Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity

Andrei Zeleneev*

January 10, 2020

Job Market Paper

[click here for the latest version]

Abstract

Homophily based on observables is widespread in networks. Therefore, homophily based on unobservables (fixed effects) is also likely to be an important determinant of the interaction outcomes. Failing to properly account for latent homophily (and other complex forms of unobserved heterogeneity, in general) can result in inconsistent estimators and misleading policy implications. To address this concern, I consider a network model with nonparametric unobserved heterogeneity, leaving the role of the fixed effects and the nature of their interaction unspecified. I argue that the outcomes of the interactions can be used to identify agents with the same values of the fixed effects. The variation in the observed characteristics of such agents allows me to identify the effects of the covariates, while controlling for the impact of the fixed effects. Building on these ideas, I construct several estimators of the parameters of interest and characterize their large sample properties. The suggested approach is not specific to the network context and applies to general two-way models with nonparametric unobserved heterogeneity, including large panels. A Monte-Carlo experiment illustrates the usefulness of the suggested approaches and supports the large sample theory findings.

Keywords: network data, homophily, fixed effects

*zeleneev@princeton.edu. Princeton University, Department of Economics, Princeton, NJ 08544. I am grateful to my advisors Bo Honoré, Ulrich Müller and especially Kirill Evdokimov for their continuous guidance and support. I also thank Michal Kolesár, Mikkel Plagborg-Møller, Christopher Sims, Mark Watson, and numerous seminar participants for valuable comments and suggestions.

1 Introduction

Unobserved heterogeneity is pervasive in economics. The importance of accounting for unobserved heterogeneity is well recognized in microeconometrics, in general, as well as in the network context, in particular. For example, since Abowd, Kramarz, and Margolis (1999), a linear regression with additive (two-way) fixed effects has become a workhorse model for analyzing interaction data. Originally employed to account for workers and firms fixed effects in the wage regression context, this technique has become a standard tool to control for two-sided unobserved heterogeneity and decompose it into agent specific effects.[1] Since the seminal work of Anderson and Van Wincoop (2003), the importance of controlling for exporters and importers fixed effects has also been well acknowledged in the context of the international trade network, including nonlinear models of Santos Silva and Tenreyro (2006) and Helpman, Melitz, and Rubinstein (2008). In other nonlinear settings, Graham (2017) argues that failing to account for agents' degree heterogeneity (captured by the additive fixed effects) in a general network formation model typically leads to erroneous inference.

While the additive fixed effects framework is commonly employed to control for unobservables in networks, it is not flexible enough to capture more complicated forms of unobserved heterogeneity, which are likely to appear in many settings. This concern can be vividly illustrated in the context of estimation of homophily effects,[2] one of the main focuses of the empirical network analysis. Since homophily (assortative matching) based on observables is widespread in networks (e.g., McPherson, Smith-Lovin, and Cook, 2001), homophily based on unobservables (fixed effects) is also likely to be an important determinant of the interaction outcomes. Since observed and unobserved characteristics (i.e., covariates and fixed effects) are typically confounded, the presence of latent homophily significantly complicates identification of the homophily effects associated with the observables (e.g., Shalizi and Thomas, 2011). Failing to properly account for homophily based on unobservables (and other complex forms of unobserved heterogeneity, in general) is likely to result in inconsistent estimators and misleading policy implications.

To address the concern discussed above, we consider a dyadic network model with a flexible (nonparametric) structure of unobserved heterogeneity, where the outcome of the interaction

[1] For example, the recent applications to employer-employee matched data feature Card, Heining, and Kline (2013); Helpman, Itskhoki, Muendler, and Redding (2017); Song, Price, Guvenen, Bloom, and Von Wachter (2019) among others. The numerous applications of this approach also include the analysis of students-teachers (Hanushek, Kain, Markman, and Rivkin, 2003; Rivkin, Hanushek, and Kain, 2005; Rothstein, 2010), patients-hospitals (Finkelstein, Gentzkow, and Williams, 2016), firms-banks (Amiti and Weinstein, 2018), and residents-counties matched data (Chetty and Hendren, 2018).

[2] The term homophily typically refers to the tendency of individuals to assortatively match based on their characteristics. For example, individuals tend to group based on gender, race, age, education level, and other socioeconomic characteristics. Similarly, countries, which share a border, have the same legal system, language or currency, are more likely to have higher trade volumes.

between agents i and j is given by

$$Y_{ij} = F(W_{ij}'\beta_0 + g(\xi_i, \xi_j)) + \varepsilon_{ij}. \tag{1.1}$$

Here, W_{ij} is a $p \times 1$ vector of pair specific observed covariates, $\beta_0 \in \mathbb{R}^p$ is the parameter of interest, and ξ_i and ξ_j are unobserved i.i.d. fixed effects, and ε_{ij} is an idiosyncratic error independent across pairs of agents. The fixed effects are allowed to interact via the coupling function $g(\cdot, \cdot)$, which is treated as *unknown*. Importantly, we do not require $g(\cdot, \cdot)$ to have any particular structure and do not specify the dimension of ξ. Finally, F is a known (up to location and scale normalizations) invertible function. The presence of F ensures that (1.1) is flexible enough to cover a broad range of the previously studied dyadic network models with unobserved heterogeneity, including nonlinear specifications such as network formation or Poisson regression models.[3]

Being agnostic about the dimensions of the fixed effects and the nature of their interactions, (1.1) allows for a wide range of forms of unobserved heterogeneity, including homophily based on unobservables.

Example (Nonparametric homophily based on unobservables)**.** Suppose $\xi = (\alpha, \nu)' \in \mathbb{R}^2$ and

$$g(\xi_i, \xi_j) = \alpha_i + \alpha_j - \psi(\nu_i, \nu_j),$$

where $\psi(\cdot, \cdot)$ is some function, which (i) satisfies $\psi(\nu_i, \nu_j) = 0$ whenever $\nu_i = \nu_j$, (ii) and is increasing in $|\nu_i - \nu_j|$ for any fixed ν_i or ν_j (e.g., $\psi(\nu_i, \nu_j) = c\,|\nu_i - \nu_j|^{\zeta}$ for some $c > 0$ and $\zeta \geqslant 1$). Here α represents the standard additive fixed effects, and ψ captures latent homophily based on ν: agents sharing the value of ν tend to interact with higher intensity compared to agents distant in terms of ν. Again, since the dimension of ξ is not specified, (1.1) can also incorporate homophily based on several unobserved characteristics (multivariate ν) in a similar manner. ■

We study identification and estimation of (1.1) under the assumption that we observe a single network of a growing size.[4] First, we focus on a simpler version of (1.1)

$$Y_{ij} = W_{ij}'\beta_0 + g(\xi_i, \xi_j) + \varepsilon_{ij}, \tag{1.2}$$

additionally assuming that the idiosyncratic errors are *homoskedastic*. We argue that the outcomes of the interactions can be used to identify agents with the same values of the unobserved fixed effects. Specifically, we introduce a certain pseudo-distance d_{ij} between a pair of agents i and j.

[3] For example, with F equal to the logistic CDF and $g(\xi_i, \xi_j) = \xi_i + \xi_j$, (1.1) corresponds to the network formation model of Graham (2017).

[4] The large single network asymptotics is standard for the literature focusing on identification and estimation of networks models with unobserved heterogeneity. See, for example, Graham (2017); Dzemski (2018); Candelaria (2016); Jochmans (2018); Toth (2017); Gao (2019).

We demonstrate that (i) $d_{ij} = 0$ if and only if $\xi_i = \xi_j$, (ii) and d_{ij} is identified and can be directly estimated from the data. Consequently, agents with the same values ξ can be identified based on the pseudo-distance d_{ij}. Then, the variation in the observed characteristics of such agents allows us to identify the parameter of interest β_0 while controlling for the impact of the fixed effects. Importantly, the identification result is not driven by the particular functional form of (1.2): a similar argument also applies to a nonparametric version of the model.

Building on these ideas, we construct an estimator of β_0 based on matching agents that are similar in terms of the estimated pseudo-distance $\hat{d}_{ij}$ (and, consequently, also similar in terms of the unobserved fixed effects). Importantly, being similar in terms ξ, the matched agents can have different observed characteristics, which allows us to estimate the parameter of interest from the pairwise differenced regressions. We demonstrate consistency of the suggested estimator and derive its rate of convergence.

Second, we extend the proposed identification and estimation strategies to cover models with general heteroskedasticity of the idiosyncratic errors. Leveraging and advancing recent developments in the matrix estimation/completion literature, we demonstrate that the error free part of the outcome $Y^*_{ij} = W'_{ij}\beta_0 + g(\xi_i, \xi_j)$ is identified and can be uniformly (across all pairs of agents) consistently estimated. In particular, working with Y^*_{ij} instead of Y_{ij} effectively reduces (1.2) to a model without the error term ε_{ij}, which can be interpreted as an extreme form of homoskedasticity. This, in turn, allows us to establish identification of β_0 by applying the same argument as in the homoskedastic model. Building on these insights, we analogously adjust the previously employed estimation approach to ensure its validity in the general heteroskedastic setting. The adjusted procedure requires preliminary estimation of the error free outcomes Y^*_{ij}, which, in turn, are used to estimate the pseudo-distances d_{ij}. Specifically, the suggested estimator $\hat{Y}^*_{ij}$ is based on the approach of Zhang, Levina, and Zhu (2017) originally employed in the context of nonparametric graphon estimation. Once $\hat{d}_{ij}$ are constructed, β_0 can be estimated as in the homoskedastic case. We show that the proposed estimator of β_0 is consistent under general forms of heteroskedasticity of the errors and establish its rate of convergence.

Third, we demonstrate how the proposed identification and estimation strategies can be naturally extended to cover model (1.1). We also argue that the pair specific fixed effects $g_{ij} = g(\xi_i, \xi_j)$ are identified for all pairs of agents i and j and can be (uniformly) consistently estimated. Identification of g_{ij} is an important result in itself since in many applications the fixed effects are the central objects of interest, rather than β_0. Moreover, this result is of special significance when F is a nonlinear function: identification of the pair specific fixed effects allows us to identify both the pair-specific and the average partial effects. Lastly, we also establish identification of the partial effects for a nonparametric version of model (1.1). This demonstrates that the previously established identification results are not driven by the parametric functional

forms of models (1.1) and (1.2).

Finally, we point out that identification of the error free outcomes Y_{ij}^* is a powerful result in itself. To the best of our knowledge, it has not been previously recognized in the econometric literature on identification of network and two-way models. In fact, the same result holds in fully nonparametric and non-separable settings, covering a wide range of dyadic network and interaction models with unobserved heterogeneity. As illustrated earlier in the context of model (1.2) with general heteroskedasticity, treating Y_{ij}^* as effectively observed substantially simplifies analysis of identification aspects of network models and, hence, provides a foundation for further results.

We also want to highlight the difference between identification of the error free outcomes $\hat{Y}_{ij}^*$, which is established in this paper, and the results previously obtained in the matrix (graphon) estimation/completion literature. The statistic literature defines consistency of matrix estimators and studies their rates of convergence in terms of the mean square error (MSE), i.e., the average of $(\hat{Y}_{ij}^* - Y_{ij}^*)^2$ taken across all matrix entries (pairs of agents).[5] However, consistency of an estimator in terms of the MSE does not necessarily imply that $\hat{Y}_{ij}^*$ is getting arbitrarily close to Y_{ij}^* (uniformly) for all pairs of agents with high probability as the sample size increases. To formally establish identification of the error free outcomes (which is econometrically important), we construct a uniformly (across all pairs of agents) consistent estimator of Y_{ij}^*. Although based on the approach of Zhang, Levina, and Zhu (2017), the estimator we propose is different (in fact, the estimator of Zhang, Levina, and Zhu (2017)[6] is not necessarily uniformly consistent for Y_{ij}^*; see Section 4.3 for a detailed comparison). Thus, our work also contributes to the matrix (graphon) estimation literature by providing an estimator and establishing its consistency in the max norm.

This paper contributes to the literature on econometrics of networks and, more generally, two-way models. The distinctive feature of our model is allowing for flexible nonparametric unobserved heterogeneity: the fixed effects can interact via the unknown coupling function $g(\cdot,\cdot)$. Importantly, we do not require $g(\cdot,\cdot)$ to have any particular structure (other than a certain degree of smoothness) and do not specify the dimensionality of the fixed effects. This is in contrast to most of the existing approaches, which either explicitly specify the form of $g(\cdot,\cdot)$ or impose restrictive assumptions on its shape.

Among explicitly specified forms of $g(\cdot,\cdot)$, the additive fixed effects structure $g(\xi_i,\xi_j)=\xi_i+\xi_j$ is by far the most popular way of incorporating unobserved heterogeneity in dyadic network models. For example, Graham (2017), Jochmans (2018), Dzemski (2018), and Yan, Jiang, Fienberg, and Leng (2019) study semiparametric network formation models treating the fixed

[5] E.g., Chatterjee (2015); Gao, Lu, and Zhou (2015); Klopp, Tsybakov, and Verzelen (2017); Zhang, Levina, and Zhu (2017); Li, Shah, Song, and Yu (2019).

[6] As well as the other approaches developed in the statistic literature.

effects as nuisance parameters to be estimated. They focus on inference on the common parameters (analogous to β_0) under the large network asymptotics. Since the number of the nuisance parameters grows with the network size, the (joint) maximum likelihood estimator of the common parameters suffers from the incidental parameter bias (Neyman and Scott, 1948). Dzemski (2018) and Yan, Jiang, Fienberg, and Leng (2019) analytically remove the bias, building on the result of Fernández-Val and Weidner (2016). Graham (2017) and Jochmans (2018) consider models with logistic errors and apply sufficiency arguments to avoid estimation of the nuisance parameters (a similar approach is also proposed in Charbonneau, 2017). Candelaria (2016) considers a version of the model of Graham (2017) relaxing the logistic distributional assumption.[7] The factor fixed effects structure of $g(\xi_i, \xi_j) = \xi_i'\xi_j$ is another specification commonly employed in panel and network models. Unlike the additive fixed effects framework, the factor structure allows for interactive unobserved heterogeneity.[8] For example, Chen, Fernández-Val, and Weidner (2014) develop estimation and inference tools for nonlinear semiparametric factor network models. We refer the interested reader to Bai and Wang (2016) and Fernández-Val and Weidner (2018) for recent reviews on large factor and panel models with additive fixed effects, respectively.[9] Notice that unlike the works mentioned above, we focus on identification and estimation of β_0 under substantially much more general structure of unobserved heterogeneity, but we do not develop inference tools.

Gao (2019) studies identification of a generalized version of the model of Graham (2017) allowing, in particular, for coupling of the (scalar) fixed effects via an unknown function. However, unlike this paper, Gao (2019) requires $g(\xi_i, \xi_j)$ to be strictly increasing (in both arguments). While being more general than the additive fixed effect structure, this form of $g(\cdot, \cdot)$ still implies that ξ can be interpreted as a popularity index (which rules out latent homophily). As a result, agents with the same values of ξ can be identified based on degree sorting (after conditioning on their covariates). Notice that no such sorting exists in the general setting when the form of $g(\cdot, \cdot)$ is not specified.

The discrete fixed effects approach has recently become another common technique to model unobserved heterogeneity in single agent (e.g., Hahn and Moon, 2010; Bonhomme and Manresa, 2015) and interactive settings (e.g., Bonhomme, Lamadon, and Manresa, 2019). With ξ being

[7] Toth (2017) also establishes identification of β_0 in a very similar semiparametric setting but does not derive the asymptotic distribution of the proposed estimators.

[8] For instance, as noted in Chen, Fernández-Val, and Weidner (2014), the factor fixed effect structure allows for certain forms of latent homophily.

[9] It is also worth noting that while this paper as well as most of the econometric literature do not specify the joint distribution of the agents' observed and unobserved characteristics (the fixed effects approach), the random effects approach is commonly employed to incorporate unobserved heterogeneity in the statistic literature (e.g., Hoff, Raftery, and Handcock, 2002; Handcock, Raftery, and Tantrum, 2007; Krivitsky, Handcock, Raftery, and Hoff, 2009). In the econometric literature, the random effects approach is utilized, for example, in Goldsmith-Pinkham and Imbens (2013), Hsieh and Lee (2016), and Mele (2017b) among others.

discrete, the considered network model (1.1) belongs to the class of stochastic block models (Holland, Laskey, and Leinhardt, 1983). While stochastic block models are routinely employed for community detection and networks/graphon estimation in statistics (see, for example, Airoldi, Blei, Fienberg, and Xing, 2008; Bickel and Chen, 2009; Amini, Chen, Bickel, Levina et al., 2013 among many others), relatively small number of works incorporate observable nodal covariates (e.g., Choi, Wolfe, and Airoldi, 2012; Roy, Atchadé, and Michailidis, 2019; Mele, Hao, Cape, and Priebe, 2019). Although our model and estimation approach are general enough to cover the stochastic block model, we focus on a case when the fixed effects are continuously distributed. In fact, the asymptotic analysis substantially simplifies when the fixed effects have finite support. In this case, the true cluster membership can be correctly determined (for example, based on the same pseudo-distance $\hat{d}_{ij}$) with probability approaching one (e.g., Hahn and Moon, 2010).

Another recent stream of the literature stresses the importance of accounting for endogeneity of the network formation process in such contexts as estimation of peer effects or, more generally, spatial autoregressive models (e.g., Goldsmith-Pinkham and Imbens, 2013; Qu and Lee, 2015; Hsieh and Lee, 2016; Arduini, Patacchini, and Rainone, 2015; Johnsson and Moon, 2019; Auerbach, 2016). Unlike in our paper, the central outcomes of interest in these works are individual whereas the network structure effectively serves as one the of explanatory (e.g., in the linear-in-means model of Manski, 1993) or control variables. The most related works are Auerbach (2016) and Johnsson and Moon (2019), where the source of endogeneity is the agents' latent characteristics (fixed effects), which affect both the links formation process as well as the individual outcomes of interest. Similarly to this paper, Auerbach (2016) considers a general network formation model leaving $g(\cdot,\cdot)$ unrestricted,[10] while Johnsson and Moon (2019) assume that $g(\cdot,\cdot)$ is strictly increasing in its arguments. Both Auerbach (2016) and Johnsson and Moon (2019) demonstrate that certain networks statistics can be employed to identify agents with the same values of ξ, which, in turn, can be used to account for endogeneity caused by the latent characteristics. The important difference between this paper and the works of Auerbach (2016) and Johnsson and Moon (2019) is that we use the same network data both to identify the parameters of interest and to control for unobserved heterogeneity. This is in contrast to the setting of the former works, which model the network formation process to tackle endogeneity in the other regression of interest.

Finally, we highlight that the considered model (1.1) does not incorporate interaction externalities. Specifically, we assume that conditional on the agents' observed and unobserved characteristics, the interactions outcomes are independent. This assumption is plausible when the interactions are primarily bilateral. Alternatively, as demonstrated by Mele (2017a), the considered model can be interpreted as a reduced form approximation of a strategic network

[10]The network formation model of Auerbach (2016), however, does not include observed covariates. More generally, the approach of Auerbach (2016) requires the variables of interest to be excluded from the network formation process.

formation game with non-negative externalities.[11] For recent reviews on (both strategic and reduced form) network formation models, we refer the interested reader to Graham (2015), Chandrasekhar (2016), and De Paula (2017).

The rest of the paper is organized as follows. In the next section we formally introduce the framework and provide (heuristic) identification arguments. We start with considering a homoskedastic version of model (1.2) and then extend the proposed identification strategy to allow for a general form of heteroskedasticity of the idiosyncratic errors. Section 3 turns the ideas of Section 2 into estimators of the parameters of interest. In Section 4 we establish consistency of the proposed estimators and derive their rates of convergence. In Section 5 we generalize the proposed identification argument to cover more general settings including model (1.1) as well as its nonparametric analogue. We also discuss extensions to directed networks and, more generally, two-way models. Section 6 illustrates the finite sample properties of the proposed estimators, and Section 7 concludes.

2 Identification of the semiparametric model

2.1 The model

We consider a network consisting of n agents. Each agent i is endowed with characteristics (X_i, ξ_i), where $X_i \in \mathcal{X}$ is observed by the econometrician whereas $\xi_i \in \mathcal{E}$ is not. We start with the following semiparametric regression model, where the (scalar) outcome of the interaction between agents i and j is given by

$$Y_{ij} = w(X_i, X_j)'\beta_0 + g(\xi_i, \xi_j) + \varepsilon_{ij}, \quad i \neq j. \tag{2.1}$$

Here, $w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^p$ is a known function, which transforms the observed characteristics of agents i and j into a pair-specific vector of covariates $W_{ij} \coloneqq w(X_i, X_j)$, $\beta_0 \in \mathbb{R}^p$ is the parameter of interest, and ε_{ij} is an unobserved idiosyncratic error. Note that unlike w, the coupling function $g : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ is *unknown*, and the dimension of the fixed effect $\xi_i \in \mathcal{E}$ is *not specified*. For simplicity of exposition, suppose that $\xi_i \in \mathbb{R}^{d_\xi}$ (the same insights apply when $\mathcal{E}$ is a general metric space).

First, we focus on an undirected model with $Y_{ij} = Y_{ji}$, so w and g are symmetric functions,

[11] The recent works studying identification and estimation of strategic network formation models also feature De Paula, Richards-Shubik, and Tamer (2018); Graham (2016); Sheng (2016); Ridder and Sheng (2015); Menzel (2015); Mele (2017b); Leung (2019); Leung and Moon (2019) among others.

and $\varepsilon_{ij} = \varepsilon_{ji}$.[12] The following assumption formalizes the sampling process.

Assumption 1.

(i) $\{(X_i, \xi_i)\}_{i=1}^n$ are i.i.d.;

(ii) conditional on $\{(X_i, \xi_i)\}_{i=1}^n$, the idiosyncratic errors $\{\varepsilon_{ij}\}_{i<j}$ are independent draws from $P_{\varepsilon_{ij}|X_i,\xi_i,X_j,\xi_j}$ with $\mathbb{E}[\varepsilon_{ij}|X_i, \xi_i, X_j, \xi_j] = 0$, and $\varepsilon_{ij} = \varepsilon_{ji}$;

(iii) the econometrician observes $\{X_i\}_{i=1}^n$ and $\{Y_{ij}\}_{i\neq j}$ determined by (2.1).

Assumption 1 is standard for the networks literature. The sampling process could be thought of as follows. First, the characteristics of agents $\{(X_i, \xi_i)\}_{i=1}^n$ are independently drawn from some population distribution. Then, conditional on the drawn characteristics, the idiosyncratic errors $\{\varepsilon_{ij}\}_{i<j}$ are independently drawn from the conditional distributions, which potentially depend on the characteristics of the corresponding agents (X_i, ξ_i) and (X_j, ξ_j).

Remark 2.1. For simplicity of exposition, we suppose that we observe Y_{ij} for all pairs of agents i and j (Assumption 1 (iii)). In Section 5.2, we discuss how to incorporate missing outcomes into the considered framework.

2.2 Identification of β_0: main insights

We study identification and estimation of β_0 under the large network asymptotics, which takes $n \to \infty$. The identification argument is based on the following observation. Suppose that we can identify two agents i and j with the same unobserved characteristics, i.e., with $\xi_i = \xi_j$. Then, for any third agent k, the difference between Y_{ik} and Y_{jk} is given by

$$Y_{ik} - Y_{jk} = \underbrace{(w(X_i, X_k) - w(X_j, X_k))'}_{(W_{ik}-W_{jk})'} \beta_0 + \varepsilon_{ik} - \varepsilon_{jk}. \tag{2.2}$$

The conditional mean independence of the regression errors now guarantees that β_0 can be identified by the regression of $Y_{ik} - Y_{jk}$ on $W_{ik} - W_{jk}$, provided that we have "enough" variation in $W_{ik} - W_{jk}$. Formally, we have

$$\beta_0 = \mathbb{E}[(W_{ik} - W_{jk})(W_{ik} - W_{jk})'|X_i, \xi_i, X_j, \xi_j]^{-1} \mathbb{E}[(W_{ik} - W_{jk})(Y_{ik} - Y_{jk})|X_i, \xi_i, X_j, \xi_j], \tag{2.3}$$

provided that $\mathbb{E}[(W_{ik} - W_{jk})(W_{ik} - W_{jk})'|X_i, \xi_i, X_j, \xi_j]$ is invertible. Since agents i and j are treated as fixed, the expectations are conditional on their characteristics (X_i, ξ_i) and (X_j, ξ_j). At

[12] In Section 5.3, we discuss how the considered identification and estimation approaches can be extended to undirected networks and, more generally, two-way models.

the same time, (X_k, ξ_k), the characteristics of agent k, and the idiosyncratic errors ε_{ik} and ε_{jk} are treated as random and integrated over. Note that the invertibility requirement insists on X_i and X_j, the observed characteristics of agents i and j, to be "sufficiently different". Indeed, if not only $\xi_i = \xi_j$ but also $X_i = X_j$, this condition is clearly violated since $W_{ik} - W_{jk} = 0$ for any agent k: in this case, β_0 can not be identified from the regression (2.2).

Hence, the problem of identification of β_0 can be reduced to the problem of identification of agents i and j with the same values of the unobserved fixed effects ($\xi_i = \xi_j$) but "sufficiently different" values of X_i and X_j.

Let Y_{ij}^* be the error free part of Y_{ij}, i.e.,

$$Y_{ij}^* := \mathbb{E}\left[Y_{ij} | X_i, \xi_i, X_j, \xi_j\right] = w(X_i, X_j)'\beta_0 + g(\xi_i, \xi_j). \tag{2.4}$$

Consider the following pseudo-distance between agents i and j

$$\begin{aligned} d_{ij}^2 &:= \min_{\beta \in \mathcal{B}} \mathbb{E}\left[(Y_{ik}^* - Y_{jk}^* - (W_{ik} - W_{jk})'\beta)^2 | X_i, \xi_i, X_j, \xi_j\right] \\ &= \min_{\beta \in \mathcal{B}} \mathbb{E}\Bigg[\underbrace{(g(\xi_i, \xi_k) - g(\xi_j, \xi_k)}_{=0, \text{ when } \xi_i = \xi_j} - (W_{ik} - W_{jk})'(\beta - \beta_0))^2 | X_i, \xi_i, X_j, \xi_j\Bigg], \end{aligned} \tag{2.5}$$

where $\mathcal{B} \ni \beta_0$ is some parameter space. Here the expectation is conditional on the characteristics of agents i and j and is taken over (X_k, ξ_k). Clearly, $d_{ij}^2 = 0$ when $\xi_i = \xi_j$: in this case, the minimum is achieved at $\beta = \beta_0$. Moreover, under a suitable (rank) condition (which we will formally discuss in Section 4.1), $d_{ij}^2 = 0$ also necessarily implies that $\xi_i = \xi_j$. Consequently, if d_{ij}^2 were available, agents with the same values of ξ could be identified based on this pseudo-distance.

However, the expectation (2.5) can not be directly computed, since the error free outcomes Y_{ij}^* are not observed. Next, we argue that the pseudo-distances d_{ij}^2 (or its close analogue) are identified for all pairs of agents i and j and, hence, can be used to identify agents with the same values of ξ (and different values of X).

2.3 Identification under conditional homoskedasticity

In this section, we consider a case when the regression errors are homoskedastic. Specifically, suppose

$$\mathbb{E}\left[\varepsilon_{ij}^2 | X_i, \xi_i, X_j, \xi_j\right] = \sigma^2 \quad \text{a.s.} \tag{2.6}$$

For a pair of agents i and j, consider the following conditional expectation

$$q_{ij}^2 := \min_{\beta \in \mathcal{B}} \mathbb{E}\left[(Y_{ik} - Y_{jk} - (W_{ik} - W_{jk})'\beta)^2 | X_i, \xi_i, X_j, \xi_j\right]. \tag{2.7}$$

Essentially, q_{ij}^2 is a feasible analogue of d_{ij}^2 with Y_{ik} and Y_{jk} replacing Y_{ik}^* and Y_{jk}^*. Importantly, unlike d_{ij}^2, q_{ij}^2 is immediately identified and can be estimated by

$$\hat{q}_{ij}^2 := \min_{\beta \in \mathcal{B}} \frac{1}{n-2} \sum_{k \neq i,j} (Y_{ik} - Y_{jk} - (W_{ik} - W_{jk})'\beta)^2. \tag{2.8}$$

Notice that since $Y_{ik} = Y_{ik}^* + \varepsilon_{ik}$ and $Y_{jk} = Y_{jk}^* + \varepsilon_{jk}$,

$$\begin{aligned} q_{ij}^2 &= \min_{\beta \in \mathcal{B}} \mathbb{E}\left[(Y_{ik}^* - Y_{jk}^* - (W_{ik} - W_{jk})'\beta + \varepsilon_{ik} - \varepsilon_{jk})^2 | X_i, \xi_i, X_j, \xi_j\right] \\ &= \min_{\beta \in \mathcal{B}} \mathbb{E}\left[(Y_{ik}^* - Y_{jk}^* - (W_{ik} - W_{jk})'\beta)^2 | X_i, \xi_i, X_j, \xi_j\right] + \mathbb{E}\left[\varepsilon_{ik}^2 + \varepsilon_{jk}^2 | X_i, \xi_i, X_j, \xi_j\right] \\ &= d_{ij}^2 + \mathbb{E}\left[\varepsilon_{ik}^2 | X_i, \xi_i\right] + \mathbb{E}\left[\varepsilon_{jk}^2 | X_j, \xi_j\right], \end{aligned} \tag{2.9}$$

where the second and the third equalities follow from Assumption 1 (ii). Hence, when the errors are homoskedastic and (2.6) holds, we have

$$q_{ij}^2 = d_{ij}^2 + 2\sigma^2. \tag{2.10}$$

Thus, for every pair of agents i and j, q_{ij}^2 differs from the pseudo-distance d_{ij}^2 by a constant term $2\sigma^2$.

Imagine that for a fixed agent i, we are looking for a match j with the same value of ξ. As discussed in Section 2.2, such an agent can be identified by minimizing d_{ij}^2. Then, (2.10) ensures that in the homoskedastic setting, such an agent can also be identified by minimizing q_{ij}^2. Hence, agents with the same values of ξ (and different values of X) can be identified based on q_{ij}^2, which can be directly estimated.

We want to stress that the identification argument provided for the homoskedastic model can be naturally extended to allow for $\mathbb{E}[\varepsilon_{ij}^2 | X_i, \xi_i, X_j, \xi_j] = \mathbb{E}[\varepsilon_{ij}^2 | X_i, X_j]$. Indeed, if the skedastic function does not depend on the unobserved characteristics, conditioning on some fixed value $X_j = x$ makes the third term $\mathbb{E}[\varepsilon_{jk}^2 | X_j = x, \xi_j] = \mathbb{E}[\varepsilon_{jk}^2 | X_j = x]$ constant again. In this case, like in the homoskedastic model, q_{ij}^2 is minimized whenever d_{ij}^2 is, which allows us to identify agents with the same values of ξ.

Remark 2.2. The identification argument provided above is heuristic and will be formalized later. Specifically, we turn these ideas into an estimator of β_0 in Section 3 and establish its rate of

convergence in Section 4.

Remark 2.3. Although homoskedasticity is considered to be an unattractive and unrealistic assumption in the modern empirical analysis, it might not necessarily be that restrictive in our context. It is common for many empirical studies that the sample variance $(n-2)^{-1}\sum_{j\neq i}(Y_{ij}-\overline{Y}_i)^2$, where $\overline{Y}_i := (n-1)^{-1}\sum_{j\neq i}Y_{ij}$, substantially varies across i, even after controlling for the observed characteristics X. This, however, does not necessarily contradicts the conditional homoskedasticity requirement (2.6). Indeed, $\varepsilon_{ij} = Y_{ij} - \mathbb{E}\left[Y_{ij}|X_i,\xi_i,X_i,\xi_j\right]$ accounts for the variation in Y_{ij} not explained by both the observable and unobservable characteristics of agents i and j. Since our model allows for a very flexible form of the interaction between the fixed effects (and their dimension is also not specified), a large part of the variation in $(n-2)^{-1}\sum_{j\neq i}(Y_{ij}-\overline{Y}_i)^2$ (after controlling on X) across agents can potentially be attributed to the difference in their unobserved characteristics ξ.

2.4 Identification under general heteroskedasticity

Under general heteroskedasticity of the errors, the identification strategy based on q_{ij}^2 no longer guarantees finding agents with the same values of ξ. Consider the same process of finding an appropriate match j for a fixed agent i. As shown in (2.9), q_{ij}^2 can be represented as a sum of three components. The first term d_{ij}^2, which we will call the signal, identifies agents with the same values of ξ. The second term $\mathbb{E}\left[\varepsilon_{ik}^2|X_i,\xi_i\right]$ does not depend on j. However, under general heteroskedasticity, the third term $\mathbb{E}[\varepsilon_{jk}^2|X_j,\xi_j]$ depends on ξ_j and distorts the signal. Hence, the identification argument provided in Section 2.3 is no longer valid in this case.

In this section, we address this issue and extend the argument of Sections 2.2 and 2.3 to a model with general heteroskedasticity. Specifically, we (heuristically) argue that the error free outcomes Y_{ij}^* are identified for all pairs of agents i and j. As a result, the pseudo-distance d_{ij}^2 introduced in (2.5) is also identified and can be directly employed to find agents with the same values of ξ (and different of X).

2.4.1 Identification of Y_{ij}^*

With Y_{ij}^* and $Y_{ij} = Y_{ij}^* + \varepsilon_{ij}$ collected as entries of $n\times n$ matrices Y^* and Y (with diagonal elements of Y missing), the problem of identification and estimation of Y^* based on its noisy proxy Y can be interpreted as a particular variation of the classical matrix estimation/completion problem. Specifically, it turns out that the considered network model (2.1) is an example of the latent space model (see, for example, Chatterjee (2015) and the references therein). Precisely, in the latent

space model, the entries of (symmetric) matrix Y should have the form of

$$Y_{ij} = f(Z_i, Z_j) + \varepsilon_{ij}, \tag{2.11}$$

where f is some symmetric function, $Z_1, \ldots, Z_n$ are some latent variables associated with the corresponding rows and columns of Y, and the errors $\{\varepsilon_{ij}\}_{i<j}$ are assumed to be (conditionally) independent.[13] Clearly, the model (2.1) fits this framework with $Z_i = (X_i, \xi_i)$. The problem of estimation of $Y^*_{ij} = f(Z_i, Z_j)$ from Y is also known as nonparametric regression without knowing the design (Gao, Lu, and Zhou, 2015) or blind regression (Li, Shah, Song, and Yu, 2019). If Y_{ij} is binary, Y can be interpreted as the adjacency matrix of a random graph. In this case, the function f is called graphon, and this problem is commonly referred to as graphon estimation (see, for example, Gao, Lu, and Zhou, 2015; Klopp, Tsybakov, and Verzelen, 2017; Zhang, Levina, and Zhu, 2017).

It turns out that the particular structure of the latent space model (2.11) allows constructing a consistent estimator of Y^* based on a single measurement Y. For example, Chatterjee (2015); Gao, Lu, and Zhou (2015); Klopp, Tsybakov, and Verzelen (2017); Zhang, Levina, and Zhu (2017); Li, Shah, Song, and Yu (2019) construct such estimators and establish their consistency in terms of the mean square error (MSE).

In particular, we build on the estimation strategy of Zhang, Levina, and Zhu (2017) to argue that the error free outcomes Y^*_{ij} are identified for all pairs of agents i and j. The proposed identification strategy consists of two main steps. First, we argue that we can identify agents with the same values of X and ξ. Then, building on this result, we demonstrate how Y^*_{ij} can be constructively identified.

Step 1: Identification of agents with the same values of X and ξ

Consider a subpopulation of agents with a fixed value of $X = x$ exclusively. Let $g_x(\xi_i, \xi_j) \coloneqq w(x,x)'\beta_0 + g(\xi_i, \xi_j)$ and $P_{\xi|X}(\xi|x)$ denote the conditional distribution of ξ given $X = x$. In this subpopulation, consider the following (squared) pseudo-distance between agents i and j

$$\begin{aligned} d^2_\infty(i,j;x) &\coloneqq \sup_{\xi_k \in \text{supp}(\xi|X=x)} \left|\mathbb{E}\left[(Y_{i\ell} - Y_{j\ell})Y_{k\ell}|\xi_i, \xi_j, \xi_k, X = x\right]\right| \\ &= \sup_{\xi_k \in \text{supp}(\xi|X=x)} \left|\mathbb{E}\left[(g_x(\xi_i, \xi_\ell) - g_x(\xi_j, \xi_\ell))g_x(\xi_k, \xi_\ell)|\xi_i, \xi_j, \xi_k, X = x\right]\right| \\ &= \sup_{\xi_k \in \text{supp}(\xi|X=x)} \left|\int (g_x(\xi_i, \xi_\ell) - g_x(\xi_j, \xi_\ell))g_x(\xi_k, \xi_\ell) dP_{\xi|X}(\xi_\ell; x)\right|, \end{aligned}$$

[13] As noted, for example, in Bickel and Chen (2009) and Bickel, Chen, and Levina (2011), the latent space model is natural in exchangeable settings due to the Aldous-Hoover theorem (Aldous, 1981; Hoover, 1979). For a detailed discussion of this result and other representation theorems for exchangeable random arrays, see, for example, Kallenberg (2005) and Orbanz and Roy (2015).

where the second equality exploits Assumption 1 (ii).

The finite sample analogue of d^2_∞ was originally proposed in Zhang, Levina, and Zhu (2017) in the context of nonparametric graphon estimation. The considered pseudo-distance is also closely related to the so-called similarity distance, a more abstract concept, which proves to be particularly useful for studying topological properties of graphons (see, for example, Lovász (2012) and the references therein).[14]

First, notice that under certain smoothness conditions, $d^2_\infty(i,j;x)$ is directly identified and, if a sample of n_x agents with $X = x$ is available, can be estimated by

$$\hat{d}^2_\infty(i,j;x) := \max_{k \neq i,j} \left| \frac{1}{n_x - 3} \sum_{\ell \neq i,j,k} (Y_{i\ell} - Y_{j\ell}) Y_{k\ell} \right|.$$

Second, note that $d^2_\infty(i,j;x) = 0$ implies that

$$\int (g_x(\xi_i,\xi_\ell) - g_x(\xi_j,\xi_\ell))\, g_x(\xi_k,\xi_\ell) dP_{\xi|X}(\xi_\ell;x) = 0$$

for almost all ξ_k.[15] In particular, we have

$$\int (g_x(\xi_i,\xi_\ell) - g_x(\xi_j,\xi_\ell))\, g_x(\xi_i,\xi_\ell) dP_{\xi|X}(\xi_\ell;x) = 0, \tag{2.12}$$

$$\int (g_x(\xi_i,\xi_\ell) - g_x(\xi_j,\xi_\ell))\, g_x(\xi_j,\xi_\ell) dP_{\xi|X}(\xi_\ell;x) = 0. \tag{2.13}$$

Hence, subtracting (2.13) from (2.12), we conclude

$$\int (g_x(\xi_i,\xi_\ell) - g_x(\xi_j,\xi_\ell))^2\, dP_{\xi|X}(\xi_\ell;x) = \int (g(\xi_i,\xi_\ell) - g(\xi_j,\xi_\ell))^2\, dP_{\xi|X}(\xi_\ell;x) = 0. \tag{2.14}$$

Thus, $d^2_\infty(i,j;x) = 0$ ensures that $g(\xi_i,\cdot)$ and $g(\xi_j,\cdot)$ are the same (in terms of the L_2 distance associated with the conditional distribution $\xi|X = x$). The following assumption guarantees that equivalence of agents i and j in terms of $g(\xi_i,\cdot)$ and $g(\xi_j,\cdot)$ also necessarily implies that $\xi_i = \xi_j$.

Assumption 2. For each $\delta > 0$, there exits $C_\delta > 0$, such that for all $x \in \text{supp}\,(X)$,

$$\|g(\xi_1,\cdot) - g(\xi_2,\cdot)\|_{2,x} := \left(\int (g(\xi_1,\xi) - g(\xi_2,\xi))^2\, dP_{\xi|X}(\xi;x) \right)^{1/2} > C_\delta$$

[14] Auerbach (2016) also utilizes another related pseudo-distance in the network formation context. Specifically, Auerbach (2016) evaluates the agents' similarity based on the L_2 distance between functions $\varphi(\xi_i,\cdot)$ and $\varphi(\xi_j,\cdot)$, where $\varphi(\xi_i,\xi_k) := \mathbb{E}\,[Y_{i\ell}Y_{k\ell}|\xi_i,\xi_j]$. At the same time, the pseudo-distance considered in this paper (and in Zhang, Levina, and Zhu, 2017) and the similarity distance of Lovász (2012) correspond to the L_∞ and L_1 distances between $\varphi(\xi_i,\cdot)$ and $\varphi(\xi_j,\cdot)$, respectively.

[15] Similar arguments are also provided in Lovász (2012) and Auerbach (2016).

for all $\xi_1, \xi_2 \in \mathcal{E}$ satisfying $\|\xi_1 - \xi_2\| \geqslant \delta$.

Assumption 2 ensures that agents i and j with different values of ξ are necessarily different in terms of $g(\xi_i, \cdot)$ and $g(\xi_j, \cdot)$, i.e., the L_2 distance (associated with the conditional distribution $\xi|X = x$) between $g(\xi_i, \cdot)$ and $g(\xi_j, \cdot)$ is bounded away from zero whenever $\|\xi_i - \xi_j\|$ is. Combined with (2.14), Assumption 2 guarantees that $d^2_\infty(i, j; x) = 0$ implies $\xi_i = \xi_j$. Consequently, we can identify agents with the same values of ξ and $X = x$ based on the pseudo-distance $d^2_\infty(i, j; x)$.

Discussion of Assumption 2.

Notice that since g is not specified, the meaningful interpretation of unobserved ξ is unclear. Assumption 2 interprets ξ is a collection of the effective unobserved fixed effects. Specifically, it means that every component of ξ affects the shape of $g(\xi, \cdot)$ in a non-trivial and unique way, so distinctively different ξ_i and ξ_j are associated with distinctively different $g(\xi_i, \cdot)$ and $g(\xi_j, \cdot)$. Assumption 2 clearly rules out a situation when some component of ξ has no actual impact on $g(\xi, \cdot)$. It also rules out a possibility that one of the components of ξ perfectly replicates or offsets the impact of the other component. For example, suppose that ξ is two dimensional and both components affect g in a pure additive way, so $g(\xi_i, \xi_j) = \xi_{1i} + \xi_{2i} + \xi_{1j} + \xi_{2j}$. Such a situation is precluded by Assumption 2 since all the agents with the same values of $\xi_1 + \xi_2$ produce exactly the same functions $g(\xi, \cdot)$. Also notice that redefining $\tilde{\xi} = \xi_1 + \xi_2$ and $\tilde{g}(\tilde{\xi}_i, \tilde{\xi}_j) = \tilde{\xi}_i + \tilde{\xi}_j$ solves the problem.

Finally, note that Assumption 2 does not rule out the possibility of the absence of unobserved heterogeneity. Since we do not restrict the distribution of ξ, it is allowed for all agents to have the same value of the fixed effect $\xi = \xi_0$ (no unobserved heterogeneity).

Step 2: Identification of Y^*_{ij}

Now, being able to identify agents with the same values of X and ξ, we can also identify the error free outcome $Y^*_{ij} = w(X_i, X_j)'\beta_0 + g(\xi_i, \xi_j)$ for any pair of agents i and j. Specifically, for a fixed agent i, we can construct a collection of agents with $X = X_i$ and $\xi = \xi_i$, i.e., $\mathcal{N}_i := \{i' : X_{i'} = X_i, \xi_{i'} = \xi_i\}$. Similarly, we construct $\mathcal{N}_j := \{j' : X_{j'} = X_j, \xi_{j'} = \xi_j\}$. Then, note that

$$\begin{aligned}
\frac{1}{n_i n_j} \sum_{i' \in \mathcal{N}_i} \sum_{j' \in \mathcal{N}_j} Y_{i'j'} &= \frac{1}{n_i n_j} \sum_{i' \in \mathcal{N}_i} \sum_{j' \in \mathcal{N}_j} \left(w(X_{i'}, X_{j'}) + g(\xi_{i'}, \xi_{j'}) + \varepsilon_{i'j'}\right) \\
&= \frac{1}{n_i n_j} \sum_{i' \in \mathcal{N}_i} \sum_{j' \in \mathcal{N}_j} \left(w(X_i, X_j) + g(\xi_i, \xi_j) + \varepsilon_{i'j'}\right) \\
&= Y^*_{ij} + \frac{1}{n_i n_j} \sum_{i' \in \mathcal{N}_i} \sum_{j' \in \mathcal{N}_j} \varepsilon_{i'j'} \\
&\xrightarrow{p} Y^*_{ij}, \qquad n_i, n_j \to \infty,
\end{aligned} \tag{2.15}$$

where n_i and n_j denote the number of elements in $\mathcal{N}_i$ and $\mathcal{N}_j$, respectively. Since in the population we can construct arbitrarily large $\mathcal{N}_i$ and $\mathcal{N}_j$, (2.15) implies that Y_{ij}^* is identified.

Remark 2.4. Although the identification argument provided above is heuristic, it captures the main insights and will be formalized. Specifically, in Section 4.3, we construct a particular estimator $\tilde{Y}_{ij}^*$ and demonstrate its uniform consistency, i.e., establish $\max_{i,j} \left|\tilde{Y}_{ij}^* - Y_{ij}^*\right| = o_p(1)$. This formally proves that Y_{ij}^* is identified for all i and j.

Identifiability of Y_{ij}^* is a strong result, which, to the best of our knowledge, is new to the econometrics literature on identification of network and, more generally, two way models. Importantly, it is not due to the specific parametric form or additive separability (in X and ξ) of the model (2.1). In fact, by essentially the same argument, the error free outcomes $Y_{ij}^* = f(X_i, \xi_i, X_j, \xi_j)$ are also identified in a fully non-separable and nonparametric model of the form

$$Y_{ij} = f(X_i, \xi_i, X_j, \xi_j) + \varepsilon_{ij}, \quad \mathbb{E}\left[\varepsilon_{ij} | X_i, \xi_i, X_j, \xi_j\right] = 0.$$

The established result implies that for studying identification aspects of a model, the noise free outcome Y_{ij}^* can be treated as directly observed. Since the noise part is removed, this greatly simplifies the analysis and provides a powerful foundation for establishing further identification results in a general context. For example, in the particular context of the model (2.1), identifiability of Y_{ij}^* implies that the pseudo-distances d_{ij}^2 are also identified for all pairs of agents i and j. Hence, as discussed in Section 2.2, agents with the same values of ξ (and different values of X) and, subsequently, β_0 can be identified based on d_{ij}^2.

3 Estimation of the Semiparametric Model

In this section, we turn the ideas of Section 2 into an estimation procedure. First, we construct an estimator of β_0 assuming that some estimator of the pseudo-distances $\hat{d}_{ij}^2$ is already available for the researcher. Then, we discuss how to construct $\hat{d}_{ij}^2$ in the homoskedastic and general heteroskedastic settings. We also briefly preview the asymptotic properties of the proposed estimators but postpone the formal analysis to Section 4.

3.1 Estimation of β_0

Suppose that we start with some estimator of the pseudo-distance $\hat{d}_{ij}^2$, which converges to d_{ij}^2 (uniformly across all pairs) at a certain rate R_n. Specifically, we assume that $\hat{d}_{ij}^2$ satisfies

$$\min_{i,j\neq i} \left| \hat{d}_{ij}^2 - d_{ij}^2 \right| = O_p(R_n^{-1}) \tag{3.1}$$

for some $R_n \to \infty$. Equipped with an estimator of the pseudo-distances $\{d_{ij}^2\}_{i\neq j}$, we propose using the following kernel based estimator of β_0

$$\hat{\beta} \coloneqq \left(\sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \sum_{k\neq i,j} (W_{ik} - W_{jk})(W_{ik} - W_{jk})' \right)^{-1} \left(\sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \sum_{k\neq i,j} (W_{ik} - W_{jk})(Y_{ik} - Y_{jk}) \right). \tag{3.2}$$

Hereafter, $\sum_{i<j} \coloneqq \sum_{i,j\in[n],i<j}$ and $\sum_{k\neq i,j} \coloneqq \sum_{k\in[n],k\neq i,j}$, where $[n] = \{1,\dots,n\}$. $K : \mathbb{R}_+ \to \mathbb{R}$ is some kernel, which is assumed to be supported on $[0,1]$. Finally, h_n is a bandwidth, which needs to satisfy $h_n \to 0$ (and some additional requirements) as $n \to \infty$.

As previously discussed, β_0 can be estimated by the regression of $Y_{ik} - Y_{jk}$ on $W_{ik} - W_{jk}$ with fixed agents i and j satisfying $\xi_i = \xi_j$, and, consequently, $d_{ij}^2 = 0$ (see Eq. (2.2) and (2.3)). However, in a finite sample, we are never guaranteed to find a pair of agents with exactly the same values of unobserved characteristics. The proposed estimator $\hat{\beta}$ addresses this issue: it combines all of the pairwise differenced regressions with the weights given by $K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right)$. Typically, the smaller $\hat{d}_{ij}^2$ is, the closer agents i and j appear to be in terms of ξ_i and ξ_j, and the higher weight is given to the corresponding pairwise differenced regression. Specifically, with probability approaching one, only the pairs that satisfy $\|\xi_i - \xi_j\| \leqslant \alpha h_n$ are given positive weights, where α is some positive constant. Since $h_n \to 0$, the quality of those matches increases and the bias introduced by the imperfect matching disappears as the sample size grows.

In Section 4.1 we provide necessary regularity conditions and formally establish the rate of convergence for $\hat{\beta}$. Specifically, we demonstrate that

$$\hat{\beta} - \beta_0 = O_p\left(h_n^2 + \frac{R_n^{-1}}{h_n} \right), \tag{3.3}$$

where R_n is as in (3.1). Here, the first term is due to the bias introduced by the imperfect matching, which is shown to be $O(h_n^2)$, and the second term captures how sampling uncertainty from the first step (estimation of d_{ij}^2) propagates to $\hat{\beta}$. As usual, under the optimal choice of $h_n \propto R_n^{-1/3}$, these

terms are of the same order, and

$$\hat{\beta} - \beta_0 = O_p\left(R_n^{-2/3}\right). \tag{3.4}$$

So, the rate of convergence for $\hat{\beta}$ crucially depends on R_n, the rate of uniform convergence for $\hat{d}_{ij}^2$.

Remark 3.1. As we will demonstrate later, R_n heavily depends on d_ξ, the dimension of the unobserved fixed effect ξ. Hence, the rate of convergence for $\hat{\beta}$ is also affected by d_ξ indirectly through R_n.

Remark 3.2. Analogously to the kernel based estimator $\hat{\beta}$, one could alternatively consider a nearest-neighbor type estimator. While $\hat{\beta}$ assigns each pair of agents the corresponding weight $K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right)$, a nearest neighbor type estimator assigns every agent i a certain (fixed or growing) number of matches closest to agent i in terms of $\hat{d}_{ij}^2$. For example, the 1 nearest neighbor estimator takes the form of

$$\hat{\beta}_{\text{NN1}} \coloneqq \left(\sum_{i=1}^{n}\sum_{k\neq i,\hat{j}(i)}(W_{ik} - W_{\hat{j}(i)k})(W_{ik} - W_{\hat{j}(i)k})'\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k\neq i,\hat{j}(i)}(W_{ik} - W_{\hat{j}(i)k})(Y_{ik} - Y_{\hat{j}(i)k})\right), \tag{3.5}$$

where $\hat{j}(i) \coloneqq \operatorname{argmin}_{j\neq i}\hat{d}_{ij}^2$ stands for the index of agent matched to agent i.[16] Although, a similar argument could be invoked to demonstrate consistency of nearest neighbor type estimators, a more detailed analysis of their asymptotic properties is intricate and out of the scope of this paper.

3.2 Estimation of d_{ij}^2

The estimator (3.2) proposed in Section 3.1 builds on the estimates of the pseudo-distances $\{\hat{d}_{ij}^2\}_{i\neq j}$. In this section, we construct particular estimators of d_{ij}^2 and briefly preview their asymptotic properties, for both homoskedastic and general heteroskedastic settings.

3.2.1 Estimation of d_{ij}^2 under homoskedasticity of the idiosyncratic errors

We start with considering the homoskedastic setting. Recall that in this case, the pseudo-distance of interest d_{ij}^2 is closely related to another quantity q_{ij}^2 defined in (2.7): specifically, $q_{ij}^2 = d_{ij}^2 + 2\sigma^2$, where σ^2 stands for the conditional variance of ε_{ij} (see Eq. (2.10)). Moreover, unlike d_{ij}^2, q_{ij}^2 can

[16] For some nearest neighbor type estimators, it also may be crucial to require the matched agents to have "sufficiently different" values of X. For example, we may require $\lambda_{min}\left((n-2)^{-1}\sum_{k\neq i,\hat{j}(i)}(W_{ik} - W_{\hat{j}(i)k})(W_{ik} - W_{\hat{j}(i)k})'\right) > \underline{\lambda_n} > 0$ for some $\underline{\lambda}_n$, which may (or may not) slowly converge to 0 as $n \to \infty$. We omit this requirement for $\hat{\beta}_{\text{NN1}}$ for the ease of notation.

be directly estimated from the raw data as in (2.8). We will demonstrate that under standard regularity condition, (2.8) is a (uniformly) consistent estimator of q_{ij}^2, which satisfies

$$\max_{i,j\neq i} \left|\hat{q}_{ij}^2 - q_{ij}^2\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Thus, as suggested by (2.10), a natural way to estimate d_{ij}^2 is to subtract $2\hat{\sigma}^2$ from $\hat{q}_{ij}^2$, where $\hat{\sigma}^2$ is a consistent estimator σ^2. One candidate for such an estimator is

$$2\hat{\sigma}^2 = \min_{i,j\neq j} \hat{q}_{ij}^2. \tag{3.6}$$

Indeed, in large samples, we expect $\min_{i,j\neq i} d_{ij}^2$ to be small since we are likely to find a pair of agents similar in terms of ξ. Hence, in large samples, $\min_{i,j\neq i} q_{ij}^2 = \min_{i,j\neq i} d_{ij}^2 + 2\sigma^2$ is expected to be close to $2\sigma^2$. Then, d_{ij}^2 can be estimated by

$$\hat{d}_{ij}^2 = \hat{q}_{ij}^2 - 2\hat{\sigma}^2 = \hat{q}_{ij}^2 - \min_{i,j\neq i} \hat{q}_{ij}^2. \tag{3.7}$$

In Section 4.2.1, we formally demonstrate that in the homoskedastic setting, this estimator satisfies

$$\max_{i,j\neq i} \left|\hat{d}_{ij}^2 - d_{ij}^2\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right)$$

when the dimension of ξ is not greater than 4, so (3.1) holds with $R_n = \left(\frac{n}{\ln n}\right)^{1/2}$. Hence, in this case, (3.4) implies that the rate of convergence for $\hat{\beta}$ is $\left(\frac{n}{\ln n}\right)^{1/3}$.

3.2.2 Estimation of d_{ij}^2 under general heteroskedasticity of the idiosyncratic errors

As suggested by Section 2.4, under general heteroskedasticity of the errors, the first step of estimation of d_{ij}^2 is to construct an estimator of Y_{ij}^*. For simplicity, we also consider a case when X is discrete and takes finitely many values. A general case and the results we provide below are also formally discussed in Section 4.2.2.

The estimator we propose is similar to the estimator of Zhang, Levina, and Zhu (2017), which was originally employed in the context of nonparametric graphon estimation. First, for all pairs of agents i and j, we estimate another pseudo-distance

$$\hat{d}_{\infty}^2(i,j) := \max_{k\neq i,j} \left|\frac{1}{n-3} \sum_{\ell\neq i,j,k} (Y_{i\ell} - Y_{j\ell}) Y_{k\ell}\right|.$$

Then, for any agent i, we define its neighborhood $\hat{\mathcal{N}}_i(n_i)$ as a collection of n_i agents closest to agent i in terms of $\hat{d}^2_\infty$ among all agents with $X = X_i$

$$\hat{\mathcal{N}}_i(n_i) \coloneqq \{i' : X_{i'} = X_i, \mathrm{Rank}(\hat{d}^2_\infty(i,i')|X = X_i) \leqslant n_i\}. \tag{3.8}$$

Also notice that by construction, $i \in \hat{\mathcal{N}}_i(n_i)$, so agent i is alway included to its neighborhood. Essentially, for any agent i, its neighborhood $\hat{\mathcal{N}}_i(n_i)$ is a collection of agents with the same observed and similar unobserved characteristics. Note that since X is discrete and takes finitely many values, we can insist on $X_{i'}$ being exactly equal to X_i. Also, note that the number of agents included in the neighborhoods should grow at a certain rate as the sample size increases. Specifically, we require $\underline{C}(n \ln n)^{1/2} \leqslant n_i \leqslant \overline{C}(n \ln n)^{1/2}$ for all i, for some positive constants $\underline{C}$ and $\overline{C}$.

Once the neighborhoods are constructed, we estimate Y^*_{ij} by

$$\hat{Y}^*_{ij} = \frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} Y_{i'j}}{n_i}, \tag{3.9}$$

where, for the ease of notation, we put $Y_{i'j} = 0$ whenever $i' = j$. Note that $\hat{Y}^*_{ij}$ is also defined for $i = j$: despite Y_{ii} is not observed (and defined), we still can estimate $Y^*_{ii} \coloneqq w(X_i, X_i) + g(\xi_i, \xi_i)$.

Remark 3.3. $\hat{Y}^*_{ij}$ defined in (3.9) is a neighborhood averaging type estimator. Another possible option is to consider a kernel based estimator of Y^*_{ij} given by

$$\hat{Y}^*_{ij} = \frac{\sum_{i'=1}^n \mathbb{1}\{X_{i'} = X_i\} \mathcal{K}\left(\frac{\hat{d}^2_\infty(i,i')}{h_n}\right) Y_{i'j}}{\sum_{i'=1}^n \mathbb{1}\{X_{i'} = X_i\} \mathcal{K}\left(\frac{\hat{d}^2_\infty(i,i')}{h_n}\right)},$$

where $\mathcal{K}$ and h_n are some kernel and bandwidth, which is supposed to go to 0 as the sample size increases. Although the kernel based estimator is a very natural generalization of (3.9), its asymptotic properties are less transparent. We do not pursue their analysis in this paper and leave for future research.

Finally, we estimate d^2_{ij} by

$$\hat{d}^2_{ij} \coloneqq \min_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{k=1}^n (\hat{Y}^*_{ik} - \hat{Y}^*_{jk} - (W_{ik} - W_{jk})'\beta)^2. \tag{3.10}$$

In Section 4.2.2, we formally establish that this estimator satisfies

$$\max_{i,j \neq i} \left|\hat{d}^2_{ij} - d^2_{ij}\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{2d_\xi}}\right),$$

where d_ξ is the dimension of ξ, so (3.1) holds with $R_n = \left(\frac{n}{\ln n}\right)^{\frac{1}{2d_\xi}}$. Specifically, when ξ is scalar (and X is discrete), $R_n = \left(\frac{n}{\ln n}\right)^{1/2}$ and, by (3.4), the rate of convergence for $\hat{\beta}$ is $\left(\frac{n}{\ln n}\right)^{1/3}$, which are exactly the same as in the homoskedastic case.

Remark 3.4. Notice that the proposed estimator (3.9) differs from the one discussed in Section 2.4.1. Specifically, (2.15) suggests using

$$\tilde{Y}_{ij}^* = \frac{1}{n_i n_j} \sum_{i' \in \hat{\mathcal{N}}_i(n_i)} \sum_{j' \in \hat{\mathcal{N}}_j(n_j)} Y_{i'j'}. \tag{3.11}$$

Recall that the rate of convergence for $\hat{\beta}$ depends on the asymptotic properties of the first step estimator $\hat{d}_{ij}^2$. While $\tilde{Y}_{ij}^*$ is a natural and (uniformly) consistent estimator of Y_{ij}^*, i.e., we have $\max_{i,j} \left|\tilde{Y}_{ij}^* - Y_{ij}^*\right| = o_p(1)$, it turns out that using $\hat{Y}_{ij}^*$ as in (3.9) theoretically guarantees a better rate of (uniform) convergence for $\hat{d}_{ij}^2$ and, consequently, for $\hat{\beta}$ too. Moreover, $\hat{Y}_{ij}^*$ is computationally more efficient.

4 Large Sample Theory

In this section, we formally study the asymptotic properties of the estimators we provided in Section 3.

The following set of basic regularity conditions will be used throughout the rest of the paper.

Assumption 3.

(i) $w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^p$ is a symmetric bounded measurable function, where $\operatorname{supp}(X) \subseteq \mathcal{X}$;

(ii) $\operatorname{supp}(\xi) \subseteq \mathcal{E}$, where $\mathcal{E}$ is a compact subset of $\mathbb{R}^{d_\xi}$;

(iii) $g : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ is a symmetric function; moreover, for some $\overline{G} > 0$, we have

$$|g(\xi_1, \xi) - g(\xi_2, \xi)| \leqslant \overline{G} \, \|\xi_1 - \xi_2\|$$

for all $\xi_1, \xi_2, \xi \in \mathcal{E}$;

(iv) for some $c > 0$, $\mathbb{E}\left[e^{\lambda \varepsilon_{ij}} | X_i, \xi_i, X_j, \xi_j\right] \leqslant e^{c\lambda^2}$ for all $\lambda \in \mathbb{R}$ a.s.

Conditions (i) and (ii) are standard. Condition (iii) requires g to be (bi-)Lipschitz continuous. Condition (iv) requires the conditional distribution of the error $\varepsilon_{ij} | X_i, \xi_i, X_j, \xi_j$ to be (uniformly over (X_i, ξ_i, X_j, ξ_j)) sub-Gaussian. It allows us to invoke certain concentration inequalities and derive rates of uniform convergence.

4.1 Rate of convergence for $\hat{\beta}$

In this section, we provide necessary regularity conditions and establish the rate of convergence for the kernel based estimator $\hat{\beta}$ introduced in (3.2). For simplicity of exposition, first, we consider a case when the unobserved fixed effect ξ is scalar.

Note that plugging $Y_{ik} - Y_{jk} = (W_{ik} - W_{jk})'\beta_0 + g(\xi_i, \xi_k) - g(\xi_j, \xi_k) + \varepsilon_{ik} - \varepsilon_{jk}$ into (3.2) gives

$$\hat{\beta} - \beta_0 = \left(\sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \sum_{k\neq i,j} (W_{ik} - W_{jk})(W_{ik} - W_{jk})' \right)^{-1} \times \left(\sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \sum_{k\neq i,j} (W_{ik} - W_{jk}) \left(g(\xi_i, \xi_k) - g(\xi_j, \xi_k) + \varepsilon_{ik} - \varepsilon_{jk}\right) \right).$$

Denote

$$\hat{A}_n := \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \frac{1}{n-2} \sum_{k\neq i,j} (W_{ik} - W_{jk})(W_{ik} - W_{jk})',$$

$$\hat{B}_n := \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \frac{1}{n-2} \sum_{k\neq i,j} (W_{ik} - W_{jk})(g(\xi_i, \xi_k) - g(\xi_j, \xi_k)),$$

$$\hat{C}_n := \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \frac{1}{n-2} \sum_{k\neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}).$$

Thus, $\hat{\beta} - \beta_0$ can be expressed as

$$\hat{\beta} - \beta_0 = \hat{A}_n^{-1}(\hat{B}_n + \hat{C}_n). \tag{4.1}$$

To derive the asymptotic properties of $\hat{A}_n$, $\hat{B}_n$, and $\hat{C}_n$, we introduce the following additional assumptions.

Assumption 4.

(i) $\xi \in \mathbb{R}$ and $\xi|X = x$ is continuously distributed for all $x \in \text{supp}\,(X)$; its conditional density $f_{\xi|X}$ (with respect to the Lebesgue measure) satisfies $\sup_{x\in\text{supp}(X)} \sup_{\xi\in\mathcal{E}} f_{\xi|X}(\xi|x) \leqslant \overline{f}_{\xi|X}$ for some constant $\overline{f}_{\xi|X} > 0$;

(ii) for all $x \in \text{supp}\,(X)$, $f_{\xi|X}(\xi|x)$ is continuous at almost all ξ (with respect to the conditional distribution of $\xi|X = x$); moreover, there exist positive constants $\overline{\delta}$ and γ such that for all $\delta \leqslant \overline{\delta}$ and for all $x \in \text{supp}\,(X)$,

$$\mathbb{P}\left(\xi_i \in \{\xi : f_{\xi|X}(\xi|x) \text{ is continuous on } B_\delta(\xi)\}|X_i = x\right) \geqslant 1 - \gamma\delta; \tag{4.2}$$

(iii) there exists $C_\xi > 0$ such that for all $x \in \text{supp}(X)$ and for any convex set $\mathcal{D} \in \mathcal{E}$ such that $f_{\xi|X}(\cdot; x)$ is continuous on $\mathcal{D}$, we have $\left|f_{\xi|X}(\xi_1|x) - f_{\xi|X}(\xi_2|x)\right| \leqslant C_\xi |\xi_1 - \xi_2|$.

Assumption 4 describes the properties of the conditional distribution of $\xi|X$. Importantly, note that we focus on a case when $\xi|X = x$ is continuously distributed for all $x \in \text{supp}(X)$. However, our framework straightforwardly allows for the (conditional) distribution of ξ to have point masses or to be discrete. In fact, the asymptotic analysis in the latter case is substantially simpler. Specifically, if ξ is discrete (and takes finitely many values), the agents can be consistently clustered based on the same pseudo-distance $\hat{d}_{ij}^2$. In this case, the second step estimator of β_0 is asymptotically equivalent to the Oracle estimator, which exploits the exact knowledge of the true cluster memberships. Moreover, β_0 can also be estimated by the pooled linear regression, which includes additional interactions of the dummy variables for the estimated cluster membership.

Conditions (ii) and (iii) are additional smoothness requirements. Condition (ii) requires the conditional density to be continuous almost everywhere. The second part of Condition (ii) bounds the probability mass of $\xi|X = x$, for which $f_{\xi|X}(\xi|X)$ is not potentially continuous on a ball $B_\delta(\xi)$. It is a weak requirement, which is immediately satisfied in many cases of interest. Condition (iii) requires the conditional density to be Lipschitz continuous whenever it is continuous.

Example (Illustration of Assumption 4 (ii))**.** Suppose $\xi|X = x$ is supported and continuously distributed on $[0, 1]$ for all $x \in \text{supp}(X)$. Then $f_{\xi|X}(\xi|x)$ is continuous on $B_\delta(\xi)$ for all $\xi \in [\delta, 1-\delta]$. Then (4.2) is satisfied with $\gamma = 2\overline{f}_{\xi|X}$, where $\overline{f}_{\xi|X}$ is as in Assumption 4 (i). ■

Assumption 5.

(i) there exist $\underline{\lambda} > 0$ and $\underline{\delta} > 0$ such that

$$\mathbb{P}\left((X_i, X_j) \in \left\{(x_1, x_2) : \lambda_{min}(\mathcal{C}(x_1, x_2)) > \underline{\lambda}, \int f_{\xi|X}(\xi|x_1) f_{\xi|X}(\xi|x_2) d\xi > \underline{\delta}\right\}\right) > 0,$$

where

$$\mathcal{C}(x_1, x_2) := \mathbb{E}\left[(w(x_1, X) - w(x_2, X))(w(x_1, X) - w(x_2, X))'\right]; \tag{4.3}$$

(ii) for each $\delta > 0$, there exists $C_\delta > 0$ such that

$$\inf_\beta \mathbb{E}\left[\left(g(\xi_i, \xi_k) - g(\xi_j, \xi_k) - (w(X_i, X_k) - w(X_j, X_k))'\beta\right)^2 |X_i, \xi_i, X_j, \xi_j\right] > C_\delta$$

a.s. for (X_i, ξ_i) and (X_j, ξ_j) satisfying $\|\xi_i - \xi_j\| \geqslant \delta$;

(iii) $d_{ij}^2 \equiv d^2(X_i, \xi_i, X_i, \xi_j) = c(X_i, X_j, \xi_i)(\xi_j - \xi_i)^2 + r(X_i, \xi_i, X_j, \xi_j)$, where the remainder satisfies $|r(X_i, \xi_i, X_j, \xi_j)| \leqslant C|\xi_j - \xi_i|^3$ a.s. for some $C > 0$, and $0 < \underline{c} < c(X_i, X_j, \xi_i) < \overline{c}$ a.s.

Assumption 5 is a collection of identification conditions. Specifically, Condition (i) is the identification condition for β_0. It ensures that in a growing sample, it is possible to find a pair of agents i and j such that (i) X_i and X_j are "sufficiently different", so $\lambda_{min}(\mathcal{C}(X_i, X_j)) > \underline{\lambda}$, (ii) and yet ξ_i and ξ_j are increasingly similar. The latter is guaranteed by $\int f_{\xi|X}(\xi|X_i) f_{\xi|X}(\xi|X_j) d\xi > \underline{\delta}$, which implies that the conditional supports of $\xi_i|X_i$ and $\xi_j|X_j$ have a non-trivial overlap. Condition (i) is crucial for establishing consistency of $\hat{\beta}$. Specifically, it ensures that $\hat{A}_n$ converges in probability to a well defined invertible matrix.

Condition (ii) ensures that d_{ij}^2 is bounded away from zero whenever $\|\xi_i - \xi_j\|$ is. Notice that it also guarantees that agents, which are close in terms of d_{ij}^2, must also be similar in terms of ξ. Hence, Condition (ii) justifies using the pseudo-distance d_{ij}^2 for finding agents with similar values of ξ in finite samples. It also can be interpreted as a rank type condition: for fixed agents i and j with $\xi_i \neq \xi_j$, $g(\xi_i, \xi_k) - g(\xi_j, \xi_k)$ can not be expressed as a linear combination of components of $W_{ik} - W_{jk}$.

Condition (iii) is a local counterpart of Condition (ii). It says that as a function of ξ_j, $d^2(X_i, X_j, \xi_i, \xi_j)$ has a local quadratic approximation around ξ_i, and the approximation remainder can be uniformly bounded as $O(|\xi_j - \xi_i|^3)$. Also notice that Condition (iii) rationalizes why we divide $\hat{d}_{ij}^2$ by h_n^2 for computing the kernel weights. Indeed, locally $d_{ij}^2 \propto (\xi_j - \xi_i)^2$, so the bandwidth h_n effectively controls how large $|\xi_j - \xi_i|$ can be for the pair of agents i and j to get a positive weight $K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right)$.

Assumption 6.

(i) $K : \mathbb{R}_+ \to \mathbb{R}$ is supported on $[0, 1]$ and bounded by $\overline{K} < \infty$. K satisfies $\mu_K := \int K(u^2) du > 0$ and $|K(z) - K(z')| \leqslant \overline{K}' |z - z'|$ for all $z, z' \in \mathbb{R}_+$ for some $\overline{K}' > 0$;

(ii) $\max_{i,j \neq i} \left| \hat{d}_{ij}^2 - d_{ij}^2 \right| = O_p(R_n^{-1})$ for some $R_n \to \infty$;

(iii) $h_n \to 0$, $n h_n / \ln n \to \infty$ and $R_n h_n^2 \to \infty$.

Assumption 6 specifies the properties of the kernel K and the bandwidth h_n. Condition (i) imposes a number of fairly standard restrictions on K including Lipschitz continuity. Condition (ii) is a high level condition, which specifies the rate of uniform convergence for the pseudo-distance estimator $\hat{d}_{ij}^2$. In Section 4.2, we formally derive R_n for the estimators (3.7) and (3.10), which are valid in the homoskedastic and in the general heteroskedastic settings, respectively. Finally, Condition (iii) restricts the rates at which the bandwidth is allowed to shrink towards zero. Requirement $n h_n / \ln n \to \infty$ ensures that we have a growing number of potential matches as the sample size increases. Additionally, to get the desired results we need $R_n h_n^2 \to \infty$: the bandwidth can not go to zero faster than $R_n^{-1/2}$. This requirement allows us to bound the effect

of the sampling variability coming from the first step (estimation of $\{d_{ij}^2\}_{i\neq j}$) on the second step (estimation of $\hat{\beta}$).

Assumption 7. There exists a bounded function $G : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ such that for all $\xi_1, \xi_2, \xi \in \mathcal{E}$

$$g(\xi_1, \xi) - g(\xi_2, \xi) = G(\xi_1, \xi)(\xi_1 - \xi_2) + r_g(\xi_1, \xi_2, \xi);$$

and there exists $C > 0$ such that for all $\delta_n \downarrow 0$

$$\limsup_{n\to\infty} \frac{\sup_\xi \sup_{\xi_1:|\xi_1-\xi|>\delta_n} \sup_{\xi_2:|\xi_2-\xi_1|\leqslant\delta_n} |r_g(\xi_1, \xi_2, \xi)|}{\delta_n^2} < C.$$

Assumption 7 is a weak smoothness requirement. It guarantees that as a function of ξ_2, the difference $g(\xi_1, \xi) - g(\xi_2, \xi)$ can be (locally) linearized around ξ_1 provided that ξ_2 is close to ξ_1 relative to the distance between ξ_1 and ξ (guaranteed by the restrictions $|\xi_1 - \xi| > \delta_n$ and $|\xi_2 - \xi_1| \leqslant \delta$). The goal of introducing these restrictions is to allow for a possibly non-differentiable g, e.g., $g(\xi_i, \xi_j) = \kappa |\xi_i - \xi_j|$. We provide an illustration of Assumption 7 in Appendix.

The following lemma establishes asymptotic properties of $\hat{A}_n$, $\hat{B}_n$, and $\hat{C}_n$.

Lemma 1. *Suppose that Assumptions 1, 3-7 hold. Then, we have:*

(i) $\hat{A}_n \xrightarrow{p} A$, *where*

$$A := \mathbb{E}\left[\lambda(X_i, X_j)\mathcal{C}(X_i, X_j)\right],$$
$$\lambda(X_i, X_j) := \int \frac{\mu_K}{\sqrt{c(X_i, X_j, \xi)}} f_{\xi|X}(\xi|X_i) f_{\xi|X}(\xi|X_j) d\xi,$$

where functions $\mathcal{C}(X_i, X_j)$ *and* $c(X_i, X_j, \xi)$ *are defined in Assumptions 6* (i) *and* (iii). *Moreover,* $\lambda_{min}(A) > C > 0$;

(ii)

$$\hat{B}_n = O_p\left(h_n^2 + \frac{R_n^{-1}}{h_n} + n^{-1}\right);$$

(iii)

$$\hat{C}_n = O_p\left(\frac{R_n^{-1}}{h_n^2}\left(\frac{\ln n}{n}\right)^{1/2} + n^{-1}\right).$$

Part (i) establishes consistency of $\hat{A}_n$ and specifies its probability limit A. Importantly, thanks to Assumption 5 (i), A is invertible, which is key for consistency of $\hat{\beta}$.

Part (ii) establishes consistency of $\hat{B}_n$ for 0 (by Assumption 6 (iii), $h_n^2 \to 0$ and $\frac{R_n^{-1}}{h_n} \to 0$) and bounds the rate of convergence. To derive the result, first, we study the asymptotic properties of

$$B_n \coloneqq \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{d_{ij}^2}{h_n^2}\right) \frac{1}{n-2} \sum_{k \neq i,j} (W_{ik} - W_{jk})(g(\xi_i, \xi_k) - g(\xi_j, \xi_k)),$$

which is the infeasible analogue of $\hat{B}_n$ based on the true pseudo-distances $\{d_{ij}^2\}_{i \neq j}$ instead of the estimates $\{\hat{d}_{ij}^2\}_{i \neq j}$. In the proof of the lemma, we demonstrate that

$$\mathbb{E}\left[B_n\right] = \mathbb{E}\left[h_n^{-1} K\left(\frac{d_{ij}^2}{h_n^2}\right)(W_{ik} - W_{jk})(g(\xi_i, \xi_k) - g(\xi_j, \xi_k))\right] = O(h_n^2).$$

This term corresponds to the bias due to the imperfect ξ-matching. It turns out that the bias part of B_n dominates the sampling variability part (up to an additional $O_p(n^{-1})$ term), so we have

$$B_n = O_p(h_n^2 + n^{-1}). \tag{4.4}$$

Second, we take into account that the true pseudo-distances $\{d_{ij}^2\}_{i \neq j}$ are not known and have to be pre-estimated by $\{\hat{d}_{ij}^2\}_{i \neq j}$ at the first step. Then, the first step sampling uncertainty propagates to the second step and its effect can be bounded as

$$\hat{B}_n - B_n = O_p\left(\frac{R_n^{-1}}{h_n}\right). \tag{4.5}$$

Combining (4.4) and (4.5) delivers the result for $\hat{B}_n$.

Part (iii) demonstrates that $\hat{C}_n \xrightarrow{p} 0$ and provides a bound on its rate of convergence. Similarly, first, we study the asymptotic properties of C_n, the infeasible analogue of $\hat{C}_n$ given by

$$C_n \coloneqq \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{d_{ij}^2}{h_n^2}\right) \frac{1}{n-2} \sum_{k \neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}).$$

We establish that $C_n = O_p(n^{-1})$, which is the standard magnitude of sampling variability in a regression with $O(n^2)$ observations. However, the first step sampling uncertainty also affects $\hat{C}_n$ and results in an additional term bounded as $\hat{C}_n - C_n = O_p\left(\frac{R_n^{-1}}{h_n^2}\left(\frac{\ln n}{n}\right)^{1/2}\right)$. Combining these results, we bound the rate of convergence for $\hat{C}_n$.

Remark 4.1. Assumption 7 is only used in the proof of Part (ii).

Lemma 1, paired with (4.1), immediately provides the rate of convergence for $\hat{\beta}$.

Theorem 1. *Under Assumptions 1, 3-7,*

$$\hat{\beta} - \beta_0 = O_p\left(h_n^2 + \frac{R_n^{-1}}{h_n} + \frac{R_n^{-1}}{h_n^2}\left(\frac{\ln n}{n}\right)^{1/2} + n^{-1}\right). \tag{4.6}$$

As pointed before, the rate of convergence for $\hat{\beta}$ crucially depends on R_n, the rate of uniform convergence for $\hat{d}_{ij}^2$. Recall that in the homoskedastic case, we have $R_n = \left(\frac{n}{\ln n}\right)^{1/2}$ (at least, when $d_\xi \leqslant 4$). In the general heteroskedastic case, we have (i) the same rate of convergence when ξ is scalar and X is discrete ; (ii) and a slower rate when $d_\xi \geqslant 2$ and/or X is continuously distributed. Hence, since Assumption 6 (iii) ensures $\frac{R_n^{-1}}{h_n^2} = o(1)$, (4.6) effectively simplifies as (3.3).

Extension to higher dimensions

All of the results presented above remain valid for $d_\xi > 1$ under (i) proper re-normalization of $\hat{A}_n$, $\hat{B}_n$, and $\hat{C}_n$ by $h_n^{-d_\xi}$ instead of h_n^{-1}, (ii) and Assumption 6 (iii) requiring $nh_n^{d_\xi}/\ln n \to \infty$ instead of $nh_n/\ln n \to \infty$ (and other conditions analogously restated in terms of multivariate ξ, if needed).

4.2 Rates of uniform convergence for $\hat{d}_{ij}^2$

Theorem 1 suggests that the asymptotic properties of the kernel based estimator $\hat{\beta}$ crucially depend on R_n, the rate of uniform convergence for $\hat{d}_{ij}^2$ defined in (3.1). In this subsection we formally establish R_n for the estimators (3.10) and (3.7). We start with considering a simpler case when the regression errors in (5.1) are homoskedastic. Then, we discuss the general heteroskedastic case.

4.2.1 Homoskedastic model

First, we consider the homoskedastic case, i.e., we assume that the idiosyncratic errors satisfy (2.6). As discussed in Section 3.2, in this case, the suggested estimator is given by $\hat{d}_{ij}^2 = \hat{q}_{ij}^2 - 2\hat{\sigma}^2$ and $d_{ij}^2 = q_{ij}^2 - 2\sigma^2$ (see Eq. (3.7) and (2.10), respectively). Then, using the triangle inequality, we obtain

$$\begin{aligned}\max_{i,j\neq i}\left|\hat{d}_{ij}^2 - d_{ij}^2\right| &= \max_{i,j\neq i}\left|(\hat{q}_{ij}^2 - q_{ij}^2) - (2\hat{\sigma}^2 - 2\sigma^2)\right| \\ &\leqslant \max_{i,j\neq i}\left|\hat{q}_{ij}^2 - q_{ij}^2\right| + \left|2\hat{\sigma}^2 - 2\sigma^2\right|. \end{aligned} \tag{4.7}$$

Hence, the rate of uniform convergence for $\hat{d}_{ij}^2$ can be bounded using the rates of (uniform) convergence for $\hat{q}_{ij}^2$ and $2\hat{\sigma}^2$. The following lemma establishes their asymptotic properties.

Lemma 2. *Suppose that (2.6) holds and $\mathcal{B}$ is compact. Then, under Assumptions 1 and 3, we have:*

(i)

$$\max_{i,j\neq i}\left|\hat{q}_{ij}^2 - q_{ij}^2\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right), \tag{4.8}$$

where $\hat{q}_{ij}^2$ *and* q_{ij}^2 *are given by* (2.8) *and* (2.7)*, respectively;*

(ii)

$$2\hat{\sigma}^2 - 2\sigma^2 = \overline{G}^2 \min_{i\neq j}\|\xi_i - \xi_j\|^2 + O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right), \tag{4.9}$$

where $2\hat{\sigma}^2$ *is given by* (3.6) *and* $\overline{G}$ *is as defined in Assumption 3* (iii)*.*

Part (i) of Lemma 2 establishes uniform consistency of $\hat{q}_{ij}^2$ for q_{ij}^2 and specifies the rate of convergence equal to $\left(\frac{n}{\ln n}\right)^{1/2}$. Notice that the rate does not depend on the dimension of ξ.

The rate of convergence for $\hat{d}_{ij}^2$ also depends on the asymptotic properties of $2\hat{\sigma}^2$. Part (ii) of Lemma 2 suggests that this rate is potentially affected by the asymptotic behavior of $\min_{i\neq j}\|\xi_i - \xi_j\|^2$, the minimal squared distance between the unobserved characteristics.

It is straightforward to show that if the dimension of ξ is less than or equal to 4,

$$\min_{i\neq j}\|\xi_i - \xi_j\|^2 = o_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right), \tag{4.10}$$

and, hence, the corresponding term does not affect the rate of uniform convergence for $\hat{d}_{ij}^2$. Indeed, since $\mathcal{E}$ is bounded (Assumption 3 (ii)), there exists $C > 0$ such that

$$\min_{i\neq j}\|\xi_i - \xi_j\| \leqslant Cn^{-1/d_\xi}$$

with probability one. Consequently, with probability one, we have

$$\min_{i\neq j}\|\xi_i - \xi_j\|^2 \leqslant C^2 n^{-2/d_\xi}. \tag{4.11}$$

This immediately implies that for $d_\xi \leqslant 4$, (4.10) trivially holds and, as a result, (4.9) simplifies as

$$2\hat{\sigma}^2 - 2\sigma^2 = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Combined with (4.7) and (4.8), this results in the following corollary.

Corollary 1. *Suppose that the hypotheses of Lemma 2 are satisfied. Then for $d_\xi \leqslant 4$, we have*

$$\max_{i,j\neq i}\left|\hat{d}_{ij}^2 - d_{ij}^2\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right),$$

where $\hat{d}_{ij}^2$ and d_{ij}^2 are given by (3.7) and (2.5), respectively.

Corollary 1 ensures that under homoskedasticity of the errors, $\hat{d}_{ij}^2$ satisfies (3.1) with $R_n = \left(\frac{n}{\ln n}\right)^{1/2}$ when the dimension of the unobserved characteristics is less than or equal to 4. Consequently, the rate for $\hat{\beta}$ given by (4.6) indeed reduces to (3.3). Moreover, when $h_n \propto R_n^{-1/3} = \left(\frac{\ln n}{n}\right)^{-1/6}$, we have

$$\hat{\beta} - \beta_0 = O_p\left(\left(\frac{\ln n}{n}\right)^{1/3}\right), \tag{4.12}$$

so the rate of convergence for $\hat{\beta}$ is $\left(\frac{n}{\ln n}\right)^{1/3}$.

Remark 4.2. As we have argued above, the term $\overline{G}^2 \min_{i\neq j}\|\xi_i - \xi_j\|^2$ in (4.9) is asymptotically negligible and does not affect R_n when $d_\xi \leqslant 4$. This argument can be straightforwardly extended to higher dimensions of ξ to obtain a *very conservative* bound on R_n. Notice that (4.11), paired with (4.9), immediately implies that for $d_\xi \geqslant 5$, we can conservatively establish

$$2\hat{\sigma}^2 - 2\sigma^2 = O_p\left(n^{-2/d_\xi}\right).$$

This, combined with (4.7) and (4.8), yields the following conservative result for $\hat{d}_{ij}^2$

$$\max_{i,j\neq i}\left|\hat{d}_{ij}^2 - d_{ij}^2\right| = O_p\left(n^{-2/d_\xi}\right).$$

We stress that these bounds are loose. With a more detailed analysis of the asymptotic behavior of $\min_{i\neq j}\|\xi_i - \xi_j\|^2$ (which is out of the scope of this paper), these results can be substantially refined.

4.2.2 Model with general heteroskedasticity

In this section, we establish the rate of uniform convergence for $\hat{d}_{ij}^2$ under general heteroskedasticity of the errors. First, we suppose that X is discrete and derive R_n for the estimator given by (3.10)-(3.9). After that, we discuss how this estimator can be modified to accommodate continuously distributed X. Finally, we also provide a generic result, which allows establishing R_n for $\hat{d}_{ij}^2$ as in (3.10) based on some denoising estimator of $\hat{Y}_{ij}^*$, potentially other than (3.9).

Estimation of d_{ij}^2 when X is discrete

Now we formally derive the rate of uniform convergence for $\hat{d}_{ij}^2$ given by (3.10)-(3.9). This rate crucially depends on the asymptotic properties of the denoising estimator $\hat{Y}_{ij}^*$.

As mentioned before, the estimator (3.10) we suggest using (when X is discrete) is similar to the estimator of Zhang, Levina, and Zhu (2017). Specifically, they consider a network formation model with $Y_{ij} = g(\xi_i, \xi_j) + \varepsilon_{ij}$, where Y_{ij} is a binary variable, which equals 1 if nodes i and j are connected by a link and 0 otherwise. $g(\xi_i, \xi_j)$ stands for the probability of i and j forming a link, and the links are formed independently conditionally on $\{\xi_i\}_{i=1}^n$, so the errors are also (conditionally) independent (as in Assumption 1 (ii)). Note that Zhang, Levina, and Zhu (2017) do not allow for observed covariates. Consequently, unlike $\hat{\mathcal{N}}_i(n_i)$ defined in (3.8), the neighborhoods constructed by Zhang, Levina, and Zhu (2017) are not conditional on $X = X_i$. Other than that, the estimator given by (3.9) is essentially the same as the estimator of Zhang, Levina, and Zhu (2017).

To derive the asymptotic properties of $\hat{Y}_{ij}^*$, we introduce the following assumption.

Assumption 8.

(i) X is discrete and takes finitely many values $\{x_1, \ldots, x_R\}$;

(ii) there exist positive constants κ and $\overline{\delta}$ such that for all $x \in \operatorname{supp}(X)$, for all $\xi' \in \operatorname{supp}(\xi|X = x)$, $\mathbb{P}(\xi \in B_\delta(\xi')|X = x) \geqslant \kappa\delta^{d_\xi}$ for all positive $\delta \leqslant \overline{\delta}$.

As pointed out before, we suppose that X is discrete and takes finitely many values. Condition (ii) is a weak high level assumption, which is easy to verify in many cases of interest. For example, it is immediately satisfied when $\xi|X$ is discrete. If $\xi|X$ is continuous, it is almost equivalent to requiring the conditional density $f_{\xi|X}(\xi|x)$ to be (uniformly) bounded away from zero.

Example (Illustration of Assumption 8 (ii))**.** Suppose that $\xi|X = x$ is supported and continuously distributed on $[0, 1]$, so $d_\xi = 1$. Then, Assumption 8 (ii) holds if, for some $c > 0$, $f_{\xi|X}(\xi|x) > c$ for all $\xi \in [0, 1]$ and $x \in \operatorname{supp}(X)$. Indeed, the length of $B_\delta(\xi') \cap [0, 1]$ is at least $\delta/2$ for all $\delta \in (0, 1]$. Consequently,

$$\mathbb{P}(\xi \in B_\delta(\xi')|X = x) \geqslant c\delta/2.$$

Hence, Assumption 8 (ii) is satisfied with $\kappa = c/2$ and $\overline{\delta} = 1$. ■

Before formally stating the result, we also introduce the following notations. For any matrix $A \in \mathbb{R}^{n\times n}$, let

$$\|A\|_{2,\infty} := \max_i \sqrt{\sum_{j=1}^n A_{ij}^2}.$$

Also let $\hat{Y}^*$ and Y^* denote $n \times n$ matrices with entries given by $\hat{Y}^*_{ij}$ and Y^*_{ij}.

Theorem 2. *Suppose that for all i, $\underline{C}(n \ln n)^{1/2} \leqslant n_i \leqslant \overline{C}(n \ln n)^{1/2}$ for some positive constants $\underline{C}$ and $\overline{C}$. Then, under Assumptions 1, 3, 8, for $\hat{Y}^*_{ij}$ given by (3.9) we have:*

(i)

$$n^{-1} \left\| \hat{Y}^* - Y^* \right\|^2_{2,\infty} = O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right); \tag{4.13}$$

(ii)

$$\max_k \max_i \left| n^{-1} \sum_\ell Y^*_{k\ell} (\hat{Y}^*_{i\ell} - Y^*_{i\ell}) \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right). \tag{4.14}$$

Theorem 2 establishes two important asymptotic properties of $\hat{Y}^*_{ij}$. In fact, both results play key role in bounding R_n, the rate of uniform convergence for $\hat{d}^2_{ij}$.

Part (i) is analogous to the result of Zhang, Levina, and Zhu (2017). Importantly, note that Zhang, Levina, and Zhu (2017) only consider scalar ξ, while Theorem 2 extends this result to the context of this paper allowing for (i) $d_\xi > 1$, (ii) possibly non-binary outcomes and unbounded errors, (iii) observed (discrete) covariates X.

Part (ii) is new. It allows us to substantially improve R_n compared to what (4.13) can guarantee individually (see also Lemma 3 and Remark 4.5 for a detailed comparison of the rates).

Note that Theorem 2 requires n_i, the number of agents included to $\hat{\mathcal{N}}_i(n_i)$, to grow at $(n \ln n)^{1/2}$ rate. As argued in Zhang, Levina, and Zhu (2017), this is the theoretically optimally rate.[17] As for $\underline{C}$ and $\overline{C}$, the authors recommend taking $n_i \approx (n \ln n)^{1/2}$ for every i (based on simulation experiments). Also notice that expectedly, the right-hand sides of (4.13) and (4.14) crucially depend on the dimension of ξ. Similarly to the standard nonparametric regression, it gets substantially harder to find agents with similar values of ξ once its dimension grows.

Building on the result of Theorem 2, the following theorem establishes the rate of uniform convergence for $\hat{d}^2_{ij}$.

Theorem 3. *Suppose that the hypotheses of Theorem 2 are satisfied and $\mathcal{B} = \mathbb{R}^p$. Then,*

$$\max_{i, j \neq i} \left| \hat{d}^2_{ij} - d^2_{ij} \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right),$$

where $\hat{d}^2_{ij}$ and d^2_{ij} are given by (3.10)-(3.9) and (2.5), respectively.

[17] The optimal choice of n_i remains the same for $d_\xi > 1$.

Theorem 3 establishes the rate of uniform convergence for the proposed estimator of d_{ij}^2. Specifically, it ensures that (3.1) holds with $R_n = \left(\frac{n}{\ln n}\right)^{\frac{1}{2d_\xi}}$. Combined with (3.4), this guarantees that under the proper choice of the bandwidth h_n, we have

$$\hat{\beta} - \beta_0 = O_p\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{3d_\xi}}\right).$$

Hence, when X is discrete, β_0 can be estimated at, at least, $\left(\frac{n}{\ln n}\right)^{\frac{1}{3d_\xi}}$ rate. Note that if ξ is scalar, then the guaranteed rate of convergence for $\hat{\beta}$ is the same as in the homoskedastic case (see Eq. (4.12)). In this case, the rate of convergence for $\hat{Y}_{ij}^*$ suggested by Theorem 2 is fast enough not to make R_n slower relative to the rate in the homoskedastic model.

Remark 4.3. Discreteness of X is needed not only for the result of Theorem 2 to hold but also plays an important role in bounding R_n. Specifically, it ensures that in (3.10), one can take $\mathcal{B} = \mathbb{R}^p$. This, in turn, combined with the result of Theorem 2, allows to achieve substantially faster rate of uniform convergence for $\hat{d}_{ij}^2$, compared to the generic result of Lemma 3, which we will also discuss later.[18]

Estimation of d_{ij}^2 when X is continuously distributed

Now we discuss how $\hat{d}_{ij}^2$ given by (3.10)-(3.9) can be modified when X is continuously distributed. Since, in this case, the probability of finding two agents with exactly the same X is zero, we have to modify how $\hat{\mathcal{N}}_i(n_i)$ is constructed (see Eq. (3.8)).

Let δ_n be a tuning parameter controlling how far from i potential neighbors can be in terms of $\|X - X_i\|$. Specifically, consider

$$\hat{\mathcal{N}}_i(n_i; \delta_n) \coloneqq \{i' : \|X_{i'} - X_i\| \leqslant \delta_n, \mathrm{Rank}(\hat{d}_\infty^2(i, i') | \, \|X - X_i\| \leqslant \delta_n) \leqslant n_i\}. \tag{4.15}$$

Note that for simplicity, we suppose that δ_n is the same for all agents, but, in principle, we can allow for agent specific $\delta_n(i)$. Suppose that X is supported on a compact set and its density bounded away from zero (possibly, after trimming). Then, it is clearly possible to choose δ_n converging to zero slowly enough such that the number of potential matches for every agent grows, i.e., $\min_i |\{i' : \|X_{i'} - X_i\| \leqslant \delta_n\}| \to \infty$ (with probability approaching one). Consequently, as the sample size increases, for each agent, we have a growing number of potential matches with increasingly similar X. Then, among them, using the same pseudo-distance $\hat{d}_\infty^2$, we can find an

[18] Moreover, taking $\mathcal{B} = \mathbb{R}^p$ substantially simplifies numerical routine, since $\hat{d}_{ij}^2$ can be analytically computed.

increasing number of agents with increasingly similar ξ. As before, we estimate Y^*_{ij} by

$$\hat{Y}^*_{ij} = \frac{1}{n_i} \sum_{i' \in \hat{\mathcal{N}}_i(n_i;\delta_n)} Y_{i'j}.$$

Properly adapting the argument of Theorem 2, it is possible to establish analogous results for this estimator. We leave this question for future research.

Remark 4.4. Another possibility to allow for continuous X is to treat it as unobserved, similarly to ξ. In this case, (X, ξ) becomes the effective latent variable and $\hat{\mathcal{N}}_i(n_i)$ is constructed without conditioning on $X = X_i$. If it also satisfies the requirements of Theorem 2 (again, possibly after X is trimmed), then the same results hold with $d_\xi + d_X$ taking the place of d_ξ, where d_X denotes the dimension of X. This is a straightforward way, to construct an estimator of Y^*_{ij} and establish its rate when X is continuous. However, since X is observed and, as discussed before, can (and should) be explicitly taken into account, such an estimator is likely to be suboptimal and, hence, is not recommended in practice.

Finally, we stress that when X (or some of its components) is continuous, the result of Theorem 3 is no longer necessarily valid. Below, we develop a generic result, which allows us to derive R_n in a general context without exploiting particular structure of the preliminary denoising estimator $\hat{Y}^*_{ij}$ or necessarily requiring X to be discrete.

Estimation of d^2_{ij} using general matrix denoising techniques

As pointed out in Section 2.4.1, the problem of construction of $\hat{Y}^*$ is highly related to the general problem of matrix estimation in the latent space model. Hence, while the considered estimator (3.9) based on Zhang, Levina, and Zhu (2017) is one of the possible solutions, a number of alternative ways of constructing $\hat{Y}^*$ are already available in the literature, both in the general (see, for example, Chatterjee (2015); Li, Shah, Song, and Yu (2019) and references therein) and graphon estimation contexts. Below we provide a generic result, which allows us to establish R_n for $\hat{d}^2_{ij}$ and, consequently, also the rate of convergence for $\hat{\beta}$ if Y^* is estimated by one of the other general techniques.

Specifically, suppose that $\hat{Y}^*$ is some estimator of Y^*, which satisfies

$$n^{-1} \left\| \hat{Y}^* - Y^* \right\|^2_{2,\infty} = O_p(\mathcal{R}_n^{-1}) \tag{4.16}$$

for some $\mathcal{R}_n \to \infty$. It turns out that the knowledge of the rate of convergence in the $(2, \infty)$ norm for $\hat{Y}^*$ allows us to bound the rate of uniform convergence of $\hat{d}^2_{ij}$.

Lemma 3. *Suppose that $\hat{Y}^*$ satisfies (4.16) for some $\mathcal{R}_n \to \infty$. Also suppose that $\mathcal{B}$ is compact. Then, under Assumptions 1, 3,*

$$\max_{i,j \neq i} \left| \hat{d}_{ij}^2 - d_{ij}^2 \right| = O_p \left(\left(\frac{\ln n}{n} \right)^{1/2} + \mathcal{R}_n^{-1/2} \right),$$

where $\hat{d}_{ij}^2$ and d_{ij}^2 are given by (3.10) and (2.5), respectively.

Lemma 3 guarantees that $\hat{d}_{ij}^2$ based on some general estimator $\hat{Y}^*$ satisfies (3.1) with

$$R_n = \left(\left(\frac{\ln n}{n} \right)^{1/2} + \mathcal{R}_n^{-1/2} \right)^{-1}.$$

provided that (4.16) holds. This, in turn, combined with (3.3) allows to establish the rate of convergence of $\hat{\beta}$. Note that unlike Theorem 3, Lemma 3 applies regardless whether X (or some of its components) is discrete or continuously distributed.

One of the important implications of Lemma 3 is that it establishes consistency of $\hat{\beta}$ and, hence, formally proves identification of β_0 in a general setting: the only necessary prerequisite is availability of some estimator $\hat{Y}^*$ consistent in the $(2, \infty)$ norm. For example, as argued in Remark 4.4, if X is continuously distributed, we still can construct $\hat{Y}^*$ such that (4.16) is satisfied with $\mathcal{R}_n = \left(\frac{n}{\ln n} \right)^{\frac{1}{2(d_\xi + d_X)}}$. Combined with the result of Lemma 3 and (3.3), this guarantees that β_0 can be consistently estimated when X is continuously distributed.

Remark 4.5. Lemma 3 can also be applied to $\hat{d}_{ij}^2$ given by (3.10)-(3.9) designed for discrete X. Specifically, it establishes that (3.1) holds with $R_n = \mathcal{R}_n^{1/2} = \left(\frac{n}{\ln n} \right)^{\frac{1}{4d_\xi}}$. This is a substantially slower rate compared to the result of Theorem 3. The difference is due to the fact that while Lemma 3 applies in a general setting, Theorem 3 additionally exploits the following features. First, it capitalizes on the result of Theorem 2 (ii), which is specifically due to the form of $\hat{Y}_{ij}^*$ given by (3.9). Second, since X is discrete, Theorem 3 allows us to use $\mathcal{B} = \mathbb{R}^p$, which also additionally facilitates the analysis.

4.3 Uniformly consistent estimation of Y_{ij}^*

One of the contributions of this paper is establishing identification of the error free outcome Y_{ij}^*. In Section 2.4.1, we heuristically argued that Y_{ij}^* is identified. In this section, we construct a uniformly (across all pairs of agents) consistent estimator of Y_{ij}^* and, hence, formally prove its identification.

As before, first, we suppose that X is discrete and takes finitely many values. The estimator we propose is an analogue of $\hat{Y}_{ij}^*$ given by (3.9), which we used before to construct $\hat{d}_{ij}^2$. It utilizes

exactly the same neighborhoods as in (3.8), but, unlike $\hat{Y}^*_{ij}$, averages over all unique outcomes $Y_{i'j'}$ with $i' \in \hat{\mathcal{N}}_i(n_i)$ and $j' \in \hat{\mathcal{N}}_j(n_j)$ (recall that in the indirected model, $Y_{ij} = Y_{ji}$, and Y_{ij} is not observed for $i = j$). For example if $\hat{\mathcal{N}}_i(n_i)$ and $\hat{\mathcal{N}}_j(n_j)$ have no elements in common, then the proposed estimator takes a simple form as in (3.11). More generally, for any i and j, let

$$\hat{\mathcal{M}}_{ij} := \{(i', j') : i' < j', (i' \in \hat{\mathcal{N}}_i(n_i), j' \in \hat{\mathcal{N}}_j(n_j)) \text{ or } (i' \in \hat{\mathcal{N}}_j(n_j), j' \in \hat{\mathcal{N}}_i(n_i))\}.$$

Essentially, $\hat{\mathcal{M}}_{ij}$ is a collection of unique unordered pairs of indexes from the Cartesian product of $\hat{\mathcal{N}}_i(n_i)$ and $\hat{\mathcal{N}}_j(n_j)$ (we suppress its dependence on n_i and n_j for brevity). Then, the estimator of Y^*_{ij} is given by

$$\tilde{Y}^*_{ij} = \frac{1}{m_{ij}} \sum_{(i',j') \in \hat{\mathcal{M}}_{ij}} Y_{i'j'}, \tag{4.17}$$

where m_{ij} denotes the number of elements in $\hat{\mathcal{M}}_{ij}$.

Theorem 4. *Suppose that the hypotheses of Theorem 2 are satisfied. Then, under Assumption 2,*

$$\max_{i,j} \left| \tilde{Y}^*_{ij} - Y^*_{ij} \right| = o_p(1),$$

where $\tilde{Y}^*_{ij}$ is given by (4.17).

Theorem 4 establishes uniform consistency of $\tilde{Y}^*_{ij}$ and, consequently, formally proves that Y^*_{ij} is identified. We also stress that the previously employed estimator $\hat{Y}^*_{ij}$ is not necessarily uniformly consistent since it averages over only a "few" (specifically, n_i) outcomes $Y_{i'j}$. At the same time, $\tilde{Y}^*_{ij}$ averages over $m_{ij} = O(n_i n_j)$ outcomes, which allows us to establish the desired result.

Remark 4.6. The purpose of Theorem 4 is to demonstrate identification of Y^*_{ij}. Although, under additional smoothness requirements, it is also possible to establish the rate of convergence for $\max_{i,j} \left| \tilde{Y}^*_{ij} - Y^*_{ij} \right|$, we do not pursue such a derivation in this paper.

Remark 4.7. Analogous uniform consistency results can also be obtained when X (or some of its components) is continuously distributed if $\tilde{Y}^*_{ij}$ is properly adjusted. As previously discussed, the possible adjustments include (i) constructing neighborhoods as in (4.15), (ii) or treating X as unobserved (see Remark 4.4). In the latter case, an argument similar to the one provided in Remark 4.4 can be invoked to establish uniform consistency of $\tilde{Y}^*_{ij}$ and, hence, prove identification of Y^*_{ij} when X is continuously distributed.

Remark 4.8. To the best of our knowledge, identifiability of the error free outcomes Y^*_{ij} is a new result to the econometrics literature on identification of network and, more generally, two

way models. Moreover, Theorem 4 also contributes to the statistics literature on graphon and, more generally, the latent space model estimation. Specifically, most of the previous work focuses on establishing the mean squared error (MSE) rate for $\hat{Y}^*$, which is essentially equivalent (up to normalization) to the rate of convergence in the Frobenius norm (see, for example, Chatterjee (2015); Gao, Lu, and Zhou (2015); Klopp, Tsybakov, and Verzelen (2017); Zhang, Levina, and Zhu (2017); Li, Shah, Song, and Yu (2019)).[19] At the same time, studying consistency of $\hat{Y}^*$ in the maximum norm has received little attention. Theorem 4 adds to the literature by establishing $\left\|\tilde{Y}^* - Y^*\right\|_\infty = o_p(1)$, i.e., demonstrating consistency of $\tilde{Y}^*$ in the maximum norm.

Another implication of Theorem 4 is that the pair-specific fixed effects $g(\xi_i, \xi_j)$ can also be consistently estimated and, hence, are identified for all pair of agents i and j. Specifically, consider

$$\hat{g}_{ij} = \tilde{Y}^*_{ij} - W'_{ij}\hat{\beta}, \tag{4.18}$$

where $\tilde{Y}^*_{ij}$ is given by (4.17). Since we have already demonstrated consistency of $\hat{\beta}$ and uniform consistency of $\tilde{Y}^*_{ij}$, $\hat{g}_{ij}$ is also uniformly consistent for $g_{ij} := g(\xi_i, \xi_j)$.

Corollary 2. *Suppose that the hypotheses of Theorem 4 are satisfied. Also suppose that* $\hat{\beta} - \beta_0 = o_p(1)$. *Then,*

$$\max_{i,j} |\hat{g}_{ij} - g_{ij}| = o_p(1),$$

where $\hat{g}_{ij}$ *is given by* (4.18).

Establishing nonparametric identification of the pair specific fixed effect g_{ij} is another contribution of the paper. This result is also of high empirical importance since in certain applications, its not β_0 but the fixed effects are the main object of interest.

5 Extensions

5.1 Identification of the partially additively separable model

In this section, we extend the identification arguments of Section 2 to cover a wide range of network models (both semiparametric and nonparametric) with nonparametric unobserved heterogeneity beyond the previously considered model (2.1).

[19] In our context, the mean squared error of $\hat{Y}^*$ is equal to $\frac{1}{n^2}\sum_{i,j}(\hat{Y}^*_{ij} - Y^*_{ij})^2 = \frac{1}{n^2}\left\|\hat{Y}^* - Y^*\right\|^2_{\mathrm{F}}$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

We start with the most general analogue of (2.1)

$$Y_{ij} = f(X_i, \xi_i, X_j, \xi_j) + \varepsilon_{ij}, \quad \mathbb{E}\left[\varepsilon_{ij} | X_i, \xi_i, X_j, \xi_j\right] = 0 \tag{5.1}$$

where $f(X_i, \xi_i, X_j, \xi_j)$ denotes the conditional mean $\mathbb{E}\left[Y_{ij} | X_i, \xi_i, X_j, \xi_j\right]$, and the sampling process still satisfies Assumptions 1 (i) and (ii). For example, if Y_{ij} is binary, (5.1) represents a network formation model, where $f(X_i, \xi_i, X_j, \xi_j)$ equals to the (conditional) probability that agents i and j form a link.

Since ξ is not observed, f (or some features of it) can not be meaningfully identified unless additional structure is imposed on it. However, as discussed in Section 2.4.1, (5.1) fits the latent space model framework (2.11). Consequently, by applying essentially the same argument as in Section 2.4.1, we conclude that the error free outcomes $Y^*_{ij} = f(X_i, \xi_i, X_j, \xi_j)$ are identified for all pairs of agents i and j.

Remark 5.1. Again, to formally prove identifiability of Y^*_{ij} for all pairs of agents i and j, it is sufficient to construct a uniformly valid estimator $\tilde{Y}^*_{ij}$ as in Section 4.3. Notice that if X is treated as unobserved, Theorem 4 similarly applies to the general model (5.1) and, hence, can be used to establish the desired result in this setting.

Notice that identification of Y^*_{ij}, *the value of* $f(X_i, \xi_i, X_j, \xi_j)$, for any pair of agents i and j is fundamentally different from identification of *function* f. Importantly, Y^*_{ij} is not a causal object and can not be directly employed in counterfactual analysis. As pointed out before, identification of counterfactually relevant features of f is not possible unless additional structure is imposed on it. For example, for the previously considered model (2.1) with $f(X_i, \xi_i, X_j, \xi_j) = w(X_i, X_j)'\beta_0 + g(\xi_i, \xi_j)$, we established identification of β_0 and the pair specific fixed effects $g_{ij} \coloneqq g(\xi_i, \xi_j)$.

In this section, we extend the semiparametric model (2.1) to the following partially additively separable model specifying

$$f(X_i, \xi_i, X_j, \xi_j) = F(h(X_i, X_j) + g(\xi_i, \xi_j)), \tag{5.2}$$

where F is a known invertible linking function and both $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ are symmetric unknown functions. Importantly, unlike in (2.1), (5.2) does not require h to have any specific parametric form.

Notice that the presence of the linking function F ensures that (5.2) is flexible enough to cover a wide range of the previously studied network formation models with unobserved heterogeneity. Specifically, a network formation model

$$Y_{ij} = \mathbb{1}\{h(X_i, X_j) + g(\xi_i, \xi_j) - U_{ij} > 0\},$$

with U_{ij} independent and identically distributed with a known invertible CDF F, can be represented as (5.2).

As before, we are interested in identification of the function h, which captures the effect of observable characteristics X_i and X_j, and g_{ij}, the values of the pair specific fixed effect for all agents i and j. Notice that since the linking function F is potentially nonlinear, identification of the pair specific fixed effects is of an additional importance for identification of the average partial effects (APE) and counterfactual analysis.

5.1.1 Nonparametric identification of h and g_{ij}

In this section, we argue that after properly normalized, h in (5.2) is nonparametrically identified. We also argue that the pair specific fixed effects g_{ij} are identified for all pair of agents i and j.

First, let us consider a simpler version of (5.2) given by

$$f(X_i, \xi_i, X_j, \xi_j) = h(X_i, X_j) + g(\xi_i, \xi_j). \tag{5.3}$$

We establish identification of h and g_{ij} in (5.2) under the following assumption.

Assumption 9. Suppose that (5.3) holds and

(i) $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric, continuous and satisfies $h(x, x) = 0$ for all $x \in \mathcal{X}$;

(ii) For any $x, \tilde{x} \in \text{supp}\,(X)$, there exists $\mathcal{E}_{x,\tilde{x}} \subseteq \mathcal{E}$ such that $\mathbb{P}(\xi \in \mathcal{E}_{x,\tilde{x}} | X = x) > 0$ and $\mathbb{P}(\xi \in \mathcal{E}_{x,\tilde{x}} | X = \tilde{x}) > 0$.

Discussion of Assumption 9 (i).

The requirement $h(x, x) = 0$ is a normalization condition. Indeed, since g and the dimension of ξ are not specified, it is without loss of generality to let $g(\xi_i, \xi_j) = \alpha_i + \alpha_j + \psi(\theta_i, \theta_j)$, where ψ is some symmetric function and $\xi = (\alpha, \theta)$. Suppose that we start with

$$f(X_i, \xi_i, X_j, \xi_j) = h(X_i, X_j) + \alpha_i + \alpha_j + \psi(\theta_i, \theta_j), \tag{5.4}$$

where h is symmetric and continuous. Then, consider a model with

$$\tilde{h}(X_i, X_j) = h(X_i, X_j) - (h(X_i, X_i) + h(X_j, X_j))/2,$$

$\tilde{\xi}_i = (\tilde{\alpha}_i, \theta_i)$ with $\tilde{\alpha}_i = \alpha_i + w(X_i, X_i)/2$, and

$$\tilde{f}(X_i, \tilde{\xi}_i, X_j, \tilde{\xi}_j) = \tilde{h}(X_i, X_j) + \tilde{\alpha}_i + \tilde{\alpha}_j + \psi(\theta_i, \theta_j). \tag{5.5}$$

First, note that $\tilde{h}$ is symmetric, continuous and satisfies $\tilde{h}(x,x) = 0$ for all $x \in \mathcal{X}$. Second, $\tilde{f}(X_i, \tilde{\xi}_i, X_j, \tilde{\xi}_j) = f(X_i, \xi_i, X_j, \xi_j)$ for any pair of agents i and j, so the normalized model (5.5) is equivalent to the original model (5.4).

Remark 5.2. The normalization introduced by Assumption 9 (i) is not the only possible. However, the requirement $h(x,x) = 0$ for all $x \in \mathcal{X}$ is natural for network models, especially when $h(X_i, X_j)$ is interpreted as the homophily index associated with the observable characteristics. For example, similar normalization requirements are also imposed in Toth (2017) and Gao (2019).

The proposed identification strategy is an extension of the arguments provided in Section 2. Again, suppose that we can find two agents i and j such that $X_i = x$, $X_j = \tilde{x}$, and $\xi_i = \xi_j$ (note that Assumption 9 (ii) guarantees that such agents exist). Also take any third agent k with $X_k = \tilde{x}$. Then $h(x, \tilde{x})$ is identified by

$$Y^*_{ik} - Y^*_{jk} = h(x,\tilde{x}) + g(\xi_i, \xi_k) - (h(\tilde{x},\tilde{x}) + g(\xi_j, \xi_k)) = h(x,\tilde{x}) \tag{5.6}$$

since $h(\tilde{x},\tilde{x}) = 0$ by Assumption 9. Notice that since Y^*_{ik} and Y^*_{jk} are identified, we treat them as effectively observed.

Consequently, the problem of identification of $h(x,\tilde{x})$ can again be reduced to the problem of identifying two agents i and j with $\xi_i = \xi_j$ (and $X_i = x$, $X_j = \tilde{x}$). We argue that such agents can be identified based on the following (squared) pseudo-distance between agents i and j

$$\begin{aligned} d^2_{ij}(x,\tilde{x}) \coloneqq \min_{\mu} \Big\{ & \mathbb{E}\left[\left(Y^*_{ik} - Y^*_{jk} + \mu\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = x \right] \\ & + \mathbb{E}\left[\left(Y^*_{ik} - Y^*_{jk} - \mu\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = \tilde{x} \right] \Big\}. \end{aligned} \tag{5.7}$$

Again, since Y^*_{ik} and Y^*_{jk} are identified, $d^2_{ij}(x,\tilde{x})$ is also identified for any pair of agents i and j and any $x, \tilde{x} \in \text{supp}\,(X)$.

$d^2_{ij}(x,\tilde{x})$ is a nonparametric analogue of the previously considered pseudo-distance d^2_{ij}. Recall that the specific parametric form $h(x,\tilde{x}) = w(x,\tilde{x})'\beta_0$ in (2.1) allows d^2_{ij} to try to fit $Y^*_{ik} - Y^*_{jk}$ *for all k at the same time* by choosing a finite dimensional β: when $\xi_i = \xi_j$, $Y^*_{ik} - Y^*_{jk}$ is perfectly explained with $\beta = \beta_0$, and $d^2_{ij} = 0$. Although h in (5.3) is no longer a finite dimensional object, a similar approach can be applied in the nonparametric context after conditioning on $X_k = x$ and $X_k = \tilde{x}$ (and fixing $X_i = x$, $X_j = \tilde{x}$). Indeed, in this case (5.3) is effectively parameterized by scalar $h(x,\tilde{x})$. Then, similarly to d^2_{ij}, the nonparametric pseudo-distance $d^2_{ij}(x,\tilde{x})$ tries to fit

$Y_{ik}^* - Y_{jk}^*$ *for all* k *with* $X_k = x$ *and* $X_k = \tilde{x}$ by choosing a scalar μ. If $\xi_i = \xi_j$,

$$d_{ij}^2(x, \tilde{x}) = \min_{\mu} \left((-h(x, \tilde{x}) + \mu)^2 + (h(x, \tilde{x}) - \mu)^2 \right) = 0,$$

where the minimum is achieved at $\mu = h(x, \tilde{x})$.

The following lemma ensures that $d_{ij}^2(x, \tilde{x}) = 0$ also guarantees that $\xi_i = \xi_j$.

Lemma 4. *Suppose that Assumptions 1* (i)*, 2, and 9 hold. Then, for any* $x, \tilde{x} \in supp\,(X)$*,* $x \neq \tilde{x}$*, for any fixed agents* i *and* j *with* $X_i = x$ *and* $X_j = \tilde{x}$*,* $d_{ij}^2(x, \tilde{x}) = 0$ *if and only if* $\xi_i = \xi_j$*.*

Lemma 4 ensures that we can identify agents i and j with $X_i = x$, $X_j = \tilde{x}$, and $\xi_i = \xi_j$ using $d_{ij}^2(x, \tilde{x})$. Importantly, Assumption 9 (ii) requires the conditional supports of $\xi|X = x$ and $\xi|X = \tilde{x}$ to have a non-trivial overlap. Hence, conditional on $X_i = x$, there is strictly positive probability that for agent i we can find a proper match among agents with $X_j = \tilde{x}$. Once such a pair of agents is found, $h(x, \tilde{x})$ is identified as in (5.6). Since this argument applies for any $x, \tilde{x} \in \text{supp}\,(X)$, $x \neq \tilde{x}$, we conclude that h is nonparametrically identified on $\text{supp}\,(X) \times \text{supp}\,(X)$.

Next, notice that since $h(X_i, X_j)$ is identified for any pair of agents i and j, we can also identify the pair specific fixed effect as

$$g_{ij} = Y_{ij}^* - h(X_i, X_j).$$

As a result, we conclude that both h and the pair specific fixed effects are nonparametrically identified in model (5.3).

Finally, we want to argue that the same results apply for the partially additively separable model (5.2). Notice that since F is known and invertible and Y_{ij}^* is identified, we can also identify $\mathcal{Y}_{ij}^* := F^{-1}(Y_{ij}^*)$. The inversion of Y_{ij}^* brings us back to the additive separable model (5.3) with effectively observed outcomes $\mathcal{Y}_{ij}^*$

$$\mathcal{Y}_{ij}^* = h(X_i, X_j) + g(\xi_i, \xi_j). \tag{5.8}$$

Consequently, applying the same argument to (5.8), we conclude that h and the pair specific fixed effects are nonparametrically identified in the partially additive separable model (5.2).

Remark 5.3. Note that identification of the pair specific fixed effects g_{ij} is fundamentally different from identification of the coupling function g. Since the fixed effects ξ_i are not observed, the coupling function g is not identified unless additional restrictions are imposed on its form. However, identification of g_{ij} for all pairs of agents is sufficient for identification of any counterfactually relevant objects including both the pair-specific and the average partial effects.

Remark 5.4. Note that identification of $\mathcal{Y}^*_{ij} = F^{-1}(Y^*_{ij})$ also implies that the identification and estimation strategies of Sections 2 and 3 apply when h in (5.2) is parameterized as $h(X_i, X_j) = w(X_i, X_j)'\beta_0$ (as in the initially considered model (2.1)). For example, as an estimator of β_0, one can take the same kernel based estimator (3.2) with $\hat{\mathcal{Y}}^*_{ij} := F^{-1}(\hat{Y}^*_{ij})$ replacing Y_{ij}, and the pseudo-distances $\hat{d}^2_{ij}$ can be estimated as in (3.10) with $\hat{\mathcal{Y}}^*_{ij}$ replacing $\hat{Y}^*_{ij}$.

5.2 Incorporating missing outcomes

From the beginning of Section 2, we assumed that $\{Y_{ij}\}_{i\neq j}$ are observed for all pairs of agents i and j in order to simplify the exposition and to facilitate the formal analysis. While this assumption is standard in the network formation context ($Y_{ij} = 1$ if agents i and j are connected, and $Y_{ij} = 0$ otherwise), in many other settings the interaction outcomes are available only for a limited number of the pairs of agents. For example, in the international trade network data of Helpman, Melitz, and Rubinstein (2008), 55% of the country pairs have zero trade flows. Moreover, unlike the international trade network, most of the other economic networks are sparse. For instance, in the employer-employee matched setting, workers typically interact with only a few firms. Hence, it is important to discuss (i) under which conditions we can incorporate missing outcomes into our framework, (ii) and how to properly adjust the proposed estimators in this case.

First, the studied procedures remain valid as long as the selection mechanism is exogenous conditional on the observed and unobserved characteristics of agents $\{(X_i, \xi_i)\}_{i=1}^n$, i.e., the structure of the observed network is (conditionally) independent of the idiosyncratic errors $\{\varepsilon_{ij}\}$.[20] This assumption is standard in the network regression literature (see, for example, Abowd, Kramarz, and Margolis, 1999 and the subsequent works). However, since we allow for an uncommonly flexible form of unobserved heterogeneity, the network exogeneity assumption is much less restrictive in the context of our model. Specifically, since we do not specify the dimensionality of the fixed effects and the nature of their interactions, our framework allows for a substantially more general selection mechanism involving both observed and unobserved characteristics. Hence, rather than treating network endogeneity as a potential threat to validity of the suggested procedure, we consider our approach to be a tool addressing this concern.

Next, we adjust the proposed estimators to allow for potentially missing outcomes. Let D_{ij} be a binary variable such that $D_{ij} = 1$ if Y_{ij} is observed and $D_{ij} = 0$ otherwise (note that $D_{ii} = 0$ by construction). Also, let

$$\mathcal{O}_{ij} := \{k : D_{ik} = 1, D_{jk} = 1\}$$

[20] This assumption is satisfied if the agents form interactions based on $\{(X_i, \xi_i)\}_{i=1}^n$ but not on the idiosyncratic errors. For example, in a structural model, it can be rationalized if the idiosyncratic errors are realized after the network is formed.

be a collection of agents k such that both Y_{ik} and Y_{jk} are observed. For simplicity, we consider the homoskedastic setting first. We adjust (2.7) and (3.2) as

$$\hat{q}_{ij}^2 = \min_{\beta \in \mathcal{B}} \frac{1}{|\mathcal{O}_{ij}|} \sum_{k \in \mathcal{O}_{ij}} (Y_{ik} - Y_{jk} - (W_{ik} - W_{jk})'\beta)^2, \tag{5.9}$$

$$\hat{\beta} = \left(\sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \sum_{k \in \mathcal{O}_{ij}} (W_{ik} - W_{jk})(W_{ik} - W_{jk})' \right)^{-1} \left(\sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) \sum_{k \in \mathcal{O}_{ij}} (W_{ik} - W_{jk})(Y_{ik} - Y_{jk}) \right), \tag{5.10}$$

where $\hat{d}_{ij}^2$ is still computed as in (3.7). In this case, the same identification and consistency arguments remain valid as long as $\min_{i,j} |\mathcal{O}_{ij}| \to \infty$ as $n \to \infty$ with probability one. This condition is satisfied for both dense and certain sparse networks, as long as the degree of sparsity is not too extreme.[21]

In the general heteroskedastic case, the first estimation step is to construct $\hat{Y}_{ij}^*$. In fact, recent developments in the matrix completion literature allow consistent estimation of Y^* even when the observed matrix Y is sparse (see, for example, Chatterjee, 2015; Klopp, Tsybakov, and Verzelen, 2017; Li, Shah, Song, and Yu, 2019).[22] Once $\hat{Y}_{ij}^*$ are constructed (for example, using one of the already developed matrix completion techniques), the rest of the estimation procedure remains the same.

5.3 Extension to directed networks and two-way models

Finally, we note that the proposed estimation procedure can be generalized to cover directed networks and, more generally, two-way models. Specifically, consider a general interaction model

$$Y_{ik} = W_{ik}'\beta_0 + g(\xi_i, \eta_k) + \varepsilon_{ik},$$

where $i \in \mathcal{I}$ and $k \in \mathcal{K}$ index senders and receivers, and ξ_i and η_k denote the sender and the receiver fixed effects.

As before, the possible identification and estimation strategies can be based on finding agents with the same/similar values of unobserved fixed effects. However, the considered interaction model consists of two types of agents, senders and receivers (e.g., firms and workers). As a result, the researcher has a flexibility to decide whether she wants to match senders or receivers depending on the context. For example, consider the sender-to-sender approach. For simplicity of exposition,

[21] Developing statistical tools valid in sparse settings is an emerging area of research in the econometric literature (e.g., Jochmans and Weidner, 2019; Verdier, 2018). We do not pursue such analysis in this paper.

[22] At least, when consistency is defined in terms of the MSE.

we also focus on the homoskedastic setting (the general estimator can be adjusted in a similar fashion). In this case, exactly the same estimator as (5.9)-(5.10) provided in Section 5.2 represents the sender-to-sender estimator of β_0.

For the sender-to-sender estimator to be consistent, we need both the number of senders $|\mathcal{I}|$ and receivers $|\mathcal{K}|$ to grow. First, we need $|\mathcal{I}| \to \infty$ in order to ensure that the sample includes pairs of senders with increasingly similar values of ξ. Second, we need $|\mathcal{K}| \to \infty$ to achieve consistency of $\hat{q}^2_{ij}$, which helps us to find senders with the same/similar values of ξ. More precisely, if some interactions are missing, we need $\min_{i,j\in\mathcal{I}} |\mathcal{O}_{ij}| \to \infty$. Notice that $\min_{i,j\in\mathcal{I}} |\mathcal{O}_{ij}| \to \infty$ means that any pair of senders has a growing number of common receivers. At the same time, receivers are allowed to participate only in a few interactions. For example, this suggests that in the employer-employee matched setting, the firm-to-firm estimator might be plausible even though the workers' mobility is typically limited.

6 Simulation Study

In this section, we illustrate the finite sample properties of the proposed estimators. Specifically, we consider the following homoskedastic variation of (2.1)

$$Y_{ij} = (X_i - X_j)^2\beta_0 - (\xi_i - \xi_j)^2 + \varepsilon_{ij}, \tag{6.1}$$

where

$$\begin{pmatrix} X_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

and $\{\varepsilon_{ij}\}_{i<j}$ are independent draws from $N(0,1)$. The true value of the parameter of interest is $\beta_0 = -1$, so model (2.1) feature homophily based on both X and ξ.

We study performance of the following estimators. The first estimator $\hat{\beta}_{\text{FE}}$ is produced by the standard linear regression with additive fixed effects. The second estimator $\hat{\beta}$ is the kernel based estimator (3.2) with $\hat{d}^2_{ij}$ computed as in (3.7) and $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}\{|u| \leqslant 1\}$ (the Epanechnikov kernel).[23] As for the bandwidth, we choose

$$h_n^2 = 0.9 \min\left\{\hat{\sigma}_{\hat{d}^2}, \text{IQR}_{\hat{d}^2}/1.349\right\} \binom{n}{2}^{-1/5}$$

following the standard (kernel density estimation) rule of thumb. Here $\hat{\sigma}_{\hat{d}^2}$ and $\text{IQR}_{\hat{d}^2}$ stand for the standard deviation of the interquartile range of the estimated pseudo-distances $\{\hat{d}^2_{ij}\}_{i<j}$, and

[23] The reported results are robust to the kernel choice.

$\binom{n}{2}$ corresponds to the number ("sample size") of the estimated pseudo-distances. Notice that since the kernel weights in (3.2) are defined as $K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right)$, it is h_n^2 (not just h_n), which corresponds to the "effective" kernel density estimation bandwidth applied to $\{\hat{d}_{ij}^2\}_{i<j}$. Finally, we also compute the 1 nearest neighbor estimator $\hat{\beta}_{\text{NN1}}$ as in (3.5).

We simulate model (6.1) for $n \in \{30, 50, 100\}$ and $\rho \in \{0, 0.3, 0.5, 0.7\}$ (the number of replications is 10,000). The simulated finite sample properties of the considered estimators are reported in Table 1 below. The naive estimator $\hat{\beta}_{\text{FE}}$ is biased whenever the observed and the unobserved characteristics of agents are correlated. The magnitude of this bias rapidly increases as ρ grows. The proposed estimators $\hat{\beta}$ and $\hat{\beta}_{\text{NN1}}$ effectively remove the bias even for networks of a moderate size (for $n = 30$ we have $\binom{30}{2} = 435$ observations in an undirected network with no missing links). Notice that the magnitudes of the bias for $\hat{\beta}$ and $\hat{\beta}_{\text{NN1}}$ are approximately the same but the kernel based estimator $\hat{\beta}$ is consistently less dispersed. This suggests that in the studied setting, the 1 nearest neighbor estimator $\hat{\beta}_{\text{NN1}}$ tends to undersmooth. Finally, notice that the proposed estimators $\hat{\beta}$ and $\hat{\beta}_{\text{NN1}}$ dominate the naive estimator $\hat{\beta}_{\text{FE}}$ not only in terms of the bias but also in terms of the standard deviation/IQR (even when $\rho = 0$). Indeed, when the fixed effects contribution to the variability in Y_{ij} is large (like in model (6.1)), controlling for unobserved heterogeneity (as both $\hat{\beta}$ and $\hat{\beta}_{\text{NN1}}$ attempt to do by differencing it out) may substantially improve precision even at the cost of significantly reducing the effective sample size.

7 Conclusion

In this paper, we study identification and estimation of network models with nonparametric unobserved heterogeneity. Importantly, we do not specify the role of the fixed effects and the nature of their interaction, allowing, for example, for homophily based on unobservables. We establish identification of all components of the model, which, in turn, allow us to identify both the pair specific and the average partial effects. In addition, leveraging and advancing recent developments in the matrix estimation/completion literature, we also demonstrate that the error free outcomes, Y_{ij}^*, are identified in a general network model. This is a powerful result in itself, which also provides a foundation for further identification results in other two-way settings including large panels. To provide a practical estimation approach, we focus on a semiparametric single index model. We construct several estimators of the parameters of interest and derive their rates of convergence. Finally, we illustrate the finite sample properties of the proposed estimators in a Monte-Carlo experiment.

Table 1: Simulation results for model (6.1)

ρ	Bias			Med Bias			Std. dev.			IQR/1.349		
	$\hat{\beta}_{\text{FE}}$	$\hat{\beta}$	$\hat{\beta}_{\text{NN1}}$	$\hat{\beta}_{\text{FE}}$	$\hat{\beta}$	$\hat{\beta}_{\text{NN1}}$	$\hat{\beta}_{\text{FE}}$	$\hat{\beta}$	$\hat{\beta}_{\text{NN1}}$	$\hat{\beta}_{\text{FE}}$	$\hat{\beta}$	$\hat{\beta}_{\text{NN1}}$
$n = 30$												
0.0	0.000	-0.001	0.010	0.010	0.000	0.006	0.063	0.028	0.050	0.044	0.025	0.036
0.3	-0.090	-0.006	0.005	-0.061	-0.005	0.004	0.124	0.033	0.060	0.106	0.029	0.044
0.5	-0.251	-0.018	-0.009	-0.225	-0.016	-0.006	0.173	0.045	0.077	0.165	0.038	0.059
0.7	-0.491	-0.052	-0.058	-0.473	-0.048	-0.048	0.196	0.080	0.109	0.191	0.063	0.092
$n = 50$												
0.0	-0.000	-0.000	0.005	0.006	-0.000	0.003	0.035	0.015	0.029	0.026	0.014	0.021
0.3	-0.090	-0.004	0.004	-0.072	-0.004	0.002	0.090	0.018	0.036	0.085	0.016	0.027
0.5	-0.251	-0.013	-0.005	-0.235	-0.012	-0.004	0.130	0.024	0.046	0.128	0.022	0.036
0.7	-0.491	-0.036	-0.038	-0.481	-0.034	-0.032	0.148	0.042	0.068	0.147	0.037	0.057
$n = 100$												
0.0	-0.000	-0.000	0.003	0.003	-0.000	0.002	0.017	0.007	0.014	0.012	0.007	0.010
0.3	-0.091	-0.003	0.002	-0.082	-0.002	0.002	0.061	0.008	0.018	0.058	0.008	0.014
0.5	-0.251	-0.008	-0.002	-0.243	-0.007	-0.002	0.089	0.011	0.024	0.087	0.010	0.019
0.7	-0.491	-0.021	-0.020	-0.486	-0.021	-0.017	0.102	0.019	0.036	0.100	0.018	0.030

This table reports the simulated bias, the median bias, the standard deviation, and the interquartile range (divided by 1.349) for the additive fixed effects estimator $\hat{\beta}_{\text{FE}}$, the kernel based estimator $\hat{\beta}$, and the 1 nearest neighbor estimator $\hat{\beta}_{\text{NN1}}$. The results are presented for different values n (size of the network) and ρ (correlation between X_i and ξ_i). The number of replications is 10,000.

Appendix

A Proofs

A.1 Proofs of the results of Section 4.1

Notation

With some abuse of notation, we denote $Z_i \coloneqq (X_i, \xi_i)$. When convenient, we use Z_i to suppress X_i and ξ_i as arguments of some function. For example, $d^2(Z_i, Z_j) \equiv d^2(X_i, \xi_i, X_j, \xi_j)$.

A.1.1 Auxiliary lemmas

Lemma A.1. *Suppose that the hypotheses of Lemma 1 are satisfied. Then, there exists $\alpha > 0$ such that:*

(i)

$$|\xi_j - \xi_i| > \alpha h_n \quad \Rightarrow \quad K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) = 0 \qquad \text{and} \qquad K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right) = 0$$

with probability one;

(ii)

$$\sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right)^2 \mathbb{1}\{|\xi_j - \xi_i| > \alpha h_n\} = 0$$

with probability approaching one;

(iii)

$$K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) = K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right) + r_K(X_i, X_j, \xi_i, \xi_j; h_n), \tag{A.1}$$

where $c_{ij} \coloneqq c(X_i, X_j, \xi_i)$*, and for some* $C > 0$*,*

$$|r_K(X_i, X_j, \xi_i, \xi_j; h_n)| \leqslant C h_n \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} \quad \text{a.s.} \tag{A.2}$$

Proof of Lemma A.1. Note that Assumption 5 (iii) guarantees that there exists some $\delta_0 > 0$ and $c^* \in (0, \underline{c})$ such that $d^2(X_i, X_j, \xi_i, \xi_j) \geqslant c^*(\xi_j - \xi_i)^2$ a.s. for $|\xi_j - \xi_i| \leqslant \delta_0$. This implies that there exists $\alpha = 1/\sqrt{c^*}$ such that $d^2(X_i, X_j, \xi_i, \xi_j) > h_n^2$ whenever $\alpha h_n < |\xi_j - \xi_i| \leqslant \delta_0$ ($\alpha h_n < \delta_0$ for large enough n). Note that since $c^* < \underline{c}$, this also immediately ensures that $c_{ij}(\xi_j - \xi_j)^2 > h_n^2$ whenever $|\xi_j - \xi_i| > \alpha h_n$ with probability one. Hence, we also conclude that

$$|\xi_j - \xi_i| > \alpha h_n \quad \Rightarrow \quad K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right) = 0 \tag{A.3}$$

with probability one. At the same time, for large enough n, $d^2(X_i, X_j, \xi_i, \xi_j) > h_n^2$ a.s. for $|\xi_i - \xi_j| \geqslant \delta_0$ (Assumption 5 (ii)). Therefore, since the kernel is supported on $[0, 1]$ (Assumption 6 (i)) for large enough n, $K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) = 0$ whenever $|\xi_j - \xi_i| > \alpha h_n$ a.s. This completes the proof of Part (i).

Clearly, we can choose α such that $d^2(X_i, X_j, \xi_i, \xi_j) > c h_n^2$ for some $c > 1$ whenever $|\xi_j - \xi_i| > \alpha h_n$. Consequently, for all pairs of agents i and j satisfying $|\xi_j - \xi_i| > \alpha h_n$, we have

$$\frac{\hat{d}_{ij}^2}{h_n^2} > c > 1.$$

Since we have

$$\frac{\hat{d}_{ij}^2}{h_n^2} \geqslant \frac{d_{ij}^2}{h_n^2} - \underbrace{\frac{\max_{i \neq j} \left| \hat{d}_{ij}^2 - d_{ij}^2 \right|}{h_n^2}}_{o_p(1)},$$

then, for all pairs of agents i and j satisfying that $|\xi_j - \xi_i| > \alpha h_n$, with probability approaching one

$$\frac{\hat{d}_{ij}^2}{h_n^2} > 1.$$

Since K is supported on $[0, 1]$ (Assumption 6 (i)), this completes the proof of Part (ii).

Finally, using the result of Part (i), we have

$$K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) = K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}.$$

Hence,

$$\begin{aligned} K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) =& K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right) \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} \\ &+ \left(K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) - K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right)\right) \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}. \end{aligned}$$

Finally, using (A.3), we also have

$$K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) = K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right) + r_K(X_i, X_j, \xi_i, \xi_j; h_n),$$

where

$$r_K(X_i, X_j, \xi_i, \xi_j; h_n) \coloneqq \left(K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) - K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right)\right) \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}.$$

Finally, note that Assumptions 5 (iii) and 6 (i) together guarantee that there exist $C > 0$ such that (A.2) holds. Q.E.D.

Lemma A.2. *Suppose that the hypotheses of Lemma 1 are satisfied. Then, for any fixed positive constant α, we have:*

(i) for some constant C_α,

$$h_n^{-1}\mathbb{E}\left[\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}|\xi_i\right] \leqslant C_\alpha \qquad \text{a.s.}$$

and, hence,

$$h_n^{-1}\mathbb{E}\left[\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}\right] \leqslant C_\alpha$$

(ii)

$$\binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} = O_p(1).$$

Proof of Lemma A.2. Let $q_{n,ij} \coloneqq h_n^{-1}\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}$. We want to argue that

$$\binom{n}{2}^{-1} \sum_{i<j} q_{n,ij} = \mathbb{E}\left[q_{n,ij}\right] + o_p(1). \tag{A.4}$$

Note that

$$\mathbb{E}\left[q_{n,ij}|\xi_i\right] = h_n^{-1} \int \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} f_\xi(\xi_j) d\xi_j \leqslant 2\alpha \overline{f}_\xi,$$

where $\overline{f}_\xi = \sup_{\xi \in \text{supp}(\xi)} f_\xi(\xi) < \infty$ (guaranteed by Assumption 4 (i)). Hence, $\mathbb{E}\left[q_{n,ij}\right] \leqslant 2\alpha\overline{f}_\xi$ is bounded. Note that taking $C_\alpha = 2\alpha\overline{f}_\xi$ completes the proof of the first part.

Similarly,

$$\mathbb{E}\left[q_{n,ij}^2\right] \leqslant \frac{2\alpha\overline{f}_\xi}{h_n} = O(h_n^{-1}) = o(n),$$

where the last equality is due to Assumption 6 (iii). Therefore, Lemma A.3 of Ahn and Powell (1993) ensures that (A.4) holds, and, as a result,

$$\binom{n}{2}^{-1} \sum_{i<j} q_{n,ij} = O_p(1).$$

Q.E.D.

A.1.2 Proof of Lemma 1 Part (i)

Proof of Lemma 1 Part (i). First, we argue that $\hat{A}_n - A_n = o_p(1)$, where

$$A_n \coloneqq \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{d_{ij}^2}{h_n^2}\right) \frac{1}{n-2} \sum_{k\neq i,j} (W_{ik} - W_{jk})(W_{ik} - W_{jk})'.$$

Then

$$\hat{A}_n - A_n = \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \left(K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) - K\left(\frac{d_{ij}^2}{h_n^2}\right)\right) \frac{1}{n-2} \sum_{k\neq i,j} (W_{ik} - W_{jk})(W_{ik} - W_{jk})'.$$

Since $\left\| \frac{1}{n-2} \sum_{k\neq i,j} (W_{ik} - W_{jk})(W_{ik} - W_{jk})' \right\|$ is uniformly bounded (W is bounded by Assumption 3 (i)), it suffices to show that

$$\binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \left| K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) - K\left(\frac{d_{ij}^2}{h_n^2}\right)\right| = o_p(1). \tag{A.5}$$

First, using Assumption 6 (i) and the result of Lemma A.1 (ii), we establish that with probability approaching one

$$\binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \left| K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) - K\left(\frac{d_{ij}^2}{h_n^2}\right)\right| \leqslant \overline{K}' \frac{\max_{i\neq j} \left|\hat{d}_{ij}^2 - d_{ij}^2\right|}{h_n^2} \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}. \tag{A.6}$$

Lemma A.2 ensures

$$\binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} = O_p(1).$$

Combining this with (A.6) and Assumption 6 (ii), we obtain

$$\binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \left| K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) - K\left(\frac{d_{ij}^2}{h_n^2}\right)\right| = O_p\left(\frac{R_n^{-1}}{h_n^2}\right). \tag{A.7}$$

Finally, Using Assumption 6 (iii), we conclude that (A.5) holds and therefore $\hat{A}_n - A_n = o_p(1)$.

Let

$$\zeta_n(Z_i, Z_j, Z_k) \coloneqq h_n^{-1} K\left(\frac{d^2(Z_i, Z_j)}{h_n^2}\right) (w(X_j, X_k) - w(X_i, X_k))(w(X_j, X_k) - w(X_i, X_k))'.$$

Then

$$A_n = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} \zeta_n(Z_i, Z_j, Z_k).$$

Note that ζ_n is symmetric in its first two arguments and, consequently, it can be symmetrized with

$$p_n(Z_i, Z_j, Z_k) \coloneqq \frac{1}{3} \left(\zeta_n(Z_i, Z_j, Z_k) + \zeta_n(Z_k, Z_i, Z_j) + \zeta_n(Z_j, Z_k, Z_i) \right).$$

Then A_n is a third order U-statistic with kernel p_n. First, we want to show that $\mathbb{E}\left[\|p_n(Z_i, Z_j, Z_k)\|^2\right] = o(n)$. Indeed, again since W is bounded, there exists $C > 0$ such that

$$\mathbb{E}\left[\|\zeta_n(Z_i, Z_j, Z_k)\|^2\right] \leqslant C\mathbb{E}\left[h_n^{-2} K\left(\frac{d^2(Z_i, Z_j)}{h_n^2}\right)^2\right].$$

Using Assumption 6 (i) and the result of Lemma A.1 (i), we have

$$K\left(\frac{d^2(Z_i, Z_j)}{h_n^2}\right)^2 \leqslant \overline{K}^2 \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}.$$

Hence, for some $C > 0$,

$$\mathbb{E}\left[\|\zeta_n(Z_i, Z_j, Z_k)\|^2\right] \leqslant C h_n^{-2} \mathbb{E}\left[\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}\right].$$

By Lemma A.2 (i), for some $C > 0$, we have

$$h_n^{-1} \mathbb{E}\left[\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}\right] \leqslant C.$$

Hence, we conclude that for some $C > 0$,

$$\mathbb{E}\left[\|\zeta_n(Z_i, Z_j, Z_k)\|^2\right] \leqslant \frac{C}{h_n},$$

and, consequently, the same bound applies to p_n, i.e., we have $\mathbb{E}\left[\|p_n(Z_i, Z_j, Z_k)\|^2\right] = O(h_n^{-1}) = o(n)$, where the last equality follows from Assumption 6 (iii). As a result, Lemma A.3 of Ahn and Powell (1993) applies and establishes that

$$A_n = \mathbb{E}\left[p_n(Z_i, Z_j, Z_k)\right] + o_p(1) = \mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)\right] + o_p(1).$$

Since we have previously established that $\hat{A}_n = A_n + o_p(1)$, we also get

$$\hat{A}_n = \mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)\right] + o_p(1).$$

The rest of the proof deals with computing $\mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)\right]$. First, note

$$\mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)|X_i, X_j\right] = \mathbb{E}\left[h_n^{-1} K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right)|X_i, X_j\right] \mathcal{C}(X_i, X_j).$$

where $\mathcal{C}(X_i, X_j)$ is as defined in (4.3)

Using the result of Lemma A.1 (iii), we obtain

$$\begin{aligned} I(X_i, X_j, \xi_i; h_n) &:= \mathbb{E}\left[h_n^{-1} K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right)|X_i, X_j, \xi_i\right] \\ &= \int h_n^{-1} K\left(\frac{d^2(X_i, X_j, \xi_i, \xi_j)}{h_n^2}\right) f_{\xi|X}(\xi_j; X_j) d\xi_j \\ &= \underbrace{\int h_n^{-1} K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right) f_{\xi|X}(\xi_j; X_j) d\xi_j}_{I_1(X_i, X_j, \xi_i; h_n)} + \underbrace{\int h_n^{-1} r_K(X_i, X_j, \xi_i, \xi_j; h_n) f_{\xi|X}(\xi_j; X_j) d\xi_j}_{I_2(X_i, X_j, \xi_i; h_n)}. \end{aligned}$$

Note that using Assumption 4 (i), we have

$$\begin{aligned} |I_2(X_i, X_j, \xi_i; h_n)| &\leqslant C \int \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} f_{\xi|X}(\xi_j; X_j) d\xi_j \\ &\leqslant 2C h_n \alpha \overline{f}_{\xi|X} \quad \text{a.s.} \end{aligned}$$

Then, we work with the first term

$$\begin{aligned} I_1(X_i, X_j, \xi_i; h_n) &:= \int h_n^{-1} K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right) f_{\xi|X}(\xi_j; X_j) d\xi_j \\ &= \frac{1}{\sqrt{c_{ij}}} \int K(u^2) f_{\xi|X}(\xi_i + h_n u/\sqrt{c_{ij}}; X_j) du, \end{aligned}$$

where the second equality follows from the change of the variable $\xi_j = \xi_i + h_n u/\sqrt{c_{ij}}$. Note that for all values of ξ_i such that $f_{\xi|X}(\xi_i|X_j)$ is continuous at ξ_i we have

$$I_1(X_i, X_j, \xi_i; h_n) \to \frac{\mu_K f_{\xi|X}(\xi_i|X_j)}{\sqrt{c(X_i, X_j, \xi_i)}}, \quad h_n \to 0,$$

and, consequently,

$$I(X_i, X_j, \xi_i; h_n) \to \frac{\mu_K f_{\xi|X}(\xi_i|X_j)}{\sqrt{c(X_i, X_j, \xi_i)}}, \quad h_n \to 0,$$

where $c_{ij} = c(X_i, X_j, \xi_i)$ and μ_K is as defined in Assumption 6 (i). By Assumption 6 (ii), this applies for almost all ξ_i. Moreover, $I(X_i, X_j, \xi_i; h_n)$ is uniformly bounded: there exists $C > 0$, such that

$$\begin{aligned} I(X_i, X_j, \xi_i; h_n) &\leqslant \overline{K} h_n^{-1} \mathbb{E}\left[\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}\right] \\ &\leqslant C, \end{aligned}$$

where the second inequality is due to Lemma A.2 (i).

Hence, the dominated convergence theorem applies and ensures that for all X_i and X_j

$$I(X_i, X_j; h_n) \coloneqq \mathbb{E}\left[I(X_i, X_j, \xi_i; h_n)|X_i, X_j\right] \to \int \frac{\mu_K}{\sqrt{c(X_i, X_j, \xi_i)}} f_{\xi|X}(\xi_i|X_i) f_{\xi|X}(\xi_i|X_j) d\xi_i, \quad h_n \to 0.$$

Therefore, for all X_i and X_j,

$$\mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)|X_i, X_j\right] \to \mathbb{E}\left[\lambda(X_i, X_j)\mathcal{C}(X_i, X_j)|X_i, X_j\right],$$

where

$$\lambda(X_i, X_j) \coloneqq \int \frac{\mu_K}{\sqrt{c(X_i, X_j, \xi)}} f_{\xi|X}(\xi|X_i) f_{\xi_i|X}(\xi|X_j) d\xi.$$

Again, since $\mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)\right] = I(X_i, X_j; h_n)\mathcal{C}(X_i, X_j)$ is uniformly bounded, the dominate convergence theorem establishes

$$\mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)\right] \to A \coloneqq \mathbb{E}\left[\lambda(X_i, X_j)\mathcal{C}(X_i, X_j)\right].$$

Hence, we conclude

$$\hat{A}_n = \mathbb{E}\left[\zeta_n(Z_i, Z_j, Z_k)\right] + o_p(1) = A + o_p(1),$$

which completes the proof of the first statement. Also note that Assumption 6 (i) also guarantees that for some $C > 0$, $\lambda_{min}(A) > C$. Q.E.D.

A.1.3 Proof of Lemma 1 Part (ii)

First, we introduce the following notations

$$q_n(Z_i, Z_j, Z_k) \coloneqq h_n^{-1} K\left(\frac{d^2(Z_i, Z_j)}{h_n^2}\right)(W(X_i, X_k) - W(X_j, X_k))(g(\xi_i, \xi_k) - g(\xi_j, \xi_k)), \tag{A.8}$$

and

$$p_n(Z_i, Z_j, Z_k) \coloneqq \frac{1}{3}(q_n(Z_i, Z_j, Z_k) + q_n(Z_k, Z_i, Z_j) + q_n(Z_j, Z_k, Z_i)). \tag{A.9}$$

Before proving Lemma 1 Part (ii), we introduce and prove the following auxiliary lemma.

Lemma A.3. *Suppose that the hypotheses of Lemma 1 are satisfied. Then,* $b_n = O(h_n^2)$ *and* $\zeta_{1,n}^2 = O(h_n^3)$.

(i) $b_n \coloneqq \mathbb{E}\left[q_n(Z_i, Z_j, Z_k)\right] = O(h_n^2)$;

(ii)

$$\zeta_{1,n} \coloneqq \mathbb{E}\left[\|\mathbb{E}\left[p_n(Z_i, Z_j, Z_k)|Z_i\right] - b_n\|^2\right] = O(h_n^3). \tag{A.10}$$

Proof of Lemma A.3. Before starting the proof, we introduce the following notations

$$\begin{aligned} s_n^{(1)}(Z_i) &\coloneqq \mathbb{E}\left[q_n(Z_i, Z_j, Z_k)|Z_i\right] - b_n, \\ s_n^{(2)}(Z_i) &\coloneqq \mathbb{E}\left[q_n(Z_k, Z_i, Z_j)|Z_i\right] - b_n, \\ s_n^{(3)}(Z_i) &\coloneqq \mathbb{E}\left[q_n(Z_j, Z_k, Z_i)|Z_i\right] - b_n, \end{aligned}$$

so

$$\mathbb{E}\left[p_n(Z_i, Z_j, Z_k)|Z_i\right] - b_n = \frac{1}{3}\left(s_n^{(1)}(Z_i) + s_n^{(2)}(Z_i) + s_n^{(3)}(Z_i)\right),$$

and

$$\zeta_{1,n}^2 = \frac{1}{9}\mathbb{E}\left[\left\|s_n^{(1)}(Z_i) + s_n^{(2)}(Z_i) + s_n^{(3)}(Z_i)\right\|^2\right]. \tag{A.11}$$

First, let us compute

$$E_q(X_i, \xi_i, X_j, Z_k; h_n) \coloneqq \mathbb{E}\left[q_n(Z_i, Z_j, Z_k)|X_i, \xi_i, X_j, Z_k\right].$$

Note that

$$E_q(X_i,\xi_i,X_j,Z_k;h_n) = \underbrace{\mathbb{E}\left[h_n^{-1}K\left(\frac{c_{ij}(\xi_j-\xi_i)^2}{h_n^2}\right)(W_{ik}-W_{jk})(g(\xi_i,\xi_k)-g(\xi_j,\xi_k))|X_i,\xi_i,X_j,Z_k\right]}_{E_{q,2}(X_i,\xi_i,X_j,Z_k;h_n)}$$
$$+\underbrace{\mathbb{E}\left[h_n^{-1}r_K(Z_i,Z_j;h_n)(W_{ik}-W_{jk})(g(\xi_i,\xi_k)-g(\xi_j,\xi_k))|X_i,X_j,\xi_i\right]}_{r_{E_q}(X_i,\xi_i,X_j,Z_k;h_n)},$$

where r_K is defined in (A.1) and can be bounded as in (A.2) (see Lemma A.1 (iii)). Using the result of Lemma A.1 (iii), boundedness of W (Assumption 3 (i)), and $|g(\xi_i,\xi_k)-g(\xi_j,\xi_k)| \leqslant \overline{G}\,|\xi_i-\xi_j|$ (Assumption 3 (iii)), we establish

$$\begin{aligned}|r_K(Z_i,Z_j;h_n)(W_{ik}-W_{jk})(g(\xi_i,\xi_k)-g(\xi_j,\xi_k))| &\leqslant Ch_n\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}\,|g(\xi_i,\xi_k)-g(\xi_j,\xi_k)|\\ &\leqslant Ch_n^2\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}\end{aligned}$$

for some generic positive constant C, which is uniform over (X_i,ξ_i,X_j,Z_k). By Lemma A.2 (i), for some $C>0$,

$$h_n^{-1}\mathbb{E}\left[\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}\right]\leqslant C.$$

Combining these results, we conclude that there exists $C>0$ such that

$$\left|r_{E_q}(X_i,\xi_i,X_j,Z_k;h_n)\right|\leqslant Ch_n^2 \quad \text{a.s.}$$

Now we compute $E_{q,2}(X_i,\xi_i,X_j,Z_k;h_n)$. Again, using the result of Lemma A.1 (i),

$$E_{q,2}(X_i,\xi_i,X_j,Z_k;h_n)=\mathbb{E}\left[h_n^{-1}K\left(\frac{c_{ij}(\xi_j-\xi_i)^2}{h_n^2}\right)(W_{ik}-W_{jk})(g(\xi_i,\xi_k)-g(\xi_j,\xi_k))\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}\right].$$

First, we consider ξ_i and ξ_k such that $|\xi_i-\xi_k|>\alpha h_n$. For such ξ_i and ξ_k, using Assumption 7, we obtain

$$\begin{aligned}(g(\xi_i,\xi_k)-g(\xi_j,\xi_k))\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\} =& G(\xi_i,\xi_k)(\xi_i-\xi_j)\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}\\ &+\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}r_g(\xi_i,\xi_j,\xi_k).\end{aligned}$$

Note that Assumption 7 guarantees that for some $C>0$ and sufficiently large n,

$$\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}\,|r_g(\xi_i,\xi_j,\xi_k)|\leqslant Ch_n^2,$$

where C is uniform over ξ_i, ξ_j, ξ_k satisfying $|\xi_i - \xi_k| > \alpha h_n$. Therefore, for ξ_i and ξ_k satisfying $|\xi_i - \xi_k| > \alpha h_n$,

$$E_{q,2}(X_i, \xi_i, X_j, Z_k; h_n) = \underbrace{\mathbb{E}\left[h_n^{-1} K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right)(W_{jk} - W_{ik})G(\xi_i, \xi_k)(\xi_j - \xi_i)|X_i, \xi_i, X_j, Z_k\right]}_{E_{q,l}(X_i, \xi_i, X_j, Z_k; h_n)} + r_{E_{q,2}}(X_i, \xi_i, X_j, Z_k; h_n),$$

where, for some $C > 0$,

$$\left|r_{E_{q,2}}(X_i, \xi_i, X_j, Z_k; h_n)\right| \leqslant C h_n^2 \quad \text{a.s.}$$

Now we work with

$$\begin{aligned} E_{q,l}(X_i, \xi_i, X_j, Z_k; h_n) &:= \mathbb{E}\left[h_n^{-1} K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right)(W_{jk} - W_{ik})G(\xi_i, \xi_k)(\xi_j - \xi_i)|X_i, \xi_i, X_j, Z_k\right] \\ &= (W_{jk} - W_{ik})G(\xi_i, \xi_k) \int h_n^{-1} K\left(\frac{c_{ij}(\xi_j - \xi_i)^2}{h_n^2}\right)(\xi_j - \xi_i) f_{\xi|X}(\xi_j; X_j) d\xi_j \\ &= \frac{h_n (W_{jk} - W_{ik})G(\xi_i, \xi_k)}{c_{ij}} \int K(u^2) u f_{\xi|X}(\xi_i + h_n u/\sqrt{c_{ij}}; X_j) du, \end{aligned}$$

where the last equality follows from the change of the variable $\xi_j = \xi_i + h_n u/\sqrt{c_{ij}}$. Note that by Assumptions 4 (iii) and 6 (i), for all ξ_i such that $f_{\xi|X}(\cdot; X_j)$ is continuous on $B_{h_n u/\sqrt{c_{ij}}}(\xi_i)$ we have

$$\left|K(u^2) u f_{\xi|X}(\xi_i + h_n u/\sqrt{c_{ij}}; X_j) - K(u^2) u f_{\xi|X}(\xi_i; X_j)\right| \leqslant \frac{\overline{K} C_\xi h_n}{\sqrt{c_{ij}}}. \tag{A.12}$$

Notice that

$$\begin{aligned} \left|\int K(u^2) u f_{\xi|X}(\xi_i + h_n u/\sqrt{c_{ij}}; X_j) du\right| &\leqslant \left|\int K(u^2) u f_{\xi|X}(\xi_i; X_j) du\right| \\ &\quad + \left|\int \left(K(u^2) u f_{\xi|X}(\xi_i + h_n u/\sqrt{c_{ij}}; X_j) - K(u^2) u f_{\xi|X}(\xi_i; X_j)\right) du\right| \\ &\leqslant \int \left|K(u^2) u f_{\xi|X}(\xi_i + h_n u/\sqrt{c_{ij}}; X_j) - K(u^2) u f_{\xi|X}(\xi_i; X_j)\right| du, \end{aligned}$$

where the first inequality is due the triangle inequality, and the second follows from $\int u K(u^2) du =$

0. Consequently, we have

$$|E_{q,l}(X_i,\xi_i,X_j,Z_k;h_n)| \leqslant \left|\frac{h_n(W_{jk}-W_{ik})G(\xi_i,\xi_k)}{c_{ij}}\right| \times \int_{-1}^{+1} \left|K(u^2)uf_{\xi|X}(\xi_i+h_nu/\sqrt{c_{ij}};X_j)-K(u^2)uf_{\xi|X}(\xi_i;X_j)\right| du.$$

Combining (A.12) with boundedness of W, G and c_{ij}^{-1} (Assumptions 3 (i), 7, and 5 (iii), respectively), we conclude that there exists a uniform constant C such that

$$|E_{q,l}(X_i,\xi_i,X_j,Z_k;h_n)| \leqslant Ch_n^2$$

almost surely for all Z_i, Z_j, and Z_k such that $|\xi_i-\xi_k|>\alpha h_n$ and $f_{\xi|X}(\cdot;X_j)$ is continuous on $B_{h_nu/\sqrt{c_{ij}}}(\xi_i)$. Combining this with the previously obtained bounds on r_{E_q} and $r_{E_{q,2}}$, we conclude that there exists a uniform constant C such that (for sufficiently large n)

$$|E_q(X_i,\xi_i,X_j,Z_k;h_n)| \leqslant Ch_n^2 \tag{A.13}$$

almost surely for all Z_i, Z_j, and Z_k such that $|\xi_i-\xi_k|>\alpha h_n$ and $f_{\xi|X}(\cdot;X_j)$ is continuous on $B_{h_nu/\sqrt{c_{ij}}}(\xi_i)$.

Second, we argue that there exists $C>0$ such that

$$|E_q(X_i,\xi_i,X_j,Z_k;h_n)| \leqslant Ch_n \tag{A.14}$$

with probability one. Indeed, this is an immediate consequence from the following observation: for some uniform constant $C>0$,

$$\|q_n(Z_i,Z_j,Z_k)\| \leqslant C\mathbb{1}\{|\xi_j-\xi_i|\leqslant \alpha h_n\},$$

which in turn is guaranteed by the result of Lemma A.1 (i) and Assumption 3 (iii).

Equipped with the bounds (A.13) and (A.14), now we want to bound $b_n=\mathbb{E}q_n(Z_i,Z_j,Z_k)$. We start with considering $\mathbb{E}\left[q_n(Z_i,Z_j,Z_k)|Z_i\right]$. Since $c_{ij}>\underline{c}>0$ (Assumption 5 (iii)), Assumption 4 (ii) guarantees that there exists $\gamma_1>0$ such that the probability mass of ξ_i such that $f_{\xi|X}(\cdot;X_j)$ is continuous on $B_{h_nu/\sqrt{c_{ij}}}(\xi_i)$ is at least $1-\gamma_1h_n$ irrespectively of the values of X_i and X_j. Also, by Assumption 4 (i), the probability mass of ξ_k such that $|\xi_k-\xi_i|>\alpha h_n$ is at least $1-\gamma_2h_n$ irrespectively of the value of ξ_i for some $\gamma_2>0$. For those values of ξ_i and ξ_k, the bound (A.13) applies. Moreover, the bound (A.14) applies with probability one. Hence, integrating $E_q(X_i,\xi_i,X_j,Z_k;h_n)$ over (X_j,Z_k) ensures that there exists $C>0$ such that $\|\mathbb{E}\left[q_n(Z_i,Z_j,Z_k)|Z_i\right]\|\leqslant Ch_n^2$ for all ξ_i such that $f_{\xi|X}(\cdot;X_j)$ is continuous on $B_{h_nu/\sqrt{c_{ij}}}(\xi_i)$, which

happens with probability $1-\gamma_1 h_n$ at least. Moreover, (A.14) immediately implies that $\|\mathbb{E}[q_n(Z_i,Z_j,Z_k)|Z_i]\| \leqslant Ch_n$ with probability one. Combining these bounds and integrating over Z_i gives

$$\begin{aligned}\|\mathbb{E}[q_n(Z_i,Z_j,Z_j)]\| &= \|\mathbb{E}[\mathbb{E}[q_n(Z_i,Z_j,Z_j)|Z_i]]\| \\ &\leqslant \mathbb{E}[\|\mathbb{E}[q_n(Z_i,Z_j,Z_j)|Z_i]\|] \\ &\leqslant (1-\gamma_1 h_n)\times Ch_n^2 + \gamma_1 h_n \times Ch_n \\ &= O(h_n^2).\end{aligned}$$

Hence, we conclude $b_n = \mathbb{E}[q_n(Z_i,Z_j,Z_k)] = O(h_n^2)$, which completes the proof of the first part.

Second, we want to bound $\zeta_{1,n}^2$ based on (A.11). Using the same bounds on $\|\mathbb{E}[q_n(Z_i,Z_j,Z_j)|Z_i]\|$ as established above and exploiting $b_n = O(h_n^2)$, for $s_n^{(1)}(Z_i) \coloneqq \mathbb{E}[q_n(Z_i,Z_j,Z_k)|Z_i] - b_n$ we have the following: (i) for some $C>0$ and $\gamma_1>0$, we have $\left\|s_n^{(1)}(Z_i)\right\|^2 \leqslant Ch_n^4$ with probability at least $1-\gamma_1 h_n$ and $\left\|s_n^{(1)}(Z_i)\right\|^2 \leqslant Ch_n^2$ with probability one. Hence, $\mathbb{E}\left[\left\|s_n^{(1)}(Z_i)\right\|\right] \leqslant Ch_n^3$ for some $C>0$. Also, since q_n is symmetric in its first two arguments, we automatically get the same result for $s_n^{(2)}$: $\mathbb{E}\left[\left\|s_n^{(2)}(Z_i)\right\|\right] \leqslant Ch_n^3$ for some $C>0$.

Now we consider $\mathbb{E}[q_n(Z_i,Z_j,Z_k)|Z_k]$. Again, the bound (A.13) applies for all ξ_i satisfying (i) $|\xi_k-\xi_i|>\alpha h_n$ and (ii) $f_{\xi|X}(\cdot;X_j)$ is continuous on $B_{h_n u/\sqrt{c_{ij}}}(\xi_i)$ (for fixed ξ_k). By the same reasoning as above, the probability mass of those ξ_i is at least $1-\gamma_3 h_n$ (irrespectively (X_i,X_j,Z_k)) for some $\gamma_3>0$. At the same time, the bound (A.14) applies with probability one. Hence, integrating $E_q(X_i,\xi_i,X_j,Z_k;h_n)$ over (X_i,ξ_i,X_j) gives $\|\mathbb{E}[q_n(Z_i,Z_j,Z_k)|Z_k]\| \leqslant Ch_n^2$ a.s. for some $C>0$. Note that this (paired with $b_n = O(h_n^2)$) automatically implies that $\mathbb{E}\left[\left\|s_n^{(3)}(Z_i)\right\|\right] < Ch_n^4$ for some $C>0$.

Finally, the bounds for $\mathbb{E}\left[\left\|s_n^{(\ell)}(Z_i)\right\|^2\right]$ for $\ell \in \{1,2,3\}$ along with (A.11) imply that $\zeta_{1,n} = O(h_n^3)$. Q.E.D.

Proof of Lemma 1 Part (ii). The first step of the proof is to bound B_n, the infeasible analogues of $\hat{B}_n$ based on $\{d_{ij}^2\}_{i\neq j}$ instead $\{\hat{d}_{ij}^2\}_{i\neq j}$. Specifically,

$$\begin{aligned}B_n &= \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{d_{ij}^2}{h_n^2}\right) \frac{1}{n-2} \sum_{k\neq i,j} (W_{ik}-W_{jk})(g(\xi_i,\xi_k)-g(\xi_j,\xi_k)) \\ &= \frac{1}{n(n-1)(n-2)} \sum_{i\neq j\neq k} q_n(Z_i,Z_j,Z_k),\end{aligned}$$

where $q_n(Z_i,Z_j,Z_k)$ is as given in (A.8). Then B_n is a third order U-statistic with the symmetrized kernel $p_n(Z_i,Z_j,Z_k)$, where p_n is given by (A.9). Note that

$b_n \coloneqq \mathbb{E}\left[q_n(Z_i, Z_j, Z_k)\right] = \mathbb{E}\left[p_n(Z_i, Z_j, Z_k)\right]$, then

$$B_n = b_n + \binom{n}{3}^{-1} \sum_{i<j<k} (p_n(Z_i, Z_j, Z_k) - b_n).$$

Now we want to bound $U_n = \binom{n}{3}^{-1} \sum_{i<j<k} (p_n(Z_i, Z_j, Z_k) - b_n)$ using a Bernstein type inequality for U-statistic developed in Arcones (1995). Specifically, Theorem 2 in Arcones (1995) guarantees that

$$U_n = O_p\left(\max\left\{\frac{\zeta_{1,n}}{n^{1/2}}, \frac{1}{n}\right\}\right),$$

where $\zeta_{1,n}$ is as define in (A.10). Lemma A.3 (ii) establishes that $\zeta_{1,n}^2 = O(h_n^3)$. Since $nh_n \to \infty$ (Assumption 6 (iii)),

$$\frac{\zeta_{1,n}}{n^{1/2}} = \frac{O(h_n^{3/2})}{n^{1/2}} = o(h_n^2).$$

Since A.3 (i) guarantees $b_n = O(h_n^2)$, we conclude

$$B_n = b_n + U_n = O_p(h_n^2 + n^{-1}). \tag{A.15}$$

The second step is to bound $\left\|\hat{B}_n - B_n\right\|$. Using Assumption 6 (i) and the result of Lemma A.1 (ii), we have, with probability approaching one,

$$\begin{aligned}\left\|\hat{B}_n - B_n\right\| \leqslant & \overline{K}' \frac{\max_{i\neq j}\left|\hat{d}_{ij}^2 - d_{ij}^2\right|}{h_n^2} \\ & \times \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} \frac{1}{n-2} \sum_{k\neq i,j} \left\|(W_{ik} - W_{jk})(g(\xi_i, \xi_k) - g(\xi_j, \xi_k))\right\|.\end{aligned} \tag{A.16}$$

Since $|g(\xi_i, \xi_k) - g(\xi_j, \xi_k)| \leqslant \overline{G}\,|\xi_j - \xi_i|$ (Assumption 3 (iii)) and W is bounded (Assumption 3 (i)), there exists $C > 0$ such that

$$\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} \frac{1}{n-2} \sum_{k\neq i,j} \left\|(W_{ik} - W_{jk})(g(\xi_i, \xi_k) - g(\xi_j, \xi_k))\right\| \leqslant C h_n \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} \quad \text{a.s.}$$

So, with probability one,

$$\binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} \frac{1}{n-2} \sum_{k \neq i,j} \|(W_{ik} - W_{jk})(g(\xi_i, \xi_k) - g(\xi_j, \xi_k))\|$$
$$\leqslant C \binom{n}{2}^{-1} \sum_{i<j} \mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\} = O_p(h_n), \tag{A.17}$$

where the last equality is due to the result of Lemma A.2 (ii). Therefore, (A.16), (A.17) and Assumption 6 (ii) together imply

$$\left\| \hat{B}_n - B_n \right\| = O_p\left(\frac{R_n^{-1}}{h_n} \right).$$

Combining this with (A.15) completes the proof. Q.E.D.

A.1.4 Proof of Lemma 1 Part (iii)

Proof of Lemma 1 Part (iii). Note that

$$\begin{aligned}
\hat{C}_n =& \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{\hat{d}_{ij}^2}{h_n^2} \right) \frac{1}{n-2} \sum_{k \neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}) \\
=& \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{d_{ij}^2}{h_n^2} \right) \frac{1}{n-2} \sum_{k \neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}) \\
&+ \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} \left(K\left(\frac{\hat{d}_{ij}^2}{h_n^2} \right) - K\left(\frac{d_{ij}^2}{h_n^2} \right) \right) \frac{1}{n-2} \sum_{k \neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}).
\end{aligned}$$

The first step is to argue that

$$C_n := \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j} K\left(\frac{d_{ij}^2}{h_n^2} \right) \frac{1}{n-2} \sum_{k \neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}) = O_p(n^{-1}).$$

Let $K_{n,ij} \coloneqq h_n^{-1} K\left(\frac{d_{ij}^2}{h_n^2}\right)$. Note

$$\begin{aligned}
\sum_{i<j} K_{n,ij} \sum_{k\neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}) &= \frac{1}{2} \sum_{i\neq j} K_{n,ij} \sum_{k\neq i,j} (W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}) \\
&= \sum_{i\neq j} K_{n,ij} \sum_{k\neq i,j} (W_{ik} - W_{jk})\varepsilon_{ik} \\
&= \sum_{i} \sum_{k\neq i} \sum_{j\neq i,k} K_{n,ij}(W_{ik} - W_{jk})\varepsilon_{ik} \\
&= \sum_{i<k} \left(\sum_{j\neq i,k} \left(K_{n,ij}(W_{ik} - W_{jk}) + K_{n,kj}(W_{ik} - W_{ji})\right) \right) \varepsilon_{ik}
\end{aligned}$$

Then

$$C_n = \binom{n}{2}^{-1} \sum_{i<k} \hat{\omega}_{n,ik} \varepsilon_{ik},$$

where

$$\hat{\omega}_{n,ik} \coloneqq \frac{1}{n-2} \sum_{j\neq i,k} \left(K_{n,ij}(W_{ik} - W_{jk}) + K_{n,kj}(W_{ik} - W_{ji})\right).$$

First, we argue that

$$\max_{i\neq k} \|\hat{\omega}_{n,ik} - \omega_{n,ik}\| = o_p(1), \tag{A.18}$$

where

$$\omega_{n,ik} \coloneqq \mathbb{E}\left[K_{n,ij}(W_{ik} - W_{jk}) + K_{n,kj}(W_{ik} - W_{ji}) | X_i, X_k, \xi_i, \xi_k\right].$$

Consider

$$\begin{aligned}
\hat{\kappa}_{n,ik} &\coloneqq \frac{1}{n-2} \sum_{j\neq i,k} K_{n,ij}(W_{ik} - W_{jk}), \\
\kappa_{n,ik} &= \mathbb{E}\left[K_{n,ij}(W_{ik} - W_{jk}) | X_i, X_k, \xi_i, \xi_k\right].
\end{aligned}$$

Conditional on X_i, X_k, ξ_i, ξ_k, $\{K_{n,ij}(W_{ik} - W_{jk})\}_{j\neq i,k}$ is a collection of bounded (given the sample size) independent variables with $\|K_{n,ij}(W_{ik} - W_{jn})\| \leqslant C h_n^{-1}$ (by boundedness of K and W).

Moreover, using boundedness of W, K, and the result of Lemma A.1 (i), we have,

$$\mathbb{E}\left[\left\|K_{n,ij}\left(W_{ik}-W_{jn}\right)\right\|^2|X_i,X_k,\xi_i,\xi_k\right]\leqslant Ch_n^{-2}\mathbb{E}\left[\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}|\xi_i\right],$$

where C is a generic positive constant, which is uniform over (X_i,X_k,ξ_i,ξ_k). Also, by Lemma A.2 (i), for some $C>0$, which is uniform over ξ_i, we have

$$h_n^{-1}\mathbb{E}\left[\mathbb{1}\{|\xi_j-\xi_i|\leqslant\alpha h_n\}|\xi_i\right]\leqslant C.$$

Hence, we conclude that there exists $C>0$, which is uniform over (X_i,X_k,ξ_i,ξ_k), such that

$$\mathbb{E}\left[\left\|K_{n,ij}\left(W_{ik}-W_{jn}\right)\right\|^2|X_i,X_k,\xi_i,\xi_k\right]\leqslant Ch_n^{-1}. \tag{A.19}$$

Hence, applying Bernstein inequality A.1, we conclude that there exist positive constants a, b, and C such that for all (X_i,X_k,ξ_i,ξ_k) and $\epsilon>0$ we have

$$\mathbb{P}\left(\|\hat{\kappa}_{n,ik}-\kappa_{n,ik}\|\geqslant\epsilon|X_i,X_k,\xi_i,\xi_k\right)\leqslant C\exp\left(-\frac{(n-2)h_n\epsilon^2}{a+b\epsilon}\right).$$

Consequently, we also get

$$\mathbb{P}\left(\|\hat{\kappa}_{n,ik}-\kappa_{n,ik}\|\geqslant\epsilon\right)\leqslant C\exp\left(-\frac{(n-2)h_n\epsilon^2}{a+b\epsilon}\right),$$

and, using the union bound, we finally obtain

$$\mathbb{P}\left(\max_{i\neq k}\|\hat{\kappa}_{n,ik}-\kappa_{n,ik}\|\geqslant\epsilon\right)\leqslant\binom{n}{2}C\exp\left(-\frac{(n-2)h_n\epsilon^2}{a+b\epsilon}\right).$$

Therefore, $nh_n/\ln n\to\infty$ (Assumption 6 (iii)) guarantees that

$$\max_{i\neq k}\|\hat{\kappa}_{n,ik}-\kappa_{n,ik}\|=o_p(1). \tag{A.20}$$

By the same reasoning,

$$\max_{i\neq k}\left\|\frac{1}{n-2}\sum_{j\neq i,k}K_{n,kj}(W_{ik}-W_{ji})-\mathbb{E}\left[K_{n,kj}(W_{ik}-W_{ji})|X_i,X_k,\xi_i,\xi_k\right]\right\|=o_p(1),$$

so combining this with (A.20) ensures that (A.18) holds.

Second, $\|\omega_{n,ik}\|$ is uniformly (over X_i, X_k, ξ_i, and ξ_k) bounded, i.e., there exists $C>0$ such that $\limsup_{n\to\infty}\max_{i\neq k}\|\omega_{n,ik}\|\leqslant C$. Indeed, using the boundedness of W and K, and the result

of Lemma A.1 (i), we conclude that there exists $C > 0$, which is uniform over X_i, X_k, ξ_i, ξ_k, such that

$$\mathbb{E}\left[\|K_{n,ij}(W_{ik} - W_{jk})\| \,|X_i, X_k, \xi_i, \xi_k\right] \leqslant Ch_n^{-1}\mathbb{E}\left[\mathbb{1}\{|\xi_j - \xi_i| \leqslant \alpha h_n\}|\xi_i\right].$$

Invoking (A.19) again, we conclude that $\mathbb{E}\left[\|K_{n,ij}(W_{ik} - W_{jk})\| \,|X_i, X_k, \xi_i, \xi_k\right]$ is uniformly (over (X_i, X_k, ξ_i, ξ_k)) bounded. By the same reasoning, $\mathbb{E}\left[\|K_{n,kj}(W_{ik} - W_{ji})\| \,|X_i, X_k, \xi_i, \xi_k\right]$ is also uniformly bounded. These results together ensure uniform boundedness of $\|\omega_{n,ik}\|$.

Combining uniform boundedness of $\|\omega_{n,ik}\|$ with (A.18) allows us to conclude that there exists $C_\omega > 0$ such that with probability approaching one $\max_{i\neq k}\|\hat{\omega}_{n,ik}\| < C_\omega$. Moreover, recall that conditional on $\{X_i, \xi_i\}_{i=1}^n$, $\{\hat{\omega}_{n,ik}\varepsilon_{ik}\}_{i<k}$ is a collection of independent vectors with zero mean, which satisfy the requirements of Theorem A.2. Therefore, Theorem A.2 guarantees that there exist some positive constants C, a, b such that for all $\{X_i, \xi_i\}_{i=1}^n$ satisfying $\max_{i\neq k}\|\hat{\omega}_{n,ik}\| < C_\omega$, for all $\epsilon > 0$,

$$\mathbb{P}\left(\|C_n\| > \epsilon|\{X_i, \xi_i\}_{i=1}^n\right) \leqslant C\exp\left(-\frac{\binom{n}{2}\epsilon^2}{a + b\epsilon}\right).$$

Since the requirement $\max_{i\neq k}\|\hat{\omega}_{n,ik}\| < C_\omega$ is satisfied with probability approaching one, we conclude that $C_n = O_p(n^{-1})$.

The second step is to bound

$$\Delta\hat{C}_n \coloneqq \hat{C}_n - C_n = \binom{n}{2}^{-1} h_n^{-1}\sum_{i<j}\left(K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) - K\left(\frac{d_{ij}^2}{h_n^2}\right)\right)\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk}).$$

First we want to bound

$$\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk})\right\|.$$

Again, since (i) conditional on $\{X_i, \xi_i\}_{i=1}^n$, $\{(W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk})\}$ is a collection of independent vectors with mean zero, which components satisfy the requirements of Theorem A.2, (ii) W is bounded; we conclude that there exist positive constants C_2, a_2 and b_2 such that for all $\{X_i, \xi_i\}_{i=1}^n$ and for all $\epsilon > 0$

$$\mathbb{P}\left(\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik} - W_{jk})(\varepsilon_{ik} - \varepsilon_{jk})\right\| > \epsilon|\{X_i, \xi_i\}_{i=1}^n\right) \leqslant C_2\exp\left(-\frac{(n-2)\epsilon^2}{a_2 + b_2\epsilon}\right).$$

Since the constants are uniform over $\{X_i, \xi_i\}_{i=1}^n$, this also holds unconditionally

$$\mathbb{P}\left(\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(\varepsilon_{ik}-\varepsilon_{jk})\right\| > \epsilon\right) \leqslant C_2 \exp\left(-\frac{(n-2)\epsilon^2}{a_2+b_2\epsilon}\right).$$

Using the union bound,

$$\mathbb{P}\left(\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(\varepsilon_{ik}-\varepsilon_{jk})\right\| > \epsilon\right) \leqslant \binom{n}{2} C_2 \exp\left(-\frac{(n-2)\epsilon^2}{a_2+b_2\epsilon}\right).$$

Note that taking $\epsilon = c\left(\frac{\ln n}{n}\right)^{1/2}$ for some $c > 0$, we have (for sufficiently large n)

$$\mathbb{P}\left(\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(\varepsilon_{ik}-\varepsilon_{jk})\right\| > c\left(\frac{\ln n}{n}\right)^{1/2}\right) \leqslant \binom{n}{2} C_2 n^{-\frac{c}{2(a_2+b_2)}}.$$

Note that we can make the right-hand side arbitrarily small by taking sufficiently large c. Therefore, we conclude

$$\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(\varepsilon_{ik}-\varepsilon_{jk})\right\| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Hence,

$$\left\|\Delta\hat{C}_n\right\| \leqslant \binom{n}{2}^{-1} h_n^{-1} \sum_{i<j}\left|K\left(\frac{\hat{d}_{ij}^2}{h_n^2}\right) - K\left(\frac{d_{ij}^2}{h_n^2}\right)\right| \times O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Combining this with (A.7), we obtain

$$\left\|\Delta\hat{C}_n\right\| = O_p\left(\frac{R_n^{-1}}{h_n^2}\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Finally, we conclude

$$\hat{C}_n = C_n + \Delta\hat{C}_n = O_p\left(\frac{R_n^{-1}}{h_n^2}\left(\frac{\ln n}{n}\right)^{1/2} + n^{-1}\right).$$

Q.E.D.

A.1.5 Proof of Theorem 1

Proof of Theorem 1. Directly follows from (4.1) and Lemma 1. Q.E.D.

A.2 Proof of the results of Section 4.2.1

A.2.1 Auxiliary Lemma A.4 and its proof

Denote

$$d^2_{ij,n-2}(\beta) \coloneqq \frac{1}{n-2}\sum_{k\neq i,j}(Y^*_{ik} - Y^*_{jk} - (W_{ik} - W_{jk})'\beta)^2,$$
$$d^2_{ij,n}(\beta) \coloneqq \frac{1}{n}\sum_{k}(Y^*_{ik} - Y^*_{jk} - (W_{ik} - W_{jk})'\beta)^2, \tag{A.21}$$

where Y^* is as defined in (2.4).

Lemma A.4. *Suppose that $\mathcal{B}$ is compact. Then, under Assumptions 1 and 3,*

$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|d^2_{ij,n-2}(\beta) - d^2_{ij}(\beta)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right),$$
$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|d^2_{ij,n}(\beta) - d^2_{ij}(\beta)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Proof of Lemma A.4. Using $Y^*_{ik} - Y^*_{jk} = (W_{ik} - W_{jk})'\beta_0 + g_{ik} - g_{jk}$, we decompose $d^2_{ij,n-2}$ and as follows

$$\begin{aligned} d^2_{ij,n-2}(\beta) =& (\beta_0 - \beta)'\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik} - W_{jk})(W_{ik} - W_{jk})'(\beta_0 - \beta) \\ &+ \frac{2}{n-2}\sum_{k\neq i,j}(W_{ik} - W_{jk})(g_{ik} - g_{jk})(\beta_0 - \beta) \\ &+ \frac{1}{n-2}\sum_{k\neq i,j}(g_{ik} - g_{jk})^2. \end{aligned}$$

Similarly,

$$\begin{aligned} d^2_{ij}(\beta) =& (\beta_0 - \beta)'\mathbb{E}\left[(W_{ik} - W_{jk})(W_{ik} - W_{jk})'|Z_i, Z_j\right](\beta_0 - \beta) \\ &+ 2\mathbb{E}\left[(W_{ik} - W_{jk})(g_{ik} - g_{jk})|Z_i, Z_j\right](\beta_0 - \beta) \\ &+ \mathbb{E}\left[(g_{ik} - g_{jk})^2|Z_i, Z_j\right]. \end{aligned}$$

Also recall

$$\mathbb{E}\left[(W_{ik}-W_{jk})(W_{ik}-W_{jk})'|Z_i,Z_j\right]=\mathbb{E}\left[(W_{ik}-W_{jk})(W_{ik}-W_{jk})'|X_i,X_j\right]=\mathcal{C}(X_i,X_j).$$

Since $\mathcal{B}$ is bounded, for some positive constants c_1 and c_2, we have

$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|d^2_{ij,n-2}-d^2_{ij}\right|\leqslant c_1S_1+c_2S_2+S_3, \tag{A.22}$$

where

$$\begin{aligned}
S_1&:=\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(W_{ik}-W_{jk})'-\mathcal{C}(X_i,X_j)\right\|,\\
S_2&:=\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(g_{ik}-g_{jk})'-\mathbb{E}\left[(W_{ik}-W_{jk})(g_{ik}-g_{jk})|Z_i,Z_j\right]\right\|,\\
S_3&:=\max_{i\neq j}\left|\frac{1}{n-2}\sum_{k\neq i,j}(g_{ik}-g_{jk})^2-\mathbb{E}\left[(g_{ik}-g_{jk})^2|Z_i,Z_j\right]\right|.
\end{aligned}$$

Now we argue that

$$S_1=O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right). \tag{A.23}$$

Indeed, conditional on Z_i and Z_j (effectively on X_i and X_j), $\{(W_{ik}-W_{jk})(W_{ik}-W_{jk})'\}_{k\neq i,j}$ is a collection of iid matrices, which are uniformly bounded over X_i and X_j. Hence, by Bernstein inequality A.1, there exist positive constants a, b, and C such that for all $\epsilon>0$ and (almost) all Z_i and Z_j, we have

$$\mathbb{P}\left(\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(W_{ik}-W_{jk})'-\mathcal{C}(X_i,X_j)\right\|>\epsilon|Z_i,Z_j\right)\leqslant C\exp\left(-\frac{(n-2)\epsilon^2}{a+b\epsilon}\right).$$

Since a, b, and C are uniform over Z_i and Z_j, we also have

$$\mathbb{P}\left(\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(W_{ik}-W_{jk})'-\mathcal{C}(X_i,X_j)\right\|>\epsilon\right)\leqslant C\exp\left(-\frac{(n-2)\epsilon^2}{a+b\epsilon}\right).$$

Finally, applying the union bound,

$$\mathbb{P}\left(S_1>\epsilon\right)\leqslant\binom{n}{2}C\exp\left(-\frac{(n-2)\epsilon^2}{a+b\epsilon}\right).$$

This guarantees that (A.23) holds. Applying the same reasoning, we obtain

$$S_2 = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right), \qquad S_3 = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Combining the rates of convergence for S_1, S_2 and S_3 with (A.22) delivers the result.

Essentially the same argument applies to demonstrate the second statement. Q.E.D.

A.2.2 Proof of Lemma 2 Part (i)

Proof of Lemma 2 Part (i). First, we argue that

$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\hat{q}_{ij}^2(\beta) - q_{ij}^2(\beta)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right),$$

where

$$\hat{q}_{ij}^2(\beta) := \frac{1}{n-2}\sum_{k\neq i,j}\left(Y_{ik} - Y_{jk} - (W_{ik} - W_{jk})'\beta\right)^2,$$

and

$$\begin{aligned} q_{ij}^2(\beta) &:= \mathbb{E}\left[(Y_{ik} - Y_{jk} - (W_{ik} - W_{jk})'\beta)^2\right] \\ &= \underbrace{\mathbb{E}\left[(Y_{ik}^* - Y_{jk}^* - (W_{ik} - W_{jk})'\beta)^2\right]}_{d_{ij}^2(\beta)} + 2\sigma^2, \end{aligned}$$

and Y^* is as defined in (2.4). We can similarly decompose $\hat{q}_{ij}^2(\beta)$

$$\begin{aligned}
\hat{q}_{ij}^2(\beta) =& \frac{1}{n-2}\sum_{k\neq i,j}\left(Y_{ik}-Y_{jk}-(W_{ik}-W_{jk})'\beta\right)^2 \\
=& \frac{1}{n-2}\sum_{k\neq i,j}\left(Y_{ik}^*-Y_{jk}^*-(W_{ik}-W_{jk})'\beta+\epsilon_{ik}-\epsilon_{jk}\right)^2 \\
=& \underbrace{\frac{1}{n-2}\sum_{k\neq i,j}\left(Y_{ik}^*-Y_{jk}^*-(W_{ik}-W_{jk})'\beta\right)^2}_{d_{ij,n-2}^2(\beta)} \\
&+\frac{2}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})'(\epsilon_{ik}-\epsilon_{jk})(\beta_0-\beta) \\
&+\frac{2}{n-2}\sum_{k\neq i,j}\left(g_{ik}-g_{jk}\right)\left(\epsilon_{ik}-\epsilon_{jk}\right) \\
&+\frac{1}{n-2}\sum_{k\neq i,j}(\varepsilon_{ik}-\varepsilon_{jk})^2.
\end{aligned}$$

Then,

$$\begin{aligned}
\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\hat{q}_{ij}^2(\beta)-q_{ij}^2(\beta)\right| \leqslant& \max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|d_{ij,n-2}^2(\beta)-d_{ij}^2(\beta)\right| \\
&+2\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})'(\epsilon_{ik}-\epsilon_{jk})(\beta_0-\beta)\right| \\
&+2\max_{i\neq j}\left|\frac{1}{n-2}\sum_{k\neq i,j}\left(g_{ik}-g_{jk}\right)\left(\epsilon_{ik}-\epsilon_{jk}\right)\right| \\
&+\max_{i\neq j}\left|\frac{1}{n-2}\sum_{k\neq i,j}\left((\varepsilon_{ik}-\varepsilon_{jk})^2-2\sigma^2\right)\right|. \qquad \text{(A.24)}
\end{aligned}$$

First, by Lemma A.4,

$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|d_{ij,n-2}^2(\beta)-d_{ij}^2(\beta)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

In the proof of Lemma 1 Part (iii), we have demonstrated that

$$\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(\varepsilon_{ik}-\varepsilon_{jk})\right\| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

By a nearly identical argument, we also have

$$\max_{i\neq j}\left|\frac{1}{n-2}\sum_{k\neq i,j}\left(g_{ik}-g_{jk}\right)\left(\epsilon_{ik}-\epsilon_{jk}\right)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Also note that since $\mathcal{B}$ is bounded, there exists $C>0$ such that

$$\begin{aligned}
&\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})'(\epsilon_{ik}-\epsilon_{jk})(\beta_0-\beta)\right| \\
\leqslant &\max_{i\neq j}\left\|\frac{1}{n-2}\sum_{k\neq i,j}(W_{ik}-W_{jk})(\varepsilon_{ik}-\varepsilon_{jk})\right\|\max_{\beta\in\mathcal{B}}\|\beta_0-\beta\| \\
= &O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).
\end{aligned}$$

The last step is to prove

$$\max_{i\neq j}\left|\frac{1}{n-2}\sum_{k\neq i,j}\left((\varepsilon_{ik}-\varepsilon_{jk})^2-2\sigma^2\right)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right). \tag{A.25}$$

Conditional on $\{Z_i, Z_j\}$, $\{\varepsilon_{ik}-\varepsilon_{jk}\}_{k\neq i,j}$ is a collection of iid sub-Gaussian random variables. Then Corollary A.1 guarantees that there exist some positive constants a, b, and C such that for all $\epsilon>0$ and (almost) all Z_i and Z_j

$$\mathbb{P}\left(\left|\frac{1}{n-2}\sum_{k\neq i,j}\left((\varepsilon_{ik}-\varepsilon_{jk})^2-2\sigma^2\right)|Z_i,Z_j\right|>\epsilon\right)\leqslant C\exp\left(-\frac{(n-2)\epsilon^2}{a+b\epsilon}\right).$$

Importantly, Assumption 3 (iv) guarantees that a, b, and C are uniform over Z_i and Z_j. Hence, we also have

$$\mathbb{P}\left(\left|\frac{1}{n-2}\sum_{k\neq i,j}\left((\varepsilon_{ik}-\varepsilon_{jk})^2-2\sigma^2\right)\right|>\epsilon\right)\leqslant C\exp\left(-\frac{(n-2)\epsilon^2}{a+b\epsilon}\right).$$

Finally, using the union bound,

$$\mathbb{P}\left(\max_{i\neq j}\left|\frac{1}{n-2}\sum_{k\neq i,j}\left((\varepsilon_{ik}-\varepsilon_{jk})^2-2\sigma^2\right)\right|>\epsilon\right)\leqslant\binom{n}{2}C\exp\left(-\frac{(n-2)\epsilon^2}{a+b\epsilon}\right),$$

which ensures that (A.25) holds.

These probability bounds together with (A.24) imply that

$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\hat{q}_{ij}^2(\beta)-q_{ij}^2(\beta)\right|=O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

This, in turn, guarantees

$$\max_{i\neq j}\left|\hat{q}_{ij}^2-q_{ij}^2\right|=O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Q.E.D.

A.2.3 Proof of Lemma 2 Part (ii)

Proof of Lemma 2 Part (ii). Note that

$$\begin{aligned}2\hat{\sigma}^2=\min_{i\neq j}\hat{q}_{ij}^2 &\leqslant \min q_{ij}^2+O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right)\\ &\leqslant 2\sigma^2+\min_{i\neq j}\mathbb{E}\left[\left(g(\xi_i,\xi_k)-g(\xi_j,\xi_k)\right)^2|\xi_i,\xi_j\right]+O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right),\end{aligned}$$

where the fist inequality follows from the result of Lemma 2 (i). Note that Assumption 3 (iii) guarantees that

$$\min_{i\neq j}\mathbb{E}\left[\left(g(\xi_i,\xi_k)-g(\xi_j,\xi_k)\right)^2|\xi_i,\xi_j\right]\leqslant\overline{G}^2\min_{i\neq j}\left\|\xi_i-\xi_j\right\|^2.$$

At the same time, using the result of Lemma 2 (i),

$$2\hat{\sigma}^2\geqslant 2\sigma^2-O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Hence,

$$2\sigma^2-O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right)\leqslant 2\hat{\sigma}^2\leqslant 2\sigma^2+\overline{G}^2\min_{i\neq j}\left\|\xi_i-\xi_j\right\|^2+O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right),$$

which completes the proof. Q.E.D.

A.3 Proofs of the results of Section 4.2.2

A.3.1 Proof of Theorem 2

First, we prove two auxiliary lemmas.

Lemma A.5. *Suppose that the hypotheses of Theorem 2 are satisfied. Then, for any $C > 0$, there exists $a_C > 0$ such that with probability approaching one, we have*

$$\min_i \frac{|\mathcal{B}_i(\delta_{n,C})|}{n-1} \geqslant C(n^{-1}\ln n)^{1/2},$$

where

$$\mathcal{B}_i(\delta) := \{i' \neq i : X_{i'} = X_i, \xi_{i'} \in B_\delta(\xi_i)\}$$

and $\delta_{n,C} := a_C(n^{-1}\ln n)^{\frac{1}{2d_\xi}}$.

Proof of Lemma A.5. Let $p_r := \mathbb{P}(X = x_r)$ and

$$p_Z(x,\xi;\delta) := \mathbb{P}(X_j = x)\mathbb{P}(\xi_j \in B_\delta(\xi)|X_j = x),$$
$$\mathcal{Q}_i(\delta) := \frac{|\mathcal{B}_i(\delta)|}{n-1}.$$

Then, by Bernstein inequality A.1, we have for all $\epsilon \in (0,1)$

$$\mathbb{P}(|\mathcal{Q}_i(\delta) - p_Z(X_i,\xi_i;\delta)| \geqslant \epsilon|X_i,\xi_i) \leqslant 2\exp\left(-\frac{\frac{1}{2}(n-1)\epsilon^2}{1+\frac{1}{3}\epsilon}\right) \leqslant 2\exp\left(-\frac{1}{4}n\epsilon^2\right).$$

Hence, we also have

$$\mathbb{P}(|\mathcal{Q}_i(\delta) - p_Z(X_i,\xi_i;\delta)| \geqslant \epsilon) \leqslant 2\exp\left(-\frac{1}{4}n\epsilon^2\right),$$

and, using the union bound,

$$\mathbb{P}(\max_i |\mathcal{Q}_i(\delta) - p_Z(X_i,\xi_i;\delta)| \geqslant \epsilon) \leqslant 2n\exp\left(-\frac{1}{4}n\epsilon^2\right).$$

Hence, for any $\tilde{C} > 0$ (and large enough n),

$$\mathbb{P}(\max_i |\mathcal{Q}_i(\delta) - p_Z(X_i,\xi_i;\delta)| \geqslant \tilde{C}(n^{-1}\ln n)^{1/2}) \leqslant 2n^{1-\tilde{C}/4}.$$

Using Assumption 8 (ii), for $\delta \leqslant \overline{\delta}$,

$$p_Z(X_i, \xi_i; \delta) \geqslant \underline{\kappa}\delta^{d_\xi} \qquad \text{a.s.}$$

where $\underline{\kappa} = \kappa \min_r p_r$. Then, for $\delta = a(n^{-1}\ln n)^{\frac{1}{2d_\xi}}$ for a fixed $a > 0$, we have (for large enough n)

$$\min p_Z(X_i, \xi_i; a(n^{-1}\ln n)^{\frac{1}{2d_\xi}}) \geqslant \underline{\kappa}a^{d_\xi}(n^{-1}\ln n)^{1/2}.$$

Hence, with probability at least $1 - 2n^{1-\tilde{C}/4}$, we have

$$\begin{aligned}\min_i \mathcal{Q}_i(a(n^{-1}\ln n)^{\frac{1}{2d_\xi}}) &\geqslant \underline{\kappa}a^{d_\xi}(n^{-1}\ln n)^{1/2} - \tilde{C}(n^{-1}\ln n)^{1/2} \\ &= (\underline{\kappa}a^{d_\xi} - \tilde{C})(n^{-1}\ln n)^{1/2}. \end{aligned} \tag{A.26}$$

Take some fixed $\tilde{C} > 4$. In this case, (A.26) holds with probability approaching one. Now take any fixed $a_C > \left(\underline{\kappa}^{-1}(C + \tilde{C})\right)^{1/d_\xi}$. For such a_C,

$$(\underline{\kappa}a^{d_\xi} - \tilde{C})(n^{-1}\ln n)^{1/2} > C(n^{-1}\ln n)^{1/2}.$$

As a result, we conclude that with probability approaching one, we have

$$\min_i \mathcal{Q}_i(a(n^{-1}\ln n)^{\frac{1}{2d_\xi}}) > C(n^{-1}\ln n)^{1/2},$$

which completes the proof. Q.E.D.

Lemma A.6. *Suppose that the hypotheses of Theorem 2 are satisfied. Then we have:*

(i) $\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})^2 = O_p\left((n^{-1}\ln n)^{\frac{1}{2d_\xi}}\right)$,

(ii) $\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \max_k |n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell} - Y^*_{i\ell})| = O_p\left((n^{-1}\ln n)^{\frac{1}{2d_\xi}}\right)$.

Proof of Lemma A.6. Proof of Part (i). The proof of the first part consists of the following two steps:

1. We argue that the object of interest $\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})^2$ can be bounded using the estimated pseudo-distance $\hat{d}^2_\infty(i, i')$. Specifically, we demonstrate

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i\ell'})^2 \leqslant 2\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i, i') + O_p\left((n^{-1}\ln n)^{\frac{1}{2d_\xi}}\right). \tag{A.27}$$

2. To complete the proof, we need to bound $\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i,i')$. Specifically, we establish

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i,i') = O_p\left((n^{-1}\ln n)^{\frac{1}{2d_\xi}}\right). \tag{A.28}$$

Notice that (A.27) and (A.28) together deliver the desired result. The rest of the proof then is to demonstrate that (A.27) and (A.28) hold.

Proof that (A.27) *holds.*

First, take some $C > \overline{C}$, where $\overline{C}$ is defined in the text of Theorem 2. Lemma A.5 guarantees, that there exist some a_C and $\delta_{n,C} = a_C(n^{-1}\ln n)^{\frac{1}{2d_\xi}}$ such that with probability approaching one, we have

$$\min_i |\mathcal{B}_i(\delta_{n,C})| \geqslant C(n\ln n)^{1/2}.$$

Since, by the hypothesis of Theorem 2, we have

$$\max_i n_i \leqslant \overline{C}(n\ln n)^{1/2} < C(n\ln n)^{1/2},$$

we conclude that $\min_i |\mathcal{B}_i(\delta_{n,C})| > n_i$ holds with probability approaching one. For any agents i and $i' \in \hat{\mathcal{N}}_i(n_i)$, let k and k' be two agents (other than i and i') such that $k \in \mathcal{B}_i(\delta_{n,C})$ and $k' \in \mathcal{M}_{i'}(\delta_{n,C})$ (by Lemma A.5, this happens simultaneously for all i and i' with probability approaching one for large enough n). Since Y^* is bounded and g is Lipschitz (Assumption 3 (iii)), for $\mathcal{G} \coloneqq \|Y^*\|_\infty \overline{G}$, we have

$$\left|n^{-1}\sum_\ell Y^{*2}_{i\ell} - n^{-1}\sum_\ell Y^*_{i\ell}Y^*_{k\ell}\right| \leqslant \mathcal{G}\delta_{n,C},$$

$$\left|n^{-1}\sum_\ell Y^*_{i'\ell}Y^*_{i\ell} - n^{-1}\sum_\ell Y^*_{i'\ell}Y^*_{k\ell}\right| \leqslant \mathcal{G}\delta_{n,C},$$

$$\left|n^{-1}\sum_\ell Y^{*2}_{i'\ell} - n^{-1}\sum_\ell Y^*_{i'\ell}Y^*_{k'\ell}\right| \leqslant \mathcal{G}\delta_{n,C},$$

$$\left|n^{-1}\sum_\ell Y^*_{i\ell}Y^*_{i'\ell} - n^{-1}\sum_\ell Y^*_{i\ell}Y^*_{k'\ell}\right| \leqslant \mathcal{G}\delta_{n,C}$$

with probability approaching one for all i, $i' \in \hat{\mathcal{N}}_i(n_i)$. Then

$$\begin{aligned} n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})^2 &\leqslant \left| n^{-1}\sum_\ell Y^{*2}_{i\ell} - n^{-1}\sum_\ell Y^*_{i'\ell}Y^*_{i\ell} \right| + \left| n^{-1}\sum_\ell Y^{*2}_{i'\ell} - n^{-1}\sum_\ell Y^*_{i\ell}Y^*_{i'\ell} \right| \\ &\leqslant \left| n^{-1}\sum_\ell Y^*_{i\ell}Y^*_{k\ell} - n^{-1}\sum_\ell Y^*_{i'\ell}Y^*_{k\ell} \right| + \left| n^{-1}\sum_\ell Y^*_{i'\ell}Y^*_{k'\ell} - n^{-1}\sum_\ell Y^*_{i\ell}Y^*_{k'\ell} \right| + 4\mathcal{G}\delta_{n,C} \\ &= \left| n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})Y^*_{k\ell} \right| + \left| n^{-1}\sum_\ell (Y^*_{i'\ell} - Y^*_{i\ell})Y^*_{k'\ell} \right| + 4\mathcal{G}\delta_{n,C}. \end{aligned}$$

Note that since Y^* is bounded, there exists $\Delta > 0$ such that for all i and i',

$$\left| n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})Y^*_{k\ell} \right| \leqslant \left| (n-3)^{-1}\sum_{\ell \neq i,i',k} (Y^*_{i\ell} - Y^*_{i'\ell})Y^*_{k\ell} \right| + \frac{\Delta}{n}.$$

So, for all i, $i' \in \hat{\mathcal{N}}_i(n_i)$, with probability approaching one we have

$$\begin{aligned} \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})^2 \leqslant & \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \max_{k \neq i,i'} \left| (n-3)^{-1}\sum_{\ell \neq i,i',k} (Y^*_{i\ell} - Y^*_{i'\ell})Y^*_{k\ell} \right| \\ & + \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \max_{k' \neq i,i'} \left| (n-3)^{-1}\sum_{\ell \neq i,i',k} (Y^*_{i'\ell} - Y^*_{i\ell})Y^*_{k'\ell} \right| \\ & + 4\mathcal{G}\delta_{n,C} + \frac{2\Delta}{n}. \end{aligned}$$

Also, by an argument which is nearly identical to the one given in the proof of Lemma 2,

$$r_n \coloneqq \max_{i \neq i' \neq k} \left| (n-3)^{-1}\sum_{\ell \neq i,i',k} (Y_{i\ell} - Y_{i'\ell})Y_{k\ell} - (n-3)^{-1}\sum_{\ell \neq i,i',k} (Y^*_{i\ell} - Y^*_{i'\ell})Y^*_{k\ell} \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{1/2} \right). \tag{A.29}$$

So, we have

$$\begin{aligned} \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1}\sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})^2 \leqslant & \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \max_{k \neq i,i'} \left| (n-3)^{-1}\sum_{\ell \neq i,i',k} (Y_{i\ell} - Y_{i'\ell})Y_{k\ell} \right| \\ & + \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \max_{k' \neq i,i'} \left| (n-3)^{-1}\sum_{\ell \neq i,i',k} (Y_{i'\ell} - Y_{i\ell})Y_{k'\ell} \right| \\ & + 4\mathcal{G}\delta_{n,C} + \frac{2\Delta}{n} + 2r_n. \end{aligned}$$

Invoking the definition of $\hat{d}^2_\infty(i,i')$, we have

$$\begin{aligned}\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1} \sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})^2 &\leqslant 2 \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i,i') + 4\mathcal{G}\delta_{n,C} + \frac{2\Delta}{n} + 2r_n \\ &= 2 \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i,i') + O_p\left((n^{-1}\ln n)^{\frac{1}{2d_\xi}}\right).\end{aligned}$$

This completes the proof that (A.27) holds.

Proof that (A.28) *holds.*

First, for any i and i', (A.29) guarantees that

$$\max_i \max_{i' \neq i} \left| \hat{d}^2_\infty(i,i') - \hat{d}^2_{\infty,*}(i,i') \right| = O_p\left((n^{-1}\ln n)^{1/2}\right),$$

where

$$\hat{d}^2_{\infty,*}(i,i') \coloneqq \max_{k \neq i,i'} \left| (n-3)^{-1} \sum_{\ell \neq i,i',k} (Y^*_{i\ell} - Y^*_{i'\ell}) Y^*_{k\ell} \right|.$$

Second, for any i and for any $i' \in \mathcal{B}_i(\delta_{n,C})$, we have for all ξ_ℓ

$$|Y^*_{i\ell} - Y^*_{i'\ell}| = |g(\xi_i, \xi_\ell) - g(\xi_{i'}, \xi_\ell)| \leqslant \overline{G}\delta_{n,C},$$

where the inequality is ensured by Assumption 3 (iii). Again, since Y^* is bounded, then, for all i and $i' \in \mathcal{B}_i(\delta_{n,C})$, we have

$$\hat{d}^2_{\infty,*}(i,i') \leqslant \|Y^*\|_\infty \overline{G}\delta_{n,C}.$$

Hence, for all i and $i' \in \mathcal{B}_i(\delta_{n,C})$, we have

$$\hat{d}^2_\infty(i,i') \leqslant \|Y^*\|_\infty \overline{G}\delta_{n,C} + O_{p,n}\left((n^{-1}\ln n)^{1/2}\right).$$

Then, by definition of $\mathcal{N}_i$ and the fact that for all i we have $n_i < |\mathcal{M}_i(\delta_{n,C})|$ with probability approaching one, we conclude

$$\begin{aligned}\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i,i') &\leqslant \|Y^*\|_\infty \overline{G}\delta_{n,C} + O_{p,n}\left((n^{-1}\ln n)^{1/2}\right) \\ &= O_{p,n}\left((n^{-1}\ln n)^{\frac{1}{2d_\xi}}\right),\end{aligned}$$

which completes the proof that (A.28) holds.

Since we have demonstrated that both (A.27) and (A.28) hold, the proof of Part (i) is complete.

Proof of Part (ii). Note that $n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell}-Y^*_{i\ell})=0$ when $i=i'$ for all k. For this reason, below $\max_{i'\in\hat{\mathcal{N}}_i(n_i)}$ should be read as $\max_{i'\in\hat{\mathcal{N}}_i(n_i):i'\neq i}$. Also, since Y^* is bounded,

$$\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \max_k \left| n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell}-Y^*_{i\ell})\right| = \frac{n-3}{n}\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \max_k \left|(n-3)^{-1}\sum_{\ell\neq i,i',k} Y^*_{k\ell}(Y^*_{i'\ell}-Y^*_{i\ell})\right| + O(n^{-1}).$$

First, consider a case when $k\neq i,i'$. Then, using (A.29), we conclude

$$\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \max_{k\neq i,i'} \left| n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell}-Y^*_{i\ell})\right| = \frac{n-3}{n}\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \max_{k\neq i,i'} \left|(n-3)^{-1}\sum_{\ell\neq i,i',k} Y_{k\ell}(Y_{i'\ell}-Y_{i\ell})\right| \tag{A.30}$$

$$+ O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right). \tag{A.31}$$

Using the definition of $\hat{d}^2_\infty(i,i')$, we obtain

$$\begin{aligned}\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \max_{k\neq i,i'} \left| n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell}-Y^*_{i\ell})\right| &= \frac{n-3}{n}\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i,i') + O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right)\\ &= O_p\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{2d_\xi}}\right)\end{aligned} \tag{A.32}$$

where the second equality follows from (A.28).

Then we consider a case when $k=i$. Then, as argued at the beginning of the proof of Part (ii), for any pair agents i and i', there exist some $\tilde{k}\in\mathcal{B}_i(\delta_{n,C})$ (other than i and i') with probability approaching one. Then, since g is Lipschitz and Y^* is bounded, we have (using (A.31))

$$\begin{aligned}\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \left| n^{-1}\sum_\ell Y^*_{i\ell}(Y^*_{i'\ell}-Y^*_{i\ell})\right| &= \frac{n-3}{n}\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \left|(n-3)^{-1}\sum_{\ell\neq i,i',k} Y_{\tilde{k}\ell}(Y_{i'\ell}-Y_{i\ell})\right|\\ &\quad + O(\delta_{n,C}) + O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).\end{aligned}$$

Since $\tilde{k}\neq i,i'$,

$$\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \left|(n-3)^{-1}\sum_{\ell\neq i,i',k} Y_{\tilde{k}\ell}(Y_{i'\ell}-Y_{i\ell})\right| \leqslant \max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \hat{d}^2_\infty(i,i'),$$

so

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \left| n^{-1} \sum_\ell Y_{i\ell}^* (Y_{i'\ell}^* - Y_{i\ell}^*) \right| \leqslant \frac{n-3}{n} \max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \hat{d}_\infty^2(i,i') + O(\delta_{n,C}) + O_p\left(\left(\frac{\ln n}{n} \right)^{1/2} \right).$$

Invoking (A.28) again delivers

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \left| n^{-1} \sum_\ell Y_{i\ell}^* (Y_{i'\ell}^* - Y_{i\ell}^*) \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right). \tag{A.33}$$

By the same argument, when $k = i'$, we have

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \left| n^{-1} \sum_\ell Y_{i'\ell}^* (Y_{i'\ell}^* - Y_{i\ell}^*) \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right). \tag{A.34}$$

Finally, (A.32), (A.33) and (A.34) together deliver the result. Q.E.D.

Proof of Theorem 2. Proof of Part (i).

$$\begin{aligned}
n^{-1} \sum_j \left(\hat{Y}_{ij}^* - Y_{ij}^* \right)^2 &= n^{-1} \sum_j \left(\frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y_{i'j} - Y_{ij}^*)}{n_i} \right)^2 \\
&\leqslant n^{-1} \sum_j \left(2 \left(\frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y_{i'j} - Y_{i'j}^*)}{n_i} \right)^2 + 2 \left(\frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y_{i'j}^* - Y_{ij}^*)}{n_i} \right)^2 \right) \\
&= \frac{2}{n} \sum_j \left(\frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y_{i'j} - Y_{i'j}^*)}{n_i} \right)^2 + \frac{2}{n} \sum_j \left(\frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y_{i'j}^* - Y_{ij}^*)}{n_i} \right)^2 \\
&= 2\mathcal{S}_i + 2\mathcal{J}_i.
\end{aligned}$$

We start with the first term

$$\mathcal{S}_i := \frac{1}{n} \sum_j \left(\frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y_{i'j} - Y_{i'j}^*)}{n_i} \right)^2.$$

Note that $Y_{i'j} - Y^*_{i'j} = \varepsilon_{i'j}$, where $\varepsilon_{i'j} = -Y^*_{i'i'}$ when $i' = j$. Then S_{1i} can be decomposed as follows

$$\begin{aligned}\mathcal{S}_i &= \frac{1}{nn_i^2}\sum_j \left(\sum_{i'\in\hat{\mathcal{N}}_i(n_i)} \varepsilon_{i'j}^2 + \sum_{i'\in\hat{\mathcal{N}}_i(n_i)} \sum_{i''\in\hat{\mathcal{N}}_i(n_i), i''\neq i'} \varepsilon_{i'j}\varepsilon_{i''j} \right) \\ &= \underbrace{\frac{1}{n_i^2} \sum_{i'\in\hat{\mathcal{N}}_i(n_i)} \frac{1}{n}\sum_j \varepsilon_{i'j}^2}_{\mathcal{S}_{1i}} + \underbrace{\frac{1}{n_i^2} \sum_{i'\in\hat{\mathcal{N}}_i(n_i)} \sum_{i''\in\hat{\mathcal{N}}_i(n_i), i''\neq i'} \frac{1}{n}\sum_j \varepsilon_{i'j}\varepsilon_{i''j}}_{\mathcal{S}_{2i}}.\end{aligned}$$

First,

$$\frac{1}{n}\sum_j \varepsilon_{i'j}^2 = \frac{1}{n}\sum_{j\neq i'} \varepsilon_{i'j}^2 - \frac{Y^*_{i'i'}}{n}.$$

Let $\sigma_i^2 \coloneqq \mathbb{E}\left[\varepsilon_{ij}^2 | Z_i\right]$. Note that

$$\max_{i'} \left| \frac{1}{n-1}\sum_{j\neq i'} \varepsilon_{i'j}^2 - \sigma_{i'}^2 \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{1/2} \right).$$

Again, this directly follows from an application of Corollary A.1 and the union bound. Corrolary A.1 applies since, conditionally on $Z_{i'}$, $\{\varepsilon_{i'j}^2\}_{j\neq i'}$ is a collection of iid random variables, which are uniformly (uniformly over $Z_{i'}$) sub-Gaussian. Assumption 3 (iv) also guarantees that $\mathbb{E}\left[\varepsilon_{ij}^2 | Z_i, Z_j\right] \leqslant C$ a.s. for some $C > 0$. Hence, we also have $\sigma_{i'}^2 \leqslant C$. Finally, since Y^* is bounded, we conclude

$$\max_{i'} \frac{1}{n}\sum_j \varepsilon_{i'j}^2 \leqslant C + O_p\left(\left(\frac{\ln n}{n} \right)^{1/2} \right),$$

and, consequently,

$$\max_i \mathcal{S}_{1i} \leqslant \frac{C}{n_i} + o_p(n_i^{-1}). \tag{A.35}$$

As for $\mathcal{S}_{2i}$,

$$\frac{1}{n}\sum_j \varepsilon_{i'j}\varepsilon_{i''j} = \frac{1}{n}\sum_{j\neq i', i''} \varepsilon_{i'j}\varepsilon_{i''j} - \frac{1}{n}(Y^*_{i'i'}\varepsilon_{i''i'} + \varepsilon_{i'i''}Y^*_{i''i''}).$$

First, as usual, applying Bernstein inequality A.2 and the union bound gives

$$\max_{i' \neq i''} \left| \frac{1}{n-2} \sum_{j \neq i', i''} \varepsilon_{i'j} \varepsilon_{i''j} \right| = O_p \left(\left(\frac{\ln n}{n} \right)^{1/2} \right).$$

Indeed, conditional on $Z_{i'}$ and $Z_{i''}$, $\{\varepsilon_{i'j}\varepsilon_{i''j}\}_{j \neq i', i''}$ is a collection of mean zero iid random variables, which satisfy the requirements of Theorem A.2. Again, since $\epsilon_{i'i''}$ is (uniformly) sub-Gaussian, we have

$$\max_{i' \neq i''} \left| \frac{1}{n} \varepsilon_{i'i''} \right| = O_p \left(n^{-1} (\ln n)^{1/2} \right).$$

Finally, since Y^* is bounded, we conclude

$$\max_{i' \neq i''} \left| \frac{1}{n} \sum_{j} \varepsilon_{i'j} \varepsilon_{i''j} \right| = O_p \left(\left(\frac{\ln n}{n} \right)^{1/2} \right),$$

and, consequently,

$$\max_{i} |\mathcal{S}_{2i}| = O_p \left(\left(\frac{\ln n}{n} \right)^{1/2} \right). \tag{A.36}$$

Combining (A.35) and (A.36) gives

$$\begin{aligned} \max_i \mathcal{S}_i &\leqslant \frac{C}{\min_i n_i} + o_p(n_i^{-1}) + O_p \left(\left(\frac{\ln n}{n} \right)^{1/2} \right) \\ &= O_p \left(\left(\frac{\ln n}{n} \right)^{1/2} \right). \end{aligned} \tag{A.37}$$

As for the second term $\mathcal{J}_i$,

$$\begin{aligned} \mathcal{J}_i \coloneqq \frac{1}{n} \sum_j \left(\frac{\sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y^*_{i'j} - Y^*_{ij})}{n_i} \right)^2 &\leqslant \frac{1}{n} \sum_j \frac{1}{n_i} \sum_{i' \in \hat{\mathcal{N}}_i(n_i)} (Y^*_{i'j} - Y^*_{ij})^2 \\ &= \frac{1}{n_i} \sum_{i' \in \hat{\mathcal{N}}_i(n_i)} \frac{1}{n} \sum_j (Y^*_{i'j} - Y^*_{ij})^2. \end{aligned}$$

Combining this with Part (i) of Lemma A.6, we obtain

$$\max_i \mathcal{J}_i = O_p \left((n^{-1} \ln n)^{\frac{1}{2d\xi}} \right). \tag{A.38}$$

Then, combining (A.37) and (A.38), ensures that

$$\max_i n^{-1}\sum_j (\hat{Y}^*_{ij} - Y^*_{ij})^2 \leqslant 2\max_i \mathcal{S}_i + 2\max_i \mathcal{J}_i = O_p\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{2d_\xi}}\right),$$

which completes the proof of the first part.

Proof of Part (ii). First,

$$\begin{aligned} n^{-1}\sum_\ell Y^*_{k\ell}(\hat{Y}^*_{i\ell} - Y^*_{i\ell}) &= n_i^{-1}\sum_{i'\in\hat{\mathcal{N}}_i(n_i)} n^{-1}\sum_\ell Y^*_{k\ell}(Y_{i'\ell} - Y^*_{i\ell}) \\ &= n_i^{-1}\sum_{i'\in\hat{\mathcal{N}}_i(n_i)}\left(n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell} - Y^*_{i\ell}) + n^{-1}\sum_\ell Y^*_{k\ell}\varepsilon_{i'\ell}\right). \end{aligned}$$

Hence,

$$\max_k\max_i\left|n^{-1}\sum_\ell Y^*_{k\ell}(\hat{Y}^*_{i\ell} - Y^*_{i\ell})\right| \leqslant \max_i\max_{i'\in\hat{\mathcal{N}}_i(n_i)}\max_k\left|n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell} - Y^*_{i\ell})\right| + \max_k\max_{i'}\left|n^{-1}\sum_\ell Y^*_{k\ell}\varepsilon_{i'\ell}\right|.$$

Applying Part (ii) of Lemma A.6, we bound the first term

$$\max_i\max_{i'\in\hat{\mathcal{N}}_i(n_i)}\max_k\left|n^{-1}\sum_\ell Y^*_{k\ell}(Y^*_{i'\ell} - Y^*_{i\ell})\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{2d_\xi}}\right). \tag{A.39}$$

Next we want to bound the second term, which can be represented as

$$n^{-1}\sum_\ell Y^*_{k\ell}\varepsilon_{i'\ell} = n^{-1}\sum_{\ell\neq k,i'} Y^*_{k\ell}\varepsilon_{i'\ell} + n^{-1}(Y^*_{kk}\varepsilon_{i'k} + Y^*_{ki'}\varepsilon_{i'i'}).$$

First, by the standard arguments (again, see, for example, the proof of Lemma 2), we have

$$\max_k\max_{i'}\left|n^{-1}\sum_{\ell\neq k,i'} Y^*_{k\ell}\varepsilon_{i'\ell}\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

Also, recall $\varepsilon_{ij} = -Y^*_{ii}$ when $i = j$. Since Y^* is bounded, we have $\max_k \max_{i'} |n^{-1}Y^*_{ki'}\varepsilon_{i'i'}| \leqslant C/n$ for some C. At the same time, since either $\varepsilon_{i'k}$ is (uniformly) sub-Gaussian (for $i' \neq k$) or is equal to $-Y^*_{kk}$ (which is bounded), we have (again, applying the union bound)

$$\max_k\max_{i'}\left|n^{-1}Y^*_{kk}\varepsilon_{i'k}\right| = O_p\left(n^{-1}(\ln n)^{1/2}\right).$$

Combining these results, we conclude

$$\max_k \max_{i'} \left| n^{-1} \sum_\ell Y_{k\ell}^* \varepsilon_{i'\ell} \right| = O_p \left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right). \tag{A.40}$$

Finally, (A.39) and (A.40) together deliver the result.

Q.E.D.

A.3.2 Proof of Theorem 3

First, we prove the following auxiliary lemma.

Lemma A.7. *Suppose that $\mathcal{B} = \mathbb{R}^p$. Then, under Assumption 1, 3, 8 (i), we have*

$$\max_{i \neq j} \left| d_{ij,n}^2 - d_{ij}^2 \right| = O_p \left(\left(\frac{\ln n}{n} \right)^{1/2} \right),$$

where

$$d_{ij,n}^2 := \min_{\beta \in \mathbb{R}_p} d_{ij,n}^2(\beta), \tag{A.41}$$

and $d_{ij,n}^2(\beta)$ is given by (A.21).

Proof of Lemma A.7. First, we argue that there exists some compact $\overline{\mathcal{B}}$ such that with probability approaching one, for all pairs of agents

$$d_{ij,n}^2 := \min_{\beta \in \mathbb{R}^p} d_{ij,n}^2(\beta) = \min_{\beta \in \overline{\mathcal{B}}} d_{ij,n}^2(\beta).$$

For all $x, \tilde{x} \in \text{supp}\,(X)$, let $\Delta w(X_k; x, \tilde{x}) := w(x, X_k) - w(\tilde{x}, X_k)$ and

$$\begin{aligned} \hat{\mathcal{C}}(x, \tilde{x}) &:= \frac{1}{n} \sum_{k=1}^n \Delta w(X_k; x, \tilde{x}) \Delta w(X_k; x, \tilde{x})' \\ &= \sum_{r=1}^R \hat{p}_r \Delta w(x_r; x, \tilde{x}) \Delta w(x_r; x, \tilde{x})' \\ &= \sum_{r=1}^R \hat{p}_r H(x_r; x, \tilde{x})', \end{aligned}$$

where $\hat{p}_r = n^{-1} \sum_{k=1}^n \mathbb{1}\{X_k = x_r\}$ and $H(x_r; x, \tilde{x}) := \Delta w(x_r; x, \tilde{x}) \Delta w(x_r; x, \tilde{x})'$.

Notice that since $\text{supp}\,(X) = \{x_1, \ldots, X_R\}$, there exists $C_\lambda > 0$ such that the minimal non-zero eigenvalue of $H(x_r; x, \tilde{x})$ is greater than C_λ uniformly over all $x_r, x, \tilde{x} \in \{x_1, \ldots, x_R\}$. Formally,

we have

$$\lambda_{min,+}(H(x_r;x,\tilde{x})) > C_\lambda \tag{A.42}$$

for all $x_r, x, \tilde{x} \in \{x_1, \ldots, x_R\}$ such that $H(x_r;x,\tilde{x})$ is non-zero, where $\lambda_{min,+}(H)$ denotes the minimal non-negative eigenvalue of a positive semidefinite matrix H.

Next, notice that there exists $C_p > 0$ such that with probability approaching one, we have $\min_r \hat{p}_r > C_p$. Combining this with (A.42), we conclude that for some $C_{\mathcal{C}} > 0$ we have, with probability approaching one,

$$\lambda_{min,+}\left(\hat{\mathcal{C}}(x,\tilde{x})\right) > C_{\mathcal{C}}$$

for all $x, \tilde{x} \in \{x_1, \ldots, x_R\}$ such that $\hat{\mathcal{C}}(x,\tilde{x})$ is non-zero.

For all $x, \tilde{x} \in \{x_1, \ldots, x_R\}$, let $\hat{O}(x,\tilde{x})$ be an orthogonal matrix, which diagonalizes $\hat{\mathcal{C}}(x,\tilde{x})$, i.e., $\hat{O}(x,\tilde{x})\hat{\mathcal{C}}(x,\tilde{x})\hat{O}(x,\tilde{x})' = \hat{\Lambda}(x,\tilde{x})$, where $\hat{\Lambda}(x,\tilde{x})$ is diagonal and its non-zero elements are bounded away from zero by $C_{\mathcal{C}}$ with probability approaching (uniformly over $x, \tilde{x} \in \{x_1, \ldots, x_R\}$).

Take any pair of agents i and j. Then

$$\begin{aligned}
d_{ij,n}^2 &= \min_{\beta \in \mathbb{R}_p} \frac{1}{n} \sum_k (Y_{ik}^* - Y_{jk}^* - (w(X_i,X_k) - w(X_j,X_k))'\beta)^2 \\
&= \min_{\beta \in \mathbb{R}_p} \frac{1}{n} \sum_k (Y_{ik}^* - Y_{jk}^* - \Delta w(X_k;X_i,X_j)'\hat{O}(X_i,X_j)'\hat{O}(X_i,X_j)\beta)^2 \\
&= \min_{b \in \mathbb{R}_p} \frac{1}{n} \sum_k (Y_{ik}^* - Y_{jk}^* - \Delta w(X_k;X_i,X_j)'\hat{O}(X_i,X_j)'b)^2,
\end{aligned}$$

where the last equality follows from the change of the variable $b = \hat{O}(X_i,X_j)\beta$. The minimum is achieved at

$$\begin{aligned}
\hat{b}_{ij} &= \left(\hat{O}(X_i,X_j) n^{-1} \sum_k \Delta w(X_k;X_i,X_j)\Delta w(X_k;X_i,X_j)'\hat{O}(X_i,X_j)'\right)^+ \\
&\quad \times \left(\hat{O}(X_i,X_j) n^{-1} \sum_k \Delta w(X_k;X_i,X_j)(Y_{ik}^* - Y_{jk}^*)\right) \\
&= \left(\hat{O}(X_i,X_j)\hat{\mathcal{C}}(X_i,X_j)\hat{O}(X_i,X_j)'\right)^+ \left(\hat{O}(X_i,X_j) n^{-1} \sum_k \Delta w(X_k;X_i,X_j)(Y_{ik}^* - Y_{jk}^*)\right) \\
&= \left(\hat{\Lambda}(X_i,X_j)\right)^+ \left(\hat{O}(X_i,X_j) n^{-1} \sum_k \Delta w(X_k;X_i,X_j)(Y_{ik}^* - Y_{jk}^*)\right),
\end{aligned}$$

where $+$ denotes the Moore–Penrose inverse. Notice that since the non-zero elements of diagonal matrix $\hat{\Lambda}(X_i, X_j)$ are bounded away from zero (uniformly in X_i and X_j) with probability approaching one, we conclude that $\max_{i\neq j} \lambda_{max}\left(\left(\hat{\Lambda}(X_i, X_j)\right)^+\right) < \overline{C}_\Lambda$ with probability approaching one. Since, for some $C > 0$, we also have $\max_{i\neq j}\left\|\hat{O}(X_i, X_j)n^{-1}\sum_k w(X_k; X_i, X_j)(Y^*_{ik} - Y^*_{jk})\right\| < C$ (w and Y^* are bounded), we conclude that for some $C_\beta > 0$, we have with probability approaching one

$$\max_{i\neq j}\left\|\hat{b}_{ij}\right\| \leqslant C_\beta.$$

Recall that $b = \hat{O}(X_i, X_j)\beta$. Hence, the minimum of $d^2_{ij,n}(\beta)$ is achieved at $\hat{\beta}_{ij} = \hat{O}(X_i, X_j)'\hat{b}_{ij}$, i.e., $d^2_{ij,n} = d^2_{ij,n}(\hat{\beta}_{ij})$. Also notice that since $\hat{O}(X_i, X_j)$ is orthogonal, with probability approaching one we also have

$$\max_{i\neq j}\left\|\hat{\beta}_{ij}\right\| \leqslant C_\beta.$$

Then, if $\overline{\mathcal{B}}$ includes all β such that $\|\beta\| \leqslant C_\beta$, we have

$$d^2_{ij,n} = d^2_{ij,n}(\hat{\beta}_{ij}) = \min_{\beta\in\overline{\mathcal{B}}} d^2_{ij,n}(\beta). \tag{A.43}$$

Hence, such a compact set $\overline{\mathcal{B}}$ clearly exists, which completes the proof of the first part.

Secondly, note that by the same argument,

$$d^2_{ij} = \min_{\beta\in\mathbb{R}_p} d^2_{ij}(\beta) = \min_{\beta\in\overline{\mathcal{B}}} d^2_{ij}(\beta), \tag{A.44}$$

where, without loss of generality, $\overline{\mathcal{B}}$ can be taken the same as before.

Third, since $d^2_{ij,n} = \min_{\beta\in\overline{\mathcal{B}}} d^2_{ij,n}(\beta)$, where $\overline{\mathcal{B}}$ is compact, we can apply the result of Lemma A.4, which guarantees that $\max_{i\neq j}\sup_{\beta\in\overline{\mathcal{B}}}\left|d^2_{ij,n}(\beta) - d^2_{ij}(\beta)\right| = O_p\left((n^{-1}\ln n)^{1/2}\right)$. This necessarily implies that

$$\max_{i\neq j}\left|\min_{\beta\in\overline{\mathcal{B}}} d^2_{ij,n}(\beta) - \min_{\beta\in\overline{\mathcal{B}}} d^2_{ij}(\beta)\right| = O_p\left((n^{-1}\ln n)^{1/2}\right).$$

Combining this with (A.43) and (A.44) completes the proof.

Q.E.D.

Proof of Theorem 3. First, denote

$$\mathbf{W}_{i-j} = \begin{pmatrix} (W_{i1} - W_{j1})' \\ (W_{i2} - W_{j2})' \\ \cdots \\ (W_{in} - W_{jn})' \end{pmatrix}, \quad \mathbf{Y}^*_{i-j} = \begin{pmatrix} Y^*_{i1} - Y^*_{j1} \\ Y^*_{i2} - Y^*_{j2} \\ \cdots \\ Y^*_{in} - Y^*_{jn} \end{pmatrix}, \quad \hat{\mathbf{Y}}^*_{i-j} = \begin{pmatrix} \hat{Y}^*_{i1} - \hat{Y}^*_{j1} \\ \hat{Y}^*_{i2} - \hat{Y}^*_{j2} \\ \cdots \\ \hat{Y}^*_{in} - \hat{Y}^*_{jn} \end{pmatrix}.$$

Then

$$\begin{aligned} \hat{d}^2_{ij} &= n^{-1} \hat{\mathbf{Y}}^{*\prime}_{i-j} \mathbf{P}_{i-j} \hat{\mathbf{Y}}^*_{i-j} \\ &= n^{-1} \left(\mathbf{Y}^{*\prime}_{i-j} \mathbf{P}_{i-j} \mathbf{Y}^*_{i-j} + \Delta\hat{\mathbf{Y}}^{*\prime}_{i-j} \mathbf{P}_{i-j} \mathbf{Y}^*_{i-j} + \mathbf{Y}^{*\prime}_{i-j} \mathbf{P}_{i-j} \Delta\hat{\mathbf{Y}}^*_{i-j} + \Delta\hat{\mathbf{Y}}^{*\prime}_{i-j} \mathbf{P}_{i-j} \Delta\hat{\mathbf{Y}}^*_{i-j} \right), \end{aligned}$$

where $\Delta\hat{\mathbf{Y}}^*_{i-j} = \hat{\mathbf{Y}}^*_{i-j} - \mathbf{Y}^*_{i-j}$ and $\mathbf{P}_{i-j} = \mathbf{I}_n - \mathbf{W}_{i-j}(\mathbf{W}'_{i-j}\mathbf{W}_{i-j})^+ \mathbf{W}'_{i-j}$, where $+$ stands for the generalized inverse. Note that $n^{-1}\mathbf{Y}^{*\prime}_{i-j}\mathbf{P}_{i-j}\mathbf{Y}^*_{i-j} = d^2_{ij,n}$, where $d^2_{ij,n}$ is as defined in (A.41). Lemma A.7 ensures $\max_{i\neq j} \left| d^2_{ij,n} - d^2_{ij} \right| = O_p((n^{-1}\ln n)^{1/2})$, so we have

$$\max_{i\neq j} \left| n^{-1} \mathbf{Y}^{*\prime}_{i-j} \mathbf{P}_{i-j} \mathbf{Y}^*_{i-j} - d^2_{ij} \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{1/2} \right). \tag{A.45}$$

Then, to complete the proof, it is sufficient to show that

$$\max_{i\neq j} \left| n^{-1} \Delta\hat{\mathbf{Y}}^{*\prime}_{i-j} \mathbf{P}_{i-j} \mathbf{Y}^*_{i-j} \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right), \tag{A.46}$$

$$\max_{i\neq j} \left| n^{-1} \Delta\hat{\mathbf{Y}}^{*\prime}_{i-j} \mathbf{P}_{i-j} \Delta\hat{\mathbf{Y}}^*_{i-j} \right| = O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right). \tag{A.47}$$

First,

$$\begin{aligned} \max_{i\neq j} \left| n^{-1} \Delta\hat{\mathbf{Y}}^{*\prime}_{i-j} \mathbf{P}_{i-j} \mathbf{Y}^*_{i-j} \right| &\leqslant \max_{i\neq j} \left| n^{-1} \Delta\hat{\mathbf{Y}}^{*\prime}_{i-j} \mathbf{Y}^*_{i-j} \right| \\ &= \max_{i\neq j} \left| n^{-1} \sum_\ell (Y^*_{i\ell} - Y^*_{j\ell})(\Delta\hat{Y}^*_{i\ell} - \Delta\hat{Y}^*_{j\ell}) \right| \\ &\leqslant 4 \max_k \max_i \left| n^{-1} \sum_\ell Y^*_{k\ell} (\hat{Y}^*_{i\ell} - Y^*_{i\ell}) \right| \\ &= O_p\left(\left(\frac{\ln n}{n} \right)^{\frac{1}{2d_\xi}} \right), \end{aligned}$$

where the last equality follows from Theorem 2 Part (ii). Similarly,

$$\begin{aligned}\max_{i\neq j}\left|n^{-1}\Delta\hat{\mathbf{Y}}_{i-j}^{*\prime}\mathbf{P}_{i-j}\Delta\hat{\mathbf{Y}}_{i-j}^{*}\right| &\leqslant \max_{i\neq j}\left|n^{-1}\Delta\hat{\mathbf{Y}}_{i-j}^{*\prime}\Delta\hat{\mathbf{Y}}_{i-j}^{*}\right| \\ &= \max_{i\neq j}\left|n^{-1}\sum_{\ell}(\Delta\hat{Y}_{i\ell}^{*}-\Delta\hat{Y}_{j\ell}^{*})^2\right| \\ &\leqslant 4\max_{i} n^{-1}\sum_{\ell}(\hat{Y}_{i\ell}^{*}-Y_{i\ell}^{*})^2 \\ &= 4n^{-1}\left\|\hat{Y}^{*}-Y^{*}\right\|_{2,\infty}^{2} = O_p\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{2d_\xi}}\right),\end{aligned}$$

where the last equality is due Theorem 2 Part (i). Combining (A.45)-(A.47) delivers the result. Q.E.D.

A.3.3 Proof of Lemma 3

Proof of Lemma 3. First, define

$$\begin{aligned}\hat{d}_{ij}^2(\beta) &\coloneqq \frac{1}{n}\sum_{k}(\hat{Y}_{ik}^{*}-\hat{Y}_{jk}^{*}-(W_{ik}-W_{jk})\beta)^2, \\ d_{ij}^2(\beta) &\coloneqq \mathbb{E}\left[(Y_{ik}^{*}-Y_{jk}^{*}-(W_{ik}-W_{jk})\beta)^2|X_i,\xi_i,X_j,\xi_j\right],\end{aligned}$$

so $\hat{d}_{ij}^2=\min_{\beta\in\mathcal{B}}\hat{d}_{ij}^2(\beta)$ and $d_{ij}^2=\min_{\beta\in\mathcal{B}}d_{ij}^2(\beta)$. Note that

$$\begin{aligned}\max_{i\neq j}\left|\hat{d}_{ij}^2-d_{ij}^2\right| &= \max_{i\neq j}\left|\min_{\beta\in\mathcal{B}}\hat{d}_{ij}^2(\beta)-\min_{\beta\in\mathcal{B}}d_{ij}^2(\beta)\right| \\ &\leqslant 2\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\hat{d}_{ij}^2(\beta)-d_{ij}^2(\beta)\right|.\end{aligned}$$

Hence, it is sufficient to show that

$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\hat{d}_{ij}^2(\beta)-d_{ij}^2(\beta)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}+\mathcal{R}_n^{-1/2}\right). \tag{A.48}$$

Define $\Delta\hat{Y}^*_{ik} \coloneqq \hat{Y}^*_{ik} - Y^*_{ik}$. We decompose $\hat{d}^2_{ij}(\beta)$ as follows

$$\begin{aligned}
\hat{d}^2_{ij}(\beta) =& \frac{1}{n}\sum_k (\hat{Y}^*_{ik} - \hat{Y}^*_{jk} - (W_{ik} - W_{jk})\beta)^2 \\
=& \frac{1}{n}\sum_k (Y^*_{ik} - Y^*_{jk} - (W_{ik} - W_{jk}) + \Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk}\beta)^2 \\
=& \frac{1}{n}\sum_k \left(Y^*_{ik} - Y^*_{jk} - (W_{ik} - W_{jk})'\beta\right)^2 \\
&+ \frac{2}{n}\sum_k \left(Y^*_{ik} - Y^*_{jk}\right)\left(\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk}\right) \\
&+ \frac{2}{n}\sum_k (W_{ik} - W_{jk})'\beta(\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk}) \\
&+ \frac{1}{n}\sum_k (\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk})^2.
\end{aligned}$$

By Lemma A.4, we have

$$\max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|\frac{1}{n}\sum_k \left(Y^*_{ik} - Y^*_{jk} - (W_{ik} - W_{jk})'\beta\right)^2 - d^2_{ij}(\beta)\right| = \max_{i\neq j}\max_{\beta\in\mathcal{B}}\left|d^2_{ij,n}(\beta) - d^2_{ij}(\beta)\right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

By the Cauchy-Schwartz inequality,

$$\max_{i\neq j}\frac{1}{n}\sum_k (\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk})^2 \leqslant 4\max_i \frac{1}{n}\sum_i (\hat{Y}^*_{ik} - Y^*_{ik})^2 = O_p(\mathcal{R}_n^{-1}).$$

Also,

$$\begin{aligned}
\max_{i\neq j}\left|\frac{1}{n}\sum_k (Y^*_{ik} - Y^*_{jk})(\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk})\right| &\leqslant \left(\max_{i\neq j}\frac{1}{n}\sum_k (Y^*_{ik} - Y^*_{jk})^2\right)^{1/2}\left(\max_{i\neq j}\frac{1}{n}\sum_k (\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk})^2\right)^{1/2} \\
&= O_p(\mathcal{R}_n^{-1/2}),
\end{aligned}$$

where the inequality is due the Cauchy-Schwartz inequality, and the equality exploits the fact that $\left(\max_{i\neq j}\frac{1}{n}\sum_k (Y^*_{ik} - Y^*_{jk})^2\right)^{1/2}$ is bounded (since Y^* is bounded). Similarly,

$$\begin{aligned}
\max_{i\neq j}\min_{\beta\in\mathcal{B}}\left|\frac{1}{n}\sum_k (W_{ik} - W_{jk})'\beta(\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk})\right| &\leqslant C\max_{i\neq j}\left\|\frac{1}{n}\sum_k (W_{ik} - W_{jk})(\Delta\hat{Y}^*_{ik} - \Delta\hat{Y}^*_{jk})\right\| \\
&= O_p(\mathcal{R}_n^{-1/2}),
\end{aligned}$$

where the inequality is guaranteed by the fact that $\mathcal{B}$ is bounded. Q.E.D.

A.4 Proof of Theorem 4

First, we prove the following auxiliary lemma.

Lemma A.8. *Suppose that the hypotheses of Theorem 2 are satisfied. Then,*

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} \|g(\xi_i, \cdot) - g(\xi_{i'}, \cdot)\|_2^2 = O_p\left((n^{-1} \ln n)^{\frac{1}{2d_\xi}}\right),$$

where

$$\|g(\xi_i, \cdot) - g(\xi_{i'}, \cdot)\|_2^2 = \int (g(\xi_i, \xi) - g(\xi_{i'}, \xi))^2 P_\xi(d\xi).$$

Proof of Lemma A.8. Lemma A.6 Part (i) guarantees that

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1} \sum_\ell (Y^*_{i\ell} - Y^*_{i'\ell})^2 = O_p\left((n^{-1} \ln n)^{\frac{1}{2d_\xi}}\right).$$

Notice that

$$Y^*_{i\ell} - Y^*_{i'\ell} = w(X_i, X_\ell) + g(\xi_i, \xi_\ell) - w(X_{i'}, X_\ell) - g(\xi_{i'}, \xi_\ell) = g(\xi_i, \xi_\ell) - g(\xi_{i'}, \xi_\ell)$$

since for $i' \in \hat{\mathcal{N}}_i(n_i)$ we have $X_{i'} = X_i$. Hence, we also have

$$\max_i \max_{i' \in \hat{\mathcal{N}}_i(n_i)} n^{-1} \sum_\ell (g(\xi_i, \xi_\ell) - g(\xi_{i'}, \xi_\ell))^2 = O_p\left((n^{-1} \ln n)^{\frac{1}{2d_\xi}}\right). \tag{A.49}$$

The next step is to note that

$$\max_{i,i'} \left| n^{-1} \sum_\ell (g(\xi_i, \xi_\ell) - g(\xi_{i'}, \xi_\ell))^2 - \|g(\xi_i, \cdot) - g(\xi_{i'}, \cdot)\|_2^2 \right| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right). \tag{A.50}$$

Indeed, treating ξ_i and $\xi_{i'}$ as non-random, we have $\mathbb{E}\left[(g(\xi_i, \xi_\ell) - g(\xi_{i'}, \xi_\ell))^2\right] = \|g(\xi_i, \cdot) - g(\xi_{i'}, \cdot)\|_2^2$. Again, the bound $O_p\left(\left(\frac{\ln n}{n}\right)^{1/2}\right)$ follows from the standard argument, which involves: (i) an application of Bernstein inequality A.1 for iid bounded random variables $\{(g(\xi_i, \xi_\ell) - g(\xi_{i'}, \xi_\ell))^2\}_\ell$ treating ξ_i and $\xi_{i'}$ as non-random; (ii) and an application of the union bound to obtain uniformity over all possible i and i'.

Finally, (A.49) and (A.50) together deliver the result. Q.E.D.

Proof of Theorem 4.

$$\begin{aligned}\tilde{Y}^*_{ij} &= \frac{1}{m_{ij}} \sum_{(i',j')\in\hat{\mathcal{M}}_{ij}} Y_{i'j'} \\ &= w(X_i, X_j)'\beta_0 + g(\xi_i,\xi_j) + \frac{1}{m_{ij}} \sum_{(i',j')\in\hat{\mathcal{M}}_{ij}} ((g(\xi_{i'},\xi_{j'}) - g(\xi_i,\xi_j)) + \varepsilon_{i'j'}) \\ &= Y^*_{ij} + \frac{1}{m_{ij}} \sum_{(i',j')\in\hat{\mathcal{M}}_{ij}} ((g(\xi_{i'},\xi_{j'}) - g(\xi_i,\xi_j)) + \varepsilon_{i'j'})\,.\end{aligned}$$

Then, it suffices to show that

$$\max_{i,j} \left| \frac{1}{m_{ij}} \sum_{(i',j')\in\hat{\mathcal{M}}_{ij}} (g(\xi_{i'},\xi_{j'}) - g(\xi_i,\xi_j)) \right| = o_p(1) \tag{A.51}$$

and

$$\max_{i,j} \left| \frac{1}{m_{ij}} \sum_{(i',j')\in\hat{\mathcal{M}}_{ij}} \varepsilon_{i'j'} \right| = o_p(1). \tag{A.52}$$

We start with establishing (A.51). First, note that for semiparametric model (**??**) the Assumption 2 implies that for any $\delta > 0$, there exists some $C_\delta > 0$ such that

$$\|g(\xi_i,\cdot) - g(\xi_{i'},\cdot)\|_2^2 > C_\delta$$

a.s. for $\|\xi_i - \xi_{i'}\| \geqslant \delta$. Combining this with the result of Lemma A.8, we establish

$$\max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \|\xi_i - \xi_{i'}\| = o_p(1).$$

Using Assumption 3 (iii),

$$|g(\xi_{i'},\xi_{j'}) - g(\xi_i,\xi_j)| \leqslant \overline{G}(\|\xi_{i'} - \xi_i\| + \|\xi_{j'} - \xi_j\|),$$

and, consequently,

$$\max_{i,j} \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \max_{j'\in\hat{\mathcal{N}}_j(n_j)} |g(\xi_{i'},\xi_{j'}) - g(\xi_i,\xi_j)| \leqslant 2\overline{G} \max_i \max_{i'\in\hat{\mathcal{N}}_i(n_i)} \|\xi_i - \xi_{i'}\| = o_p(1).$$

Consequently, we conclude

$$\max_{i,j}\left|\frac{1}{m_{ij}}\sum_{(i',j')\in\hat{\mathcal{M}}_{ij}}(g(\xi_{i'},\xi_{j'})-g(\xi_i,\xi_j))\right| \leqslant \max_{i,j}\max_{i'\in\hat{\mathcal{N}}_i(n_i)}\max_{j'\in\hat{\mathcal{N}}_j(n_j)}|g(\xi_{i'},\xi_{j'})-g(\xi_i,\xi_j)| = o_p(1),$$

so (A.51) holds.

To complete the proof we also need to demonstrate that (A.52) holds. First, note that

$$m_{ij} \geqslant \underline{m} := \min\{n_i,n_j\}(\min\{n_i,n_j\}-1)/2, \tag{A.53}$$

with the equality when $\hat{\mathcal{N}}_i(n_i)$ and $\hat{\mathcal{N}}_j(n_j)$ are the same sets. Since, conditional on $\{X_i,\xi_i\}_{i=1}^n$, $\{\varepsilon_{i'j'}\}$ are independent and sub-Gaussian (Assumption 3 (iv)), applying Bernstein inequality A.2, we conclude that there exist some positive constants C, a, and b such that for all $\epsilon > 0$ and for (almost) all $\{X_i,\xi_i\}_{i=1}^n$,

$$\mathbb{P}\left(\left|\frac{1}{m_{ij}}\sum_{(i',j')\in\hat{\mathcal{M}}_{ij}}\varepsilon_{i'j'}\right| > \epsilon|\{X_i,\xi_i\}_{i=1}^n\right) \leqslant C\exp\left(-\frac{m_{ij}\epsilon^2}{a+b\epsilon}\right),$$

and, consequently,

$$\mathbb{P}\left(\left|\frac{1}{m_{ij}}\sum_{(i',j')\in\hat{\mathcal{M}}_{ij}}\varepsilon_{i'j'}\right| > \epsilon\right) \leqslant C\exp\left(-\frac{m_{ij}\epsilon^2}{a+b\epsilon}\right) \leqslant C\exp\left(-\frac{\underline{m}\epsilon^2}{a+b\epsilon}\right).$$

To get uniformity over i and j, note that it is not more than $\binom{n}{n_i}$ and $\binom{n}{n_j}$ possible combinations to form $\hat{\mathcal{N}}_i(n_i)$ and $\hat{\mathcal{N}}_j(n_j)$. Recall that

$$\underline{C}(n\ln n)^{1/2} \leqslant n_i \leqslant \overline{C}(n\ln n)^{1/2}. \tag{A.54}$$

Since $\binom{n}{k}$ is monotone in k when $k \leqslant \lceil n/2\rceil$, we have $\binom{n}{n_i} \leqslant \binom{n}{\overline{n}}$, where $\overline{n} = \max n_i$. Consequently, it is not more than $\binom{n}{\overline{n}}^2$ possible combinations to form $\hat{\mathcal{M}}_{ij}$. Consequently, using the union bound,

$$\mathbb{P}\left(\max_{i,j}\left|\frac{1}{m_{ij}}\sum_{(i',j')\in\hat{\mathcal{M}}_{ij}}\varepsilon_{i'j'}\right| > \epsilon\right) \leqslant \binom{n}{\overline{n}}^2 C\exp\left(-\frac{\underline{m}\epsilon^2}{a+b\epsilon}\right). \tag{A.55}$$

Since $\binom{n}{\overline{n}} < (\frac{n\times e}{\overline{n}})^{\overline{n}}$, we have $\ln\left(\binom{n}{\overline{n}}\right) < \overline{n}\ln(n)$ for sufficiently large $\overline{n}$. Hence, using (A.54), we

obtain

$$\ln\left(\binom{n}{\overline{n}}\right) < \overline{C} n^{1/2} (\ln n)^{3/2}. \tag{A.56}$$

Using (A.54) again and recalling the definition of $\underline{m}$ in (A.53), we also conclude that there exists some $C_m > 0$ such that

$$\underline{m} \geqslant C_m n \ln n. \tag{A.57}$$

Finally, notice that (A.55), (A.56), and (A.57) together imply that for any fixed $\epsilon > 0$,

$$\mathbb{P}\left(\max_{i,j} \left| \frac{1}{m_{ij}} \sum_{(i',j') \in \hat{\mathcal{M}}_{ij}} \varepsilon_{i'j'} \right| > \epsilon \right) \to 0,$$

which implies that (A.52) holds.

Since we have demonstrated that both (A.51) and (A.52) hold, the proof is complete. Q.E.D.

A.5 Proofs of the results of Section 5.1

Proof of Lemma 4. Denote

$$d_{ij}^2(\mu; x, \tilde{x}) := \mathbb{E}\left[\left(Y_{ik}^* - Y_{jk}^* + \mu\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = x\right] + \mathbb{E}\left[\left(Y_{ik}^* - Y_{jk}^* - \mu\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = \tilde{x}\right],$$

and

$$\mu_{ij}^*(x, \tilde{x}) := \operatorname*{argmin}_{\mu} d_{ij}^2(\mu; x, \tilde{x}),$$

so $d_{ij}^2(x, \tilde{x}) = \min_\mu d_{ij}^2(\mu; x, \tilde{x}) = d_{ij}^2(\mu_{ij}^*(x, \tilde{x}); x, \tilde{x})$.

Notice that if $\xi_i = \xi_j$, then

$$d_{ij}^2(x, \tilde{x}) = \min_{\mu} \left((-h(x, \tilde{x}) + \mu)^2 + (h(x, \tilde{x}) - \mu)^2\right) = 0,$$

where the minimum is achieved at $\mu_{ij}^*(x, \tilde{x}) = h(x, \tilde{x})$.

Now suppose $d_{ij}^2(x, \tilde{x}) = 0$. First, we argue that this necessarily implies that $\mu_{ij}^*(x, \tilde{x}) = h(x, \tilde{x})$. The prove is by contradiction. Suppose that $d_{ij}^2(x, \tilde{x}) = d_{ij}^2(\mu_{ij}^*(x, \tilde{x}), x, \tilde{x}) = 0$ for some $\mu_{ij}^*(x, \tilde{x}) \neq$

$h(x, \tilde{x})$. Then,

$$\begin{aligned} 0 &= \mathbb{E}\left[\left(Y_{ik}^* - Y_{jk}^* + \mu_{ij}^*(x, \tilde{x})\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = x\right] \\ &= \mathbb{E}\left[\left(-h(x, \tilde{x}) + g(\xi_i, \xi_k) - g(\xi_j, \xi_k) + \mu_{ij}^*(x, \tilde{x})\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = x\right]. \end{aligned}$$

Since $\mathcal{E}_{x,\tilde{x}} \subseteq \text{supp}\,(\xi_k | X_k = x)$, we conclude that

$$g(\xi_i, \xi_k) - g(\xi_j, \xi_k) = h(x, \tilde{x}) - \mu_{ij}^*(x, \tilde{x}) \neq 0 \tag{A.58}$$

for all $\xi_k \in \text{supp}\,(\xi_k | X_k = x)$.

Now note that

$$\begin{aligned} &\mathbb{E}\left[\left(Y_{ik}^* - Y_{jk}^* - \mu_{ij}^*(x, \tilde{x})\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = \tilde{x}\right] \\ =&\mathbb{E}\left[\left(h(x, \tilde{x}) + g(\xi_i, \xi_k) - g(\xi_j, \xi_k) - \mu_{ij}^*(x, \tilde{x})\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = \tilde{x}\right] \\ \geqslant&\mathbb{P}(\xi_k \in \mathcal{E}_{x,\tilde{x}} | X_k = \tilde{x}) \mathbb{E}\left[\left(h(x, \tilde{x}) + g(\xi_i, \xi_k) - g(\xi_j, \xi_k) - \mu_{ij}^*(x, \tilde{x})\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = \tilde{x}, \xi_k \in \mathcal{E}_{x,\tilde{x}}\right] \\ =&4\mathbb{P}(\xi_k \in \mathcal{E}_{x,\tilde{x}} | X_k = \tilde{x})(h(x, \tilde{x}) - \mu_{ij}^*(x, \tilde{x}))^2 > 0, \end{aligned} \tag{A.59}$$

where the last equality follows from (A.58), and the last (strict) inequality follows from (A.58) and Assumption 9 (ii). Finally, notice that (A.59) implies that $d_{ij}^2(x, \tilde{x}) = d_{ij}^2(\mu_{ij}^*(x, \tilde{x}), x, \tilde{x}) > 0$, which contradicts the initial hypothesis.

Consequently, $d_{ij}^2(x, \tilde{x}) = 0$ necessarily implies that $\mu_{ij}^*(x, \tilde{x}) = h(x, \tilde{x})$. Hence,

$$\begin{aligned} 0 &= \mathbb{E}\left[\left(Y_{ik}^* - Y_{jk}^* + \mu_{ij}^*(x, \tilde{x})\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = x\right] \\ &= \mathbb{E}\left[\left(g(\xi_i, \xi_k) - g(\xi_j, \xi_k)\right)^2 | X_i, \xi_i, X_j, \xi_j, X_k = x\right] \\ &= \int \left(g(\xi_i, \xi_k) - g(\xi_j, \xi_k)\right)^2 P_{\xi|X}(d\xi_k; x). \end{aligned}$$

This, paired with Assumption 2, guarantees that $\xi_i = \xi_j$, which completes the proof. Q.E.D.

A.6 Bernstein inequalities

In this section we specify the Bernstein inequalities, which we refer to in the proofs.

A.6.1 Bernstein inequality for bounded random variables

Theorem A.1 (Bernstein inequality; see, for example, Bennett (1962))**.** *Let $Z_1, \ldots, Z_n$ be mean zero independent random variables. Assume there exists a positive constant M such that $|Z_i| \leqslant M$*

with probability one for each i. Also let $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i^2]$. Then for all $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{n\epsilon^2}{2\left(\sigma^2 + \frac{1}{3}M\epsilon\right)}\right).$$

A.6.2 Bernstein inequality for unbounded random variables

Lemma A.9 (Moments of a sub-Gaussian random variable)**.** *Let Z be a mean zero random variable satisfying $\mathbb{E}\left[e^{\lambda Z}\right] \leqslant e^{v\lambda^2}$ for all $\lambda \in \mathbb{R}$, for some $v > 0$. Then for every integer $q \geqslant 1$,*

$$\mathbb{E}\left[Z^{2q}\right] \leqslant q!(4v)^q.$$

Proof of Lemma A.9. See Theorem 2.1 in Boucheron, Lugosi, and Massart (2013). Q.E.D.

Theorem A.2 (Bernstein inequality for unbounded random variables)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables. Assume that there exist some positive constants ν and c such that $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i^2] \leqslant \nu$ such that for all integers $q \geqslant 3$*

$$\frac{1}{n}\sum \mathbb{E}\left[|Z_i|^q\right] \leqslant \frac{q!c^{q-2}}{2}\nu.$$

Then, for all $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{n\epsilon^2}{2(\nu + c\epsilon)}\right).$$

Proof of Theorem A.2. See Corollary 2.11 in Boucheron, Lugosi, and Massart (2013). Q.E.D.

Specifically, we make use of the following corollary.

Corollary A.1. *Let $Z_1, \ldots, Z_n$ be mean zero independent random variables. Assume that there exists some $v > 0$ such that $\mathbb{E}\left[e^{\lambda Z_i}\right] \leqslant e^{v\lambda^2}$ for all $\lambda \in \mathbb{R}$ and for all $i \in \{1, \ldots, n\}$. Then, there exist some positive constants C, a, and b such that for all constants $\alpha_1, \ldots, \alpha_n$ satisfying $\max_i |\alpha_i| < \overline{\alpha}$ and for all $\epsilon > 0$,*

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n \alpha_i Z_i\right| \geqslant \epsilon\right) \leqslant C\exp\left(-\frac{n\epsilon^2}{a + b\epsilon}\right)$$

and

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n \alpha_i \left(Z_i^2 - \mathbb{E}\left[Z_i^2\right]\right)\right| \geqslant \epsilon\right) \leqslant C\exp\left(-\frac{n\epsilon^2}{a + b\epsilon}\right).$$

Proof of Corollary A.1. Follows from Lemma A.9 and Theorem A.2. Q.E.D.

Remark A.1. Note that the constants C, a, and b depend on v and $\overline{\alpha}$ only.

B Illustration of Assumption 7

Suppose ξ is scalar and $g(\xi_i, \xi_k) = \kappa |\xi_i - \xi_k|$. Note that, as a function of ξ_i, $g(\xi_i, \xi_k)$ is non-differentiable at $\xi_i = \xi_k$. When $\xi_i, \xi_j \leqslant \xi_k$, we have

$$g(\xi_i, \xi_k) - g(\xi_j, \xi_k) = \kappa(\xi_k - \xi_i) - \kappa(\xi_k - \xi_j) = -\kappa(\xi_i - \xi_j).$$

If $\xi_i, \xi_j \geqslant \xi_k$,

$$g(\xi_i, \xi_k) - g(\xi_j, \xi_k) = \kappa(\xi_i - \xi_k) - \kappa(\xi_j - \xi_i) = \kappa(\xi_i - \xi_j).$$

So, we can take

$$G(\xi_i, \xi_k) = \begin{cases} -\kappa, & \xi_i < \xi_k \\ \kappa, & \xi_i \geqslant \xi_k \end{cases}.$$

Then, the remainder $r_g(\xi_i, \xi_j, \xi_k) = 0$ when $\xi_i, \xi_j \leqslant \xi_k$ or $\xi_i, \xi_j \geqslant \xi_k$. However, if, for example, $\xi_i \leqslant \xi_k \leqslant \xi_j$,

$$g(\xi_i, \xi_k) - g(\xi_j, \xi_k) = \kappa(2\xi_k - \xi_i - \xi_j).$$

Since $G(\xi_i, \xi_k) = -\kappa$,

$$g(\xi_i, \xi_k) - g(\xi_j, \xi_k) = -\kappa(\xi_i - \xi_j) + 2\kappa(\xi_k - \xi_j),$$

so $r_g(\xi_i, \xi_j, \xi_k) = 2\kappa(\xi_k - \xi_j)$ when $\xi_i \leqslant \xi_k \leqslant \xi_j$. Clearly, in this case, the linearization remainder is no longer $O(|\xi_i - \xi_j|^2)$. Assumption 7 allows for this possibility. The remainder r_g is bounded by $C\delta_n^2$ only when $|\xi_i - \xi_k| > \delta_n$ and $|\xi_j - \xi_i| \leqslant \delta_n$. Under these restrictions, ξ_k can not lie between ξ_i and ξ_j.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67, 251–333.

Ahn, H. and J. L. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29.

Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing (2008): "Mixed membership stochastic blockmodels," *Journal of machine learning research*, 9, 1981–2014.

Aldous, D. J. (1981): "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, 11, 581–598.

Amini, A. A., A. Chen, P. J. Bickel, E. Levina, et al. (2013): "Pseudo-likelihood methods for community detection in large sparse networks," *The Annals of Statistics*, 41, 2097–2122.

Amiti, M. and D. E. Weinstein (2018): "How much do idiosyncratic bank shocks affect investment? Evidence from matched bank-firm loan data," *Journal of Political Economy*, 126, 525–587.

Anderson, J. E. and E. Van Wincoop (2003): "Gravity with gravitas: A solution to the border puzzle," *American Economic Review*, 93, 170–192.

Arcones, M. A. (1995): "A Bernstein-type inequality for U-statistics and U-processes," *Statistics & Probability Letters*, 22, 239–247.

Arduini, T., E. Patacchini, and E. Rainone (2015): "Parametric and semiparametric iv estimation of network models with selectivity," Working paper, Einaudi Institute for Economics and Finance (EIEF).

Auerbach, E. (2016): "Identification and Estimation of a Partially Linear Regression Model using Network Data," Working paper, Northwestern University.

Bai, J. and P. Wang (2016): "Econometric Analysis of Large Factor Models," *Annual Review of Economics*, 8, 53–80.

Bennett, G. (1962): "Probability inequalities for the sum of independent random variables," *Journal of the American Statistical Association*, 57, 33–45.

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

Bickel, P. J., A. Chen, and E. Levina (2011): "The method of moments and degree distributions for network models," *The Annals of Statistics*, 39, 2280–2301.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739.

Bonhomme, S. and E. Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83, 1147–1184.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press.

Candelaria, L. E. (2016): "A Semiparametric Network Formation Model with Multiple Linear Fixed Effects," Working paper, Duke University.

Card, D., J. Heining, and P. Kline (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality," *The Quarterly Journal of Economics*, 128, 967–1015.

Chandrasekhar, A. (2016): "Econometrics of network formation," *The Oxford Handbook of the Economics of Networks*, 303–357.

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20, S1–S13.

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

Chen, M., I. Fernández-Val, and M. Weidner (2014): "Nonlinear factor models for network and panel data," Cemmap working paper CWP38/18.

Chetty, R. and N. Hendren (2018): "The impacts of neighborhoods on intergenerational mobility II: County-level estimates," *The Quarterly Journal of Economics*, 133, 1163–1228.

Choi, D. S., P. J. Wolfe, and E. M. Airoldi (2012): "Stochastic blockmodels with a growing number of classes," *Biometrika*, 99, 273–284.

De Paula, Á. (2017): "Econometrics of network models," in *Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress*, Cambridge University Press Cambridge, 268–323.

De Paula, Á., S. Richards-Shubik, and E. Tamer (2018): "Identifying preferences in networks with bounded degree," *Econometrica*, 86, 263–288.

Dzemski, A. (2018): "An Empirical Model of Dyadic Link Formation in a Network with Unobserved Heterogeneity," *The Review of Economics and Statistics*, 1–14.

Fernández-Val, I. and M. Weidner (2016): "Individual and time effects in nonlinear panel models with large N, T," *Journal of Econometrics*, 192, 291–312.

——— (2018): " Fixed Effects Estimation of Large- T Panel Data Models ," *Annual Review of Economics*, 10, 109–138.

Finkelstein, A., M. Gentzkow, and H. Williams (2016): "Sources of geographic variation in health care: Evidence from patient migration," *Quarterly Journal of Economics*, 131, 1681–1726.

Gao, C., Y. U. Lu, and H. H. Zhou (2015): "Rate-optimal graphon estimation," *The Annals of Statistics*, 43, 2624–2652.

Gao, W. Y. (2019): "Nonparametric identification in index models of link formation," *Forthcoming in Journal of Econometrics*.

Goldsmith-Pinkham, P. and G. W. Imbens (2013): "Social networks and the identification of peer effects," *Journal of Business & Economic Statistics*, 31, 253–264.

Graham, B. S. (2015): "Methods of identification in social networks," *Annual Review of Economics*, 7, 465–485.

——— (2016): "Homophily and transitivity in dynamic network formation," Working paper, University of California - Berkeley.

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

Hahn, J. and H. R. Moon (2010): "Panel data models with finite number of multiple equilibria," *Econometric Theory*, 26, 863–881.

Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007): "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 301–354.

Hanushek, E. A., J. F. Kain, J. M. Markman, and S. G. Rivkin (2003): "Does peer ability affect student achievement?" *Journal of Applied Econometrics*, 18, 527–544.

Helpman, E., O. Itskhoki, M. A. Muendler, and S. J. Redding (2017): "Trade and inequality: From theory to estimation," *Review of Economic Studies*, 84, 357–405.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating trade flows: Trading partners and trading volumes," *Quarterly Journal of Economics*, 123, 441–487.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002): "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983): "Stochastic blockmodels: First steps," *Social networks*, 5, 109–137.

Hoover, D. N. (1979): "Relations on probability spaces and arrays of random variables," Preprint, Institute for Advanced Study, Princeton, NJ.

Hsieh, C.-S. and L. F. Lee (2016): "A Social Interactions Model with Endogenous Friendship Formation and Selectivity," *Journal of Applied Econometrics*, 31, 301–319.

Jochmans, K. (2018): "Semiparametric Analysis of Network Formation," *Journal of Business and Economic Statistics*, 36, 705–713.

Jochmans, K. and M. Weidner (2019): "Fixed-Effect Regressions on Network Data," *Econometrica*, 87, 1543–1560.

Johnsson, I. and H. R. Moon (2019): "Estimation of peer effects in endogenous social networks: Control function approach," *Forthcoming in Review of Economics and Statistics*.

Kallenberg, O. (2005): *Probabilistic symmetries and invariance principles*, Springer, New York.

Klopp, O., A. B. Tsybakov, and N. Verzelen (2017): "Oracle inequalities for network models and sparse graphon estimation," *The Annals of Statistics*, 45, 316–354.

Krivitsky, P. N., M. S. Handcock, A. E. Raftery, and P. D. Hoff (2009): "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, 31, 204–213.

Leung, M. and H. R. Moon (2019): "Normal Approximation in Large Network Models," Working paper, University of Southern California.

Leung, M. P. (2019): "A weak law for moments of pairwise stable networks," *Journal of Econometrics*, 210, 310–326.

Li, Y., D. Shah, D. Song, and C. L. Yu (2019): "Nearest Neighbors for Matrix Estimation Interpreted as Blind Regression for Latent Variable Model," *Forthcoming in IEEE Transactions on Information Theory*.

Lovász, L. (2012): *Large Networks and Graph Limits*, vol. 60, American Mathematical Soc.

Manski, C. F. (1993): "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies*, 60, 531–542.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001): "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444.

Mele, A. (2017a): "A structural model of dense network formation," *Econometrica*, 85, 825–850.

——— (2017b): "A structural model of homophily and clustering in social networks," Working paper, Johns Hopkins University - Carey Business School.

Mele, A., L. Hao, J. Cape, and C. E. Priebe (2019): "Spectral inference for large Stochastic Blockmodels with nodal covariates," *arXiv preprint arXiv:1908.06438*.

Menzel, K. (2015): "Strategic network formation with many agents," Working paper, NYU.

Neyman, J. and E. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica*, 16, 1–32.

Orbanz, P. and D. M. Roy (2015): "Bayesian models of graphs, arrays and other exchangeable random structures," *IEEE transactions on pattern analysis and machine intelligence*, 37, 437–461.

Qu, X. and L.-f. Lee (2015): "Estimating a spatial autoregressive model with an endogenous spatial weight matrix," *Journal of Econometrics*, 184, 209–232.

Ridder, G. and S. Sheng (2015): "Estimation of large network formation games," Working paper, UCLA.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005): "Teachers, schools, and academic achievement," *Econometrica*, 73, 417–458.

Rothstein, J. (2010): "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *The Quarterly Journal of Economics*, 125, 175–214.

Roy, S., Y. Atchadé, and G. Michailidis (2019): "Likelihood Inference for Large Scale Stochastic Blockmodels with Covariates based on a Divide-and-Conquer Parallelizable Algorithm with Communication," *Journal of Computational and Graphical Statistics*, 1–22.

Santos Silva, J. M. and S. Tenreyro (2006): "The log of gravity," *Review of Economics and Statistics*, 88, 641–658.

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

Sheng, S. (2016): "A structural econometric analysis of network formation games," Working paper, UCLA.

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. Von Wachter (2019): "Firming up inequality," *Quarterly Journal of Economics*, 134, 1–50.

Toth, P. (2017): "Semiparametric Estimation in Networks with Homophily and Degree Heterogeneity," Working paper, University of Nevada.

Verdier, V. (2018): "Estimation and inference for linear models with two-way unobserved heterogeneity and sparsely matched data," *Forthcoming in Review of Economics and Statistics*.

Yan, T., B. Jiang, S. E. Fienberg, and C. Leng (2019): "Statistical Inference in a Directed Network Model With Covariates," *Journal of the American Statistical Association*, 114, 857–868.

Zhang, Y., E. Levina, and J. Zhu (2017): "Estimating network edge probabilities by neighbourhood smoothing," *Biometrika*, 104, 771–783.