, there are few, if any, studies that explicitly consider the transferability of causal effects under network interference.

The purpose of this paper is to fill this gap. Specifically, we propose a framework for inferring causal policy effects in target network data by transferring results obtained from source network data. In particular, we focus on the situation in which only individual covariates (or their distributions) are available for the target data but its network structure is completely unknown. Such situations are typical. For example, when evaluating infection prevention policies such as mandatory face-mask wearing or vaccination, the target population of interest for policymakers is the entire country. Collecting detailed network information for all citizens would be prohibitively costly, whereas demographic variables are often readily available from surveys and the census. As another example, suppose a financial company wishes to promote its insurance or savings products for farmers. Using an RCT among Chinese rice farmers, Cai *et al.* (2015) show that holding detailed information sessions significantly increases insurance take-up through social networks in each village. Given this evidence, the insurer might wish to scale up the same sessions nationwide. In that case, the target population is all farmers in the country, but information on the social networks in all villages is usually unavailable.

In order to transfer results from one sample to another, it is generally necessary to impose some similarity (or transferability) condition that links the two samples. A common condition of this kind is the so-called *covariate shift*, which assumes that the two datasets share common conditional mean potential outcome functions, while the covariate distributions may differ. When the objective is merely to estimate the conditional mean potential outcome, as is often the case in the causal inference literature, the covariate-shift assumption alone suffices.

However, from a policymaker's perspective, the goal is often to assess the expected social impact of a specific policy, rather than to estimate the conditional mean function itself. Motivated by this, we focus on the policy that assigns treatment to every unit in the dataset. Then, the causal parameter of interest in this context is the average total treatment effect (ATTE) over the target data. The total treatment effect is defined as the difference in potential outcomes when all units are assigned to treatment versus when all units are assigned to control; this is also referred to as the global treatment effect (e.g., Chin, 2019; Ugander and Yin, 2023; Faridani and Niehaus, 2024). In policy settings such as nationwide infection-prevention campaigns or the promotion of insurance services to all farmers, as in the examples above, the ATTE should be a natural target parameter.

In the absence of network information in the target data, the covariate-shift assumption alone is not sufficient to point estimate the target ATTE. To address this issue, we propose to conduct a sensitivity analysis with respect to the target network's degree distribution. Specifically, following the idea of Wasserstein distributionally robust optimization (e.g., Blanchet and Murthy, 2019; Blanchet et al., 2021; Gao and Kleywegt, 2023), we quantify the uncertainty in the target degree distribution using the Wasserstein distance from a given reference distribution. While in the literature of sensitivity analysis on distributional uncertainty, the Kullback-Leibler divergence is more commonly used (e.g., Duchi and Namkoong, 2021; Spini, 2021; Christensen and Connault, 2023), the Wasserstein distance offers several practical merits, such as allowing non-overlapping supports and computational simplicity. We show that the resulting bound estimator for the target ATTE can be obtained by solving a set of simple linear programming problems. Under regularity conditions, we derive the limiting distribution of the bound at each Wasserstein radius via the functional delta method (Fang and Santos, 2019). Moreover, we propose a dependent wild bootstrap method to approximate the distribution of the bound, following the approach of Conley et al. (2023).

As an empirical illustration, we apply our method to data from a field experiment conducted by Cai et al. (2015), which investigated how social networks affect the adoption of a weather insurance product among rural Chinese rice farmers. To evaluate the performance of our sensitivity analysis framework, we randomly partition the villages into source and target groups and estimate the bound on the ATTE for the target group. Since the full network information is available for all villages, we are able to compute the point estimate of the target ATTE and examine how the choice of the Wasserstein radius influences the coverage of the ATTE. In this setting, since both groups are essentially drawn from a common population (i.e., they are all rice farmers in the same Chinese province), the resulting bounds are shown to be very informative even under small Wasserstein radius.

Paper organization The remainder of the paper is organized as follows. Section 2 formally presents our problem setup and the parameter of primary interest, the ATTE. Section 3 provides a linear-programming characterization of the Wasserstein bound on the target ATTE. The estimation of the bound and its asymptotic properties are discussed in Section 4, where we also introduce a wild bootstrap procedure for inference. In Section 5, we conduct a set of Monte Carlo simulations to examine the finite sample performance of the proposed inference method. Section 6 presents an empirical illustration based on the data from Cai et al. (2015). Section 7 concludes. The appendix contains proofs and additional technical supplementary material.

2 Problem Setup

Consider a set of observations $\mathcal{I}$ of size $n^{\mathcal{I}} = |\mathcal{I}|$, which we label the experimental sample. Here, "experimental" is for terminology purposes only, and we do not strictly require that an experiment is performed on $\mathcal{I}$, provided that a suitable independence condition (given in Assumption 2.2) is satisfied. We also refer to $\mathcal{I}$ interchangeably as the source sample, source data, etc.

For each unit $i \in \mathcal{I}$, we observe $(Y_i, D_i, X_i, A_{i1}^{\mathcal{I}}, \ldots, A_{in^{\mathcal{I}}}^{\mathcal{I}})$, where $Y_i \in \mathbb{R}$ is the outcome of interest, $D_i \in \{0, 1\}$ is the treatment indicator, and $X_i \in \mathcal{X}$ is the vector of covariates. We assume that $\mathcal{X}$ is a finite set with $d_x = |\mathcal{X}|$. Here, $A_{ij}^{\mathcal{I}}$ denotes the (i, j)-th element of $n^{\mathcal{I}} \times n^{\mathcal{I}}$ adjacency matrix $\boldsymbol{A}^{\mathcal{I}}$. We assume that $\boldsymbol{A}^{\mathcal{I}}$ is fixed during the experiment and treat it as a non-stochastic object. In addition, $\boldsymbol{A}^{\mathcal{I}}$ does not have self-loops and may or may not be directed; i.e., $A_{ij}^{\mathcal{I}} \neq A_{ji}^{\mathcal{I}}$ is allowed. The *degree* of i, the number of network connections of i, is denoted as $G_i = \sum_{j \in \mathcal{I}} A_{ij}^{\mathcal{I}}$. When $\boldsymbol{A}^{\mathcal{I}}$ is directed, G_i is interpreted as the out-degree of i. Because $\boldsymbol{A}^{\mathcal{I}}$ is treated as non-stochastic, so is the degree G_i. Let $\mathcal{G}$ denote the finite set of possible degree values and write $d_g = |\mathcal{G}|$. For a generic $n^{\mathcal{I}}$-dimensional treatment assignment $\boldsymbol{d}^{\mathcal{I}} \in \{0, 1\}^{n^{\mathcal{I}}}$, define the number of treated peers for i by $S_i(\boldsymbol{d}^{\mathcal{I}}) := \sum_{j \in \mathcal{I}} A_{ij}^{\mathcal{I}} d_j$ and denote its realized value by $S_i = S_i(\boldsymbol{D}^{\mathcal{I}})$, where $\boldsymbol{D}^{\mathcal{I}} = \{D_i : i \in \mathcal{I}\}$.

Next, we introduce the *exposure mapping* $E_i = e(S_i, G_i)$, where $e : \mathcal{G} \times \mathcal{G} \to \mathcal{E}$ is a deterministic function chosen by the researcher. The exposure mapping summarizes peer-treatment impacts into lower dimensional statistics. Since individuals usually have their own unique social networks, it is generally impossible to identify meaningful causal parameters without dimensionality reduction of the interference structure, and exposure mapping is the standard approach in the literature (e.g., Aronow and Samii, 2017; Aronow *et al.*, 2021; Leung, 2024). Typical choices of exposure mapping include $e(S, G) = S$, $e(S, G) = S/G$, and $e(S, G) = \mathbf{1}\{S > 0\}$. In this study, we assume that the exposure mapping is correctly specified (in the sense of Assumption 2.1(i) below) as a function of (S, G).[1]

The individuals of primary interest are not those in the experimental sample but those in the target data $\mathcal{J}$, whose size is $n^{\mathcal{J}} := |\mathcal{J}|$. We assume that either the same set of covariates $X_j \in \mathcal{X}$ as in $\mathcal{I}$ is observed for all $j \in \mathcal{J}$, or the distribution of X over $\mathcal{J}$ is known (i.e., the proportion of each $x \in \mathcal{X}$ is observed). The former is reasonable if X consists of socioeconomic characteristics that policymakers can access through official surveys. Miao *et al.* (2024) consider transfer learning when only a subset of X is observed for the target data. The latter case corresponds to situations, for example, where X_j contains private information or where $\mathcal{J}$ is so large that individual-level data cannot be collected and only their distributions are publicly available to researchers. In both cases, the covariates for the target data are treated as given. As mentioned in the introduction, the network $\boldsymbol{A}^{\mathcal{J}}$ in the target data is assumed to be unobserved. Although partial knowledge of $\boldsymbol{A}^{\mathcal{J}}$ can help tighten the bounds on policy effects, we focus on the case where $\boldsymbol{A}^{\mathcal{J}}$ is completely unknown for clarity of presentation.

Now, we introduce the following transferability assumption.

---

[1] When the exposure mapping is misspecified, the resulting estimates generally exhibit biased causal interpretations; see Leung (2024). To mitigate the misspecification problem, Hoshino and Yanagi (2023) propose a specification test for the exposure mapping.

Assumption 2.1 (Transferability).   (i) For both data $\mathcal{I}$ and $\mathcal{J}$, the outcome Y is generated as

$$Y = y(D, E, G, X, \epsilon), \tag{2.1}$$

where ϵ is an unobserved disturbance term of arbitrary dimension.

(ii) For all $i \in \mathcal{I}$, $j \in \mathcal{J}$, and $(d, e, g, x) \in \{0, 1\} \times \mathcal{E} \times \mathcal{G} \times \mathcal{X}$,

$$\mu(d, e, g, x) := \mathbb{E}^{\mathcal{I}}[Y_i(d, e) \mid G_i = g, X_i = x] = \mathbb{E}^{\mathcal{J}}[Y_j(d, e) \mid G_j = g, X_j = x],$$

where $Y(d, e) := y(d, e, G, X, \epsilon)$ denotes the potential outcome when $(D, E) = (d, e)$.

Assumption 2.1 requires a certain degree of similarity between the source and target data to ensure the transferability of results. Specifically, condition (i) imposes two main restrictions. First, the exposure mapping $E = e(S, G)$ must be correctly specified.[2] Note that a correct exposure mapping is generally not unique; any mapping consistent with (2.1) can be employed. However, for estimation efficiency, it is preferable to use a "coarser" mapping (in the sense of Hoshino and Yanagi (2023)). Second, the outcome may depend on the network structure, but only through the degree G. This is motivated by possible heterogeneity in treatment and spillover effects with respect to G. For example, when $e(S, G) = S/G$ is used, we wish to distinguish between having exactly one friend, who is treated, and having many friends, all of whom are treated. In addition to G, if desired, our approach allows to include other "node-level" network covariates (e.g., centrality, local clustering) in the model, as in Lin and Xu (2017); however, the resulting prediction bounds will be much larger relative to the present specification. Also note that the model cannot incorporate "network-level" statistics as covariates; an extreme case is $Y = y_{\boldsymbol{A}}(D, E, G, X, \epsilon)$. In such cases, it is impossible to generalize results from a single network to another network without additional structural assumptions.

Condition (ii) is our main transferability assumption and parallels the covariate-shift assumption in the transfer learning literature. It states that the relationship between the potential outcomes and the covariates (G, X) is the same in the source and target data, although the distributions of these variables may differ. Given condition (i), condition (ii) holds if the conditional distributions of ϵ_i ($i \in \mathcal{I}$) and ϵ_j ($j \in \mathcal{J}$) given (G, X) are identical. This is plausible when $\mathcal{I}$ and $\mathcal{J}$ are drawn from the same population; for example, $\mathcal{I}$ is a village where an experiment was conducted, and $\mathcal{J}$ comprises all other villages in the same province, as in Cai *et al.* (2015). If we can additionally assume the additive separability: $y(D, E, G, X, \epsilon) = \mu(D, E, G, X) + \epsilon$, then condition (ii) reduces to requiring only that the error terms have mean zero conditional on (G, X).

It is important to note that merely estimating $\mu(d, e, g, x)$ may not necessarily be informative for evaluating the social impact of a specific policy (i.e., a treatment rule) among the target data. This is because in order to evaluate a given treatment rule, we need to determine not only each unit's own treatment status d, but also the exposure value e. However, the exposure e is not identifiable in the absence of network information, in general.

---

[2]Since E is fully determined by (S, G), we may rewrite (2.1) as $Y = y(D, S, G, X, \epsilon)$. A similar model specification can be found in Leung (2020). However, except when $\mathcal{G}$ is a very small set, a fully nonparametric regression on (S, G) is unrealistic, so the use of an exposure mapping will eventually be required in practice. We express our model in the form of (2.1) to highlight this point.

With this in mind, we now introduce our main causal parameter of interest. Let $Y_i(\boldsymbol{d}^{\mathcal{I}})$ denote the potential outcome when $\boldsymbol{D}^{\mathcal{I}} = \boldsymbol{d}^{\mathcal{I}}$. Observe that the two potential outcome notations are related in the following manner: $Y_i(\boldsymbol{d}^{\mathcal{I}}) = Y_i(d_i, e(S_i(\boldsymbol{d}^{\mathcal{I}}), G_i))$. Then, the total treatment effect (TTE) for unit $i \in \mathcal{I}$ is defined as

$$
\begin{aligned}
\tau_i &:= Y_i(\mathbf{1}_{n^{\mathcal{I}}}) - Y_i(\mathbf{0}_{n^{\mathcal{I}}}) \\
&= Y_i(1, e(G_i, G_i)) - Y_i(0, e(0, G_i)).
\end{aligned}
$$

The TTE for $j \in \mathcal{J}$ is similarly defined. The TTE is interpreted as the individual-level effect of a policy that assigns all units in the same network to treatment. There is a large literature on statistical inference for parameters related to the TTE (e.g., Chin, 2019; Yu *et al.*, 2022; Ugander and Yin, 2023; Faridani and Niehaus, 2024). In particular, Yu *et al.* (2022) is conceptually related to our study in that they also consider estimation under unknown networks.[3] A notable fact is that the TTE depends on the degree of i but not on the other network statistics.

Because we can observe only one potential outcome for each individual, individual TTEs are not computable. Hence, we adopt the average TTE conditioned on the degree and covariates, which we refer to as the ATTE, as our main parameter of interest:

$$
\kappa^{\mathcal{J}} := \frac{1}{n^{\mathcal{J}}} \sum_{j \in \mathcal{J}} \mathbb{E}^{\mathcal{J}}[\tau_j \mid G_j, X_j].
$$

By Assumption 2.1(ii),

$$
\begin{aligned}
\mathbb{E}^{\mathcal{J}}[\tau_j \mid G_j, X_j] &= \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} \mathbb{E}^{\mathcal{J}}[\tau_j \mid G_j = g, X_j = x]\mathbf{1}\{G_j = g, X_j = x\} \\
&= \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} (\mu(1, e(g, g), g, x) - \mu(0, e(0, g), g, x))p^{\mathcal{J}}(x)\frac{\mathbf{1}\{G_j = g, X_j = x\}}{p^{\mathcal{J}}(x)},
\end{aligned}
$$

where $p^{\mathcal{J}}(x)$ is the proportion of units with covariate value x in the target data, which is assumed to be known. When we can observe X_j for all $j \in \mathcal{J}$, we set $p^{\mathcal{J}}(x) = (n^{\mathcal{J}})^{-1} \sum_{j \in \mathcal{J}} \mathbf{1}\{X_j = x\}$. Moreover, letting $\pi^{\mathcal{J}}(g, x)$ be the conditional degree distribution given $X = x$:

$$
\pi^{\mathcal{J}}(g, x) := \frac{1}{n^{\mathcal{J}}} \sum_{j \in \mathcal{J}} \frac{\mathbf{1}\{G_j = g, X_j = x\}}{p^{\mathcal{J}}(x)},
$$

we can write the ATTE as

$$
\kappa^{\mathcal{J}} = \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} (\mu(1, e(g, g), g, x) - \mu(0, e(0, g), g, x))p^{\mathcal{J}}(x)\pi^{\mathcal{J}}(g, x). \tag{2.2}
$$

As shown here, if $p^{\mathcal{J}}$ is known, we do not need to collect individual covariates to compute $\kappa^{\mathcal{J}}$. However, since

---

[3]Their method, like ours, does not require knowledge of the network structure. However, it assumes that the direct and interference effects are additively separable and that researchers have prior knowledge of the average baseline outcome. The approach of Faridani and Niehaus (2024) also allows for settings without precise information about network connections. However, they assume that there is a known distance measure, such that spillover effects decay in a power of this distance.

G_j's are unobserved, $\pi^{\mathcal{J}}(g, x)$ is also unknown, so $\kappa^{\mathcal{J}}$ cannot be computed directly. For this issue, the next section introduces a sensitivity analysis framework with respect to the uncertainty of $\pi^{\mathcal{J}}$.

Remark 2.1 (Separating the direct and spillover effects). It is easy to see that the ATTE can be decomposed into a direct effect and a spillover effect: $\kappa^{\mathcal{J}} = \kappa_{\text{direct}}^{\mathcal{J}} + \kappa_{\text{spill}}^{\mathcal{J}}$, where

$$\kappa_{\text{direct}}^{\mathcal{J}} = \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} \big(\mu(1, e(g,g), g, x) - \mu(0, e(g,g), g, x) \big) p^{\mathcal{J}}(x) \pi^{\mathcal{J}}(g, x),$$

$$\kappa_{\text{spill}}^{\mathcal{J}} = \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} \big(\mu(0, e(g,g), g, x) - \mu(0, e(0,g), g, x) \big) p^{\mathcal{J}}(x) \pi^{\mathcal{J}}(g, x).$$

Applying our proposed method, we can construct bounds for $\kappa_{\text{direct}}^{\mathcal{J}}$ and $\kappa_{\text{spill}}^{\mathcal{J}}$ separately. However, caution is needed in interpreting these quantities. Note that $\sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} \mu(0, e(g,g), g, x) p^{\mathcal{J}}(x) \pi^{\mathcal{J}}(g, x)$ represents the average of conditional mean outcomes when all units are untreated but at the same time all of their peers are treated, which is a logical contradiction. Therefore, $\kappa_{\text{direct}}^{\mathcal{J}}$ and $\kappa_{\text{spill}}^{\mathcal{J}}$ are not, by themselves, representing "policy effects" of any implementable policy.

Lastly in this section, we discuss the identification of $\mu(d, e, g, x)$. The following assumption is plausible when the source data are obtained through an RCT.

Assumption 2.2 (Unconfoundedness). $\epsilon_i \perp\!\!\!\perp \boldsymbol{D}^{\mathcal{I}} \mid G_i, X_i$ for all $i \in \mathcal{I}$.

Assumption 2.2 ensures that, conditional on (G_i, X_i), the potential outcomes $\{Y_i(d, e)\}$ are independent of the realized (D_i, E_i). Since $\mu(d, e, g, x)$ is estimated using only the source data, this assumption is not required for the target data. Under Assumption 2.2,

$$\mathbb{E}^{\mathcal{I}}\left[Y_i \mid D_i = d, E_i = e, G_i = g, X_i = x\right] = \mathbb{E}^{\mathcal{I}}\left[Y_i(d, e) \mid D_i = d, E_i = e, G_i = g, X_i = x\right]$$
$$= \mu(d, e, g, x).$$

This implies that $\mu(d, e, g, x)$ is nonparametrically identifiable when the event $\{D_i = d, E_i = e, G_i = g, X_i = x\}$ occurs with positive probability.

3 The Linear Programming Problem

3.1 A linear-programming characterization of ATTE

As shown in (2.2), in order to compute the ATTE $\kappa^{\mathcal{J}}$ directly, we need the information of $\pi^{\mathcal{J}}(g, x)$, which is unavailable by assumption. Instead, suppose the researcher has a candidate baseline conditional degree distribution $\pi_x^* \in \mathcal{P}(\mathcal{G})$, where $\mathcal{P}(\mathcal{G})$ is the set of probability distributions whose support is a subset of $\mathcal{G}$. There are several reasonable choices for the baseline distribution. The most natural option would be to use the degree distribution in the source data $\pi_x^*(g) = \pi^{\mathcal{I}}(g, x) := (n^{\mathcal{I}})^{-1} \sum_{i \in \mathcal{I}} \mathbf{1}\{G_i = g, X_i = x\} / p^{\mathcal{I}}(x)$, where $p^{\mathcal{I}}(x) := (n^{\mathcal{I}})^{-1} \sum_{i \in \mathcal{I}} \mathbf{1}\{X_i = x\}$. This choice is particularly advocated when the source and target data come from the same population. Another possibility is to learn a link-prediction model using any method with the source data $\{(X_i, A_{i1}^{\mathcal{I}}, \ldots, A_{in^{\mathcal{I}}}^{\mathcal{I}}) : i \in \mathcal{I}\}$, obtain a predicted adjacency matrix for $\mathcal{J}$, $\widehat{\boldsymbol{A}}^{\mathcal{J}}$, and set π_x^* to

the conditional degree distribution on $\widehat{\boldsymbol{A}}^{\mathcal{J}}$. If the researcher has background knowledge about the target data from previous studies and observations, π_x^* may instead be specified a priori.

To quantify the distance between distribution functions in $\mathcal{P}(\mathcal{G})$, this paper uses the Wasserstein distance.

Definition 3.1 (q-Wasserstein distance). The q-Wasserstein distance between $\pi \in \mathcal{P}(\mathcal{G})$ and $\pi^* \in \mathcal{P}(\mathcal{G})$ is given as follows ($q \in [1, \infty)$):

$$\mathcal{W}_q(\pi, \pi^*) := \left(\min_{\Gamma \in \Pi(\pi, \pi^*)} \sum_{(u,v) \in \mathcal{G}^2} \Gamma(u,v) |u - v|^q \right)^{1/q},$$

where $\Pi(\pi, \pi^*)$ consists of all nonnegative matrices $\Gamma(u, v)$ satisfying

$$\sum_{v \in \mathcal{G}} \Gamma(u,v) = \pi^*(u), \quad \sum_{u \in \mathcal{G}} \Gamma(u,v) = \pi(v).$$

The Kullback-Leibler divergence is a popular choice in sensitivity analysis for quantifying the distance to a reference distribution (e.g., Duchi and Namkoong, 2021; Spini, 2021; Christensen and Connault, 2023). However, its greatest limitation lies in the requirement of absolute continuity, which significantly restricts the choice of reference degree distribution π_x^*. For example, if we set $\pi_x^*(g) = \pi^{\mathcal{I}}(g, x)$, then, because $\mathcal{I}$ is typically smaller than $\mathcal{J}$, the support of $\pi^{\mathcal{I}}(g, x)$ is likely to be strictly contained in that of $\pi^{\mathcal{J}}(g, x)$. Consequently, $\pi^{\mathcal{J}}$ is not absolutely continuous with respect to $\pi^{\mathcal{I}}$ and is therefore excluded from the candidate set of distributions.[4] In contrast, the Wasserstein distance can be computed for essentially any pair of distributions. Moreover, using the Wasserstein distance allows us to characterize the bounds on $\kappa^{\mathcal{J}}$ through a set of linear programming problems. For a more detailed discussion of the advantages of the Wasserstein distance over the Kullback–Leibler divergence in the context of distributionally robust optimization, see Gao and Kleywegt (2023).

Next, we define the (δ, q)-Wasserstein ball centered at π^*:

$$\mathbb{B}(\pi^*, \delta, q) := \{\pi \in \mathcal{P}(\mathcal{G}) : \mathcal{W}_q(\pi, \pi^*) \leqslant \delta\},$$

for a radius $\delta \in (0, \infty)$. Then, the lower and the upper bounds for $\kappa^{\mathcal{J}}$ at a given Wasserstein radius δ can be formulated as follows, respectively:

$$
\begin{aligned}
\underline{\kappa}_{\delta,q} &:= \sum_{x \in \mathcal{X}} \left[\min_{\pi_x \in \mathbb{B}(\pi_x^*, \delta, q)} \sum_{g \in \mathcal{G}} m(g, x) \pi_x(g) \right] \\
\overline{\kappa}_{\delta,q} &:= \sum_{x \in \mathcal{X}} \left[\max_{\pi_x \in \mathbb{B}(\pi_x^*, \delta, q)} \sum_{g \in \mathcal{G}} m(g, x) \pi_x(g) \right]
\end{aligned}
\tag{3.1}
$$

---

[4]Note that if the support of $\pi^{\mathcal{I}}(g, x)$ is a strict subset of that of $\pi^{\mathcal{J}}(g, x)$, then it is impossible to nonparametrically estimate $\mu(d, e, g, x)$ on those (g, x) values. In such cases, one eventually needs to perform inter- or extrapolation of the estimates by assuming a functional form such as in (4.1).

where

$$m(g, x) := (\mu(1, e(g, g), g, x) - \mu(0, e(0, g), g, x)) \, p^{\mathcal{J}}(x).$$

Clearly, if $\pi^{\mathcal{J}}(\cdot, x) \in \mathbb{B}(\pi_x^*, \delta, q)$ for every $x \in \mathcal{X}$, $\kappa^{\mathcal{J}} \in [\underline{\kappa}_{\delta,q}, \overline{\kappa}_{\delta,q}]$ holds. In addition, $\pi^{\mathcal{J}}(\cdot, x) \in \mathbb{B}(\pi_x^*, \delta, q)$ holds for any baseline π_x^* if we take sufficiently large δ.

Example 3.1. To illustrate the bounds (3.1), we provide a toy example here. Suppose that there are no covariates and there are only two support points for the degree distribution: $\mathcal{G} = \{0, 1\}$. m depends only on g, and we assume $m(0) \leqslant m(1)$. For the baseline degree distribution, we set $\pi^*(g) = (\alpha^*)^g(1 - \alpha^*)^{1-g}$. Then, for any Bernoulli distribution $\pi(g) = (\alpha)^g(1 - \alpha)^{1-g}$, setting $q = 1$, $\mathcal{W}_1(\pi, \pi^*) = |\alpha - \alpha^*|$ holds. Under this setup, the lower and the upper bounds can be obtained as follows:

$$\underline{\kappa}_{\delta,1} = \min_{\alpha \in [0,1] \, : \, |\alpha - \alpha^*| \leqslant \delta} (1 - \alpha)m(0) + \alpha m(1)$$

$$= m(0) + \max\{0, \alpha^* - \delta\}(m(1) - m(0))$$

$$\overline{\kappa}_{\delta,1} = \max_{\alpha \in [0,1] \, : \, |\alpha - \alpha^*| \leqslant \delta} (1 - \alpha)m(0) + \alpha m(1)$$

$$= m(0) + \min\{1, \alpha^* + \delta\}(m(1) - m(0)).$$

Hence, if the chosen δ is large enough, we will have the trivial bounds $\underline{\kappa}_{\delta,1} = m(0)$ and $\overline{\kappa}_{\delta,1} = m(1)$.

Figure 3.1 presents the areas of $[\underline{\kappa}_{\delta,1}, \overline{\kappa}_{\delta,1}]$ when $m(0) = 0$, $m(1) = 1$, and $\pi^{\mathcal{J}}(g) = (0.4)^g(0.6)^{1-g}$. The dotted horizontal line corresponds to the target parameter $\kappa^{\mathcal{J}} = 0.4$. It is evident from the left panel that, when the Wasserstein ball is centered at the true $\pi^{\mathcal{J}}(g)$, the interval $[\underline{\kappa}_{\delta,1}, \overline{\kappa}_{\delta,1}]$ contains $\kappa^{\mathcal{J}}$ for any value of $\delta > 0$. As the middle and right panels illustrate, even when the reference probability distribution π^* deviates from the true $\pi^{\mathcal{J}}$, increasing δ sufficiently large still ensures the coverage of $\kappa^{\mathcal{J}}$.
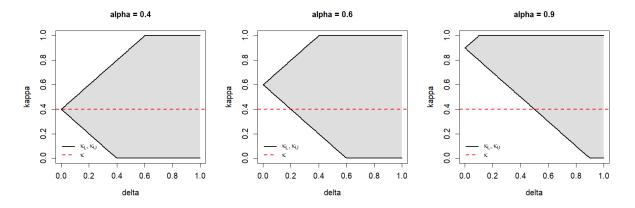

Figure 3.1: Upper and lower bounds of $\kappa^{\mathcal{J}}$

Notes: $\pi^{\mathcal{J}} = \text{Bernoulli}(0.4)$, $m(1) = 1$, and $m(0) = 0$. (Left) $\alpha^* = 0.4$. (Middle) $\alpha^* = 0.6$. (Right) $\alpha^* = 0.9$.

Hereinafter, since every minimization problem can be converted into a maximization problem by changing the sign of the objective function, we mainly focus on the computation of the upper bound $\overline{\kappa}_{\delta,q}$. For

completeness, the estimation and inference for the lower bound $\underline{\kappa}_{\delta,q}$ are summarized in Appendix B.

Our goal is to maximize the following objective function: $\sum_{x\in\mathcal{X}}\sum_{g\in\mathcal{G}} m(g,x)\pi_x(g)$ with respect to π_x subject to $\pi_x \in \mathbb{B}(\pi_x^*,\delta,q)$ for each $x \in \mathcal{X}$. Note that since $\sum_{u\in\mathcal{G}}\Gamma_x(u,v) = \pi_x(v)$, we may write $\sum_{g\in\mathcal{G}} m(g,x)\pi_x(g) = \sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)m(v,x)$. In addition, restricting the parameter space to the Wasserstein ball $\mathbb{B}(\pi_x^*,\delta,q)$ is equivalent to satisfying the following set of linear equalities and inequalities: $\sum_{v\in\mathcal{G}}\Gamma_x(u,v) = \pi_x^*(u)$, $\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)|u-v|^q \leqslant \delta^q$, and $\Gamma_x(u,v)\geqslant 0$. Consequently, the upper bound $\overline{\kappa}_{\delta,q}$ corresponds the objective value of the following linear program:

$$\text{maximize} \sum_{x\in\mathcal{X}}\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)m(v,x)$$
$$\text{subject to} \sum_{v\in\mathcal{G}}\Gamma_x(u,v) = \pi_x^*(u), \sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)\big|u-v\big|^q \leqslant \delta^q, \Gamma_x(u,v)\geqslant 0, \forall\,(x,u,v)\in\mathcal{X}\times\mathcal{G}^2 \tag{3.2}$$

Remark 3.1 (Non-uniqueness of the solution)**.** Since the parameter space for Γ_x is a compact convex subset of the probability simplex, the optimal value in the problem (3.2) exists uniquely. However, as in typical linear programming problems, the solution that attains the optimal value is not unique in general.[5] Note that once the target distribution π_x is fixed for each $x \in \mathcal{X}$, the optimal transference plan can be found uniquely when $q > 1$ (see, e.g., Theorem 1.5.1 in Panaretos and Zemel (2020)). For the uniqueness of π_x, since the objective function is linear, the solution π_x will be unique if the Wasserstein ball $\mathbb{B}(\pi_x^*,\delta,q)$ were strictly convex, which is not true in general in our setting.

Despite the non-uniqueness of the solution to problem (3.2), one might still wish to exemplify specific network structures that attain the maximum or minimum objective value. Note, however, that the degree distribution obtained from (3.2) need not be *graphic*; that is, it is not always possible to realize an arbitrary degree distribution with a simple graph. In graph theory, the Erdös–Gallai theorem provides a simple necessary and sufficient condition for a sequence of positive integers to be graphic (see, e.g., Tripathi *et al.*, 2010). When this condition is met, one can generate such graphs using some computational algorithms.[6] Meanwhile, even when the obtained degree distribution is not graphic, it is still possible to construct a graph whose expected degree sequence matches the given degree distribution, for instance, by employing the Chung-Lu model (see, e.g., 4.1.5 of Jackson, 2008).

Remark 3.2 (Interpretation of δ)**.** Interpreting the Wasserstein neighbourhood size δ in practice is a central issue in sensitivity analysis. One possible approach is to split the source data into disjoint networks. For example, in a school experiment, students' friendship networks are often disjoint across grades. Then, by computing the Wasserstein distance between the degree distribution of one grade and that of another, we obtain a typical discrepancy value $\widehat{\delta}$ between degree distributions drawn from the same population. If the source and target data are believed to come from a similar population, we may then set δ, conservatively, for example $\delta \in (0, 2\widehat{\delta}\,]$.

---

[5]For example, consider the following setup without covariates: $\mathcal{G} = \{1,2,3\}$, $(m(1),m(2),m(3)) = (1,2,3)$, $(\pi^*(1),\pi^*(2),\pi^*(3)) = (1/2,1/2,0)$, $\delta = 1$, and $q = 1$. Then, the optimal π is given for example by $(\pi(1),\pi(2),\pi(3)) = (0,1/2,1/2)$, which can be achieved by two different transference plans: $\Gamma^{(1)} = \begin{bmatrix} 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $\Gamma^{(2)} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}$.

[6]For example, the `igraph` R package offers the function `realize_degseq` that can be used for this purpose.

The number of variables in the linear program (3.2) is $d_x d_g^2$. Although the problem can be simplified by decomposing it into d_x sub-linear programs, some computational effort may still be required when d_g is large. Fortunately, the dual problem of (3.2) can be easily derived.

Proposition 3.1 (Dual problem)**.** Suppose that $m(v, x)$ is bounded uniformly in $(v, x) \in \mathcal{G} \times \mathcal{X}$. Then, for any $q \in [1, \infty)$ and $\delta > 0$,

$$\overline{\kappa}_{\delta,q} = \sum_{x \in \mathcal{X}} \left[\min_{\lambda_x \geqslant 0} \left\{ \lambda_x \delta^q + \sum_{u \in \mathcal{G}} \max_{v \in \mathcal{G}} \{ m(v, x) - \lambda_x |u - v|^q \} \pi_x^*(u) \right\} \right]. \tag{3.3}$$

Proposition 3.1 shows that the optimal value of the primal linear program (3.2) can be obtained by solving d_x separate univariate minimization problems. The derivation of (3.3) is provided in Appendix A.1. For a formal proof in a more general setting, see Theorem 1 of Blanchet and Murthy (2019) or Theorem 1 of Gao and Kleywegt (2023). As an illustration, the dual problem for Example 3.1 is presented in Appendix A.2.

3.2 Examples of degree distributions

When the researcher has prior knowledge about the network structure in the target data, the baseline π_x^* can be chosen based on it. For example, when links are believed to exist independently with each other with equal probability (i.e., an Erdős–Rényi graph), the degree distribution of a large network can be approximated by a Poisson distribution. However, many empirical networks are known to deviate substantially from the Poisson distribution (e.g., Albert and Barabási, 2002). For example, across a wide range of scientific areas, a power-law distribution (i.e., $\pi(g) \sim g^{-c}$ for some $c > 0$) often serves as a good approximation of the observed degree distribution (e.g., Kolaczyk, 2009).
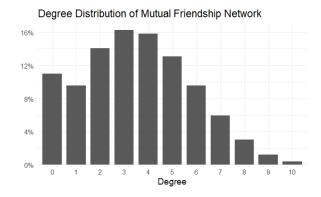
Meanwhile, in social relationship networks, extremely large degrees are rarely observed in practice. Figure 3.2 shows the degree distributions of a mutual friendship network among students and a bilateral information-exchange network among farmers, created from Paluck *et al.* (2016) and Cai *et al.* (2015), respectively. In both cases, we assume that there is a link only when the two individuals nominate each other as partners. As indicated in the left panel, the number of closest school friends peaks at about three or four. In the information exchange network among farmers, a large share of farmers has no such partner.

These observations suggest that, depending on the type of data, its degree distribution may follow a typical shape pattern such as unimodality, monotonicity, or symmetry. Explicitly imposing the shape restrictions on the candidate degree distributions can yield tighter prediction bounds. For example, in the case of monotonicity as in the right panel of Figure 3.2, we can add the linear inequality constraints $\pi_x(g_1) \geqslant \pi_x(g_2)$ for all $g_1 < g_2$ directly into the linear program (3.2).

4 Estimation and Asymptotic Properties

4.1 Estimation

The linear program in (3.2) is not feasible because $m(g, x) = (\mu(1, e(g, g), g, x) - \mu(0, e(0, g), g, x)) p^{\mathcal{J}}(x)$ is unknown. Nonparametrically estimating $\mu(d, e, g, x)$ is unrealistic due to the curse of dimensionality, except
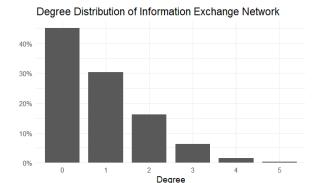
Figure 3.2: Real data examples of degree distribution

Notes: (Left) Mutual friendship network among students: data source Paluck *et al.* (2016). (Right) Mutual information exchange network among farmers: data source Cai *et al.* (2015).

when the sample size $n^{\mathcal{I}}$ is extremely large. Therefore, we would need to introduce additional functional-form restrictions on the outcome equation $y(d, e, g, x, \epsilon)$ in most applications. Although many specifications could be considered, we adopt the following varying-coefficient model as a typical candidate:

$$y(d, e, g, x, \epsilon) = w(d, e, g)^{\top} \beta(x) + \epsilon, \tag{4.1}$$

where $w : \{0, 1\} \times \mathcal{E} \times \mathcal{G} \to \mathbb{R}^{d_w}$ is a pre-specified basis function, and ϵ is a scalar error term. Then, under this specification, we only need to estimate the coefficient functions $\beta(x)$ to recover $m(g, x)$.

For the estimation of $\beta(x)$, we adopt the kernel weighted regression approach proposed by Li and Racine (2010). Recalling that the covariates X are discrete variables, we partition X into d_c-dimensional categorical variables X^c and d_o-dimensional ordered variables X^o ($d_c + d_o = d_x$). Then, define the kernel weight function for discrete covariates as follows: $L_{i,b}(x) := \prod_{j=1}^{d_c} L_{ji,b}^c(x^c) \prod_{k=1}^{d_o} L_{ki,b}^o(x^o)$, where

$$L_{ji,b}^c(x^c) := \mathbf{1}\{X_{ji}^c = x_j^c\} + \mathbf{1}\{X_{ji}^c \neq x_j^c\} b_c$$
$$L_{ki,b}^o(x^o) := \mathbf{1}\{X_{ki}^o = x_k^o\} + \mathbf{1}\{X_{ki}^o \neq x_k^o\} b_o^{|X_{ki}^o - x_k^o|},$$

$x = (x^c, x^o), x^c = (x_1^c, \dots, x_{d_c}^c), x^o = (x_1^o, \dots, x_{d_o}^o)$, with bandwidths $b = (b_c, b_o) \equiv (b_{c,n^{\mathcal{I}}}, b_{o,n^{\mathcal{I}}}) \in [0, 1]^2$. Our estimator of $\beta(x)$ is given by

$$\widehat{\beta}(x) := \left(\frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i W_i^{\top} L_{i,b}(x) \right)^{-1} \frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i Y_i L_{i,b}(x), \tag{4.2}$$

where $W_i = w(D_i, E_i, G_i)$. Then, $m(g, x)$ can be estimated by $\widehat{m}(g, x) := z(g, x)^{\top} \widehat{\beta}(x)$, where

$$z(g, x) := p^{\mathcal{J}}(x)\{w(1, e(g, g), g) - w(0, e(0, g), g)\}.$$

12

Finally, by replacing m in (3.2) by $\widehat{m}$, we can estimate $\overline{\kappa}_{\delta,q}$ by

$$\widehat{\overline{\kappa}}_{\delta,q} := \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v)\widehat{m}(v,x) \right]$$

$$\text{subject to } \sum_{v \in \mathcal{G}} \Gamma_x(u,v) = \pi_x^*(u), \ \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v)\big|u-v\big|^q \leqslant \delta^q, \Gamma_x(u,v) \geqslant 0, \forall\, (x,u,v) \in \mathcal{X} \times \mathcal{G}^2$$

Of course, one may alternatively solve the dual problem (3.3) by putting $\widehat{m}$ in the place of m. We can similarly obtain $\widehat{\underline{\kappa}}_{\delta,q}$, whose definition should be clear from the context.

4.2 Asymptotic properties

In this subsection, we derive the asymptotic distribution of $\widehat{\overline{\kappa}}_{\delta,q}$ and present a wild bootstrap procedure for approximating the distribution. We begin by stating the asymptotic distributions of $\widehat{\beta}$ and $\widehat{m}$ in the next proposition. Since these results are not quite new and depend heavily on the model specification in (4.1), all assumptions and detailed discussion are relegated to Appendix A.3. The definitions of the asymptotic covariance matrices are also provided there.

Proposition 4.1 (Asymptotic normality of $\widehat{\beta}$ and $\widehat{m}$)**.** Suppose that Assumption A.1 in Appendix A.3 holds. Then,

(i) $\sqrt{n^{\mathcal{I}}}\left(\widehat{\beta}(x) - \beta(x)\right) \xrightarrow{d} N\left(\mathbf{0}_{d_w}, (\Sigma_{\mathcal{I}}(x))^{-1}\Omega_{\mathcal{I}}(x)(\Sigma_{\mathcal{I}}(x))^{-1}\right)$ for each $x \in \mathcal{X}$,

(ii) $\sqrt{n^{\mathcal{I}}}(\widehat{\boldsymbol{m}} - \boldsymbol{m}) \xrightarrow{d} N\left(\mathbf{0}_{d_x d_g}, \boldsymbol{Z}\boldsymbol{J}_{\mathcal{I}}\boldsymbol{\Omega}_{\mathcal{I}}\boldsymbol{J}_{\mathcal{I}}\boldsymbol{Z}^{\top}\right),$

where $\boldsymbol{m} = (m(v_1, x_1), \ldots, m(v_{d_g}, x_1), \ldots, m(v_1, x_{d_x}), \ldots, m(v_{d_g}, x_{d_x}))^{\top}$, and $\widehat{\boldsymbol{m}}$ is defined similarly.

We now turn to the asymptotic distribution of $\widehat{\overline{\kappa}}_{\delta,q}$. By the fundamental theorem of linear programming, an optimal Γ_x for each $x \in \mathcal{X}$ can be found among the set of basic feasible solutions of (3.2); that is, the "corners" of the feasible set for Γ_x satisfying all equality and inequality constraints in (3.2). We denote this set by $\mathcal{B}_{\delta,q,x}$. Let $\mathcal{S}_{\delta,q,x}^*$ denote the set of maximizers:

$$\mathcal{S}_{\delta,q,x}^* := \operatorname*{argmax}_{\Gamma \in \mathcal{B}_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)m(v,x).$$

Furthermore, define $\mathbb{G} = (\mathbb{G}(v_1, x_1), \ldots, \mathbb{G}(v_{d_g}, x_1), \ldots, \mathbb{G}(v_1, x_{d_x}), \ldots, \mathbb{G}(v_{d_g}, x_{d_x}))$ as a $d_x d_g$-dimensional multivariate normal random variable with mean zero and covariance matrix $\boldsymbol{Z}\boldsymbol{J}_{\mathcal{I}}\boldsymbol{\Omega}_{\mathcal{I}}\boldsymbol{J}_{\mathcal{I}}\boldsymbol{Z}^{\top}$.

Theorem 4.1 (Asymptotic distribution of $\widehat{\overline{\kappa}}_{\delta,q}$)**.** Suppose that Assumption A.1 in Appendix A.3 holds. Then,

$$\sqrt{n^{\mathcal{I}}}\left(\widehat{\overline{\kappa}}_{\delta,q} - \overline{\kappa}_{\delta,q}\right) \xrightarrow{d} \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x \in \mathcal{S}_{\delta,q,x}^*} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v)\mathbb{G}(v,x) \right].$$

Theorem 4.1 states that the limiting distribution of the upper-bound estimator is not pivotal, but can be numerically simulated through $\mathbb{G}$ to estimate the asymptotic critical values. A similar result to our theorem

13

can be found in Bhattacharya (2009).

To estimate the critical value at a given significance level, a natural approach would proceed as follows. First, we estimate $\mathcal{S}^*_{\delta,q,x}$ by

$$\widehat{\mathcal{S}}^*_{\delta,q,x} := \left\{ \Gamma \in \mathcal{B}_{\delta,q,x} : \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)\widehat{m}(v,x) \geqslant \widehat{\overline{\kappa}}_{\delta,q,x} - a \right\}, \tag{4.3}$$

for some threshold parameter $a \equiv a_{n_{\mathcal{I}}}$ tending to zero, where $\widehat{\overline{\kappa}}_{\delta,q,x} := \max_{\Gamma \in \mathcal{B}_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)\widehat{m}(v,x)$. Second, generate independent draws $\mathbb{G}^{(r)} \sim N\left(\mathbf{0}_{d_x d_g}, \mathbf{Z}\mathbf{J}_{\mathcal{I}}\mathbf{\Omega}_{\mathcal{I}}\mathbf{J}_{\mathcal{I}}\mathbf{Z}^\top\right)$ for $r = 1, \ldots, R$, with sufficiently large R. For each draw, compute $\xi^{(r)}_{\delta,q} := \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x \in \widehat{\mathcal{S}}^*_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v)\mathbb{G}^{(r)}(v,x)\right]$. Finally, the α-level critical value is estimated by the $(1 - \alpha)$ empirical quantile of $\{\xi^{(r)}_{\delta,q} : r = 1, \ldots, R\}$.

This approach is straightforward, and a similar method has been considered in Bhattacharya (2009). However, note that to implement the second step above, we must consistently estimate the covariance matrix $\mathbf{Z}\mathbf{J}_{\mathcal{I}}\mathbf{\Omega}_{\mathcal{I}}\mathbf{J}_{\mathcal{I}}\mathbf{Z}^\top$, which typically requires a heteroscedasticity and autocorrelation consistent (HAC) estimator. In general, the accuracy of the normal approximation with a HAC-estimated covariance matrix is limited when the sample size is not large.

Alternatively to the normal approximation with a HAC-estimated covariance, following Fang and Santos (2019), this paper considers a bootstrap procedure. In particular, since the data may exhibit cross-sectional dependence, we adopt the wild bootstrap approach by Conley *et al.* (2023). Specifically, to capture the dependence among units, we consider a setup similar to Kelejian and Prucha (2007), Kim and Sun (2011), and Conley *et al.* (2023). That is, we assume that there is a socio-economic distance measure Δ_{ij} such that the dependence between i and j becomes stronger as Δ_{ij} becomes smaller.[7] Although Δ_{ij} may be unobservable, an approximation $\widetilde{\Delta}_{ij} = \Delta_{ij} + \nu_{ij}$ is available, where ν_{ij} is a measurement error. Let $K : \mathbb{R} \to [-1, 1]$ be a real-valued kernel function, and define the matrix $\mathbb{K}_{\mathcal{I}} := (K(\widetilde{\Delta}_{ij}/d))_{i,j \in \mathcal{I}}$, where $d \equiv d_{n_{\mathcal{I}}}$ is a bandwidth parameter. Further, assuming that $\mathbb{K}_{\mathcal{I}}$ is positive semidefinite,[8] obtain its eigen-decomposition $\mathbb{K}_{\mathcal{I}} = \Phi_{\mathcal{I}}\Lambda_{\mathcal{I}}\Phi_{\mathcal{I}}^\top$, where $\Lambda_{\mathcal{I}}$ is a diagonal matrix of the nonnegative eigenvalues of $\mathbb{K}_{\mathcal{I}}$, and the columns of $\Phi_{\mathcal{I}}$ are the corresponding orthonormal eigenvectors. Now, we are ready to present our bootstrap procedure.

---

[7]If it is believed that the dependence is only through network link connections, we can alternatively use Kojevnikov (2021)'s network wild bootstrap approach. The socio-economic distance-based approach considered here has the advantage of flexibility in the choice of distance measure, so that we can allow dependence of individuals even when they are apart in the network.

[8] The positive semidefinite-ness of $\mathbb{K}_{\mathcal{I}}$ is not always guaranteed and heavily depends on the choice of the kernel function K. For more detailed discussion on this issue, see, for example, Kelejian and Prucha (2007) and Conley *et al.* (2023).

Algorithm 4.1 Wild bootstrap procedure for inference on $\overline{\kappa}_{\delta,q}$

1: Estimate $\widehat{\beta}(x)$ for all $x \in \mathcal{X}$ using (4.2)

2: Compute the residual $\widehat{\epsilon}_i := Y_i - W_i^\top \widehat{\beta}(X_i)$ for all $i \in \mathcal{I}$

3: **for** $b = 1$ to B **do**

4:     Draw $\boldsymbol{\eta}^{(b)} = (\eta_1^{(b)}, \ldots, \eta_{n^{\mathcal{I}}}^{(b)}) \sim \Phi_{\mathcal{I}} \Lambda_{\mathcal{I}}^{1/2} N(\mathbf{0}_{n^{\mathcal{I}}}, I_{n^{\mathcal{I}}})$

5:     Generate a bootstrap sample $\{(W_i, Y_i^{*(b)}) : i \in \mathcal{I}\}$, where $Y_i^{*(b)} := W_i^\top \widehat{\beta}(X_i) + \eta_i^{(b)} \widehat{\epsilon}_i$

6:     Obtain $\widehat{\beta}^{*(b)}(x)$ by the kernel weighted regression of $Y_i^{*(b)}$ on W_i for all $x \in \mathcal{X}$

7:     Compute $\widehat{\overline{\kappa}}_{\delta,q}^{*(b)} := \sqrt{n^{\mathcal{I}}} \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x \in \widehat{\mathcal{S}}_{\delta,q,x}^*} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v) z(v,x)^\top (\widehat{\beta}^{*(b)}(x) - \widehat{\beta}(x)) \right]$

8: **end for**

9: Compute the empirical α quantile $\widehat{\chi}_{B,\alpha}$ of $\left\{ \sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q}^{*(b)} - \widehat{\overline{\kappa}}_{\delta,q}) : b = 1, \ldots, B \right\}$

The validity of this bootstrap procedure is stated in the next proposition. Again, the assumptions used here are all relegated to Appendix A.3.

Theorem 4.2 (Validity of the wild bootstrap). Suppose that Assumptions A.1 and A.2 in Appendix A.3 hold. Then,

$$\mathrm{Pr}^* \left(\sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q}^* - \widehat{\overline{\kappa}}_{\delta,q}) \leqslant s \right) = \mathrm{Pr} \left(\sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q} - \overline{\kappa}_{\delta,q}) \leqslant s \right) + o_P(1)$$

uniformly in $s \in \mathbb{R}$, where Pr^* denotes the conditional probability given the source data.

Theorem 4.2 implies that $\widehat{\chi}_{B,\alpha}$ is a consistent estimator for the α quantile of $\sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q} - \overline{\kappa}_{\delta,q})$ as $B \to \infty$. Therefore, the asymptotic $100(1-\alpha)\%$ confidence interval (CI) for $\overline{\kappa}_{\delta,q}$ can be obtained by

$$\mathcal{C}_{1-\alpha}(\overline{\kappa}_{\delta,q}) := \left[\widehat{\overline{\kappa}}_{\delta,q} - \frac{\widehat{\chi}_{B,1-\alpha/2}}{\sqrt{n^{\mathcal{I}}}}, \ \widehat{\overline{\kappa}}_{\delta,q} - \frac{\widehat{\chi}_{B,\alpha/2}}{\sqrt{n^{\mathcal{I}}}} \right].$$

5   Monte Carlo Simulation

In this section, we examine the finite sample performance of the proposed wild bootstrap procedure through Monte Carlo simulations. We consider the following data generating process:

$$Y_i = \sum_{\ell=1}^{6} W_{i\ell}\, b_\ell(X_{1i}, X_{2i}) + \epsilon_i, \quad i \in \mathcal{I},$$

where $(W_{i1}, \ldots, W_{i6})^\top = w(D_i, e(S_i, G_i), G_i)$, $w(d, e, g) = (1, d, e, de, \log(g+1), e\log(g+1))$, $e(s,g) = s/g$, and

$$
\begin{aligned}
& b_1(x_1, x_2) = 1, && b_2(x_1, x_2) = 1 + 0.5\Phi(x_1 + x_2), \\
& b_3(x_1, x_2) = \Phi(x_1) + \Phi(x_2), && b_4(x_1, x_2) = \Phi(x_1) + \Phi(x_2), \\
& b_5(x_1, x_2) = 0.5\exp\{-0.5(x_1 + x_2)\}, && b_6(x_1, x_2) = 0.5\exp\{-0.5(x_1 + x_2)\}.
\end{aligned}
$$

The sample size is either $n^{\mathcal{I}} = 400$ or 1200. The treatment variable and covariates are generated as follows: $D_i \sim \text{Bernoulli}(0.5)$, $X_{1i} \sim \text{Bernoulli}(0.5)$, $X_{2i} \sim \text{Unif}\{-1, 0, 1\}$, and $X_{3i} \sim N(0, 1)$. Supposing that the target population shares the same distribution of (X_1, X_2) as the source population, we set $p^{\mathcal{J}}(x) = 1/6$ for all $x \in \mathcal{X}$. In addition, we create a mismeasured version of X_{3i} as $X_{3i}^{\text{er}} = X_{3i} + \nu_i$, where $\nu_i \sim \text{Unif}[-0.3, 0.3]$.

The network $\boldsymbol{A}^{\mathcal{I}}$ is generated as follows. We first draw each unit's degree G_i independently from $\mathcal{G} = \{0, 1, 2, 3, 4\}$. Then, for each $j \neq i$, we set $A_{ij} = \boldsymbol{1}\{\text{dist}_{ij} \leqslant \overline{g}_i\}$, where $\overline{g}_i$ is the G_i-th smallest element of $\{\text{dist}_{ij} : j \in \mathcal{I} \backslash \{i\}\}$, and dist_{ij} is the Mahalanobis distance based on (X_2, X_3). We also define $\widetilde{\text{dist}}_{ij}$ as the Mahalanobis distance based on (X_2, X_3^{er}), which serves as the proxy of dist_{ij}. The error term follows a network autoregressive process $\epsilon_i = \rho \sum_{j \neq i} \widetilde{A}_{ij}^{\mathcal{I}} \epsilon_j + u_i$, where $u_i \sim N(0, 1)$ and $\widetilde{A}_{ij}^{\mathcal{I}}$ denotes the (i, j)-th element of the row-normalized version of $\boldsymbol{A}^{\mathcal{I}}$. The network autoregressive parameter is chosen from $\rho \in \{0.3, 0.5\}$.

To implement our inferential procedure, several functions and parameters need to be specified. First, the bandwidth $b = (b_c, b_o)$ in the discrete kernel regression is set as $b = c_b \cdot \widehat{b}_{n^{\mathcal{I}}}$, where c_b is a scaling constant chosen from $c_b \in \{0.5, 1, 2\}$, and $\widehat{b}_{n^{\mathcal{I}}}$ are optimal bandwidths estimated via leave-one-out cross validation in the kernel regression.[9] The solution set $\mathcal{S}_{\delta,q,x}^*$ is estimated according to (4.3), with $a = \widehat{\kappa}_{\delta,q,x} \cdot (n^{\mathcal{I}})^{-2/5}$. To assess the impact of estimating $\mathcal{S}_{\delta,q,x}^*$ on the precision of inference, we also consider an infeasible estimator that employs the true $\mathcal{S}_{\delta,q,x}^*$ in line 7 of Algorithm 4.1. For the kernel function used in the wild bootstrap, we set $K(u) = \boldsymbol{1}\{|u| \leqslant 1\}(1 - u)^2$. As a distance measure that combines information of both the covariate distance (which is mismeasured) and network proximity, we consider the following network weighted Mahalanobis distance

$$\widetilde{\Delta}_{ij} = \gamma_{ij}\widetilde{\text{dist}}_{ij}, \quad \text{where} \quad \gamma_{ij} = \boldsymbol{1}\{j \neq i\}\Phi\left(1 - \frac{1}{\text{path}_{ij} - 1}\right),$$

and path_{ij} denotes the shortest-path distance between units i and j on $\boldsymbol{A}^{\mathcal{I}}$. For example, $\gamma_{ii} = 0$, $\gamma_{ij} = 0$ if $A_{ij}^{\mathcal{I}} = 1$ (i.e., $\text{path}_{ij} = 1$), $\gamma_{ij} = 0.5$ if $\text{path}_{ij} = 2$, and so forth. Defining the distance in this way may be seen as a combination of Kojevnikov (2021) and Conley *et al.* (2023). The bandwidth d is set to be the $(c_d \max_{i \in \mathcal{I}} G_i/n^{\mathcal{I}})$ quantile of $\{\widetilde{\Delta}_{ij} : i, j \in \mathcal{I}, i \neq j\}$, where c_d is chosen from $c_d \in \{2, 4, 6\}$. For comparison, we also compute the empirical coverage for the estimator that ignores cross-sectional dependence (i.e., setting $\mathbb{K}_{\mathcal{I}} = I_{n^{\mathcal{I}}}$).

In this analysis, we perform the wild bootstrap to simulate the distribution of $T_{\delta,q} := \sqrt{n^{\mathcal{I}}}(\widehat{\kappa}_{\delta,q} - \overline{\kappa}_{\delta,q})$, where we consider the Wasserstein ball with $q = 2$ and four radius values $\delta \in \{0.05, 0.1, 0.2, 0.5\}$ centered at the uniform reference distribution $\pi_x^*(g) = 1/5$, for all $g \in \mathcal{G}$ and $x \in \mathcal{X}$. The number of bootstrap replications is set to $B = 500$, and we compute the 95% and 99% bootstrap CIs for $T_{\delta,q}$, checking whether it is contained in each case. This procedure is repeated for 500 Monte Carlo replications to compute the empirical coverage probabilities.

Tables 1 and 2 report the results for $\rho = 0.3$ and $\rho = 0.5$, respectively. The main findings are as follows. When the dependence of the error terms is relatively weak ($\rho = 0.3$), our wild bootstrap method performs well overall. In particular, when the sample size is large, the empirical coverage rates are satisfactorily close to the nominal levels in almost all cases. The choice of the two bandwidths, one in the discrete kernel regression

---

[9]We used the `npscoef` function in the `np` package. Since performing the cross validation in every iteration is computationally too demanding, we computed the optimal bandwidths for 20 burn-in samples in each setting and used their averages as $\widehat{b}_{n^{\mathcal{I}}}$.

and the other in the dependent wild bootstrap, has relatively a small influence on the results. Moreover, the effect of estimating $\mathcal{S}^*_{\delta,q,x}$ on the coverage accuracy is almost negligible, which is consistent with our theory. By contrast, when the network dependence among error terms is ignored, the resulting CIs are clearly too narrow, especially for smaller samples. As the magnitude of network dependence increases ($\rho = 0.5$), this undercoverage becomes more serious. Even for our dependent wild bootstrap, a slight loss of coverage is observed for smaller samples, but the accuracy improves as the sample size grows. Similar results to ours have been reported in previous studies, such as Kim and Sun (2011). Overall, these results confirm that the proposed wild bootstrap procedure performs reliably and is relatively insensitive to the choice of tuning parameters, at least for this particular DGP. Ideally, a fully data-driven method for selecting these factors could be developed, but we leave this for future research.

6 An Empirical Illustration

As an empirical illustration, we apply our bound estimator and wild bootstrap method to the data on farmers' insurance adoption in Cai *et al.* (2015). Cai *et al.* (2015) conducted a field experiment to estimate the effect of providing intensive information sessions about the weather insurance on farmers' insurance take-up decisions. In the experiment, four types of sessions were provided: first round simple, first round intensive, second round simple, and second round intensive. In each round, the simple sessions only explain the insurance contract, while intensive sessions cover all information provided in simple sessions and additionally provide financial education to help farmers understand how the insurance works and its benefits. The farmers were randomly assigned to each session according to household size and area of rice production per capita, which we denote by *hhsize* and *rice*, respectively.

In this analysis, the outcome variable is $Y_i \in \{0,1\}$, which indicates whether farmer i decided to buy the weather insurance after attending the session. Let $int_i \in \{0,1\}$ denote whether i was assigned to an intensive session, and *first*$_i \in \{0,1\}$ denote whether i was assigned to the first round session. The spillover effects matter only for the second round participants, as they can receive information from the first round participants. Then, as own treatment indicator, we set $D_i = int_i$. Meanwhile, reflecting the experimental design, the exposure variable is defined as follows:

$$E_i = (1 - \textit{first}_i) \sum_{j \in \mathcal{I}} A^{\mathcal{I}}_{ij} \, int_j \, \textit{first}_j / G_i,$$

where $A^{\mathcal{I}}_{ij}$ indicates whether i and j are mutual information-exchange partners. The TTE in this context is given by $\tau_i = Y_i(1,1) - Y_i(0,0)$. τ_i interpreted as the individual policy effect for the policy that provides intensive session for all farmers, and they all have enough time to exchange their information with their partners.

As the covariates, we use $X_{1i} = \mathbf{1}\{hhsize_i \geqslant \text{Med}[hhsize_i]\}$ and $X_{2i} = \mathbf{1}\{rice_i \geqslant \text{Med}[rice_i]\}$, where Med denotes the empirical median. For the basis function w, we consider the following form: $w(d,e,g) = (1, d, e, de, \log(g+1), e\log(g+1))$.

To evaluate the performance of our proposed method in a realistic setting, we randomly divide the original data into two groups. Specifically, since the data consist of 47 administrative villages, we randomly select 17 villages as the source sample ($n^{\mathcal{I}} = 1514$) and use the remaining 30 villages as the target sample ($n^{\mathcal{J}} = 3351$).

Table 1: Empirical coverage probabilities: $\rho = 0.3$

$n^{\mathcal{I}}$	c_b	$\mathcal{S}^*$	$\mathbb{K}$	95% CI				99% CI			
				$\delta = 0.05$	0.1	0.2	0.5	$\delta = 0.05$	0.1	0.2	0.5
400	0.5	est	$c_d = 2$	0.942	0.942	0.934	0.930	0.982	0.982	0.978	0.980
			$c_d = 4$	0.942	0.940	0.940	0.926	0.984	0.984	0.984	0.982
			$c_d = 6$	0.934	0.934	0.930	0.926	0.978	0.976	0.978	0.976
			$\mathbb{K} = I$	0.854	0.852	0.854	0.856	0.946	0.948	0.948	0.944
		true	$c_d = 2$	0.944	0.944	0.944	0.942	0.982	0.982	0.980	0.984
			$c_d = 4$	0.942	0.938	0.940	0.938	0.984	0.984	0.984	0.984
			$c_d = 6$	0.936	0.934	0.934	0.936	0.978	0.978	0.980	0.982
			$\mathbb{K} = I$	0.854	0.854	0.854	0.866	0.946	0.944	0.940	0.946
	1.0	est	$c_d = 2$	0.948	0.946	0.938	0.940	0.978	0.978	0.982	0.986
			$c_d = 4$	0.954	0.954	0.942	0.948	0.986	0.986	0.990	0.984
			$c_d = 6$	0.942	0.942	0.940	0.938	0.984	0.984	0.978	0.982
			$\mathbb{K} = I$	0.862	0.862	0.860	0.856	0.950	0.946	0.942	0.954
		true	$c_d = 2$	0.948	0.948	0.946	0.950	0.978	0.978	0.980	0.984
			$c_d = 4$	0.952	0.952	0.948	0.956	0.986	0.986	0.986	0.986
			$c_d = 6$	0.942	0.942	0.942	0.946	0.984	0.984	0.984	0.988
			$\mathbb{K} = I$	0.862	0.864	0.862	0.866	0.952	0.952	0.952	0.950
	2.0	est	$c_d = 2$	0.946	0.946	0.940	0.948	0.986	0.986	0.986	0.988
			$c_d = 4$	0.960	0.958	0.954	0.954	0.988	0.988	0.988	0.986
			$c_d = 6$	0.952	0.952	0.942	0.946	0.980	0.980	0.980	0.984
			$\mathbb{K} = I$	0.876	0.874	0.870	0.864	0.952	0.952	0.952	0.954
		true	$c_d = 2$	0.946	0.946	0.946	0.952	0.986	0.986	0.988	0.986
			$c_d = 4$	0.960	0.960	0.960	0.960	0.988	0.988	0.988	0.990
			$c_d = 6$	0.950	0.950	0.948	0.954	0.980	0.980	0.980	0.982
			$\mathbb{K} = I$	0.876	0.874	0.874	0.874	0.952	0.954	0.954	0.960
1200	0.5	est	$c_d = 2$	0.954	0.954	0.952	0.956	0.992	0.992	0.992	0.988
			$c_d = 4$	0.960	0.960	0.956	0.954	0.994	0.994	0.994	0.988
			$c_d = 6$	0.964	0.962	0.958	0.962	0.992	0.992	0.992	0.990
			$\mathbb{K} = I$	0.890	0.890	0.878	0.884	0.962	0.962	0.950	0.954
		true	$c_d = 2$	0.954	0.954	0.956	0.964	0.992	0.992	0.992	0.990
			$c_d = 4$	0.960	0.960	0.960	0.960	0.994	0.994	0.994	0.990
			$c_d = 6$	0.966	0.966	0.966	0.964	0.992	0.992	0.992	0.990
			$\mathbb{K} = I$	0.890	0.888	0.888	0.892	0.962	0.962	0.960	0.964
	1.0	est	$c_d = 2$	0.956	0.954	0.952	0.956	0.992	0.992	0.992	0.990
			$c_d = 4$	0.962	0.960	0.956	0.952	0.994	0.994	0.994	0.992
			$c_d = 6$	0.964	0.962	0.960	0.964	0.992	0.992	0.992	0.992
			$\mathbb{K} = I$	0.894	0.892	0.884	0.888	0.966	0.964	0.958	0.954
		true	$c_d = 2$	0.956	0.956	0.960	0.964	0.992	0.992	0.992	0.992
			$c_d = 4$	0.962	0.962	0.960	0.958	0.994	0.994	0.994	0.992
			$c_d = 6$	0.964	0.964	0.966	0.962	0.992	0.992	0.992	0.990
			$\mathbb{K} = I$	0.894	0.896	0.892	0.900	0.966	0.966	0.964	0.964
	2.0	est	$c_d = 2$	0.962	0.962	0.956	0.958	0.992	0.992	0.990	0.988
			$c_d = 4$	0.966	0.968	0.958	0.954	0.992	0.992	0.992	0.992
			$c_d = 6$	0.962	0.962	0.962	0.956	0.992	0.992	0.992	0.990
			$\mathbb{K} = I$	0.898	0.900	0.890	0.890	0.970	0.970	0.966	0.968
		true	$c_d = 2$	0.962	0.962	0.960	0.966	0.992	0.992	0.992	0.994
			$c_d = 4$	0.966	0.968	0.970	0.964	0.992	0.994	0.994	0.994
			$c_d = 6$	0.962	0.962	0.962	0.964	0.994	0.992	0.992	0.992
			$\mathbb{K} = I$	0.900	0.902	0.898	0.902	0.970	0.970	0.970	0.968

NOTE: "est" and "true" in the column $\mathcal{S}^*$ indicate that the estimated and true $\mathcal{S}^*_{\delta,q,x}$ are used, respectively. In the column $\mathbb{K}$, "$\mathbb{K} = I$" indicates that network dependence is ignored in this case.

Table 2: Empirical coverage probabilities: $\rho = 0.5$

$n^{\mathcal{I}}$	c_b	$\mathcal{S}^*$	Estimator	95% CI				99% CI			
				$\delta = 0.05$	0.1	0.2	0.5	$\delta = 0.05$	0.1	0.2	0.5
400	0.5	est	$c_d = 2$	0.934	0.932	0.922	0.918	0.976	0.976	0.974	0.978
			$c_d = 4$	0.942	0.940	0.930	0.924	0.984	0.984	0.980	0.978
			$c_d = 6$	0.930	0.930	0.914	0.912	0.974	0.974	0.972	0.972
			$\mathbb{K} = I$	0.804	0.802	0.788	0.788	0.902	0.896	0.892	0.902
		true	$c_d = 2$	0.934	0.934	0.932	0.938	0.976	0.976	0.976	0.976
			$c_d = 4$	0.942	0.942	0.944	0.938	0.984	0.984	0.982	0.980
			$c_d = 6$	0.930	0.930	0.930	0.924	0.974	0.974	0.974	0.976
			$\mathbb{K} = I$	0.804	0.802	0.804	0.802	0.902	0.902	0.904	0.916
	1.0	est	$c_d = 2$	0.940	0.934	0.930	0.932	0.976	0.976	0.974	0.978
			$c_d = 4$	0.946	0.940	0.930	0.934	0.984	0.984	0.982	0.978
			$c_d = 6$	0.932	0.930	0.922	0.920	0.978	0.976	0.978	0.974
			$\mathbb{K} = I$	0.810	0.808	0.804	0.788	0.904	0.900	0.896	0.906
		true	$c_d = 2$	0.940	0.940	0.936	0.946	0.976	0.976	0.974	0.976
			$c_d = 4$	0.944	0.946	0.944	0.940	0.984	0.984	0.984	0.980
			$c_d = 6$	0.932	0.932	0.934	0.936	0.978	0.978	0.978	0.980
			$\mathbb{K} = I$	0.810	0.812	0.816	0.804	0.904	0.904	0.906	0.920
	2.0	est	$c_d = 2$	0.942	0.944	0.936	0.938	0.982	0.982	0.980	0.978
			$c_d = 4$	0.946	0.944	0.940	0.936	0.984	0.986	0.988	0.988
			$c_d = 6$	0.940	0.940	0.934	0.928	0.980	0.980	0.980	0.980
			$\mathbb{K} = I$	0.830	0.826	0.818	0.808	0.922	0.922	0.912	0.914
		true	$c_d = 2$	0.942	0.944	0.942	0.950	0.982	0.982	0.982	0.984
			$c_d = 4$	0.946	0.948	0.948	0.948	0.984	0.986	0.988	0.988
			$c_d = 6$	0.940	0.940	0.938	0.946	0.980	0.980	0.980	0.978
			$\mathbb{K} = I$	0.830	0.830	0.824	0.824	0.922	0.922	0.922	0.922
1200	0.5	est	$c_d = 2$	0.946	0.944	0.942	0.944	0.988	0.988	0.988	0.990
			$c_d = 4$	0.952	0.952	0.946	0.942	0.990	0.990	0.990	0.988
			$c_d = 6$	0.958	0.956	0.948	0.952	0.994	0.994	0.992	0.986
			$\mathbb{K} = I$	0.844	0.836	0.834	0.822	0.932	0.932	0.920	0.930
		true	$c_d = 2$	0.952	0.948	0.946	0.958	0.988	0.988	0.988	0.988
			$c_d = 4$	0.952	0.952	0.952	0.958	0.990	0.990	0.990	0.988
			$c_d = 6$	0.958	0.956	0.958	0.960	0.994	0.994	0.994	0.986
			$\mathbb{K} = I$	0.844	0.844	0.844	0.828	0.934	0.934	0.940	0.938
	1.0	est	$c_d = 2$	0.950	0.946	0.940	0.942	0.988	0.988	0.988	0.990
			$c_d = 4$	0.950	0.950	0.948	0.944	0.990	0.990	0.990	0.990
			$c_d = 6$	0.962	0.960	0.954	0.950	0.994	0.994	0.992	0.986
			$\mathbb{K} = I$	0.850	0.848	0.840	0.832	0.936	0.936	0.928	0.928
		true	$c_d = 2$	0.952	0.952	0.952	0.960	0.988	0.988	0.988	0.988
			$c_d = 4$	0.952	0.954	0.956	0.960	0.990	0.990	0.990	0.990
			$c_d = 6$	0.962	0.960	0.960	0.964	0.994	0.994	0.994	0.986
			$\mathbb{K} = I$	0.850	0.848	0.842	0.840	0.936	0.938	0.944	0.936
	2.0	est	$c_d = 2$	0.960	0.958	0.946	0.944	0.988	0.988	0.988	0.988
			$c_d = 4$	0.954	0.954	0.954	0.948	0.992	0.990	0.990	0.990
			$c_d = 6$	0.960	0.960	0.958	0.952	0.994	0.994	0.994	0.990
			$\mathbb{K} = I$	0.858	0.860	0.848	0.846	0.940	0.938	0.932	0.930
		true	$c_d = 2$	0.960	0.960	0.954	0.962	0.988	0.988	0.988	0.988
			$c_d = 4$	0.954	0.954	0.958	0.960	0.992	0.992	0.992	0.990
			$c_d = 6$	0.960	0.960	0.962	0.962	0.994	0.994	0.994	0.990
			$\mathbb{K} = I$	0.858	0.856	0.850	0.844	0.940	0.940	0.944	0.934

NOTE: "est" and "true" in the column $\mathcal{S}^*$ indicate that the estimated and true $\mathcal{S}^*_{\delta,q,x}$ are used, respectively. In the column $\mathbb{K}$, "$\mathbb{K} = I$" indicates that network dependence is ignored in this case.

We then compute the ATTE bounds for the target sample by transferring the estimates obtained from the source sample. In this analysis, because the network structure in the target data is actually known, we can directly compute $\pi^{\mathcal{J}}(g, x)$ for each (g, x). This enables us to approximately assess the coverage property of our bound estimator under different choices of the Wasserstein radius δ. In addition, to illustrate the effect of increasing the size of source sample, we also consider a case in which five additional villages are included in the source sample ($n^{\mathcal{I}} = 1812$).

To determine a plausible range for δ, we compute the 2-Wasserstein distance between the degree distributions of the 17 source villages and 30 target villages for each covariate group (throughout this analysis, we use the 2-Wasserstein distance). The results are reported in Figure 6.1. In the figure, "LL" stands for the subsample with $(X_1 = 0, X_2 = 0)$, "LU" for $(X_1 = 0, X_2 = 1)$, and so on. The number shown at the top of each panel indicates the computed 2-Wasserstein distance. From these results, we observe that when the source and target data are drawn from the same population, the typical 2-Wasserstein distance is roughly 0.25 or so.

Figure 6.1: Conditional degree distributions

Based on the above finding, we slightly conservatively set the region for δ as $\delta \in (0, 0.6]$. As the baseline conditional degree distribution, we set $\pi_x^*(g) = \pi^{\mathcal{I}}(g, x)$ (see Figure 6.1). The distance measure $\widetilde{\Delta}_{ij}$ is computed using the Mahalanobis distance based on age, gender, acreage of rice production, and household size, weighted by the path length as in Section 5. Furthermore, when i and j belong to different villages, we set $\widetilde{\Delta}_{ij} = \infty$. All other setups for estimation and bootstrap inference follow those used in the simulation analysis in Section 5.

We report our bound estimation results in Figure 6.2. In the figure, the left and right panels correspond to the cases with 17 and 22 villages in the source sample, respectively. The upper shaded area represents the upper half of the 95% CI for the upper bound, and the lower shaded area shows the lower half of the 95% CI

20

for the lower bound. The dashed line indicates the (infeasible) point estimate of $\kappa^{\mathcal{J}}$, computed using the true conditional degree distribution $\pi^{\mathcal{J}}$ in the target data. Since the source and target datasets come from essentially the same population and $\pi^{\mathcal{I}} \approx \pi^{\mathcal{J}}$ holds, as shown in Figure 6.1, our ATTE bound is highly informative, successfully covering the estimated $\kappa^{\mathcal{J}}$ even for small values of δ. In addition, the estimated worst case bound does not fall below zero for any $\delta \leqslant 0.6$ for both sample sizes. Regarding the impact of increasing the size of the source sample, we can observe that the length of the CI for each δ is significantly narrower in the right panel than in the left. When 22 villages are used for the source sample, the lower 95% bound remains positive for almost the entire range of δ values considered here. From these findings, we may state that the ATTE is likely positive for the target data with a certain degree of confidence.

Figure 6.2: Estimated coefficient functions

(a) Sensitivity analysis result (17 villages: $n^{\mathcal{I}} = 1514$)   (b) Sensitivity analysis result (22 villages: $n^{\mathcal{I}} = 1812$)

7   Conclusion

This paper proposes a transfer learning framework for policy evaluation in settings where the network structure of the target data is unobserved. Following the existing literature, we adopt a covariate-shift type assumption to estimate conditional mean potential outcomes using experimental source data. However, in the presence of spillover effects, this assumption alone is insufficient to evaluate a specific policy in the target data due to the lack of network information. To address this issue, we propose a sensitivity analysis approach that quantifies the uncertainty in the unobserved target network using the Wasserstein distance between degree distributions. The resulting bounds on the policy effect can be computed by solving a set of linear programming problems. We derive the asymptotic distribution of the bound estimator via the functional delta method and develop a wild bootstrap procedure for inference. As an empirical application, we use the experimental data from Cai et al. (2015) to illustrate the practical implementation and empirical usefulness of the proposed method.

Several limitations should be noted. First, the covariate-shift assumption may be violated if the source and target data are too dissimilar. Second, the current model specification assumes that network effects can be entirely captured by node-level covariates, not allowing any network-level heterogeneity. Third, the proposed framework cannot be directly applied to the evaluation of more complex policies that assign treatment based

on individual characteristics or network positions. Finally, as with any sensitivity analysis, the interpretation and selection of the uncertainty parameter (δ in our context) remain open questions.

Acknowledgments

Appendix

A Technical Appendix

A.1 Derivation of the dual problem (3.3)

Recall that our primal linear program is formulated as follows:

maximize $\displaystyle\sum_{x\in\mathcal{X}}\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)m(v,x)$

subject to $\displaystyle\sum_{v\in\mathcal{G}}\Gamma_x(u,v)=\pi_x^*(u),\ \sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)\big|u-v\big|^q\leqslant\delta^q,\Gamma_x(u,v)\geqslant 0,\forall\,(x,u,v)\in\mathcal{X}\times\mathcal{G}^2$

Now, introduce dual variables $n(u,x)$ for the equality constraint $\sum_{v\in\mathcal{G}}\Gamma_x(u,v)=\pi_x^*(u)$ for each $(x,u)\in\mathcal{X}\times\mathcal{G}$ and $\lambda_x\geqslant 0$ for the inequality constraint $\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)\big|u-v\big|^q\leqslant\delta^q$ for each $x\in\mathcal{X}$. Then, the Lagrangian function is given by

$$L(\Gamma,n,\lambda)=\sum_{x\in\mathcal{X}}\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)m(v,x)-\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{G}}n(u,x)\left(\sum_{v\in\mathcal{G}}\Gamma_x(u,v)-\pi_x^*(u)\right)$$

$$-\sum_{x\in\mathcal{X}}\lambda_x\left(\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)\big|u-v\big|^q-\delta^q\right)$$

$$=\sum_{x\in\mathcal{X}}\lambda_x\delta^q+\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{G}}n(u,x)\pi_x^*(u)+\sum_{x\in\mathcal{X}}\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)\left\{m(v,x)-n(u,x)-\lambda_x|u-v|^q\right\}.$$

Define the dual function by

$$D(n,\lambda):=\sup_{\Gamma\geqslant 0}L(\Gamma,n,\lambda)$$

$$=\sum_{x\in\mathcal{X}}\lambda_x\delta^q+\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{G}}n(u,x)\pi_x^*(u)+\sup_{\Gamma\geqslant 0}\sum_{x\in\mathcal{X}}\sum_{u,v\in\mathcal{G}^2}\Gamma_x(u,v)\left\{m(v,x)-n(u,x)-\lambda_x|u-v|^q\right\}.$$

If the following inequality is not satisfied

$$m(v,x)-n(u,x)-\lambda_x|u-v|^q\leqslant 0 \tag{A.1}$$

for some (x,u,v), then we can set the corresponding element of $\Gamma_x(u,v)$ arbitrarily large, resulting in an unbounded $D(n,\lambda)$. Thus, whenever (A.1) is satisfied, we must have

$$D(n,\lambda)=\sum_{x\in\mathcal{X}}\left(\lambda_x\delta^q+\sum_{u\in\mathcal{G}}n(u,x)\pi_x^*(u)\right).$$

To minimize the dual function $D(n, \lambda)$, in view of (A.1), we can profile out n from $D(n, \lambda)$ by setting

$$n(u, x) = \max_{v \in \mathcal{G}}\{m(v, x) - \lambda_x |u - v|^q\}.$$

Plugging this into $\lambda_x \delta^q + \sum_{u \in \mathcal{G}} n(u, x)\pi_x^*(u)$ gives the objective function in (3.3).

A.2 The dual problem of Example 3.1

The dual problem of Example 3.1 is as follows: $\min_{\lambda \geqslant 0} D(\lambda)$, where

$$D(\lambda) := \left\{ \lambda\delta + \sum_{u \in \mathcal{G}} \left[\max_{v \in \mathcal{G}}\{m(v) - \lambda|u - v|\} \right]\pi^*(u) \right\}.$$

By direct calculation,

$$\sum_{u \in \mathcal{G}} \left[\max_{v \in \mathcal{G}}\{m(v) - \lambda|u - v|\} \right]\pi^*(u) = \max\{m(0) - \lambda, m(1)\}\alpha^* + \max\{m(0), m(1) - \lambda\}(1 - \alpha^*)$$

$$= m(1)\alpha^* + \max\{m(0), m(1) - \lambda\}(1 - \alpha^*).$$

Now, when $\lambda > m(1) - m(0)$,

$$\min_{\lambda > m(1) - m(0)} D(\lambda) = \min_{\lambda > m(1) - m(0)} \{\lambda\delta + m(1)\alpha^* + m(0)(1 - \alpha^*)\}$$

$$> m(0) + (m(1) - m(0))(\alpha^* + \delta).$$

Meanwhile, if $\lambda \leqslant m(1) - m(0)$,

$$\min_{0 \leqslant \lambda \leqslant m(1) - m(0)} D(\lambda) = \min_{0 \leqslant \lambda \leqslant m(1) - m(0)} \{\lambda\delta + m(1) - \lambda(1 - \alpha^*)\}.$$

Hence, if $\delta \geqslant (1 - \alpha^*)$, we should set $\lambda = 0$, leading to $\min_{0 \leqslant \lambda \leqslant m(1) - m(0)} D(\lambda) = m(1)$. On the other hand, if $\delta < (1 - \alpha^*)$, the optimal λ is given by $m(1) - m(0)$, leading to $\min_{0 \leqslant \lambda \leqslant m(1) - m(0)} D(\lambda) = m(0) + (m(1) - m(0))(\alpha^* + \delta)$. Then, it is straightforward to see that $\min_{\lambda \geqslant 0} D(\lambda) = \overline{\kappa}_{\delta,1}$ holds.

A.3 Proofs of Proposition 4.1, Theorem 4.1, and Theorem 4.2

Throughout the proofs, we use the following notations:

$$\Sigma_{n^{\mathcal{I}}}(x) := \frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} \mathbb{E}[W_i W_i^\top \mathbf{1}\{X_i = x\}]$$

$$\Omega_{n^{\mathcal{I}}}(x) := \frac{1}{n^{\mathcal{I}}} \sum_{i,i' \in \mathcal{I}} \mathbb{E}\left[W_i W_{i'}^\top \epsilon_i \epsilon_{i'} \mathbf{1}\{X_i = X_{i'} = x\}\right]$$

$$\Omega_{n^{\mathcal{I}}}(x_1, x_2) := \frac{1}{n^{\mathcal{I}}} \sum_{i,i' \in \mathcal{I}} \mathbb{E}\left[W_i W_{i'}^\top \epsilon_i \epsilon_{i'} \mathbf{1}\{X_i = x_1, X_{i'} = x_2\}\right]$$

$$\widehat{\boldsymbol{J}}_{n^{\mathcal{I}}} := \begin{pmatrix} \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}}W_iW_i^\top L_{i,b}(x_1)\right)^{-1} & \mathbf{0}_{d_w\times d_w} & \cdots & \mathbf{0}_{d_w\times d_w} \\ \mathbf{0}_{d_w\times d_w} & \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}}W_iW_i^\top L_{i,b}(x_2)\right)^{-1} & \cdots & \mathbf{0}_{d_w\times d_w} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d_w\times d_w} & \mathbf{0}_{d_w\times d_w} & \cdots & \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}}W_iW_i^\top L_{i,b}(x_{d_x})\right)^{-1} \end{pmatrix}$$

$$\boldsymbol{J}_{n^{\mathcal{I}}} := \begin{pmatrix} (\Sigma_{n^{\mathcal{I}}}(x_1))^{-1} & \mathbf{0}_{d_w\times d_w} & \cdots & \mathbf{0}_{d_w\times d_w} \\ \mathbf{0}_{d_w\times d_w} & (\Sigma_{n^{\mathcal{I}}}(x_2))^{-1} & \cdots & \mathbf{0}_{d_w\times d_w} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d_w\times d_w} & \mathbf{0}_{d_w\times d_w} & \cdots & (\Sigma_{n^{\mathcal{I}}}(x_{d_x}))^{-1} \end{pmatrix}$$

$$\boldsymbol{\Omega}_{n^{\mathcal{I}}} := \begin{pmatrix} \Omega_{n^{\mathcal{I}}}(x_1) & \Omega_{n^{\mathcal{I}}}(x_1,x_2) & \cdots & \Omega_{n^{\mathcal{I}}}(x_1,x_{d_x}) \\ \Omega_{n^{\mathcal{I}}}(x_2,x_1) & \Omega_{n^{\mathcal{I}}}(x_2) & \cdots & \Omega_{n^{\mathcal{I}}}(x_2,x_{d_x}) \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{n^{\mathcal{I}}}(x_{d_x},x_1) & \Omega_{n^{\mathcal{I}}}(x_{d_x},x_2) & \cdots & \Omega_{n^{\mathcal{I}}}(x_{d_x}) \end{pmatrix}$$

and

$$\underbrace{Z(x)}_{d_g\times d_w} := \begin{pmatrix} z(g_1,x)^\top \\ z(g_2,x)^\top \\ \vdots \\ z(g_{d_g},x)^\top \end{pmatrix}, \qquad \underbrace{\boldsymbol{Z}}_{d_gd_x\times d_wd_x} := \begin{pmatrix} Z(x_1) & \mathbf{0}_{d_g\times d_w} & \cdots & \mathbf{0}_{d_g\times d_w} \\ \mathbf{0}_{d_g\times d_w} & Z(x_2) & \cdots & \mathbf{0}_{d_g\times d_w} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d_g\times d_w} & \mathbf{0}_{d_g\times d_w} & \cdots & Z(x_{d_x}) \end{pmatrix}.$$

Moreover, we write $\boldsymbol{\beta} = (\beta(x_1)^\top, \beta(x_2)^\top, \ldots, \beta(x_{d_x})^\top)^\top$, $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}(x_1)^\top, \widehat{\beta}(x_2)^\top, \ldots, \widehat{\beta}(x_{d_x})^\top)^\top$, $m(x) := (m(g_1,x), \ldots, m(g_{d_g},x))^\top$, and $\boldsymbol{m} := (m(x_1)^\top, \ldots, m(x_{d_x})^\top)^\top$.

Assumption A.1.   1. $||w(d,e,g)|| \leqslant c_w < \infty$ a.s. uniformly in $(d,e,g) \in \{0,1\} \times \mathcal{E} \times \mathcal{G}$.

2. For all $i \in \mathcal{I}$,

$$\epsilon_i = \sum_{j\in\mathcal{I}} r_{ij}\varepsilon_j,$$

where r_{ij} is a non-stochastic possibly unknown weight; $\{\varepsilon_i\}$ are independently and identically distributed over $\mathcal{I}$, independent of $\{(W_i,X_i)\}$, with mean zero and variance σ_ε^2; $\mathbb{E}|\varepsilon_i|^4 < \infty$; and $\max\{\max_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}|r_{ij}|, \max_{j\in\mathcal{I}}\sum_{i\in\mathcal{I}}|r_{ij}|\} \leqslant c_r < \infty$, uniformly in $n^{\mathcal{I}}$.

3. $\Sigma_{n^{\mathcal{I}}}(x)$, $\Omega_{n^{\mathcal{I}}}(x)$, and $\boldsymbol{\Omega}_{n^{\mathcal{I}}}$ are positive definite for all sufficiently large $n^{\mathcal{I}}$.

4. For all $x \in \mathcal{X}$,

$$\left\|\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}}W_iW_i^\top \mathbf{1}\{X_i = x\} - \Sigma_{n^{\mathcal{I}}}(x)\right\| = O_P\left(1/\sqrt{n^{\mathcal{I}}}\right)$$

5. There exists $b \in (0,1)$ such that $(b_c, b_o) \asymp b$ and $\sqrt{n^{\mathcal{I}}}b \to 0$.

Assumption A.1.1 is standard in applications. A.1.2 allows for cross-sectional dependence in the error terms. For example, if $r_{ij} = A_{ij}^{\mathcal{I}}$, the last condition implies that each individual has only finitely many interacting partners. The same type of error structure has often been considered in the literature (e.g., Kelejian and Prucha, 2007; Conley et al., 2023). A.1.3 is a standard non-singularity condition. It also requires that the proportion of each x-value is nondegenerate. A.1.4 is high-level but can be satisfied under appropriate weak dependence conditions on $\{(W_i, X_i)\}$. Finally, A.1.5 is a technical condition to eliminate the bias in the kernel regression.

Next, we introduce assumptions used to establish the validity of the wild bootstrap procedure. Let

$$\mathcal{B}_{i,\mathcal{I}} := \{j \in \mathcal{I} : \widetilde{\Delta}_{ij} \leqslant d\}, \quad \lambda_{i,\mathcal{I}} := |\mathcal{B}_{i,\mathcal{I}}|, \quad \lambda_{\mathcal{I}} := \frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} \lambda_{i,\mathcal{I}}, \quad V_i := \begin{pmatrix} W_i \mathbf{1}\{X_i = x_1\} \\ W_i \mathbf{1}\{X_i = x_2\} \\ \vdots \\ W_i \mathbf{1}\{X_i = x_{d_x}\} \end{pmatrix} \epsilon_i,$$

where d is the bandwidth parameter used in the kernel function K.

Assumption A.2.   1. $K(s) = K(-s)$ for all $s \in \mathbb{R}$ and $K(0) = 1$; $\sup_{i \in \mathcal{I}} \mathbb{E}(\sum_{j \notin \mathcal{B}_{i,\mathcal{I}}} |K(\widetilde{\Delta}_{ij}/d)|)/\mathbb{E}\lambda_{\mathcal{I}} = O(1)$; $\sup_{i \in \mathcal{I}} \sum_{j \notin \mathcal{B}_{i,\mathcal{I}}} |K(\widetilde{\Delta}_{ij}/d)|/\mathbb{E}\lambda_{\mathcal{I}} = O_P(1)$; $\mathbb{K}_{\mathcal{I}}$ is symmetric and positive semidefinite a.s.

2. There exists $c_{q_0} > 0$ such that $(n^{\mathcal{I}})^{-1} \sum_{i,j \in \mathcal{I}} ||\mathbb{E}[V_i V_j^{\top}]|| \Delta_{ij}^{q_0} < c_{q_0}$, where q_0 denotes the Parzen characteristic exponent of the kernel function K.

3. $\{\nu_{ij}\}$ are independent of $\{(W_i, X_i, \varepsilon_i)\}$ and are uniformly bounded in $i, j \in \mathcal{I}$.

4. For all $i \in \mathcal{I}$, $\lambda_{i,\mathcal{I}} \leqslant c\mathbb{E}\lambda_{\mathcal{I}}$ a.s., for some $c > 0$.

5. $d \to \infty$, and $\mathbb{E}\lambda_{\mathcal{I}} \to \infty$ such that $\mathbb{E}\lambda_{\mathcal{I}}/\sqrt{n^{\mathcal{I}}} \to 0$.

6. $a \downarrow 0$ such that $\sqrt{n^{\mathcal{I}}}a \to \infty$.

Assumptions A.2.1, A.2.2, A.2.3, and A.2.4 correspond, respectively, to Assumptions 1, 3, 4, and 5 in Conley et al. (2023). Specifically, A.2.1 collects the conditions on the kernel weight function. As noted in Footnote 8, the positive semidefinite-ness of $\mathbb{K}_{\mathcal{I}}$ is a high-level condition. Conley et al. (2023) provide an alternative bootstrap procedure for situations where this condition fails. A.2.2 requires that the dependence between i and j decays as the true distance Δ_{ij} increases. The formal definition of the Parzen characteristic exponent q_0, along with related discussion, can be found for example in Andrews (1991) and Conley et al. (2023). A.2.3 requires that the measurement errors are independent and uniformly bounded, which is standard in the HAC estimation literature. A.2.4 restricts the number of neighbors each unit can have to be of the same order. A.2.5 imposes conditions on the bandwidth d. Finally, A.2.6 is a technical condition needed to ensure the consistency of $\widehat{\mathcal{S}}^*_{\delta,q,x}$ for $\mathcal{S}^*_{\delta,q,x}$.

Proof of Proposition 4.1

(i) Let $\ell_{i,b}(x) := L_{i,b}(x) - \mathbf{1}\{X_i = x\}$. It is easy to see that $\ell_{i,b}(x) \leqslant c \cdot b$. To see this, for example, suppose that $d_o = d_c = 1$. Then,

$$\ell_{i,b}(x) = \mathbf{1}\{X_i^c \neq x^c, X_i^o = x^o\}b_c + \mathbf{1}\{X_i^c = x^c, X_i^o \neq x^o\}b_o^{|X_i^o - x^o|} + \mathbf{1}\{X_i^c \neq x^c, X_i^o \neq x^o\}b_c b_o^{|X_i^o - x^o|}.$$

With this and Assumptions A.1.1 and A.1.4,

$$
\begin{aligned}
\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top L_{i,b}(x) &= \frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top \mathbf{1}\{X_i = x\} + \frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top \ell_{i,b}(x) \\
&= \frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top \mathbf{1}\{X_i = x\} + O(b) \\
&= \Sigma_{n^{\mathcal{I}}}(x) + O_P(1/\sqrt{n^{\mathcal{I}}} + b).
\end{aligned}
\tag{A.2}
$$

Next, write

$$
\begin{aligned}
\sqrt{n^{\mathcal{I}}}\left(\widehat{\beta}(x) - \beta(x)\right) &= \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top L_{i,b}(x)\right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}} W_i(Y_i - W_i^\top \beta(x))L_{i,b}(x) \\
&= A_1(x) + A_2(x) + A_3(x),
\end{aligned}
$$

where

$$A_1(x) := \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top L_{i,b}(x)\right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}} W_i W_i^\top \left\{\beta(X_i) - \beta(x)\right\} L_{i,b}(x)$$

$$A_2(x) := \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top L_{i,b}(x)\right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}} W_i \epsilon_i \mathbf{1}\{X_i = x\}$$

$$A_3(x) := \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top L_{i,b}(x)\right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}} W_i \epsilon_i \ell_{i,b}(x).$$

Observe that, for all $x \in \mathcal{X}$,

$$
\begin{aligned}
A_1(x) &= \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top L_{i,b}(x)\right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}} W_i W_i^\top \left\{\beta(X_i) - \beta(x)\right\} \mathbf{1}\{X_i = x\} \\
&\quad + \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}} W_i W_i^\top L_{i,b}(x)\right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}} W_i W_i^\top \left\{\beta(X_i) - \beta(x)\right\} \ell_{i,b}(x) \\
&= O_P\left(\sqrt{n^{\mathcal{I}}} b\right).
\end{aligned}
$$

For $A_3(x)$,

$$\mathbb{E}\left\|\frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}} W_i \epsilon_i \ell_{i,b}(x)\right\|^2 = \frac{1}{n^{\mathcal{I}}}\sum_{i,i'\in\mathcal{I}} \mathbb{E}\left[W_i^\top W_{i'} \epsilon_i \epsilon_{i'} \ell_{i,b}(x)\ell_{i',b}(x)\right]$$

27

$$= \frac{1}{n^{\mathcal{I}}} \sum_{i,i',j,j' \in \mathcal{I}} \mathbb{E}\left[W_i^\top W_{i'} r_{ij} r_{i'j'} \varepsilon_j \varepsilon_{j'} \ell_{i,b}(x) \ell_{i',b}(x)\right]$$

$$= \frac{\sigma_\varepsilon^2}{n^{\mathcal{I}}} \sum_{i,i',j \in \mathcal{I}} r_{ij} r_{i'j} \mathbb{E}\left[W_i^\top W_{i'} \ell_{i,b}(x) \ell_{i',b}(x)\right]$$

$$\leqslant \frac{\sigma_\varepsilon^2 c^2 b^2}{n^{\mathcal{I}}} \sum_{i,i',j \in \mathcal{I}} |r_{ij}| \cdot |r_{i'j}| = O(b^2).$$

Hence, by (A.2) and Markov's inequality, we have $A_3(x) = O_P(b)$. Hence, we have $\sqrt{n^{\mathcal{I}}}(\widehat{\beta}(x) - \beta(x)) = A_2(x) + o_P(1)$ by Assumption A.1.5.

To apply the central limit theorem to $A_2(x)$, define

$$a_j := \boldsymbol{c}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1/2} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i \mathbf{1}\{X_i = x\} r_{ij} \varepsilon_j$$

where $\boldsymbol{c} \in \mathbb{R}^{d_w}$ satisfying $||\boldsymbol{c}|| = 1$. Note that $\mathbb{E}[a_j] = 0$ and $\sum_{j \in \mathcal{I}} \mathbb{E}[a_j^2] = 1$ hold:

$$\sum_{j \in \mathcal{I}} \mathbb{E}[a_j^2] = \boldsymbol{c}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1/2} \frac{1}{n^{\mathcal{I}}} \sum_{i,i',j \in \mathcal{I}} \mathbb{E}[W_i W_{i'}^\top \mathbf{1}\{X_i = x, X_{i'} = x\} r_{ij} r_{i'j} \varepsilon_j \varepsilon_j] \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1/2} \boldsymbol{c} = 1$$

by Assumption A.1.2. Moreover, by Assumptions A.1.1, Assumptions A.1.2, and Assumptions A.1.3,

$$\sum_{j \in \mathcal{I}} \mathbb{E}[a_j^4] = \frac{1}{(n^{\mathcal{I}})^2} \sum_{j \in \mathcal{I}} \sum_{i_1,i_2,i_3,i_4 \in \mathcal{I}} \mathbb{E}\Big[r_{i_1 j} r_{i_2 j} W_{i_1}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1/2} \boldsymbol{cc}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1/2} W_{i_2}$$

$$\times r_{i_3 j} r_{i_4 j} W_{i_3}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1/2} \boldsymbol{cc}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1/2} W_{i_4} \cdot \mathbf{1}\{X_{i_1} = X_{i_2} = X_{i_3} = X_{i_4} = x\} \varepsilon_j^4\Big]$$

$$\leqslant \frac{c}{(n^{\mathcal{I}})^2} \sum_{j \in \mathcal{I}} \sum_{i_1,i_2,i_3,i_4 \in \mathcal{I}} |r_{i_1 j}| \cdot |r_{i_2 j}| \cdot |r_{i_3 j}| \cdot |r_{i_4 j}| \mathbb{E}\left[W_{i_1}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1} W_{i_2} W_{i_3}^\top \left(\Omega_{n^{\mathcal{I}}}(x)\right)^{-1} W_{i_4}\right]$$

$$\leqslant \frac{c c_w^4 c_r^4}{n^{\mathcal{I}}} \to 0.$$

Then, by Lyapunov's central limit theorem, we obtain $\sum_{j \in \mathcal{I}} a_j \xrightarrow{d} N(0,1)$. Finally, by (A.2) and Slutsky's theorem,

$$\sqrt{n^{\mathcal{I}}} \left(\widehat{\beta}(x) - \beta(x)\right) \xrightarrow{d} N\left(\boldsymbol{0}_{d_w}, (\Sigma_{\mathcal{I}}(x))^{-1} \Omega_{\mathcal{I}}(x)(\Sigma_{\mathcal{I}}(x))^{-1}\right),$$

where $\Sigma_{\mathcal{I}}(x) := \lim_{n^{\mathcal{I}} \to \infty} \Sigma_{n^{\mathcal{I}}}(x)$, and $\Omega_{\mathcal{I}}(x) := \lim_{n^{\mathcal{I}} \to \infty} \Omega_{n^{\mathcal{I}}}(x)$.

(ii) By definition,

$$\sqrt{n^{\mathcal{I}}}(\widehat{\boldsymbol{m}} - \boldsymbol{m}) = \sqrt{n^{\mathcal{I}}} \boldsymbol{Z} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

$$= \boldsymbol{Z} \widehat{\boldsymbol{J}}_{n^{\mathcal{I}}} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} \begin{pmatrix} W_i \mathbf{1}\{X_i = x_1\} \\ W_i \mathbf{1}\{X_i = x_2\} \\ \vdots \\ W_i \mathbf{1}\{X_i = x_{d_x}\} \end{pmatrix} \epsilon_i + o_P(1),$$

28

where the last equality follows from the same argument as above. Note that we have $\widehat{\boldsymbol{J}}_{n^{\mathcal{I}}} \overset{p}{\to} \boldsymbol{J}_{n^{\mathcal{I}}}$ by Assumption A.1.4. Similarly as above, we define

$$
\boldsymbol{a}_j := \boldsymbol{c}^\top \left(\boldsymbol{\Omega}_{n^{\mathcal{I}}}\right)^{-1/2} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} \begin{pmatrix} W_i \mathbf{1}\{X_i = x_1\} \\ W_i \mathbf{1}\{X_i = x_2\} \\ \vdots \\ W_i \mathbf{1}\{X_i = x_{d_x}\} \end{pmatrix} r_{ij} \varepsilon_j,
$$

for any $\boldsymbol{c} \in \mathbb{R}^{d_w d_x}$ satisfying $||\boldsymbol{c}|| = 1$. Then, by verifying the Lyapunov condition, we obtain $\sum_{j \in \mathcal{I}} \boldsymbol{a}_j \overset{d}{\to} N(0, 1)$, which implies the desired result:

$$
\sqrt{n^{\mathcal{I}}}(\widehat{\boldsymbol{m}} - \boldsymbol{m}) \overset{d}{\to} N\left(\boldsymbol{0}_{d_x d_g}, \boldsymbol{Z} \boldsymbol{J}_{\mathcal{I}} \boldsymbol{\Omega}_{\mathcal{I}} \boldsymbol{J}_{\mathcal{I}} \boldsymbol{Z}^\top\right)
$$

by Slutsky's theorem, where $\boldsymbol{J}_{\mathcal{I}} := \lim_{n^{\mathcal{I}} \to \infty} \boldsymbol{J}_{n^{\mathcal{I}}}$, and $\boldsymbol{\Omega}_{\mathcal{I}} := \lim_{n^{\mathcal{I}} \to \infty} \boldsymbol{\Omega}_{n^{\mathcal{I}}}$.

$\square$

Proof of Theorem 4.1

Define

$$
\phi(\boldsymbol{m}) := \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x \in \mathcal{B}_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u, v) m(v, x) \right].
$$

Then, we can write concisely $\overline{\kappa}_{\delta,q} = \phi(\boldsymbol{m})$ and $\widehat{\overline{\kappa}}_{\delta,q} = \phi(\widehat{\boldsymbol{m}})$.

By Theorem 2.1 of Fang and Santos (2019) (see also Shapiro (1991)), we know that

$$
\sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q} - \overline{\kappa}_{\delta,q}) = \sqrt{n^{\mathcal{I}}}(\phi(\widehat{\boldsymbol{m}}) - \phi(\boldsymbol{m}))
$$
$$
= \phi'_{\boldsymbol{m}}(\sqrt{n^{\mathcal{I}}}(\widehat{\boldsymbol{m}} - \boldsymbol{m})) + o_P(1),
$$

and therefore $\sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q} - \overline{\kappa}_{\delta,q}) \overset{d}{\to} \phi'_{\boldsymbol{m}}(N(\boldsymbol{0}_{d_x d_g}, \boldsymbol{Z} \boldsymbol{J}_{\mathcal{I}} \boldsymbol{\Omega}_{\mathcal{I}} \boldsymbol{J}_{\mathcal{I}} \boldsymbol{Z}^\top))$, where $\phi'_{\boldsymbol{m}}(\boldsymbol{h})$ is the Hadamard directional derivative of ϕ at $\boldsymbol{m}$ in the direction $\boldsymbol{h} \in \mathbb{R}^{d_g d_x}$.

The explicit form of ϕ' can be derived as follows. Let us denote

$$
\langle \Gamma, f \rangle := \sum_{u,v \in \mathcal{G}^2} \Gamma(u, v) f(v)
$$
$$
\phi(x, f) := \max_{\Gamma \in \mathcal{B}_{\delta,q,x}} \langle \Gamma, f \rangle,
$$

so that $\phi(\boldsymbol{m}) = \sum_{x \in \mathcal{X}} \phi(x, m(\cdot, x))$. Define

$$
\mathcal{S}^*_{\delta,q,x}(f) := \underset{\Gamma \in \mathcal{B}_{\delta,q,x}}{\mathrm{argmax}} \langle \Gamma, f \rangle.
$$

Consider any sequence $h_t \to h \in \mathbb{R}^{d_g}$ as $t \downarrow 0$. Observe that, for $\Gamma_t \in \mathcal{S}^*_{\delta,q,x}(f + th_t)$ and $\Gamma_0 \in \mathcal{S}^*_{\delta,q,x}(f)$,

$$\frac{\phi(x, f + th_t) - \phi(x, f)}{t} = \frac{\langle \Gamma_t, f + th_t \rangle - \langle \Gamma_0, f \rangle}{t} = \frac{\langle \Gamma_t, f \rangle + t\langle \Gamma_t, h_t \rangle - \langle \Gamma_0, f \rangle}{t}$$

$$\leqslant \langle \Gamma_t, h_t \rangle,$$

where the last inequality follows because $\langle \Gamma_t, f \rangle \leqslant \langle \Gamma_0, f \rangle$. Since Γ_t is a sequence in $\mathcal{B}_{\delta,q,x}$ and $\mathcal{B}_{\delta,q,x}$ is compact, the right-hand side converges to $\langle \Gamma_0, h \rangle$, leading to

$$\limsup_{t \downarrow 0} \frac{\langle \Gamma_t, f + th_t \rangle - \langle \Gamma_0, f \rangle}{t} \leqslant \max_{\Gamma \in \mathcal{S}^*_{\delta,q,x}(f)} \langle \Gamma, h \rangle. \tag{A.3}$$

Meanwhile,

$$\frac{\langle \Gamma_t, f + th_t \rangle}{t} \geqslant \frac{\langle \Gamma_0, f + th_t \rangle}{t}$$

$$= \frac{\langle \Gamma_0, f \rangle}{t} + \langle \Gamma_0, h_t \rangle.$$

Since the above holds for all $\Gamma_0 \in \mathcal{S}^*_{\delta,q,x}(f)$,

$$\liminf_{t \downarrow 0} \frac{\langle \Gamma_t, f + th_t \rangle - \langle \Gamma_0, f \rangle}{t} \geqslant \max_{\Gamma \in \mathcal{S}^*_{\delta,q,x}(f)} \langle \Gamma, h \rangle. \tag{A.4}$$

From (A.3) and (A.4), we can find that $\phi'_f(x, h) = \max_{\Gamma \in \mathcal{S}^*_{\delta,q,x}} \langle \Gamma, h \rangle$.

Hence, in our context, writing $\mathcal{S}^*_{\delta,q,x} := \operatorname{argmax}_{\Gamma \in \mathcal{B}_{\delta,q,x}} \langle \Gamma, m(\cdot, x) \rangle$,

$$\phi'_{\boldsymbol{m}}(\boldsymbol{h}) = \lim_{t \downarrow 0} \frac{\sum_{x \in \mathcal{X}} \phi(x, m(\cdot, x) + th_t(\cdot, x)) - \phi(x, m(\cdot, x))}{t}$$

$$= \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x \in \mathcal{S}^*_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u, v) h(v, x) \right].$$

Consequently,

$$\sqrt{n^{\mathcal{I}}} \left(\widehat{\overline{\kappa}}_{\delta,q} - \overline{\kappa}_{\delta,q} \right) = \phi'_{\boldsymbol{m}}(\sqrt{n^{\mathcal{I}}}(\widehat{\boldsymbol{m}} - \boldsymbol{m})) + o_P(1)$$

$$= \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x \in \mathcal{S}^*_{\delta,q,x}} \sqrt{n^{\mathcal{I}}} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u, v) \left(\widehat{m}(v, x) - m(v, x) \right) \right] + o_P(1)$$

$$\xrightarrow{d} \sum_{x \in \mathcal{X}} \left[\max_{\Gamma_x \in \mathcal{S}^*_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u, v) \mathbb{G}(v, x) \right]$$

by Proposition 4.1(ii).

$\square$

Proof of Theorem 4.2

Let $\epsilon_i^* := \eta_i \widehat{\epsilon}_i$. Write

$$\sqrt{n^{\mathcal{I}}} \left(\widehat{\beta}^*(x) - \widehat{\beta}(x) \right) = \left(\frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i W_i^\top L_{i,b}(x) \right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i Y_i^* L_{i,b}(x) - \sqrt{n^{\mathcal{I}}} \widehat{\beta}(x)$$

$$= A_1^*(x) + A_2^*(x) + A_3^*(x),$$

where

$$A_1^*(x) := \left(\frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i W_i^\top L_{i,b}(x) \right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i W_i^\top \left\{ \widehat{\beta}(X_i) - \widehat{\beta}(x) \right\} L_{i,b}(x)$$

$$A_2^*(x) := \left(\frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i W_i^\top L_{i,b}(x) \right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i \epsilon_i^* \mathbf{1}\{X_i = x\}$$

$$A_3^*(x) := \left(\frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i W_i^\top L_{i,b}(x) \right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i \epsilon_i^* \ell_{i,b}(x).$$

Analogously to the proof of Proposition 4.1, we can easily find that $A_1^*(x) = O_P\left(\sqrt{n^{\mathcal{I}}} b\right)$. For $A_3^*(x)$, decompose $A_3^*(x) = A_{31}^*(x) + A_{32}^*(x)$, where

$$A_{31}^*(x) := \left(\frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i W_i^\top L_{i,b}(x) \right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i (\epsilon_i^* - \eta_i \epsilon_i) \ell_{i,b}(x)$$

$$A_{32}^*(x) := \left(\frac{1}{n^{\mathcal{I}}} \sum_{i \in \mathcal{I}} W_i W_i^\top L_{i,b}(x) \right)^{-1} \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i \eta_i \epsilon_i \ell_{i,b}(x).$$

Noting that $\widehat{\epsilon}_i - \epsilon_i = Y_i - W_i^\top \widehat{\beta}(X_i) - \epsilon_i = W_i^\top (\beta(X_i) - \widehat{\beta}(X_i))$, write

$$W_i(\epsilon_i^* - \eta_i \epsilon_i) \ell_{i,b}(x) = W_i \eta_i (\widehat{\epsilon}_i - \epsilon_i) \ell_{i,b}(x)$$

$$= W_i W_i^\top (\beta(X_i) - \widehat{\beta}(X_i)) \eta_i \ell_{i,b}(x)$$

$$=: c_i \eta_i \ell_{i,b}(x),$$

where $c_i = O_P(1/\sqrt{n^{\mathcal{I}}})$ by Proposition 4.1(i). Further,

$$\mathbb{E}^* \left\| \frac{1}{\sqrt{n^{\mathcal{I}}}} \sum_{i \in \mathcal{I}} W_i(\epsilon_i^* - \eta_i \epsilon_i) \ell_{i,b}(x) \right\|^2 = \frac{1}{n^{\mathcal{I}}} \sum_{i,i' \in \mathcal{I}} \mathbb{E}^* \left[c_i^\top c_{i'} \eta_i \eta_{i'} \ell_{i,b}(x) \ell_{i',b}(x) \right]$$

$$= \frac{1}{n^{\mathcal{I}}} \sum_{i,i' \in \mathcal{I}} \mathbb{E}^* [\eta_i \eta_{i'}] c_i^\top c_{i'} \ell_{i,b}(x) \ell_{i',b}(x).$$

Recall that $\mathbb{E}^*[\boldsymbol{\eta}\boldsymbol{\eta}^\top] = \mathbb{K}_{\mathcal{I}}$, and hence $\mathbb{E}^*[\eta_i \eta_{i'}] = K(\widetilde{\Delta}_{i,i'}/d)$. Then,

$$\frac{1}{n^{\mathcal{I}}} \sum_{i,i' \in \mathcal{I}} c_i^\top c_{i'} \ell_{i,b}(x) \ell_{i',b}(x) K\left(\frac{\widetilde{\Delta}_{i,i'}}{d} \right) \leqslant O_P(b^2) \frac{1}{(n^{\mathcal{I}})^2} \sum_{i,i' \in \mathcal{I}} \left| K\left(\frac{\widetilde{\Delta}_{i,i'}}{d} \right) \right|$$

31

$$= o_P(b^2),$$

where the last line is due to Lemma A.1 of Conley *et al.* (2023). Hence, by Assumption A.1.5 and Markov's inequality with (A.2), we have $A_{31}^*(x) = o_{P*}(1)$ in probability. Similarly,

$$\mathbb{E}\left(\mathbb{E}^*\left\|\frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}}W_i\eta_i\epsilon_i\ell_{i,b}(x)\right\|^2\right) = \frac{1}{n^{\mathcal{I}}}\sum_{i,i'\in\mathcal{I}}\mathbb{E}\left[W_i^\top W_i\epsilon_i\epsilon_{i'}\mathbb{E}^*[\eta_i\eta_{i'}]\ell_{i,b}(x)\ell_{i',b}(x)\right]$$
$$= \frac{1}{n^{\mathcal{I}}}\sum_{i,i'\in\mathcal{I}}\mathbb{E}\left[W_i^\top W_i\epsilon_i\epsilon_{i'}\ell_{i,b}(x)\ell_{i',b}(x)K\left(\frac{\widetilde{\Delta}_{i,i'}}{d}\right)\right] = O(b^2),$$

implying that $A_{32}^*(x) = o_{P*}(1)$ in probability.

We apply the same decomposition to $A_2^*(x)$: $A_2^*(x) = A_{21}^*(x) + A_{22}^*(x)$,

$$A_{21}^*(x) := \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}}W_iW_i^\top L_{i,b}(x)\right)^{-1}\frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}}W_i(\epsilon_i^* - \eta_i\epsilon_i)\mathbf{1}\{X_i = x\}$$

$$A_{22}^*(x) := \left(\frac{1}{n^{\mathcal{I}}}\sum_{i\in\mathcal{I}}W_iW_i^\top L_{i,b}(x)\right)^{-1}\frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}}W_i\eta_i\epsilon_i\mathbf{1}\{X_i = x\}.$$

Then, by the same argument as in the evaluation of $A_{31}^*(x)$, it is straightforward to see that $A_{21}^*(x) = o_{P*}(1)$ in probability.

Since the above discussion applies to all $x \in \mathcal{X}$, consequently, we have

$$\sqrt{n^{\mathcal{I}}}(\widehat{m}^* - \widehat{m}) = \sqrt{n^{\mathcal{I}}}\boldsymbol{Z}\left(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}\right)$$
$$= \boldsymbol{Z}\widehat{\boldsymbol{J}}_{n^{\mathcal{I}}}\frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}}\begin{pmatrix} W_i\mathbf{1}\{X_i = x_1\} \\ W_i\mathbf{1}\{X_i = x_2\} \\ \vdots \\ W_i\mathbf{1}\{X_i = x_{d_x}\} \end{pmatrix}\eta_i\epsilon_i + o_{P*}(1),$$

with probability approaching one, where the definitions of $\widehat{m}^*$ and $\widehat{\boldsymbol{\beta}}^*$ should be clear from the context. Furthermore, following the same argument as in the proof of Theorem 3.1 (equation (20)) of Conley *et al.* (2023), we obtain

$$(\boldsymbol{\Omega}_{n^{\mathcal{I}}})^{-1/2}\frac{1}{\sqrt{n^{\mathcal{I}}}}\sum_{i\in\mathcal{I}}\begin{pmatrix} W_i\mathbf{1}\{X_i = x_1\} \\ W_i\mathbf{1}\{X_i = x_2\} \\ \vdots \\ W_i\mathbf{1}\{X_i = x_{d_x}\} \end{pmatrix}\eta_i\epsilon_i \xrightarrow{d*} N(\mathbf{0}_{d_wd_x}, I_{d_wd_x})$$

in probability, and hence

$$\Pr^*\left(\sqrt{n^{\mathcal{I}}}(\widehat{m}^* - \widehat{m}) \leqslant s\right) = \Pr\left(\sqrt{n^{\mathcal{I}}}(\widehat{m} - m) \leqslant s\right) + o_P(1)$$

uniformly in $s \in \mathbb{R}$. Then, in view of Proposition 4.1(ii) and Theorem 4.1, we can see that $\sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q}^* - \widehat{\overline{\kappa}}_{\delta,q})$ and $\sqrt{n^{\mathcal{I}}}(\widehat{\overline{\kappa}}_{\delta,q} - \overline{\kappa}_{\delta,q})$ share the same asymptotic distribution conditional on the event $\{\widehat{\mathcal{S}}_{\delta,q,x}^* = \mathcal{S}_{\delta,q,x}^*\}$.

In view of the proof of Theorem 4.1, we can see that $\widehat{\overline{\kappa}}_{\delta,q,x} = \overline{\kappa}_{\delta,q,x} + O_P(1/\sqrt{n^{\mathcal{I}}})$, where $\overline{\kappa}_{\delta,q,x} := \max_{\Gamma \in \mathcal{B}_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v) m(v,x)$. Suppose that $\Gamma \in \mathcal{S}_{\delta,q,x}^*$. Then,

$$\sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)\widehat{m}(v,x) = \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)m(v,x) + \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)(\widehat{m}(v,x) - m(v,x))$$

$$= \overline{\kappa}_{\delta,q,x} + O_P(1/\sqrt{n^{\mathcal{I}}})$$

$$= \widehat{\overline{\kappa}}_{\delta,q,x} + O_P(1/\sqrt{n^{\mathcal{I}}})$$

$$\geqslant \widehat{\overline{\kappa}}_{\delta,q,x} - a$$

with probability approaching one under Assumption A.2.6. This implies that $\Pr(\mathcal{S}_{\delta,q,x}^* \subseteq \widehat{\mathcal{S}}_{\delta,q,x}^*) \to 1$ as $n^{\mathcal{I}} \to \infty$. On the other hand, suppose that $\Gamma \in \widehat{\mathcal{S}}_{\delta,q,x}^*$. Then,

$$\sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)m(v,x) = \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)\widehat{m}(v,x) + \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)(m(v,x) - \widehat{m}(v,x))$$

$$\geqslant \widehat{\overline{\kappa}}_{\delta,q,x} - a + O_P(1/\sqrt{n^{\mathcal{I}}})$$

$$= \overline{\kappa}_{\delta,q,x} - a + O_P(1/\sqrt{n^{\mathcal{I}}}).$$

Here, note that if $\Gamma \notin \mathcal{S}_{\delta,q,x}^*$, then the strict inequality $\sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)m(v,x) < \overline{\kappa}_{\delta,q,x}$ must hold. Hence, since $a + O_P(1/\sqrt{n^{\mathcal{I}}})$ converges to zero in probability as $n^{\mathcal{I}}$ increases, the above inequality implies that $\Gamma \in \mathcal{S}_{\delta,q,x}^*$ holds with probability approaching one; that is, $\Pr(\widehat{\mathcal{S}}_{\delta,q,x}^* \subseteq \mathcal{S}_{\delta,q,x}^*) \to 1$. Hence, $\Pr(\mathcal{S}_{\delta,q,x}^* = \widehat{\mathcal{S}}_{\delta,q,x}^*) \to 1$. $\qquad \square$

B Estimation and Bootstrap Inference for the Lower Bound

The estimation of the lower bound $\underline{\kappa}_{\delta,q}$ can be performed by solving the following linear programming:

$$\widehat{\underline{\kappa}}_{\delta,q} := \sum_{x \in \mathcal{X}} \left[\min_{\Gamma_x} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v)\widehat{m}(v,x) \right]$$

$$\text{subject to } \sum_{v \in \mathcal{G}} \Gamma_x(u,v) = \pi_x^*(u), \ \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v)|u-v|^q \leqslant \delta^q, \Gamma_x(u,v) \geqslant 0, \forall (x,u,v) \in \mathcal{X} \times \mathcal{G}^2,$$

where $\widehat{m}$ is obtained through the varying-coefficient estimation as in Subsection 4.1.

To describe the wild bootstrap procedure for the lower bound, let

$$\widehat{\mathcal{T}}_{\delta,q,x}^* := \left\{ \Gamma \in \mathcal{B}_{\delta,q,x} : \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)\widehat{m}(v,x) \leqslant \widehat{\underline{\kappa}}_{\delta,q,x} + a \right\},$$

which is considered as the estimator of $\mathcal{T}_{\delta,q,x}^* := \operatorname{argmin}_{\Gamma \in \mathcal{B}_{\delta,q,x}} \sum_{u,v \in \mathcal{G}^2} \Gamma(u,v)m(v,x)$. Then, the distribution of $\sqrt{n^{\mathcal{I}}}(\widehat{\underline{\kappa}}_{\delta,q} - \underline{\kappa}_{\delta,q})$ can be simulated in the following manner.

Algorithm B.1 Wild bootstrap procedure for inference on $\underline{\kappa}_{\delta,q}$

1: Estimate $\widehat{\beta}(x)$ for all $x \in \mathcal{X}$ using (4.2)

2: Compute the residual $\widehat{\epsilon}_i := Y_i - W_i^\top \widehat{\beta}(X_i)$ for all $i \in \mathcal{I}$

3: **for** $b = 1$ to B **do**

4:     Draw $\boldsymbol{\eta}^{(b)} = (\eta_1^{(b)}, \ldots, \eta_{n^{\mathcal{I}}}^{(b)}) \sim \Phi_{\mathcal{I}} \Lambda_{\mathcal{I}}^{1/2} N(\mathbf{0}_{n^{\mathcal{I}}}, I_{n^{\mathcal{I}}})$

5:     Generate a bootstrap sample $\{(W_i, Y_i^{*(b)}) : i \in \mathcal{I}\}$, where $Y_i^{*(b)} := W_i^\top \widehat{\beta}(X_i) + \eta_i^{(b)} \widehat{\epsilon}_i$

6:     Obtain $\widehat{\beta}^{*(b)}(x)$ by the kernel weighted regression of $Y_i^{*(b)}$ on W_i for all $x \in \mathcal{X}$

7:     Compute $\widehat{\underline{\kappa}}_{\delta,q}^{*(b)} := \sqrt{n^{\mathcal{I}}} \sum_{x \in \mathcal{X}} \left[\min_{\Gamma_x \in \widehat{\mathcal{T}}_{\delta,q,x}^*} \sum_{u,v \in \mathcal{G}^2} \Gamma_x(u,v) z(v,x)^\top (\widehat{\beta}^{*(b)}(x) - \widehat{\beta}(x)) \right]$

8: **end for**

9: Compute the empirical α quantile $\widehat{\omega}_{B,\alpha/2}$ of $\left\{ \sqrt{n^{\mathcal{I}}}(\widehat{\underline{\kappa}}_{\delta,q}^{*(b)} - \widehat{\underline{\kappa}}_{\delta,q}) : b = 1, \ldots, B \right\}$

Further, the asymptotic $100(1 - \alpha)\%$ CI for $\underline{\kappa}_{\delta,q}$ can be obtained by

$$\mathcal{C}_{1-\alpha}(\underline{\kappa}_{\delta,q}) := \left[\widehat{\underline{\kappa}}_{\delta,q} - \frac{\widehat{\omega}_{B,1-\alpha/2}}{\sqrt{n^{\mathcal{I}}}}, \ \widehat{\underline{\kappa}}_{\delta,q} - \frac{\widehat{\omega}_{B,\alpha/2}}{\sqrt{n^{\mathcal{I}}}} \right].$$

References

Albert, R. and Barabási, A.L., 2002. Statistical mechanics of complex networks, *Reviews of Modern Physics*, 74 (1), 47.

Andrews, D.W., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, 817–858.

Aronow, P.M., Eckles, D., Samii, C., and Zonszein, S., 2021. Spillover effects in experimental data, *in:* J. Druckman and D.P. Green, eds., *Advances in Experimental Political Science*, Cambridge University Press, chap. 16, 289–319.

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

Bhattacharya, D., 2009. Inferring optimal peer assignment from experimental data, *Journal of the American Statistical Association*, 104 (486), 486–500.

Blanchet, J. and Murthy, K., 2019. Quantifying distributional model risk via optimal transport, *Mathematics of Operations Research*, 44 (2), 565–600.

Blanchet, J., Murthy, K., and Nguyen, V.A., 2021. *Statistical Analysis of Wasserstein Distributionally Robust Estimators*, INFORMS, chap. 8, 227–254.

Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D., Marlow, C., Settle, J.E., and Fowler, J.H., 2012. A 61-million-person experiment in social influence and political mobilization, *Nature*, 489 (7415), 295–298.

Buchanan, A.L., Hudgens, M.G., Cole, S.R., Mollan, K.R., Sax, P.E., Daar, E.S., Adimora, A.A., Eron, J.J., and Mugavero, M.J., 2018. Generalizing evidence from randomized trials using inverse probability of sampling weights, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181 (4), 1193–1209.

Cai, J., Janvry, A.D., and Sadoulet, E., 2015. Social networks and the decision to insure, *American Economic Journal: Applied Economics*, 7 (2), 81–108.

Carter, M., Laajaj, R., and Yang, D., 2021. Subsidies and the African green revolution: direct effects and social network spillovers of randomized input subsidies in Mozambique, *American Economic Journal: Applied Economics*, 13 (2), 206–229.

Chin, A., 2019. Regression adjustments for estimating the global treatment effect in experiments with interference, *Journal of Causal Inference*, 7 (2), 20180026.

Christensen, T. and Connault, B., 2023. Counterfactual sensitivity and robustness, *Econometrica*, 91 (1), 263–298.

Conley, T.G., Gonçalves, S., Kim, M.S., and Perron, B., 2023. Bootstrap inference under cross-sectional dependence, *Quantitative Economics*, 14 (2), 511–569.

Dahabreh, I.J., Robertson, S.E., Steingrimsson, J.A., Stuart, E.A., and Hernan, M.A., 2020. Extending inferences from a randomized trial to a new target population, *Statistics in Medicine*, 39 (14), 1999–2014.

Degtiar, I. and Rose, S., 2023. A review of generalizability and transportability, *Annual Review of Statistics and Its Application*, 10 (1), 501–524.

Duchi, J.C. and Namkoong, H., 2021. Learning models with uniform performance via distributionally robust optimization, *The Annals of Statistics*, 49 (3), 1378–1406.

Fang, Z. and Santos, A., 2019. Inference on directionally differentiable functions, *The Review of Economic Studies*, 86 (1), 377–412.

Faridani, S. and Niehaus, P., 2024. Linear estimation of global average treatment effects, *NBER Working Paper*, 33319.

Gao, R. and Kleywegt, A., 2023. Distributionally robust stochastic optimization with Wasserstein distance, *Mathematics of Operations Research*, 48 (2), 603–655.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J.S., 2015. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178 (3), 757–778.

Hoshino, T. and Yanagi, T., 2023. Randomization test for the specification of interference structure, *arXiv preprint*, 2301.05580.

Jackson, M.O., 2008. *Social and Economic Networks*, Princeton University Press.

Kelejian, H.H. and Prucha, I.R., 2007. HAC estimation in a spatial framework, *Journal of Econometrics*, 140 (1), 131–154.

Kim, M.S. and Sun, Y., 2011. Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix, *Journal of Econometrics*, 160 (2), 349–371.

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

Kolaczyk, E.D., 2009. *Statistical Analysis of Network Data: Methods and Models*, Springer.

Leung, M.P., 2020. Treatment and spillover effects under network interference, *Review of Economics and Statistics*, 102 (2), 368–380.

Leung, M.P., 2024. Identifying treatment and spillover effects using exposure contrasts, *arXiv preprint*, 2403.08183.

Li, Q. and Racine, J.S., 2010. Smooth varying-coefficient estimation and inference for qualitative and quantitative data, *Econometric Theory*, 26 (6), 1607–1637.

Lin, Z. and Xu, H., 2017. Estimation of social-influence-dependent peer pressure in a large network game, *The Econometrics Journal*, 20 (3), S86–S102.

Miao, X., Zhao, J., and Kang, H., 2024. Transfer learning between US presidential elections: How should we learn from a 2020 ad campaign to inform 2024 ad campaigns?, *arXiv preprint*, 2411.01100.

Paluck, E.L., Shepherd, H., and Aronow, P.M., 2016. Changing climates of conflict: A social network experiment in 56 schools, *Proceedings of the National Academy of Sciences*, 113 (3), 566–571.

Panaretos, V.M. and Zemel, Y., 2020. *An Invitation to Statistics in Wasserstein Space*, Springer.

Shapiro, A., 1991. Asymptotic analysis of stochastic programs, *Annals of Operations Research*, 30 (1), 169–186.

Spini, P.E., 2021. Robustness, heterogeneous treatment effects and covariate shifts, *arXiv preprint*, 2112.09259.

Stuart, E.A., Cole, S.R., Bradshaw, C.P., and Leaf, P.J., 2011. The use of propensity scores to assess the generalizability of results from randomized trials, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174 (2), 369–386.

Tripathi, A., Venugopalan, S., and West, D.B., 2010. A short constructive proof of the Erdős–Gallai characterization of graphic lists, *Discrete Mathematics*, 310 (4), 843–844.

Ugander, J. and Yin, H., 2023. Randomized graph cluster randomization, *Journal of Causal Inference*, 11 (1), 20220014.

Wu, L. and Yang, S., 2023. Transfer learning of individualized treatment rules from experimental to real-world data, *Journal of Computational and Graphical Statistics*, 32 (3), 1036–1045.

Yu, C.L., Airoldi, E.M., Borgs, C., and Chayes, J.T., 2022. Estimating the total treatment effect in randomized experiments with unknown network structure, *Proceedings of the National Academy of Sciences*, 119 (44), e2208975119.