Title
Semiparametric and Nonparametric Methods for Network Data

Author
Auerbach, Eric James

Publication Date
2017

Peer reviewed|Thesis/dissertation

Semiparametric and Nonparametric Methods for Network Data

by

Eric James Auerbach

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor James Powell, Co-chair
Professor Bryan Graham, Co-chair
Professor Demian Pouzo
Professor Adityanand Guntuboyina

Summer 2017

Semiparametric and Nonparametric Methods for Network Data

Abstract

Semiparametric and Nonparametric Methods for Network Data

by

Eric James Auerbach

Doctor of Philosophy in Economics

University of California, Berkeley

Professor James Powell, Co-chair

Professor Bryan Graham, Co-chair

This dissertation studies two frameworks for incorporating network data into economic modeling.

In the first chapter I consider the latent space framework of Holland and Leinhardt (1981) in which the existence of a link between two agents depends on their position in a latent space. I use this framework to estimate the parameters of a linear model in which the regressors and errors covary with the agents latent positions. Neither the endogenous relationship between the regressors and errors nor the distribution of network links are restricted parametrically. Instead, the model is identified by variation in the regressors unexplained by the agents latent positions. I first demonstrate that agents with similar columns of the squared adjacency matrix, the ijth entry of which contains the number of other agents linked to both agents i and j, necessarily have a similar distribution of network links. I then propose a semi parametric estimator based on matching pairs of agents with similar columns of the squared adjacency matrix. I find sufficient conditions for the estimator to be consistent and asymptotically normal, and provide a consistent estimator for its asymptotic variance.

In the second chapter I consider the rooted network framework of Aldous and Steele (2004). I use this framework to specify a nonparametric regression of a scalar outcome on a sparse network. The main assumption is that the outcome depends predominately on the configuration of agents and links nearby a distinguished agent. I first establish notion of distance between such configurations and then use it to construct a nearest-neighbor estimator of the regression function.

In the third chapter I revisit the latent space setting of the first chapter. I first specify a semi parametric model of link formation in which the existence of a link between a pair of agents depend on their positions in some latent space, an idiosyncratic error, and some linear combination of observed link covariates. I then proposes an estimator for the infinite-dimensional component of the model using a variation on the matching strategy outlined

in the first chapter ands characterize the rate of convergence of the estimator using large-deviation arguments.

To Christina. Her love of economics made this possible.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisors Jim and Bryan for their advice and support. I would also thank my friends and family for their encouragement, in particular my brother, Jonathan, and my parents, Cary and Paul. My wife Christina was a constant source of sanity and reason throughout the entire process. The least I can do is dedicate this thesis to her.

Chapter 1

Identification and Estimation of Models with Endogenous Network Formation

1.1 Introduction

In many social networks, linked agents make similar decisions. One explanation for this phenomenon is peer effects, in which agents are influenced by or choose to imitate the behavior of their peers. Another is latent homophily, in which linked agents have underlying characteristics that generate correlated though otherwise unrelated behaviors. Distinguishing between peer effects and latent homophily matters because the former often suggests that a policy maker can efficiently influence mass behavior by manipulating only a small number of key agents or links.[1] However, recent work has questioned not only the existence of network peer effects, but the extent to which they can be identified in nonexperimental settings at all.[2]

This chapter considers network peer effects as part of a broader study about the identification and estimation of models with endogenous network formation.[3] In this chapter, I address two fundamental questions. First, when are models with endogenous networks identified? Second, how can data on network links be used to control for this sort of endogeneity in estimation?

[1]Recent examples include Ballester, Calvó-Armengol, and Zenou (2006); Christakis and Fowler (2007); Calvó-Armengol, Patacchini, and Zenou (2009); Banerjee, Chandrasekhar, Duflo, and Jackson (2013), and Elliott, Golub, and Jackson (2014)

[2]For instance, Shalizi and Thomas (2011); Carrell, Sacerdote, and West (2013); Angrist (2014); Jackson (2014), and Graham (2015)

[3]Endogeneity refers to models in which the regressors and errors are correlated. A network represents a collection of pairs of agents that are distinguished in some economically meaningful way (ie, the pairs are "linked," "connected," "friends," etc.). Network endogeneity refers to models in which the correlation between the regressors and errors is explained by latent factors that influence link formation in a network.

I study these questions in the context of a linear model in which a correlation between the regressors and errors is caused by an omitted vector of unobserved social characteristics. I do not assume that the researcher has access to instrument or control variables for the endogenous regressors. Instead, relevant features of the social characteristics are to be inferred by how agents link in a network. To do this, I consider a nonparametric model of link formation in which the probability that two agents link is some unknown function of their social characteristics. The model admits a basic random utility interpretation and is consistent with a number of network formation models from the literature, including Chandrasekhar and Jackson (2014); Graham (2014); Leung (2015); Ridder and Sheng (2015), and Menzel (2015).

In recent work, Goldsmith-Pinkham and Imbens (2013); Hsieh and Lee (2014); Johnsson and Moon (2015), and Arduini, Patacchini, and Rainone (2015) all consider related models with endogenous network formation. These chapters all impose strong parametric assumptions on the network formation model to identify and estimate the parameters of interest. The performance of their estimators, however, depend on the accuracy of these assumptions which may potentially fail to capture the full heterogeneity in linking behavior underlying many real world networks.

The first contribution of this chapter is to provide identification conditions that do not require parametric restrictions on the network model. The idea behind these conditions is familiar: the model is identified if conditional on the distribution of network links, the regressors and errors are uncorrelated and the distribution of the regressors is nondegenerate. A key feature of this chapter is that it introduces new tools to formalize these conditions and make them straightforward to apply in practice.

For instance, I demonstrate that the linear peer effects model of Bramoullé, Djebbari, and Fortin (2009) is not generally identified when the network is endogenous. In particular, the nondegeneracy condition is violated because the explanatory variable of interest (an agent's expected peers' characteristics) is completely determined by the distribution of network links. Similar non-identification results are found in the related grouped peer effects literature (for instance, Manski (1993); Graham and Hahn (2005); Graham (2008)), and I discuss how strategies from this literature might be used to restore identification in the network setting.

The second contribution of this chapter is to propose a new matching procedure to estimate models with endogenous networks. Specifically, I propose matching pairs of agents with similar columns of the squared adjacency matrix.[4] The idea follows from a new result I derive in this setting that agents with similar columns of this matrix necessarily have a similar distribution of network links. The logic is related to recent arguments from the link prediction literature (for example Bickel, Chen, and Levina (2011); Zhang, Levina, and Zhu (2015)), though to my knowledge the results of this chapter and its application to the study of network endogeneity are original.

[4]The adjacency matrix of a network is a matrix with the number of rows and columns equal to the number of agents that contains a 1 in the ijth entry if agents i and j are linked and a 0 otherwise. The squared adjacency matrix refers to the matrix square of the adjacency matrix and agent i's column of the squared adjacency matrix is the ith column of this matrix.

The proposed estimator resembles other matching estimators from the literature (for instance Powell (1987); Heckman, Ichimura, and Todd (1998); Abadie and Imbens (2006)) and is similarly straightforward to implement and interpret. However, its large sample properties are nonstandard when compared to this literature for two reasons.

The first reason concerns the dimension of the matching variable. The above literature makes asymptotic approximations that require the density function of the matching variable to exist and be bounded away from zero. In this chapter, the matching variable is a column vector of length equal to the sample size. Since the usual notion of a density function does not exist in this setting, these asymptotic approximations are generally inapplicable. I sidestep the issue by appealing to arguments from the functional nonparametrics literature (for example, Ferraty and Vieu (2006); Hong and Linton (2016)) in which the density function is replaced by the more general notion of a small ball probability. I then adapt tools from the literature on dense graph limits (for instance, Lovász (2012)) to characterize this probability and find sufficient conditions for consistency and asymptotic normality. As is common in the matching literature, the bias of my estimator is potentially large relative to its variance. Accurate inference requires a bias correction and I propose a variation on the jackknife technique proposed by Powell, Stock, and Stoker (1989).

The second reason this estimator is nonstandard is that even though the matching variable is generated in the sense that its entries are sample averages with variances on the order of the inverse of the sample size, this variation does not influence the asymptotic distribution of the estimator. This result is unusual because it seemingly contradicts a developed literature on asymptotic variance formulas for semiparametric estimators (for instance Newey (1994); Chen, Linton, and Van Keilegom (2003); Hahn and Ridder (2013)). The intuition behind this result is that the average squared difference between two agents' matching variables estimates a particular measure of network distance between the agents. Evaluating the variance of my estimator does not require bounding the sampling variation of all of these estimated distances, but only those that correspond to pairs of matched agents. Since the estimated distances between matched agents is small by construction, their means and variances must also be small, and under certain regularity conditions the total variation is small enough to be asymptotically negligible. As a result, the asymptotic variance of my estimator does not have the usual correction term for a first stage estimation error.

The matching logic extends to various nonlinear and nonparametric settings, or to allow for weghted, directed, bipartite, multiple, sampled, or higher-order networks. I explore some of these extensions in an appendix to this chapter (though formal results are left to future work). The method also has important limitations. The model and estimator generally require the network to be dense (the number of links is proportional to the square of the sample size) and that the network links are exchangeable (the distribution of network links does not depend on how the agents are indexed). Some sparsity can be accomodated by letting the link probabilities decrease with the sample size (as in Bickel and Chen (2009)), and although the rate of convergence is likely to be affected, this may be unimportant if the total number of agents is large. The assumption of conditional link independence can also be weakened. For instance, it can be replaced with the conditional independence of some higher-order network event, such as the formation of cliques of a particular size, along the lines proposed by Chandrasekhar and Jackson (2014).

The structure of this chapter is as follows. Section 2 introduces the model, identification conditions, and proposed estimator. Section 3 contains the main results of the chapter. Section 3.2 provides the main identification results and section 3.3 the main asymptotic results: sufficient conditions for consistency and asymptotic normality. Section 4 provides simulation evidence and Section 5 concludes. Proofs of the various lemmas and theorems are collected in Appendix A and some extensions to the proposed model and estimator can be found in Appendix B. Appendices C and D contain additional context for the results. Appendix C illustrates the proposed matching strategy using three example parametric link distributions from the literature. Appendix D provides details about a behavioral interpretation for the model and estimator.

1.2 Model and Estimator

Model

Let $\{y_i, x_i\}_{i=1}^n$ be an independent and identically distributed sequence of data for n agents with $y_i \in \mathbb{R}, x_i \in \mathbb{R}^k$ for some positive integer k, and D be an $n \times n$ stochastic binary adjacency matrix corresponding to an unlabelled, unweighted, and undirected random network between the n agents. The joint distribution of $\{y_i, x_i\}_{i=1}^n$ and D is determined by the following semiparametric model

$$y_i = x_i\beta + \lambda(w_i) + \varepsilon_i \tag{1.1}$$

$$D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\}\mathbb{1}\{i \neq j\} \tag{1.2}$$

in which $\{w_i\}_{i=1}^n$ is an independent and identically distributed sequence of unobserved social characteristics, λ and f are unknown Lebesgue measurable functions with the latter symmetric in its arguments, and $\{\eta_{ij}\}_{i,j=1}^n$ is a symmetric matrix of unobserved scalar disturbances with independent and identically distributed upper diagonal entries that are mutually independent of $\{x_i, w_i, \varepsilon_i\}_{i=1}^n$. I suppose for the sake of exposition that $E[\varepsilon_i|x_i, w_i] = 0$, although the main results of this chapter will be derived under a weaker uncorrelatedness assumption. It is generally without loss to normalize the marginal distributions of w_i and η_{ij} to be standard uniform.

In this model, endogeneity takes the form of a dependence between x_i and the unobserved error $u_i = \lambda(w_i) + \varepsilon_i$ through w_i. Network formation is represented by $\binom{n}{2}$ conditionally independent Bernoulli trials in which the probability that agents i and j link is proportional to $f(w_i, w_j)$. Parametric examples of (1.2) in the network formation literature include Holland and Leinhardt (1981); Duijn, Snijders, and Zijlstra (2004); Krivitsky, Handcock, Raftery, and Hoff (2009); Dzemski (2014); Graham (2014) and Nadler (2016) (see section 3 of Graham (2015) for a review). Leung (2015); Ridder and Sheng (2015) and Menzel (2015) also consider network formation games with strategic interactions that imply equation (1.2) as a reduced form distribution of links. More details about a behavioral interpretation for this model can be found in Appendix D.

Example 1 (Network Peer Effects): Let y_i be student GPA, x_i be a vector of student characteristics (age, grade, gender, etc.), and $D_{ij} = 1$ if students i and j are friends and 0 otherwise. One extension of the Manski (1993) linear-in-means peer effects model of student achievement to the network setting is

$$\begin{aligned} y_i &= x_i\beta + E[x_j|D_{ij}=1,w_i]\rho_1 + E[y_j|D_{ij}=1,w_i]\rho_2 + \lambda(w_i) + \varepsilon_i \\ D_{ij} &= \mathbb{1}\{\eta_{ij} \le f(w_i,w_j)\}\mathbb{1}\{i \ne j\} \end{aligned}$$

in which $E[x_j|D_{ij}=1,w_i]$ denotes the mean characteristics and $E[y_j|D_{ij}=1,w_i]$ the mean GPA of agent i's friends, conditional on agent i's social characteristics w_i. Bramoullé, Djebbari, and Fortin (2009) consider a similar model in which the network is exogenous ($\lambda(w_i) = 0$) and Goldsmith-Pinkham and Imbens (2013); Hsieh and Lee (2014); Johnsson and Moon (2015), and Arduini, Patacchini, and Rainone (2015) consider related models with additional parametric assumptions on λ or f.[5]

Example 2 (Information Diffusion) Banerjee, Chandrasekhar, Duflo, and Jackson (2013) model household participation in a microfinance program in which information about the program diffuses over a social network. The authors control for household-level heterogeneity in program information by specifying and simulating a joint model of information diffusion and program participation. Ignoring for now that their outcome is binary,[6] I propose a semiparametric alternative

$$\begin{aligned} y_i &= x_i\beta + E[y_j|D_{ij}=1,w_i]\rho + \lambda(w_i) + \varepsilon_i \\ D_{ij} &= \mathbb{1}\{\eta_{ij} \le f(w_i,w_j)\}\mathbb{1}\{i \ne j\} \end{aligned}$$

In this linear example, $i = 1, ..., n$ indexes households with program participants, y_i is a measure of the intensity of participation (for exmaple, the number of loans or the amount of money borrowed), x_i is a vector of observed household characteristics (caste, religion, wealth, etc.), $D_{ij} = 1$ if households i and j have a social connection, and w_i are characteristics that influence social network formation. $\lambda(w_i)$, the probability that household i is informed about the program given their social characteristics, is a correction term for selection into the program due to heterogeneous information.

[5] The use of the expected peer outcomes $E[y_j|D_{ij} = 1, w_i]$ intead of their empirical couterparts $\sum_j y_j D_{ij} / \sum_j D_{ij}$ masks another endogeneity issue generated by having dependent variables on the right hand side of the outcome equation. Bramoullé, Djebbari, and Fortin (2009) resolve this issue by using functions of D and $\{x_i\}_{i=1}^n$ as instruments for $\sum_j y_j D_{ij} / \sum_j D_{ij}$. I ignore the issue here because the simultaneity issue is unrelated to the unobserved heterogeneity focus of this chapter.

[6] In future work I plan to demonstrate how the results of this chapter can be extended to certain nonlinear and nonparametric models along the lines of Manski (1987) and Honoré and Powell (1997).

Example 3 (Job Mobility): Schmutte (2014) studies a bipartite[7] labor market network in which worker i and industry-occupation j are linked if worker i works in industry-occupation j at some point in time. He identifies several clusters of highly connected workers and industry-occupations in the labor market network and uses the clusters as proxy variables for unobserved worker and industry-occupation heterogeneity in a linear model of labor market earnings. Using the network formation model of this chapter to directly characterize the relationship between this unobserved heterogeneity and the observed network clusters, I characterize his model as a model with an enodgneous network along the lines of

$$\log(y_{it}) = x_{it}\beta + \theta(\phi_1(w_i)) + \psi(\phi_2(w_{j(i,t)})) + \varepsilon_{it}$$
$$D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(\phi_1(w_i), \phi_2(w_j))\}$$

in which y_{it} is the earnings of worker i in time period t, x_{it} are worker characteristics (age, gender, race, etc.), $j(i,t)$ indexes the industry-occupation of worker i in period t, w_i and $w_{j(i,t)}$ denote unobserved worker and industry-occupation characteristics (for instance, ability or productivity), and ϕ_1 and ϕ_2 map worker and industry-occupation characteristics to the network clusters.

Example 4 (Research Productivity): Ductor, Fafchamps, Goyal, and van der Leij (2014) study a model of research productivity in which a researcher's current publication quality depends on past quality, researcher characteristics, and a vector of network statistics derived from a coauthorship network (in which two researchers are linked if they have previously been coauthors) including agent degree, eigenvector centrality, betweeness centrality, etc. The authors experiment with several different models of productivity, including various combinations of network statistics. I propose a semiparametric alternative that treats the unknown combination of network statistics as unobserved network heterogeneity

$$y_i = x_i\beta + \lambda(w_i) + \varepsilon_i$$
$$D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\}\mathbb{1}\{i \neq j\}$$

in which w_i might characterize the academic community of researcher i (for instance, a field of study) and $\lambda(w_i)$ indexes heterogeneity in research productivity due to this community. A key feature of this model is that the estimation of β does not depend on correctly identifying the relevant features of the network that make up $\lambda(w_i)$

In many cases, the function λ (or the functions ϕ and ψ in Example 3) is not a nuisance parameter, but also an object of interest in the analysis. In future work I plan to demonstrate how the tools of this chapter can be extended to estimate and conduct inference about features of these parameters as well.

[7] In Appendix B I define a bipartite network and describe how one might extend the methods of this chapter to the bipartite setting.

Estimator

Estimation is complicated by the fact that the social charactersitics $\{w_i\}_{i=1}^n$ are unobserved. If the social characteristics were observed, (1.1) corresponds to the partially linear regression of Engle, Granger, Rice, and Weiss (1986), and many tools exist to estimate β (for example, Chamberlain (1986); Powell (1987); Newey (1988); Robinson (1988)). If the social characteristics were unobserved but identified by the distribution of D, one can extend these methods by replacing the social characteristics with empirical analogs as in Ahn and Powell (1993); Ahn (1997), and Hahn and Ridder (2013). This particular approach is taken by Arduini, Patacchini, and Rainone (2015) and Johnsson and Moon (2015).

However, in many empirical applications the social characteristics are neither observed nor identified by the distribution of D. This chapter demonstrates that identifying, estimating, and conducting inference about β is still possible without imposing parametric restrictions on either f or λ by matching pairs of agents with similar link distributions. The result is motivated by two key insights.

One insight concerns the identification of β, which holds if two conditions are satisfied. The first condition is that $\lambda(w_i)$ depends on w_i only through the schedule of linking probabilities $f(w_i, \cdot) : [0, 1] \to [0, 1]$. The second is that there is excess variation in the distirbution of x_i that is not explained by $f(w_i, \cdot)$. Formally, consider the pseudometric on the space of social characteristics defined by

$$d(u, v) = ||f(u, \cdot) - f(v, \cdot)||_2 = \left(\int (f(u, \tau) - f(v, \tau))^2 \, d\tau \right)^{1/2}$$

Here, the linking function $f(u, \cdot)$ gives the probability that an agent with social characteristics u links with agents of every other social characteristic in $[0, 1]$, and $d(u, v)$ is the integrated squared difference in the linking functions of agents u and v. The identification conditions are that β is identified if $E[(x_i - x_j)'(\lambda(w_i) - \lambda(w_j))|d(w_i, w_j) = 0] = 0$ and $E[(x_i - x_j)'(x_i - x_j)|d(w_i, w_j) = 0]$ is positive definite. These conditions are similar to the usual identification conditions for linear models with unobserved heterogeneity in the panel data setting (see, for example Wooldridge (2010) Chapter 10): it is the notion of the network distance measure d used to partial out the endogenous variation that is different.

The logic behind the first identification condition is that d describes the totality of information that the distribution of D contains about w_i. That is, if $d(w_i, w_j) = 0$ then there is no feature of the network that can distinguish between agents i and j. They will have the same probability of being connected in any particular configuration of links, and thus will have the same distribution of degrees, eigenvector centralities, average peer characteristics, and any other agent-level statistic of D. If $E[(x_i - x_j)'(\lambda(w_i) - \lambda(w_j))|d(w_i, w_j) = 0] \neq 0$, then matching agents with similar link distributions will not control for all of the unobserved heterogeneity in (1.1), but under (1.2) there is no further information in the distribution of D that can identify it. Additionally, when w_i is identified by the distribution of D, $d(w_i, w_j) = 0$ implies $|w_i - w_j| = 0$, so that $E[(x_i - x_j)'(\lambda(w_i) - \lambda(w_j))|d(w_i, w_j) = 0] = 0$ holds trivially. As a consequence, this first identification condition is more general than those imposed by

Goldsmith-Pinkham and Imbens (2013); Hsieh and Lee (2014); Johnsson and Moon (2015), and Arduini, Patacchini, and Rainone (2015).

A sufficient condition for $E[(x_i - x_j)'(\lambda(w_i) - \lambda(w_j))|d(w_i, w_j) = 0] = 0$ is for $\lambda(w_i)$ to be continuous in d (ie, if $\{w^t\}_{t=1}^{\infty}$ such that $d(w_i, w^t) \to 0$ then $|\lambda(w_i) - \lambda(w^t)| \to 0$). I prefer the former condition because in some cases there is variation in $\lambda(w_i)$ that is not continuous in d but is uncorrelated with x_i so that the first identification condition is still valid. For instance, suppose the omitted function is an indicator for whether or not an agent is linked to agent 1, or $\lambda(w_i) = D_{i1}$. Then $\lambda(w_i)$ is not continuous with respect to d, but $D_{i1} = E[D_{i1}|w_i] + (D_{i1} - E[D_{i1}|w_i])$ in which the first summand is continuous with respect to d and the second is uncorrelated with x_i.

The logic behind the second identification condition is that matching agents with similar link distributions only identifies β if there is excess variation in the distribution of x_i not explained by the linking function $f(w_i, \cdot)$. Otherwise there is a dimension of the covariate space such that all of the variation in y_i can be explained by w_i regardless of the magnitude of β. One example of this is when x_i contains agent-level statistics of the adjacency matrix. Another is the case of linear-in-means network peer effects. I discuss these cases in more detail below.

The second insight is that the average squared difference in the ith and jth columns of the squared adjacency matrix ($D \times D$) can be used to bound $d(w_i, w_j)$. The logic has two steps. First, there exists another pseudometric δ on $[0, 1]^2$ such that $d(w_i, w_j)$ can be bounded in terms of $\delta(w_i, w_j)$. Second, $\delta(w_i, w_j)$ can be consistently estimated by the root average squared difference in the ith and jth columns of the squared adjacency matrix

$$\hat{\delta}_{ij} = \left(n^{-1} \sum_{t=1}^{n} \left((n-2)^{-1} \sum_{s=1}^{n} D_{ts}(D_{is} - D_{js}) \right)^2 \right)^{1/2} \tag{1.3}$$

Here, the codegree $\sum_{s=1}^{n} D_{ts} D_{is}$ gives the number of other agents that are linked to both agents i and t, $\{\sum_{s=1}^{n} D_{ts} D_{is}\}_{t=1}^{n}$ is the collection of codegrees between agent i and the other agents in the sample, and $\hat{\delta}_{ij}$ gives the root average squared difference in i's and j's collection of codegrees. Similar relationships between configurations of network moments and the distribution of links have also been exploited by Lovász and Szegedy (2007, 2010); Bickel, Chen, and Levina (2011); Lovász (2012), and Zhang, Levina, and Zhu (2015).

The two insights indicate that when the ith and jth columns of the squared adjacency matrix are similar and the identification conditions for β hold then $(y_i - y_j)$ and $(x_i - x_j)\beta + (\varepsilon_i - \varepsilon_j)$ are approximately equal. This result is limited in the sense that it is insufficient to estimate λ by a series approximation as in Newey (1988) and Ai and Chen (2003) because w_i is not necessarily identified. However, one can recover β by matching pairs of agents with d-similar social characteristics. This chapter demonstrates that under certain regularity conditions β is consistently estimated by a pairwise difference estimator

$$\hat{\beta} = \left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(x_i - x_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)\right)^{-1}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(y_i - y_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)\right) \tag{1.4}$$

in which K is a kernel density function and h_n a bandwidth parameter depending on the sample size.

The estimator has a form similar to established pairwise difference estimators from the literature (in particular, Ahn and Powell (1993)). However, the large sample properties of $\hat{\beta}$ are not typical of this literature. For example, unless the researcher is willing to put substantial structure on the unknown linking function f, the distribution of $\hat{\delta}_{ij}$ can be difficult to characterize near 0, complicating the usual balancing of asymptotic bias and variance. The problem is related to the small ball problem in the functional nonparametrics literature (see for example Masry (2005); Ferraty and Vieu (2006); Hong and Linton (2016)) and can severely amplify the usual curse of dimensionality. Of particular concern is the possibility that the quantity of matches shrinks to zero quicker than the averages in (1.4) converge, though in the proofs of this chapter I demonstrate how the structure of the network model sufficiently mitigates this problem such that under certain regularity conditions the proposed estimator is consistent and asymptotically normal.

Example 1 (Network Peer Effects) In the network peer effects model

$$\begin{aligned} y_i &= x_i\beta + E[x_j|D_{ij} = 1, w_i]\rho_1 + E[y_j|D_{ij} = 1, w_i]\rho_2 + \lambda(w_i) + \varepsilon_i \\ D_{ij} &= \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\}\mathbb{1}\{i \neq j\} \end{aligned}$$

the parameters ρ_1 and ρ_2 are not identified since $E[x_j|D_{ij} = 1, w_i] = E[x_jD_{ij}|w_i]/E[D_{ij}|w_i]$ is a fixed function of w_i that is indistinguishable from $\lambda(w_i)$. In particular, the model violates the nondegeneracy identification condition since

$$E[x_j|D_{ij} = 1, w_i] = \int E[x_j|w_j = w]f(w_i, w)dw \Big/ \int f(w_i, w)dw$$

and $d(w_i, w_{i'}) = ||f(w_i, \cdot) = f(w_{i'}, \cdot)||_2 = 0$ implies

$$E[(E[x_j|D_{ij} = 1, w_i] - E[x_j|D_{i'j} = 1, w_{i'}])^2 \,|d(w_i, w_{i'}) = 0] = 0$$

It is helpful to contrast this result with that of Goldsmith-Pinkham and Imbens (2013), who study the model

$$\begin{aligned} y_i &= x_i\beta + E[x_j|D_{ij} = 1, w_j, Z_{ij}]\rho_1 + E[y_j|D_{ij} = 1, w_j, Z_{ij}]\rho_2 + w_i\rho_3 + \varepsilon_i \\ D_{ij} &= \mathbb{1}\{\eta_{ij} \leq |w_i - w_j|\gamma_1 + Z_{ij}\gamma_2\}\mathbb{1}\{i \neq j\} \end{aligned}$$

Their model is identified by two restrictions. The first is the functional form restriction on the network heterogeneity $\lambda(w_i) = w_i\rho_3$. The second is the introduction of exogenous link covariates Z_{ij}, assumed to be independent of w_i and w_j.[8]

Example 2 (Information Diffusion) In the microfinance program participation model

$$\begin{aligned} y_i &= x_i\beta + E[y_j|D_{ij} = 1, w_i]\rho + \lambda(w_i) + \varepsilon_i \\ D_{ij} &= \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\}\mathbb{1}\{i \neq j\} \end{aligned}$$

the parameter ρ is not identified following previous arguments. The parameter β is identified if two households with the same distribution of links have the same probability of being informed about the program and a household's covariates are not completely determined by their distribution of links. For example, if households only link to other households of the same religion or caste, then the second condition is violated. In contrast to Banerjee, Chandrasekhar, Duflo, and Jackson (2013), the estimation of β does not require many-networks asymptotics.

Example 3 (Job Mobility): In the labor market earnings model

$$\begin{aligned} \log(y_{it}) &= x_{it}\beta + \theta(\phi_1(w_i)) + \psi(\phi_2(w_{j(i,t)})) + \varepsilon_{it} \\ D_{ij} &= \mathbb{1}\{\eta_{ij} \leq f(\phi_1(w_i), \phi_2(w_j))\} \end{aligned}$$

β is identified if agents in different network clusters have a different distribution of network links and there is excess variation in the worker and industry-occupation covariates that are not explained by the network links. The first is satisfied by construction since Schmutte (2014) defines the clusters as functions of the network links. The second is satisfied if the covariates have overlapping support across clusters.

Example 4 (Research Productivity): In the research productivity model

$$\begin{aligned} y_i &= x_i\beta + \lambda(w_i) + \varepsilon_i \\ D_{ij} &= \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\}\mathbb{1}\{i \neq j\} \end{aligned}$$

β is identified if there is excess variation in the covariates that is not explained by the network links. This may not be satisfied if reasearchers only coauthor with other researchers with similar publication histories.

[8] It is also possible to incorporate link covariates into the framework of this chapter by replacing equation (1.2) with $D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j, Z_{ij})\}$. In the appendix, I demonstrate how the estimator of this chapter can be extended to models with link covariates by matching on conditional codegree vectors, although a formal study of the asymptotic properties of such an estimator is left to future work.

1.3 Main Results

Terminology and Notation

This section details additional constructions required for the lemmas, theorems and proofs. I define agent i's network type to be the projection of the link function f onto his social characteristics: $f_{w_i}(\cdot) := f(w_i, \cdot) : [0,1] \to [0,1]$. In words, it is the collection of probabilities that agent i links to agents with each social characteristic in $[0,1]$. I consider network types to be elements of $L^2([0,1])$, the usual inner product space of square integrable functions on the unit interval. As suggested by the notation of the previous section, $d(w_i, w_j) = ||f_{w_i} - f_{w_j}||_2$ is the L^2 metric on the space of network types.

I require two network theoretic constructions: (average) agent degrees and (average) agent-pair codegrees. The degree of agent i is the fraction of other agents linked to agent i in D, or $(n-1)^{-1} \sum_{t \neq i} D_{it}$. Under (1.2), that $(n-1)^{-1} \sum_{t \neq i} D_{it} \to_{a.s.} \int f_{w_i}(\tau) d\tau$ follows from the usual strong law of large numbers. Similarly, for $i \neq j$ the codegree of agent pair (i,j) is the fraction of other agents linked to both agent i and agent j, or $(n-2)^{-1} \sum_{t \neq i,j} D_{it} D_{jt}$. Again, under (1.2), $(n-2)^{-1} \sum_{t \neq i,j} D_{it} D_{jt} \to_{a.s.} \int f_{w_i}(\tau) f_{w_j}(\tau) d\tau = \langle f_{w_i}, f_{w_j} \rangle_{L^2}$. For reference, I denote this codegree by $\hat{p}_{ij}$ and its almost sure limit with $p(w_i, w_j)$. I emphasize that $p(w_i, w_i)$ refers to the limiting codegree of two distinct agents with social characteristics equal to w_i and *not* to the limiting degree of agent i. That is $p(w_i, w_i) := \int f_{w_i}(\tau)^2 d\tau = ||f_{w_i}||_2^2 \neq \int f_{w_i}(\tau) d\tau$.

Notice that p also defines a link function, in which $p(w_i, w_j)$ gives the probability that agents i and j have a link in common, as opposed to $f(w_i, w_j)$, which gives the probability that they are directly linked themselves. To distinguish p from f I refer to it as the codegree link function (associated with f), and the function $p_{w_i}(\cdot) := p(w_i, \cdot) : [0,1] \to [0,1]$ as agent i's codegree type. I also take codegree types to be elements of $L^2([0,1])$. I refer to the pseudometric on $[0,1]$ induced by L^2-differences in codegree types with δ, so that

$$\begin{aligned}
\delta(u,v) = ||p(u,\cdot) - p(v,\cdot)||_2 &= \left(\int \left(p(u,\tau) - p(v,\tau)\right)^2 d\tau \right)^{1/2} \\
&= \left(\int \left(\int f(\tau,s) \left(f(u,s) - f(v,s)\right) ds \right)^2 d\tau \right)^{1/2}
\end{aligned}$$

for any pair of social characteristics u and v. Under (1.2), my Lemma 1 demonstrates that the root average squared difference in the ith and jth columns of the squared adjacency matrix (given by (1.3)) provides a uniformly consistent estimator for $\delta(w_i, w_j)$.

I use two different conditional expectations defined over events on the network types. Let Z_i and Z_{ij} be arbitrary random matrices indexed at the agent and agent-pair level respectively. Then $E[Z_{ij}|\ ||f_{w_i} - f_{w_j}||_2 = x]$ refers to the conditional expectation

$$\lim_{h \to 0} E[Z_{ij}|\ (w_i, w_j) \in \{(u,v) : x \leq ||f_u - f_v||_2 \leq x + h\}]$$

and $E\left[Z_i | f_{w_i} = f\right]$ refers to the conditional expectation

$$\lim_{h \to 0} E\left[Z_i | \, w_i \in \{w : ||f_w - f||_2 \leq h\}\right]$$

Though f_{w_i} is a random function, these conditional expectations implicitly refer to the measure induced by the random variable w_i. Conditional means with respect to the agent codegree differences or types are defined in an analogous way.

Let $u_i = \lambda(w_i) + \varepsilon_i$. I use the functional $\lambda(f)$ to denote $E[u_i | f_{w_i} = f]$ and ν_i for the associated residual $u_i - \lambda(f_{w_i})$. This allows me to rewrite equations (1.1) and (1.2) in a way that emphasizes the identification and estimation strategy described in the previous section.

$$y_i = x_i\beta + \lambda(f_{w_i}) + \nu_i \tag{1.5}$$

$$D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\} \tag{1.6}$$

Model Identification

This section gives conditions for agents with similar network types but different regressors to identify β.

Assumption 1: The random sequence $\{x_i, \nu_i, w_i\}_{i=1}^n$ is independent and identically distributed with entries mutually independent of $\{\eta_{ij}\}_{j>i=1}^n$, a symmetric random array with independent and identically distributed entries above the diagonal. The variables w_i and η_{ij} have standard uniform marginals. The conditional distributions of $\{y_i\}_{i=1}^n$ and D are given by equations (1.5) and (1.6) respectively, for some Lebesgue-measurable and symmetric link function $f : [0,1]^2 \to [0,1]$.

Assumption 1 is a restatement of the discussed model and is included primarily as a reference. Since the marginal distributions of w_i and η_{ij} are not seperately identified from f, the assumption of standard uniform marginals is without loss of generality (see Bickel and Chen (2009) for a discussion).

Assumption 2: The variables x_i and u_i both have finite sixth moments and $E[(x_i - x_j)'(u_i - u_j)| \, ||f_{w_i} - f_{w_j}||_2 = 0] = 0$.

The second part of Assumption 2 is satisfied if x_i and u_i are uncorrelated conditional on f_{w_i}.

Assumption 3: The conditional covariance matrix $\Gamma_0 = E\left[(x_i - x_j)'(x_i - x_j) \, | \, ||f_{w_i} - f_{w_j}||_2 = 0\right]$ is positive definite.

Assumption 3 states that there is some independent variation in each of the regressors that is not explained by the network types. Section 2 explores cases when it is unrealistic, for example when the regressors include functions of the adjacency matrix. The assumption can be weakened in cases when the researcher has some additional information about the network formation process (for example, exogenous link covariates) or structure on the endogenous covariation in equation (1.5).

Theorem 1: Suppose Assumptions 1-3 hold. Then β is the unique minimizer of $E\left[((y_i - y_j) - (x_i - x_j)b)^2 \mid ||f_{w_i} - f_{w_j}||_2 = 0\right]$ over $b \in \mathbb{R}^k$.

Model Estimation

This section characterizes the large sample properties of $\hat{\beta}$. The first part provides sufficient conditions for consistency. The second part provides sufficient conditions for the limiting distribution to be normal. Accurate inference may require a bias correction and the third part demonstrates how a variation on the jackknife method proposed by Powell, Stock, and Stoker (1989) can be used for this purpose. The fourth part provides a consistent estimator for the asymptotic variance.

Consistency

Consistency of $\hat{\beta}$ requires an additional continuity condition on the conditional expectation functions from Assumptions 2 and 3, and restrictions on the bandwidth sequence and kernel density function.

Assumption 4: The conditional expectation functions satisfy
$\lim_{h\to 0} E[(x_i - x_j)'(u_i - u_j)| \, ||f_{w_i} - f_{w_j}||_2 = h] = 0$ and
$\lim_{h\to 0} E[(x_i - x_j)'(x_i - x_j)| \, ||f_{w_i} - f_{w_j}||_2 = h] = \Gamma_0$.

Assumption 4 is satisfied if Assumptions 2 and 3 hold and the conditional expectation functionals $E[x_i'u_i|f_{w_i}]$ and $E[x_i'x_i|f_{w_i}]$ as defined in Section 3.1 are continuous with respect to f_{w_i} in the L^2-sense. This condition might not be satisfied if the network is sparse, because f_{w_i} may be uniformly close to zero so that small variations in f_{w_i} correspond to large variation in x_i and u_i. In the appendix, I discuss how the estimator can be altered to mitigate this problem by allowing the magnitude of f to change with the sample size.

Assumption 5: The bandwidth sequence $h_n \to 0$, $n^{1-\gamma}h_n^2 \to \infty$ for some $\gamma > 0$, and $nr_n \to \infty$ for $r_n = E\left[K\left(\frac{||p_{w_i} - p_{w_j}||_2}{h_n}\right)\right]$. K is supported, bounded, and differentiable on $[0, 1]$, and strictly positive on $[0, 1)$.

The first two restrictions on the bandwidth sequence are standard. The third condition, that $nr_n \to \infty$ is not. This condition is required to ensure that the number of matches used to estimate $\hat{\beta}$ is increasing with n. If p_{w_i} was a d-dimensional random vector with compact support and a strictly positive density function, $P(||p_{w_i} - p_{w_j}||_2 \leq h_n)$ would be on the order of h_n^d. The number of agent-pairs with similar codegree types would then be on the order of nh_n^d, which increases with n if the second bandwidth condition were changed to

$n^{1-\gamma}h_n^d \to \infty$. Since p_{w_i} is infinite dimensional, $P(||p_{w_i} - p_{w_j}||_2 \leq h_n)$ cannot necessarily be approximated by a polynomial of h_n of known order and so this third bandwidth condition is required.

The conditions on the kernel density function K are satisfied by a type-II kernel density function (examples include the Epanechnikov, Biweight, and Bartlett kernels). It is possible to extend this proof to include the standard uniform kernel density function, although kernels supported on all of $\mathbb{R}$ (for example the Gaussian kernel) may potentially cause problems in this setting (see Hong and Linton (2016) for a discussion).

If the collection of network differences between agents $\{||f_{w_i} - f_{w_j}||_2\}_{i\neq j}$ were observed and used to construct the matches in $\hat{\beta}$, the arugments for consistency would be similar to those of Ahn and Powell (1993), though with some alterations to accomodate the dimensionality of f_{w_i}. That the estimator is still consistent when $||f_{w_i} - f_{w_j}||_2$ is replaced by $\hat{\delta}_{ij}$ follows from two arguments. First, $\{\hat{\delta}_{ij}\}_{i\neq j}$ converges uniformly to the codegree differences $\{||p_{w_i} - p_{w_j}||_2\}_{i\neq j}$. Second, agent-pairs with small codegree differences have small network differences. These results are stated in Lemmas 1 and 2 respectively.

Lemma 1: Suppose Assumptions 1 and 5 hold. Then

$$\max_{(i\neq j)} \left| \hat{\delta}_{ij} - ||p_{w_i} - p_{w_j}||_2 \right| = o_{a.s.}\left(n^{-\gamma/4}h_n\right)$$

in which γ refers to the exponent from Assumption 5.

Lemma 1 demonstrates that the collection of $\binom{n}{2}$ empirical codegree differences observed by the researcher converges uniformly to their population analogs at a rate slightly slower than $n^{-1/2}$ (since h_n can be taken to be arbitrarily close to $n^{-1/2}$ by taking γ close to 0). The proof involves repeated applications of Bernstein's Inequality and the union bound over the $\binom{n}{2}$ distinct empirical codegrees that make up $\{\hat{\delta}_{ij}\}_{i\neq j}$

Lemma 2: Suppose Assumption 1 holds. Then for every $\epsilon > 0$ there exists a $\delta > 0$ such that with probability at least $1 - \epsilon^2/4$

$$||p_{w_i} - p_{w_j}||_2 \leq \delta \implies ||f_{w_i} - f_{w_j}||_2 \leq \epsilon$$

Lemma 2 is the main justification for the matching strategy of this chapter. The result is somewhat unexpected since $||p_{w_i} - p_{w_j}||_2 \leq ||f_{w_i} - f_{w_j}||_2$ is almost an immediate consequence of Jensen's inequality.[9] Nevertheless, pairs of agents with similar codegree types have similar network types with high probability.

[9] To see this, note $||p_{w_i} - p_{w_j}||_2^2 = \int \left(\int f(t,s)\left(f(w_i,s) - f(w_j,s)\right)ds\right)^2 dt \leq \int \left(\int \left(f(t,s)\left(f(w_i,s) - f(w_j,s)\right)\right)^2 ds\right) dt \leq \int \left(f(w_i,s) - f(w_j,s)\right)^2 ds = ||f_{w_i} - f_{w_j}||_2^2$, where the first inequality is due to Jensen and the second due to the fact that f is bounded between 0 and 1.

The lemma is related to Theorem 13.27 of Lovász (2012), which demonstrates that $||p_{w_i} - p_{w_j}||_2 = 0$ implies $||f_{w_i} - f_{w_j}||_2 = 0$ when f is continuous. The logic of his result is illustrated below.

$$
\begin{aligned}
&||p_{w_i} - p_{w_j}||_2^2 = 0 \implies \int \left(\int f(\tau, s) \left(f(w_i, s) - f(w_j, s) \right) ds \right)^2 d\tau = 0 \\
&\implies \int f(\tau, s) \left(f(w_i, s) - f(w_j, s) \right) ds = 0 \text{ for every } \tau \\
&\implies \int f(w_i, s) \left(f(w_i, s) - f(w_j, s) \right) ds = 0 \text{ and } \int f(w_j, s) \left(f(w_i, s) - f(w_j, s) \right) ds = 0 \\
&\implies \int \left(f(w_i, s) - f(w_j, s) \right)^2 ds = 0 \implies ||f_{w_i} - f_{w_j}||_2^2 = 0
\end{aligned}
$$

Essentially, the result follows from the fact that if agents i and j have identical codegree types, then the difference in their network types $(f_{w_i} - f_{w_j})$ must be uncorrelated with each other network type in the population, as indexed by τ. In particular, the difference is uncorrelated with f_{w_i} and f_{w_j}, the network types of agents i and j. However, this can only be the case if the network types of i and j are perfectly correlated.

Lovász's theorem demonstrates that agent-pairs with identical codegree types also have identical network types. However, consistency of $\hat{\beta}$ requires a stronger result, that agent-pairs with similar but not necessarily equivalent codegree types have similar network types. This is the statement of Lemma 2. Unfortunately the above proof cannot simply be extended by replacing each occurance of 0 with some function of a small $\epsilon > 0$, because the third implication relies on $\int f(\tau, s) \left(f(w_i, s) - f(w_j, s) \right) ds = 0$ for exactly all τ, which is not guaranteed by the condition $||p_{w_i} - p_{w_j}||_2^2 \leq \epsilon$ for any $\epsilon > 0$. Despite this, the proof of Lemma 2 demonstrates that the two notions of distance are similar in enough places that matching agents with similar codegree types is sufficient to partial out $\lambda(f_{w_i})$ in equation (1.5) and consistently estimate β.

Theorem 2: Suppose Assumptions 1-5 hold. Then $\left(\hat{\beta} - \beta \right) \to_p 0$.

Theorem 2 is almost a direct consequence of Lemmas 1 and 2, several applications of the continuous mapping theorem, and Lemma 3.1 from Powell, Stock, and Stoker (1989).

Asymptotic Normality

I provide two asymptotic normality results. The first result concerns the case when the support of the agent linking function f_{w_i} is finite, so that $P(||f_{w_i} - f_{w_j}||_2 = 0) = P(||p_{w_i} - p_{w_j}||_2 = 0) > 0$ and there exists an $\epsilon > 0$ such that $P(0 < ||f_{w_i} - f_{w_j}||_2 < \epsilon) = P(0 < ||p_{w_i} - p_{w_j}||_2 < \epsilon) = 0$.

Theorem 3: Suppose Assumptions 1-5 hold. Further suppose the support of f_{w_i} is finite. Then

$$V_{3,n}^{-1/2} \left(\hat{\beta} - \beta \right) \to_d \mathcal{N} \left(0, I_k \right)$$

where $V_{3,n} = \Gamma_0^{-1}\Omega_0\Gamma_0^{-1} \times s/n$, Γ_0 is as defined in Assumption 3, I_k is the $k \times k$ identity matrix, and

$$s = P(||p_i - p_j||_2 = 0, ||p_i - p_k||_2 = 0)/P(||p_i - p_j||_2 = 0)^2$$
$$\Omega_0 = E\left[(x_i - x_j)'(x_i - x_k)(u_i - u_j)(u_i - u_k)|\, ||p_i - p_j||_2 = 0, ||p_i - p_k||_2 = 0\right]$$

When the support of network types is finite, pairs of agents with similar codegree types have identical network types with high probability, and so the proof of Theorem 3 follows from Assumptions 1-5, Lemmas 1 and 2, and standard arguments. This theorem is included for three reasons. First, it adds to a literature noting that the adverse effects of unobserved heterogeneity can be mild when the support of this variation is finite (for example Hahn and Moon (2010); Bonhomme and Manresa (2015)). Second, the assumption of discrete heterogeneity is not uncommon in empirical work (for instance, Schmutte (2014); Bonhomme, Lamadon, and Manresa (2015)). Third, it provides an easy to interpret condition such that $\hat{\beta}$ is consistent and asymptotically normal at the $\sqrt{n}$-rate.

The second result concerns the more general case when the support of f_{w_i} is not necessarily finite. In this case, the proof of asymptotic normality requires additional structure on f and the conditional expectations from Assumption 4, which is given in the following Assumptions 6 and 7. Assumption 8 modifies the bandwidth sequence accordingly.

Assumption 6: There exists an integer K and a partition of $[0, 1)$ into K equally spaced, adjacent, and non-intersecting intervals $\cup_{t=1}^{K}[x_t^1, x_t^2)$ such that for any $t \in \{1, ..., K\}$ and almost every $x, y \in [x_t^1, x_t^2)$ and $s \in [0, 1]$, $|f(x, s) - f(y, s)| \leq C_6|x - y|^{\alpha}$, for some $C_6 \geq 0$ and $\alpha > 0$.

Assumption 6 imposes that the space of social characteristics can be partitioned into K segments such that on each partition segment the link function f is almost everywhere Hölder continuous of some order. The partition allows for discrete jumps of the link function as to include discrete models such as the stochastic blockmodel (see Appendix C for a definition and discussion) as a special case. The restriction that the partition is uniformly sized is without loss, and the results can also be extended to let $K_n \to 0$ slowly with n.[10]

Assumption 7: The conditional expectation
$E[(x_i - x_j)'(u_i - u_j)|\, ||f_{w_i} - f_{w_j}||_2 = h] \leq C_7 h^{\zeta}$ for some C_7, $\zeta > 0$ and all h in a neighborhood to the right of 0.

Assumption 7 stengthens Assumption 4 so that the slope of the conditional expectaton $E[(x_i - x_j)'(u_i - u_j)|\, ||f_{w_i} - f_{w_j}||_2]$ is bounded by a fractional polynomial to the right of 0.

[10] This corresponds to a stochastic blockmodel with a growing number of blocks as in Wolfe and Olhede (2013). A similar condition is used by Zhang, Levina, and Zhu (2015)

Assumption 8: The bandwidth sequence $h_n = C_8 \times n^{-\rho}$ for $\rho \in \left(\frac{\alpha}{4+8\alpha}, \frac{\alpha}{2+4\alpha}\right)$ and some $C_8 > 0$. $K(\sqrt{u})$ is supported, bounded, and twice differentiable on $[0,1]$, and strictly positive on $[0,1)$.

The rate of convergence of the bandwidth sequence depends on the exponent from Assumption 6. When $\alpha = 1$ this bandwidth choice is approximately on the order of magnitude considered by Ahn and Powell (1993). The proof of Theorem 4 is simplified by requiring the composition of K and $\sqrt{\cdot}$ to be twice differentiable at 0, and all of the kernel density functions in the discussion of Assumption 5 satisfy this additional condition.

The second asymptotic normality proof uses Assumption 6 to strengthen Lemma 2 in the following way.

Lemma 3: Suppose Assumptions 1 and 6 hold. Then for almost every (w_i, w_j) pair

$$||p_{w_i} - p_{w_j}||_2 \leq ||f_{w_i} - f_{w_j}||_2 \leq 32\, C_6^{\frac{1}{2+4\alpha}} \left(||p_{w_i} - p_{w_j}||_2\right)^{\frac{\alpha}{1+2\alpha}}$$

so long as $||p_{w_i} - p_{w_j}||_2 < \sqrt{8C_6}K^{-\alpha}$, where C_6 and α are the constants from Assumption 6.

Theorem 4: Suppose Assumptions 1-3 and 6-8 hold. Further suppose $\alpha \times \zeta > 1/2$. Then

$$V_{4,n}^{-1/2}\left(\hat{\beta} - \beta_{h_n}\right) \to_d \mathcal{N}\left(0, I_k\right)$$

where $V_{4,n} = \Gamma_0^{-1}\Omega_n\Gamma_0^{-1}/n$, Γ_0 is as defined in Assumption 3, r_n is as defined in Assumption 5, and I_k is the $k \times k$ identity matrix, and

$$\beta_{h_n} = \beta + \left(\Gamma_0\right)^{-1} E\left[(x_i - x_j)'(u_i - u_j)K\left(\frac{||p_i - p_j||_2}{h_n}\right)\right] / \left(2r_n\right)$$

$$\Omega_n = E\left[(x_i - x_j)'(x_i - x_k)(u_i - u_j)(u_i - u_k)K\left(\frac{||p_i - p_j||_2}{h_n}\right) K\left(\frac{||p_i - p_k||_2}{h_n}\right)\right] / \left(r_n^2\right)$$

The statement of Theorem 4 warrants three remarks. First, the variance is not necesarily on the order of the inverse of the sample size. This is because the variance of the kernel $r_n^{-2}E\left[K\left(\frac{||p_i-p_j||_2}{h_n}\right)K\left(\frac{||p_i-p_k||_2}{h_n}\right)\right]$ can potentially diverge with n. When this variance converges to a limit, then $\left(\hat{\beta} - \beta_{h_n}\right)$ is asymptotically normal with variance $\Gamma_0\Omega_0\Gamma_0 \times \sigma/n$ where $\sigma = \lim_{n\to\infty} r_n^{-2}E\left[K\left(\frac{||p_i-p_j||_2}{h_n}\right)K\left(\frac{||p_i-p_k||_2}{h_n}\right)\right]$ and Ω_0 is as defined in Theorem 3. Even when this variance diverges, Assumptions 6-8 and Lemma 3 ensure that the rate of convergence for $V_{4,n}$ is on the order of at least $n^{-1/2}$ and is close to n^{-1} when α is close to 1. In the appendix, I propose an adaptive bandwidth procedure that requires each agent to belong to the same number of matches, which normalizes $r_n^{-2}E\left[K\left(\frac{||p_i-p_j||_2}{h_n}\right)K\left(\frac{||p_i-p_k||_2}{h_n}\right)\right] = 1$. Though this choice of bandwdith potentially inflates the bias of the estimator relative to

$\hat{\beta}$, simulation evidence suggests that this inflation is often small relative to the reduction in variance.

Second, the estimator has an oracle property in the sense that the estimation error of $\hat{\delta}_{ij}$ around $\delta(w_i, w_j)$ is asymptotically negligable, so that the researcher may conduct inference as though the codegree differences between agents were known. The intuition in this case is that conditional on (w_i, w_j), the asymptotic variance of $\sqrt{n}(\hat{\delta}_{ij} - \delta(w_i, w_j))$ is bounded from above by $d(w_i, w_j)$. Since the estimator is premised on $d(w_i, w_j)$ being close to zero, it follows that the variance of $\sqrt{n}(\hat{\delta}_{ij} - \delta(w_i, w_j))$ converges in probability to 0. When $\alpha \times \zeta > 1/2$ this rate of convergence is sufficiently fast to not impact the asymptotic variance of $\hat{\beta}$. This is distinct from the results of Ahn and Powell (1993). Their approach would roughly correspond to matching agents based on $\delta(\hat{w}_i, \hat{w}_j)$, where $\hat{w}_i$ is a consistent estimator for w_i. In their case, the variation of $\hat{w}_i$ around w_i and $\hat{w}_j$ around w_j is completely unrelated to $d(w_i, w_j)$, and so this variation does inflate the asymptotic variance of their estimator.

Third, the asymptotic distribution $\hat{\beta}$ is not centered at β, but at the pseudo-truth β_{h_n}. Though β_{h_n} converges to β, the rate of convergence can be slow depending on the size of α and ζ. This problem is common with matching estimators, although it is exacerbated here by the relatively weak relationship between the codegree and network distances as demonstrated by Lemma 3. In particular, Assumptions 6-8 and Lemma 3 only imply that $|\beta_{h_n} - \beta| = O_p\left(n^{\frac{-\zeta\alpha^2}{2(1+2\alpha)^2}}\right)$ which can imply a worst-case scenario bias on the order of $n^{-1/36}$.

Bias Correction

Inferences about β based on the asymptotic distribution provided by Theorem 4 will only be valid if $V_{4,n}^{-1/2}(\beta_{h_n} - \beta) = o_p(1)$. Otherwise, accurate inference requires a bias correction. The technique proposed in this chapter requires an additional smoothness condition

Assumption 9: The pseudo-truth function β_h satisfies $\beta_h = \sum_{l=1}^{L} C_l h^{l/\theta} + O\left(h^{(L+1)/\theta}\right)$ for some positive integer $L > \zeta\alpha/(2\theta(1+2\alpha))$, k-dimensional constants $C_1, C_2, ..., C_L$, $\theta > 0$, and h in a fixed open neighborhood to the right of 0.

Assumption 9 requires that the asymptotic bias from Theorem 4 is sufficiently smooth with respect to the bandwidth choice.

I propose the following jackknife bias corrected estimator $\bar{\beta}_L$. For an arbitrary sequence of distinct positive numbers $\{c_1, c_2, ..., c_L\}$ with $c_1 = 1$, $\bar{\beta}_L$ is defined to be

$$\bar{\beta}_L = \sum_{l=1}^{L} a_l \hat{\beta}_{c_l h_n} \tag{1.7}$$

in which $\hat{\beta}_{c_l h_n}$ refers to the pairwise difference estimator (1.4) with the choice of bandwidth $c_l \times h_n$ and the sequence$\{a_1, a_2, ...a_L\}$ satisfies

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & c_2^{2/\theta} & \dots & c_L^{2/\theta} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & c_2^{L/\theta} & \dots & c_L^{L/\theta} \end{pmatrix} \times \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_L \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Theorem 5: Suppose Assumptions 1-3 and 6-9 hold, and $L > \zeta\alpha/(2\theta(1+2\alpha))$. Then

$$V_{5,n}^{-1/2}\left(\bar{\beta}_L - \beta\right) \to_d \mathcal{N}\left(0, I_k\right)$$

where $V_{5,n} = \sum_{l_1=1}^{L}\sum_{l_2=1}^{L} a_{l_1} a_{l_2} \Gamma_0^{-1}\Omega_{n,l_1 l_2}\Gamma_0^{-1}/n$, Γ_0 is as defined in Assumption 3, r_n is as defined in Assumption 5, I_k is the $k \times k$ identity matrix, and

$$\Omega_{n,l_1 l_2} = E\left[(x_i - x_j)'(x_i - x_k)(u_i - u_j)(u_i - u_k) K\left(\frac{||p_i - p_j||_2}{h_n}\right) K\left(\frac{||p_i - p_k||_2}{h_n}\right)\right] / \left(r_n^2\right)$$

Variance Estimation

The asymptotic variances from Theorems 3-5 can be consistently estimated using the sample analogs of Γ_0 and $\Omega_{n,l_1 l_2}$. That is, let $\hat{u}_i = y_i - \hat{\beta} x_i$,

$$\hat{\Gamma}_h = \binom{n}{2}^{-1} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \left(x_i - x_j\right)'\left(x_i - x_j\right) K\left(\frac{||\hat{p}_i - \hat{p}_j||_2}{h}\right)$$

and $\hat{\Omega}_{h_1,h_2} =$

$$\binom{n}{3}^{-2} \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n} \left(x_i - x_j\right)'\left(x_i - x_k\right)\left(\hat{u}_i - \hat{u}_j\right)\left(\hat{u}_i - \hat{u}_k\right) K\left(\frac{||\hat{p}_i - \hat{p}_j||_2}{h_1}\right) K\left(\frac{||\hat{p}_i - \hat{p}_k||_2}{h_2}\right)$$

then

Theorem 6: Suppose Assumptions 1-5 hold. Then $\left(\hat{\Gamma}_{h_n}^{-1}\hat{\Omega}_{h_n,h_n}\hat{\Gamma}_h^{-1} - nV_{4,n}\right) \to_p 0$ and $\left(\sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\hat{\Gamma}_{c_{l_1}h_n}^{-1}\hat{\Omega}_{c_{l_1}h_n, c_{l_2}h_n}\hat{\Gamma}_{c_{l_2}h_n}^{-1} - nV_{5,n}\right) \to_p 0$

A corollary to Theorem 6 is that $\hat{\Gamma}_{h_n}^{-1}\hat{\Omega}_{h_n,h_n}\hat{\Gamma}_h^{-1}$ also consistently estimates $nV_{3,n}$ under the hypothesis of Theorem 3. These statistics can be used to build confidence intervals or test hypotheses about β under the relevant assumptions in the usual way. Asymptotic theory has little to say about the actual choices of bandwidths and constants used in the construction of the estimators in this section. The setting potentially allows for choices based on cross validation which I leave to future work.

1.4 Simulations

This section presents simulation evidence for three types of network formation models: a stochastic blockmodel, a beta model, and a homophily model. To simplify the exposition, a detailed explanation of the models is deferred to Appendix C. For each of R simulations, I draw a random sample of n observations $\{\xi_i, \varepsilon_i, \omega_i\}_{i=1}^n$ from a trivariate normal distribution with mean 0 and covariance given by the identity matrix and a random symmetric matrix $\{\eta_{ij}\}_{i,j=1}^n$ with independent and identically distributed upper diagonal entries with standard uniform marginals. For each of the following link functions f, the adjacency matrix D is formed by $D = \mathbb{1}\{\eta_{ij} \leq f(\Phi(\omega_i), \Phi(\omega_j))\}$ where Φ is the cummulative distribution function for the standard univariate normal distribution.

The first design draws D from a stochastic blockmodel where

$$f_1(u,v) = \begin{cases} 1/3 & \text{if } u \leq 1/3 \text{ and } v > 1/3 \\ 1/3 & \text{if } 1/3 < u \leq 2/3 \text{ and } v \leq 2/3 \\ 1/3 & \text{if } u > 2/3 \text{ and } (v > 2/3 \text{ or } v \leq 1/3) \\ 0 & \text{otherwise} \end{cases}$$

The linking function f_1 generates network types with finite support as in the hypothesis of Theorem 3. For this model, I take $\lambda(\omega_i) = \lceil 3\Phi(\omega_i) \rceil$, $x_i = \xi_i + \lambda(\omega_i)$, and $y_i = \beta x_i + \gamma\lambda(\omega_i) + \varepsilon_i$. The second and third designs draw D from the beta model and homophily model where

$$f_2(u,v) = \frac{\exp(u+v)}{1+\exp(u+v)} \text{ and } f_3(u,v) = 1 - (u-v)^2$$

For these models, $\lambda(\omega_i) = \omega_i$, $x_i = \xi_i + \lambda(\omega_i)$ and $y_i = \beta x_i + \gamma\lambda(\omega_i) + \varepsilon_i$.

Let x and y to denote the stacked n-dimensional vector of observations $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, and Z_1 for the $(n \times 2)$ matrix $\{x_i, \lambda(\omega_i)\}_{i=1}^n$. I use c_i to denote a vector of network statitics for agent i based on D containing agent degree $n^{-1}\sum_{j=1}^n D_{ij}$, eigenvector centrality,[11] and average peer covariates $\sum_{j=1}^n D_{ij}x_j / \sum_{j=1}^n D_{ij}$. Z_2 denotes the stacked vector $\{x_i, c_i\}_{i=1}^n$.

For each design, I evaluate the performance of six estimators. The benchmark is $\hat{\beta}_1 = (Z_1'Z_1)^{-1}(Z_1'y)$, the infeasible OLS regression of y on x and $\lambda(\omega_i)$. $\hat{\beta}_2 = (x'x)^{-1}(x'y)$ is the naïve OLS regression of y on x. $\hat{\beta}_3 = (Z_2'Z_2)^{-1}(Z_2'y)$ is the OLS regression of y on x and the vector of network controls c. $\hat{\beta}_4$ is the proposed pairwise difference estimator given in (1.4) without bias correction, $\hat{\beta}_5$ is the bias corrected estimator (1.7), and $\hat{\beta}_6$ is the pairwise difference estimator with an adaptive bandwidth but without bias correction (see Appendix A for more details). The pairwise difference estimators all use the Epanechnikov kernel $K(u) = 3(1-u^2)\mathbb{1}\{u^2 < 1\}/4$. Estimators $\hat{\beta}_4$ and $\hat{\beta}_5$ use the bandwidth sequence $n^{-1/9}/10$ and the estimator $\hat{\beta}_6$ uses the bandwidth sequence $n^{-1/9}/5$. Since $n^{1/9}$ is roughly equal to 2

[11] Agent i's eigenvector centrality statistics refers to the ith entry of the eigenvector of D associated with the largest eigenvalue.

for the sample sizes considered in this section, the results are close to a constant bandwidth choice of $h_n = .05$ and .1 respectively.

Tables 1-3 demonstrates the results for $R = 1000$, $\beta = \gamma = 1$ and for each n in $\{50, 100, 200, 500, 800\}$. For each model, estimator and sample size, the first row gives the mean, the second gives the mean absolute error of the simulated estimators around β, the third gives the mean absolute error divided by that of $\hat{\beta}_1$, and the fourth gives the proportion of the simulation draws that fall outside of a 0.95 confidence interval based on the asymptotic distributions derived in the previous section.

Table 1 contains results for the stochastic blockmodel. The naïve estimator $\hat{\beta}_2$ has a large and stable positive bias that is not reduced as n is increased. The OLS estimator with network controls $\hat{\beta}_3$ is not asymptotically well defined in this example because the network statistics converge to constants. The results in Table 1 instead demonstrate a common "fix" in the literature, which is to instead calculate $(Z_2'Z_2)^{+}(Z_2'y)$ where $+$ refers to the Moore-Penrose pseudo-inverse. The results for this estimator indicate that adding network controls mitigates some of the bias in β_1 (due to sampling variation in the number of agents in each block), however the estimator is otherwise poorly behaved. Notice this bias returns when the block sizes stabilize (in particular when $n = 800$).

The results for the pairwise difference estimators illustrate the content of Theorem 3, that when the unobserved heterogeneity is discrete, the proposed estimator identifies pairs of agents of the same type with high probability. As a result, the pairwise difference estimators $\hat{\beta}_4$ and $\hat{\beta}_6$ behave similar to the infesible $\hat{\beta}_2$. For the stochastic blockmodel, Assumption 9 is not valid, and so the jackknife bias correction actually inflates both the bias and variance of $\hat{\beta}_4$. Looking at the relative mean absolue error for this estimator, it is clear that the relative performance of the error is deteriorating as n increases (though the bias and variance of this estimator is still on the order of $1/\sqrt{n}$).

Table 2 contains results for the beta model. Relative to the stochastic blockmodel, all of the estimators for the beta model (except infeasible OLS) have large biases. This is because the link function f_2 is very flat, so that the variation in linking probabilities that identifies the network positions is relatively small (see also Section 5 of Johnsson and Moon, 2015). In appendix C I demonstrate that the social characteristics are identified by the distribution of D (they are consistently estimated by the order statistics of the degree distribution), but the bound on the deviation of the social characteristics given by the network metric is large: $|u - v| \leq 20 \times d(u, v)$.

Still, the proposed pairwise difference estimator offers a substantial improvement in performence relative to both the naïve estimator $\hat{\beta}_2$ and the estimator with network controls $\hat{\beta}_3$. For example, when $n = 100$, $\hat{\beta}_5$ has approximately half the bias and mean absolute error of $\hat{\beta}_1$ while $\hat{\beta}_3$ offers a reduction of less than ten percent. When $n = 800$ the reduction in bias is over three times as large (75% relative to 23%).

Table 3 contains results for the homophily model. As in the case of the beta model, I demonstrate in Appendix C that the social characteristics are also identified in the homophily model. Unlike the beta model, there is a relatively large amount of information about the network positions in the linking probabilities so that all of the estimators in Table 3 are

n		Infeasible OLS $\hat{\beta}_1$	Naïve OLS $\hat{\beta}_2$	OLS with Controls $\hat{\beta}_3$	Pairwise Difference $\hat{\beta}_4$	Bias Corrected $\hat{\beta}_5$	Adaptive Bandwidth $\hat{\beta}_6$
50							
	bias	0.004	0.829	0.268	0.060	0.022	0.106
	MAE	0.116	0.829	0.274	0.224	0.240	0.150
	rMAE	1.000	7.147	2.362	1.931	2.069	1.293
	size	0.057	0.063	0.072	0.115	0.123	0.067
100							
	bias	0.003	0.829	0.226	0.021	-0.022	0.019
	MAE	0.083	0.829	0.229	0.089	0.094	0.084
	rMAE	1.000	9.988	2.759	1.072	1.133	1.012
	size	0.064	0.053	0.108	0.053	0.058	0.056
200							
	bias	0.001	0.823	0.180	0.004	-0.040	0.002
	MAE	0.056	0.823	0.183	0.058	0.069	0.058
	rMAE	1.000	14.696	3.268	1.036	1.232	1.036
	size	0.049	0.044	0.215	0.045	0.064	0.058
500							
	bias	0.000	0.824	0.172	0.006	0.038	0.001
	MAE	0.035	0.824	0.174	0.035	0.048	0.035
	rMAE	1.000	23.543	4.971	1.000	1.371	1.000
	size	0.033	0.061	0.777	0.037	0.047	0.044
800							
	bias	0.001	0.823	0.314	0.008	-0.036	0.000
	MAE	0.029	0.823	0.314	0.029	0.043	0.029
	rMAE	1.000	28.379	10.828	1.000	1.483	1.000
	size	0.057	0.038	0.127	0.054	0.068	0.062

Table 1.1: This table contains simulation results for 1000 replications and a sample size of $n = 100, 200, 500$. Bias gives the mean estiamtor minus 1. MAE gives the mean absolute error of the estimator around 1. rMAE gives the mean absolute error relative to the benchmark $\hat{\beta}_1$. Size gives the proportion of draws that fall outside the asymptotic 0.95 confidence interval.

n		Infeasible OLS $\hat{\beta}_1$	Naïve OLS $\hat{\beta}_2$	OLS with Controls $\hat{\beta}_3$	Pairwise Difference $\hat{\beta}_4$	Bias Corrected $\hat{\beta}_5$	Adaptive Bandwidth $\hat{\beta}_6$
50							
	bias	0.000	0.496	0.462	0.379	0.335	0.365
	MAE	0.119	0.496	0.463	0.381	0.341	0.366
	rMAE	1.000	4.168	3.891	3.202	2.866	3.076
	size	0.064	0.063	0.075	0.049	0.066	0.070
100							
	bias	0.006	0.501	0.462	0.336	0.269	0.298
	MAE	0.082	0.501	0.462	0.336	0.270	0.299
	rMAE	1.000	6.110	5.634	4.098	3.293	3.646
	size	0.055	0.053	0.055	0.039	0.062	0.081
200							
	bias	0.002	0.501	0.444	0.290	0.200	0.231
	MAE	0.058	0.501	0.444	0.290	0.200	0.231
	rMAE	1.000	8.638	7.655	5.000	3.448	3.983
	size	0.050	0.041	0.036	0.033	0.054	0.070
500							
	bias	0.003	0.499	0.403	0.246	0.136	0.151
	MAE	0.036	0.499	0.403	0.246	0.136	0.151
	rMAE	1.000	13.861	11.194	6.833	3.778	4.194
	size	0.049	0.042	0.054	0.022	0.033	0.076
800							
	bias	0.000	0.500	0.385	0.237	0.122	0.122
	MAE	0.028	0.500	0.385	0.237	0.122	0.122
	rMAE	1.000	17.857	13.750	8.464	4.357	4.357
	size	0.050	0.054	0.078	0.037	0.050	0.062

Table 1.2: This table contains simulation results for 1000 replications and a sample size of $n = 100, 200, 500$. Bias gives the mean estiamtor minus 1. MAE gives the mean absolute error of the estimator around 1. rMAE gives the mean absolute error relative to the benchmark $\hat{\beta}_1$. Size gives the proportion of draws that fall outside the asymptotic 0.95 confidence interval.

n		Infeasible OLS $\hat{\beta}_1$	Naïve OLS $\hat{\beta}_2$	OLS with Controls $\hat{\beta}_3$	Pairwise Difference $\hat{\beta}_4$	Bias Corrected $\hat{\beta}_5$	Adaptive Bandwidth $\hat{\beta}_6$
50							
	bias	0.007	0.505	0.269	0.128	0.087	0.140
	MAE	0.120	0.505	0.274	0.108	0.121	0.211
	rMAE	1.000	4.208	2.283	0.900	1.008	1.758
	size	0.068	0.051	0.063	0.062	0.068	0.132
100							
	bias	0.005	0.502	0.162	0.100	0.057	0.089
	MAE	0.081	0.502	0.167	0.124	0.108	0.116
	rMAE	1.000	6.198	2.062	1.531	1.333	1.432
	size	0.049	0.059	0.061	0.053	0.066	0.083
200							
	bias	0.001	0.503	0.095	0.085	0.039	0.055
	MAE	0.057	0.503	0.100	0.097	0.075	0.077
	rMAE	1.000	8.825	1.754	1.702	1.316	1.351
	size	0.054	0.059	0.054	0.050	0.057	0.069
500							
	bias	0.000	0.501	0.047	0.074	0.028	0.035
	MAE	0.035	0.501	0.053	0.077	0.048	0.046
	rMAE	1.000	14.314	1.514	2.200	1.371	1.314
	size	0.043	0.059	0.039	0.045	0.058	0.051
800							
	bias	0.000	0.501	0.034	0.070	0.023	0.030
	MAE	0.028	0.501	0.086	0.072	0.039	0.038
	rMAE	1.000	17.893	3.071	2.571	1.392	1.357
	size	0.039	0.040	0.041	0.038	0.050	0.047

Table 1.3: This table contains simulation results for 1000 replications and a sample size of $n = 100, 200, 500$. Bias gives the mean estiamtor minus 1. MAE gives the mean absolute error of the estimator around 1. rMAE gives the mean absolute error relative to the benchmark $\hat{\beta}_1$. Size gives the proportion of draws that fall outside the asymptotic 0.95 confidence interval.

much better behaved. In fact, for this model $|u - v| \leq d(u, v)$.

In this example, the OLS estimator with network controls actually performs comparably to the uncorrected pairwise difference estimator $\hat{\beta}_4$. This is because the peer characteristics variable $\sum_{j=1}^{n} D_{ij} x_j / \sum_{j=1}^{n} D_{ij}$ is a good approximation of w_i when n is large. However, the bias corrected estimator $\hat{\beta}_5$ outperforms both estimators over all of the sample sizes considered.

1.5 Directions for Future Work

I highlight two directions for future work. The first is to consider models in which the parameter of interest depends on the distribution of network links. For example, one might be interested in the functions $\beta(w_i)$ and $\lambda(w_i)$ in the model $y_i = x_i\beta(w_i) + \lambda(w_i) + \varepsilon_i$. To see why, suppose that x_i is an indicator for the adoption of some treatment. Then the function β describes how the treatment effect varies over the network, which intuitively might be nonconstant if the impact of treatment for a particular agent depends on the proportion of his social connections that have been similarly treated. Estimating $\beta(w_i)$ potentially allows the researcher to determine which positions in the network are associated with, for example, the largest or smallest treatment effects. I plan to demonstrate how the tools of this chapter might be used to estimate these and other features of both $\beta(w_i)$ and $\lambda(w_i)$ in future work.

The second direction for future work concerns a behavioral motivation for the model and estimator of this chapter. In Appendix D, I provide a basic random utility interpretation for the network model along the lines of Graham (2014). However, the discussion is otherwise largely divorced from a developed literature on economic models with strategic link formation. In future work, I hope to explore more connections between the setting of this chapter and that literature.

One connection is potentially provided by the literature on network formation games with private information. Recent work in this literature employs a similar network formation model as a within-equilibrium reduced form characterization of linking behavior (see for example Leung (2015); Ridder and Sheng (2015); Menzel (2015)). Here the social characteristics are public information about individual agents and the linking probabilities are conditionally independent given these characteristics and some equilibirium selection process (in this setting the link errors $\{\eta_{ij}\}_{i \neq j}$ constitute private information about the quality of individual links).

Understanding the mapping between structural models of network formation and this reduced-form representation might be mutually beneficial for both the network formation and network endogeneity literatures. For instance, the tools of this chapter could be used to fit models of network formation in which not all of the public information that informs linking decisions is observed by the researcher. At the same time, a deeper understanding of network formation is important to help researchers fitting models with endogenous network formation identify and control for the many types of unobserved heterogeneity lurking in the errors.

Chapter 2

A Sparse Network Regression

2.1 Introduction

A growing literature in economics studies how networks influence a variety of behaviors and outcomes. For instance, a firm's decision to enter a market may depend on the proximity of consumers and other firms, a student's scholastic achievement may depend on the student's ability, study habits, and relationships with other students, and the probability that a financial institution suffers from a financial crisis may depend on borrowing and lending relationships between other institutions. A common empirical strategy in the literature is to model the outcome of interest as a function of a low-dimensional vector of network statistics, such as the average number of links.[1] This strategy has the benefit of being parsimonious and simple to implement. However, it is limited in the sense that these network statistics only characterize a small number of potentially relevant features of the network. When economic theory is ambivalent as to the correct choice of network statistics to include in the model, an alternative approach may be more appropriate.

This chapter considers a setting in which the researcher observes data on a sequence of outcomes, each associated with a sparse network.[2] The goal is to specify and estimate a nonparametric regression of the outcome given its associated network. If the data consists of many independent draws of the outcome for a small number of possible networks, the problem can be solved by averaging over the outcomes associated with each network. In many cases of interest, however, the observed outcomes all correspond to different networks. Estimation is complicated by the fact that it is not clear how an outcome corresponding to one network can be used to estimate the regression function evaluated at a different network. This chapter applies a notion of network distance based on local approximations to estimate the regression function using observations with similar but not identical networks. I sketch

[1] Examples include Bramoullé, Djebbari, and Fortin (2009); Goldsmith-Pinkham and Imbens (2013); Banerjee, Chandrasekhar, Duflo, and Jackson (2013); Ductor, Fafchamps, Goyal, and van der Leij (2014).

[2] A network consists of a set of agents, a set of links between pairs of agents, and a set of covariates assigned to each agent and link. A network is sparse when the number of agents linked any particular agent is finite.

the premise of the local approximation here and provide a formal discussion in the second section of this chapter.

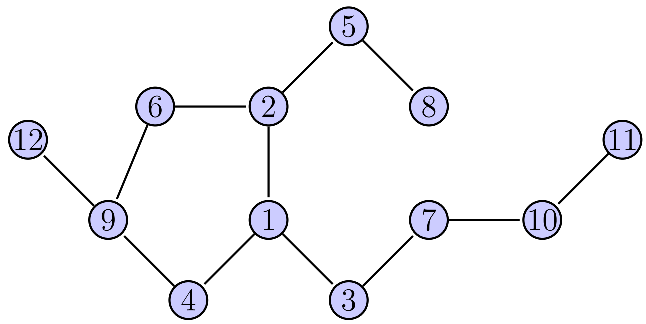
The driving assumption behind the local approximation is that the outcome of interest is predominantly explained by the configuration of agents, links, and covariates nearby a particular agent in the network. For instance, in many models of classroom peer effects, a student's scholastic achievement (the outcome of interest) is explained by the student's covariates, friends and their covariates, and to a lesser extent the student's friends of friends and their covariates. As this radius of consideration expands to the student's friends of friends of friends and their friends, the marginal impact of these additional students, friendships and covariates on the initial student's outcome is likely to be smaller. The idea behind the local approximation is that there exists a radius around the initial student such that ignoring the configuration of students, friendships, and covariates beyond this radius has only a negligible impact on the initial student's scholastic achievement.

An illustration of the local approximation can be found in Figures 2.1 and 2.2. The top panels of both figures depict a network defined on twelve agents with no covariates. The middle panel of Figure 2.1 depicts the configuration of agents and links within a radius of 1 around agent 1. This is the network formed by agent 1 and the agents linked to agent 1. The bottom panel of Figure 2.1 depicts the configuration of agents and links within a radius of 2 around agent 1. This is the network formed by agent 1, the agents linked to agent 1, and the agents linked to the agents linked to agent 1. Similarly, the middle and bottom panels of Figure 2.2 shows the configurations of agents and links within a radius of 1 and 2 respectively around agent 9.

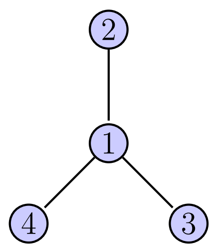
The main idea of this chapter is model the outcome of interest using the low dimensional configuration of agents and links within a radius of the distinguished agent. When the network is sparse the number of possible configurations within a finite radius is finite. As a result, estimating a regression function using local approximations can be done using standard tools (for instance, a linear regression) so long as the radius is sufficiently small relative to the sample size.

The quality of this estimator depends on the choice of radius used to generate the local approximation. If this radius is small, then the number of possible local approximations is small, the fraction of observed networks with local approximations equivalent to the network of interest is large, and thus the variance of the estimator will be small. However, there may be a sizeable bias due to the fact that networks associated with different outcomes may have similar local approximations. Increasing the radius decreases this bias, but at the cost of inflating the variance. The idea is to choose a radius that balances both the bias and variance of the estimator. The logic is entirely similar to that of sieve estimation in which a high-dimensional parameter space is replaced by a low-dimensional approximation of complexity depending on the sample size. Here it is the size of the radius used the generate the local approximation that characterizes the complexity of this sieve approximation.

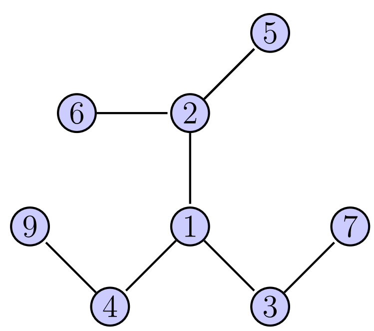
The remainder of this chapter is as follows. The second section of this chapter formalizes the notion of the local approximation. The third section of this chapter uses the notion of the local approximation to define a nonparametric regression of a scalar outcome on a



(a) A network with twelve agents.

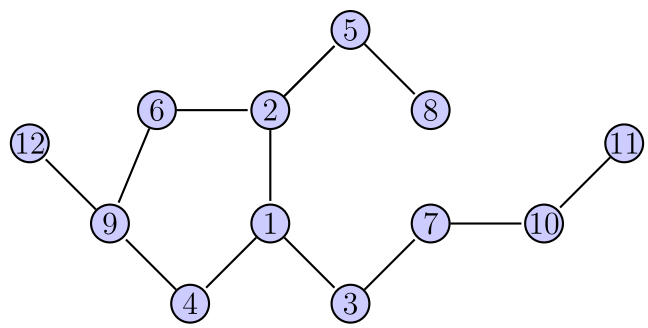


(b) The network formed by agent 1 and the three agents linked to agent 1.

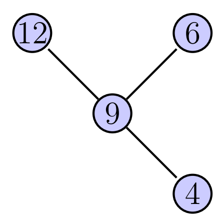


(c) The network formed by agent 1, the three agents linked to agent 1, and the four agents linked to the three agents linked to agent 1.

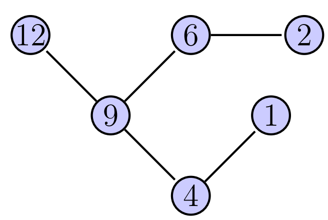
Figure 2.1: The top network depicts a configuration of links connecting twelve agents. The middle network depicts the network induced by the agents linked to agent 1. The bottom network depicts the network induced by the agents linked to the agents linked to agent 1. The premise of this chapter is that in order to predict an outcome associated with agent 1, the middle and bottom networks can be used as a low-dimensional approximation to the top network.



(a) A network with twelve agents.



(b) The network formed by agent 1 and the three agents linked to agent 9.



(c) The network formed by agent 9, the three agents linked to agent 9, and the two agents linked to the three agents linked to agent 1.

Figure 2.2: The top network depicts a configuration of links connecting twelve agents. The middle network depicts the network induced by the agents linked to agent 9. The bottom network depicts the network induced by the agents linked to the agents linked to agent 9. The premise of this chapter is that in order to predict an outcome associated with agent 9, the middle and bottom networks can be used as a low-dimensional approximation to the top network.

sparse network. The section also provides an estimator and characterizes its large sample properties. Simulation evidence and an empirical application to firm entry in networked markets is forthcoming.

2.2 The Local Approximation

The local approximation of this chapter is a variation on the local network topology as described by Aldous and Steele (2004). To minimize notation, I focus on a special case of undirected random networks with discretely valued link weights and agent-specific covariates.

An economic network $X = (V(X), W(X), C(X))$ is a collection of agents (V), a $|V| \times |V|$ symmetric matrix of finitely-supported off-diagonal link weights (W), and a $|V| \times K$ matrix of discretely distributed covariates (C). Specifically, the ijth entry of W, w_{ij} denotes the (possibly infinite) weight of the link between agents i and j, and the ikth entry of C contains the kth covariate of agent i. The number of agents $|V|$ may be countable infinite, and the network X is said to be random when the link weights W and agent covariates C are stochastic.

Given a network X and a pair of agents i and j in $V(X)$, the agent distance $d_V(i,j)$ is defined to be the length of the shortest path connecting agents i and j through X, where the length contributed by link kl to a path is given by its weight w_{kl}.[3] In this setting, higher weights indicate less important links in that they correspond to greater agent distances. This is without loss of generality, since any function of w_{ij} can also be used to construct d_V. However, this chapter does require that d_V is chosen such that X is *locally finite* under d_V. That is, for each $i \in V$ the set $\{j : d_V(i,j) \leq r\}$ is almost surely finite for any $r > 0$.

A *rooted network* $X_\rho = (X, \rho)$ is a network with a distinguished agent $\rho \in V$, called the root. It is helpful to think of X_ρ as the network X "from the point of view" of agent ρ. For any $r > 0$, X_ρ^r denotes the r-neighborhood of X_ρ: the rooted subnetwork induced by the agents, weights, and covariates within agent distance r of ρ.[4] Two rooted networks X_{ρ_1} and Y_{ρ_2} are *isomorphic* ($X_{\rho_1} \simeq Y_{\rho_2}$) if all of their r-neighborhoods are equivalent up to a relabeling of the non-rooted agents. That is, for any $r > 0$ there exists a bijection $f : V(X)^r_{\rho_1} \to V(Y)^r_{\rho_2}$ such that $f(\rho_1) = \rho_2$, $w_{ij} = w_{f(i)f(j)}$, and $c_{ik} = c_{f(i)k}$ for any $i, j \in V(X)^r_{\rho_1}$ and covariate k. X^0_ρ, the rooted network with radius 0 around ρ, is equivalent to $\left(\{\rho\}, \emptyset, \{c_{\rho k}\}_{k=1}^K\right)$ which is a network with one agent and the vector of covariates associated with that agent $\{c_{\rho k}\}_{k=1}^K$.

If two rooted networks are not isomorphic, they can be assigned a strictly positive network distance inversely proportional to the maximum r such that they have equivalent

[3] Formally, a path is just a finite sequence of agents. Let P_{ij} denote the set of all paths beginning at agent i and ending at agent j. For $p \in P_{ij}$, I write $|p|$ for the total number of agents in the path and p_t denote identity (in V) of the agent at the tth position of the path. Then the shortest path metric $d_V(i,j) := \inf_{p \in P_{ij}} \sum_{t=1}^{|p|-1} w_{p_t p_{t+1}}$

[4] Formally, $X^r_\rho := (V^r_\rho, W^r_\rho, C^r_\rho)$, where $V^r_\rho := \{i \in V : d_V(\rho, i) \leq r\}$, $W^r_\rho := \{w_{ij} \in W : i, j \in V^r_\rho\}$, and $C^r_\rho := \{c_{ik} \in C : i \in V^r_\rho\}$.

r-neighborhoods under $\simeq$. Formally, d_X defines a metric on the set of rooted networks

$$d_X(X_{\rho_1}, X_{\rho_2}) := \inf\{(1+r)^{-1} : X^r_{\rho_1} \simeq X^r_{\rho_2}\} \times 1\{X^0_{\rho_1} \simeq X^0_{\rho_2}\} + 2 \times 1\{X^0_{\rho_1} \not\simeq X^0_{\rho_2}\} \quad (2.1)$$

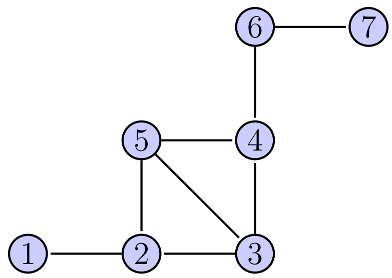
where two networks have a network distance of 2, for example, if their roots have different vectors of covariates ($c_{\rho_1 k} \neq c_{\rho_2 k}$ for some k), which implies they have different r-neighborhoods for any positive r. Notice $d_X(X_\rho, X^r_\rho) \leq (1+r)^{-1}$ by construction.

Let $\mathcal{X}$ denote the set of $\simeq$-equivalence classes of rooted networks equipped with the metric d_X. $\mathcal{X}$ is a Polish space: a separable and complete metric space. The topology on $\mathcal{X}$ induced by d_X is called the *local topology* and $\mathcal{P}(\mathcal{X})$ denotes the space of probability measures based on this topology. Weak convergence for a sequence of measures $\pi_n \in \mathcal{P}(\mathcal{X})$ is called *local weak convergence*, which is defined in the usual way $\pi_n \to_w \pi \iff \int_{x\in\mathcal{X}} f(x)\pi_n(dx) \to \int_{x\in\mathcal{X}} f(x)\pi(dx)$ for all bounded, continuous $f : \mathcal{X} \to \mathbb{R}$. For any random rooted network, X_ρ, $X^r_\rho \to_p X_\rho$ as $r \to \infty$ by construction, where two random rooted networks converge in probability if their laws are local weak convergent.

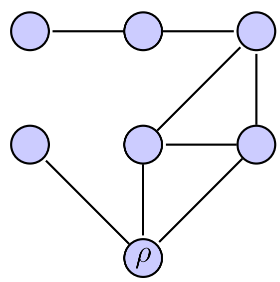
Two comments are necessary. First, to avoid ambiguity when referring to elements or distributions on $\mathcal{X}$, an equivalence class of networks is represented by its canonical form: an unlabeled configuration of agents, weights, and covariates surrounding a root. Second, $\mathcal{X}$ is easily extended to include networks rooted at multiple agents by defining the r-neighborhood around some subset of agents S to be all of the agents, weights, and covariates within agent distance r of any agent in S and extending the notion of the network isomorphism in an analogous way. Two rooted networks with a different number of roots cannot have isomorphic r-neighborhoods for any r, and so under the network metric d_X, their network distance is set equal to 2. Otherwise, the two settings are essentially equivalent.

The logic behind the local approximation is illustrated by two examples in Figures 2.3 and 2.4. The top panel of Figure 2.3 depicts a network with seven agents with no agent covariates. A link between a pair of agents indicates a link weight of 1 while the absence of a link between a pair of agents indicates a link weight of infinity. The middle panel depicts a network that is equivalent to the network in the top panel, rooted at the second agent. The root of the network is denoted by a ρ. The bottom panel depicts a network that is of network distance 1/3 from the network in the top panel rooted at the second agent. Notice that the networks formed by rooting the network in the top panel at agents 3 and 5 are isomorphic. The networks formed by rooting the network in the top panel at agents 2 and 4 have a network distance of 1/2.

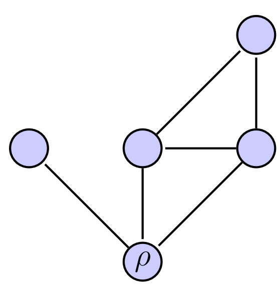
The top panel of Figure 2.4 similarly depicts a network with eight agents. The middle panel depicts the network formed by rooting the network from the top panel at any of the eight agents. The bottom panel depicts the rooted network formed by truncating the network in the middle panel at a radius of 2.



(a) A network with seven agents.

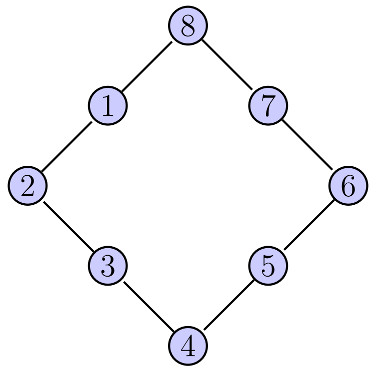


(b) This network is equivalent to the above network rooted at agent 2.

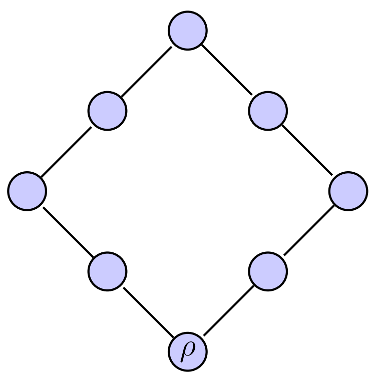


(c) This network is equivalent to the above network rooted at agent 2 and truncated at a radius of 2.

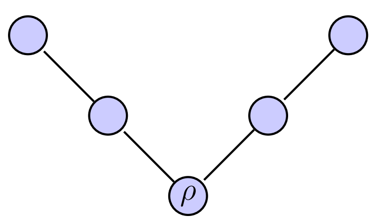
Figure 2.3:



(a) A network with eight agents.



(b) This network is equivalent to the above network rooted at any of the eight agents.



(c) This network is equivalent to the above network rooted at any of the eight agents and truncated at a radius of 2.

Figure 2.4:

2.3 The Nonparametric Network Regression

This section applies the local approximation to the problem of specifying and estimating a nonparametric regression of a scalar outcome on a sparse network. Let (y_ρ, X_ρ) be a random element of $\mathbb{R} \times \mathcal{X}$. The pair are related by the following model

$$y_\rho = m(X_\rho) + \varepsilon_\rho \tag{2.2}$$

where $m : \mathcal{X} \to \mathbb{R}$ is the unknown regression function such that for a given $x \in \mathcal{X}$, $m(x) := E[y_\rho | d(X_\rho, x) = 0]$ such that $E\left[\varepsilon_\rho | d(X_\rho, x) = 0\right] = 0$.

Let $\{y_i, X_i\}_{i=1}^n$ be a sequence of independent observations with the same distribution as (y_ρ, X_ρ) and x an arbitrarily element of $\mathcal{X}$. This section considers the estimation $m(x)$. One way to interpret x is as a network for which an outcome has not yet been observed.

This section studies the following smoothed nearest-neighbor estimator for $m(x)$

$$\hat{m}(x) = (nq_n)^{-1} \sum_{i=1}^{n} y_i K\left(F_n\left(d_X(X_i, x)\right)/q_n\right)$$

where F_n is the empirical distribution function of $d_X(X_i, x)$ (that is, $F_n\left(d_X(X_i, x)\right) = n^{-1} \sum_{j=1}^n \mathbb{1}\{d_X(X_j, x) \leq d_X(X_i, x)\}$), K is a kernel density function, and q_n a bandwidth sequence chosen by the researcher. This section demonstrates consistency and asymptotic normality of $\hat{m}(x)$ under the following assumptions.

Assumption 1 *The sequence of data $\{y_i, X_i\}_{i=1}^n$ is independent and identically distributed. F_X, the marginal distribution of $d(X_i, x)$ is strictly monotonic and smooth.*

Assumption 2 *The outcome variable has finite second moments: $E[y_i^2] \leq \infty$. The regression function $E[y_i | d_X(X_i, x)] = E[m(X_i) | d_X(X_i, x)]$ is smooth to the right of* 0

Assumption 3 *The kernel density function K is twice differentiable and satisfies $\int K(u)du = 1$ and $\int uK(u)du = \int u^2 K(u)du = 0$. The bandwidth sequence q_n satisfies $q_n^2 n \to \infty$ and $q_n^7 n \to 0$.*

Assumption 1 as states is particularly strong and requires a discussion. The independent and identically distributed assumption on the sequence of rooted networks can be weakened to infinite exchangeability and F_X redefined to be the (possibly random) limit of the empirical distribution function $F_{i,n}$, see for instance Proposition 1.4 of Kallenberg (2006). I am currently working on extending the proof of Theorem 1 to allow for spatially correlated rooted networks and errors. This is for the special case in which the data is generated by taking one large network on n agents and generating n different networks by rooting the network at each of the n distinct agents. Agents who are nearby in the network (in the sense of the metric d_V) are likely to have related rooted networks (in the d_X snese) and errors. Under the assumption of local finiteness, this sort of dependence can be incorporated by substituting the central limit theorem applied below for one that accommodates weak dependence.

The main result is given as Theorem 1

Theorem 1 *Suppose Assumptions 1 through 3 hold. Then*

$$\sqrt{nq_n/V_n}\,(\hat{m}_n(x) - m(x)) \to_d \mathcal{N}(0,1)$$

where $V_n = V_{1,n} + V_{2,n} + 2C_n$ *and*

$$\begin{aligned}
V_{1,n} &= q_n^{-3} E\left[y_{i_1} y_{i_2} K'\left(\frac{F_X(d_X(X_{i_1},x))}{q_n}\right) K'\left(\frac{F_X(d_X(X_{i_2},x))}{q_n}\right) F_X(d_X(X_{i_1},x) \wedge d_X(X_{i_2},x))\right] \\
V_{2,n} &= q_n^{-1} E\left[y_i^2 K\left(\frac{F_X(d_X(X_i,x)}{q_n}\right)^2\right] \\
C_n &= q_n^{-2} E\left[y_i K'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right) E\left[y_j K\left(\frac{F_X(d_X(X_j,x))}{q_n}\right)(\mathbb{1}\{d_X(X_j,x) \le d_X(X_i,x)\}\right.\right. \\
&\qquad \left.\left. -\ F_X(d_X(X_i,x)))\,|X_i\right]\right]
\end{aligned}$$

Proof 1 (of Theorem 1) *Let* $m_{q_n}(x) = q_n^{-1} E\left[y_i K\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\right]$ *and* $\tilde{m}_n(x) = (nq_n)^{-1}\sum_{i=1}^n y_i K\left(\frac{F_X(d_X(X_i,x)}{q_n}\right)$. *Decompose* $(\hat{m}_n(x) - m(x)) = (\hat{m}_n(x) - \tilde{m}_n(x)) + (\tilde{m}_n(x) - m_{q_n}(x)) + (m_{q_n}(x) - m(x)) = I_1 + I_2 + I_3$.

First, $\sqrt{nq_n} I_1 =$ $(nq_n^3)^{-1/2}\sum_{i=1}^n y_i K'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)[F_n(d_x(X_i,x)) - F_X(d_X(X_i,x))] + o_p(1)$ *since*

$$\begin{aligned}
I_1 &= (nq_n)^{-1}\sum_{i=1}^n y_i\left[K\left(\frac{F_n(d_X(X_i,x))}{q_n}\right) - K\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\right] \\
&= (nq_n^2)^{-1}\sum_{i=1}^n y_i K'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)[F_n(d_x(X_i,x)) - F_X(d_X(X_i,x))] \\
&\quad + (nq_n^2)^{-1}\sum_{i=1}^n y_i K''\left(\frac{\iota_i}{q_n}\right)[F_n(d_X(X_i,x)) - F_X(d_X(X_i,x))]^2
\end{aligned}$$

where ι_i *is an intermediate value between* $F_n(d_x(X_i,x))$ *and* $F_X(d_X(X_i,x)$. *Since*

$$\max_{i=1,\ldots,n}[F_n(d_x(X_i,x)) - F_X(d_X(X_i,x))]^2\,\mathbb{1}\{F_X(d_X(X_i,x)) \le q_n\} = o_p(n^{\gamma-1}q_n)$$

for any $\gamma > 0$ *by Lemma 1, the second term is* $o_p(1/\sqrt{nq_n})$.

Second, $\sqrt{nq_n}I_3 = o_p(1)$ by Lemma 2. As a result, $I_1+I_2+I_3$ is asymptotically equivalent to

$$I_4 = (nq_n)^{-1}\sum_{i=1}^{n} y_i \left[K\left(\frac{F_n\left(d_X(X_i,x)\right)}{q_n}\right) - K\left(\frac{F_X\left(d_X(X_i,x)\right)}{q_n}\right)\right]$$
$$+(nq_n)^{-1}\sum_{i=1}^{n}\left(y_i K\left(\frac{F_X(d_X(X_i,x)}{q_n}\right) - E\left[y_i K\left(\frac{F_X\left(d_X(X_i,x)\right)}{q_n}\right)\right]\right)$$

$\sqrt{nq_n/V_n}I_4 \to_d \mathcal{N}(0,1)$ *follows from Lemma 3, which completes the proof.* $\square$

The proof of Theorem 1 is essentially just a decomposition of the difference between the prediction $\hat{m}_n(x)$ and the conditional mean $m(x)$ into three parts. The first part concerns the deviation of F_n around its probability limit F_x. The second part concerns the deviation of the average outcome around its mean. The third part is a bias term that is asymptotically negligible due to the smoothness assumption on m given by Assumption 2 and the choice of kernel density function and bandwidth sequence given by Assumption 3. The theorem then follows from the usual central limit theorem for V-Statistics (see, for example, Chapter 5 of Serfling (2009)). The estimator and proof are along the lines of those considered by Stute (1984).

2.4 Lemmas and Proofs

Lemma 1 *For any $\gamma > 0$*

$$\max_{i=1,\ldots,n}\left[F_n(d_X(X_i,x)) - F_X(d_X(X_i,x))\right]\mathbb{1}\{F_X(d_X(X_i,x)) \le q_n\} = o_p\left(n^{(\gamma-1)/2}q_n^{1/2}\right)$$

Proof 2 (of Lemma 1) *First fix agent i and some $\epsilon > 0$. Then Bernstein's Inequality implies $P\left(|F_n(d_X(X_i,x)) - F_X(d_X(X_i,x))|\,\mathbb{1}\{F_X(d_X(X_i,x)) \le q_n\} \ge \epsilon\right) \le 2\exp\left(\frac{-n\epsilon^2}{2(\sigma^2+\epsilon/3)}\right)$ where $\sigma^2 = n^{-1}\sum_{i=1}^{n}F_X(d_X(X_i,x))(1-F_x(d_X(X_i,x))\mathbb{1}\{F_X(d_X(X_i,x)) \le q_n\} \le q_n$. The union bound then implies for a fixed $\epsilon > 0$ that*

$$P\left(\max_{i=1,\ldots,n}|F_n(d_X(X_i,x)) - F_X(d_X(X_i,x))|\,\mathbb{1}\{F_X(d_X(X_i,x)) \le q_n\} \ge q_n^{1/2}n^{(\gamma-1)/2}\epsilon\right)$$
$$\le 2n\exp\left(\frac{-n^{\gamma}\epsilon^2}{2(1+q_n^{1/2}n^{(\gamma-1)/2}\epsilon/3}\right) \le 2n\exp\left(-n^{\gamma}\epsilon^2/3\right)$$

which converges to 0 for any $\gamma > 0$. $\square$

Lemma 2 *The conditional expectation*

$$\sqrt{nq_n}E\left[(m(X_i)-m(x))\,K\left(\frac{F_X\left(d_X(X_i,x)\right)}{q_n}\right)\right] = o\left(1\right)$$

Proof 3 (of Lemma 2)

$$E\left[(m(X_i) - m(x))\, K\left(\frac{F_X\left(d_X(X_i, x)\right)}{q_n}\right)\right] = E\left[\left(\mu\left(F_X^{-1}(q_n u_i)\right) - \mu(0)\right) K\left(u_i\right)\right]$$
$$= (\mu \circ F_X^{-1})'(0) q_n E\left[u_i K\left(u_i\right)\right] + (\mu \circ F_X^{-1})''(0) q_n^2 E\left[u_i^2 K\left(u_i\right)\right] + O\left(q_n^3\right)$$

where $u_i = F_X\left(d_X(X_i, x)\right)/q_n$ *has a standard uniform distribution. Since* $E\left[u_i K\left(u_i\right)\right]$ *and* $E\left[u_i^2 K\left(u_i\right)\right]$ *are* 0 *by choice of kernel, the result follows from* $n^{1/7} q_n \to 0$. $\square$

Lemma 3

$$\sqrt{nq_n}\begin{pmatrix} V_{1,n} & C_n \\ C_n & V_{2,n}\end{pmatrix}^{-1/2}\begin{pmatrix} \frac{1}{nq_n^2}\sum_{i=1}^n y_i K'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\left[F_n\left(d_x(X_i, x)\right) - F_X(d_X(X_i, x))\right] \\ \frac{1}{nq_n}\sum_{i=1}^n\left(y_i K\left(\frac{F_X(d_X(X_i,x)}{q_n}\right) - E\left[y_i K\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\right]\right)\end{pmatrix}$$

converges in distribution to a $\mathcal{N}(0, I_2)$ *where* I_2 *is the* 2*-dimensional identity matrix.*

Proof 4 (of Lemma 3) $(nq_n^2)^{-1}\sum_{i=1}^n y_i K'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\left[F_n\left(d_x(X_i, x)\right) - F_X(d_X(X_i, x))\right] = (nq_n)^{-2}\sum_{i=1}^n\sum_{j=1}^n y_i K'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\left[\mathbb{1}\{d_X(X_j, x) \le d_X(X_i, x)\} - F_X(d_X(X_i, x))\right]$. *As a result,*

$$\begin{pmatrix} \frac{1}{nq_n^2}\sum_{i=1}^n y_i K'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\left[F_n\left(d_x(X_i, x)\right) - F_X(d_X(X_i, x))\right] \\ \frac{1}{nq_n}\sum_{i=1}^n\left(y_i K\left(\frac{F_X(d_X(X_i,x)}{q_n}\right) - E\left[y_i K\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\right]\right)\end{pmatrix}$$

is a 2*-dimensional V-statistic with a non-symmetric kernel function and mean* $O_p(1/n^{1/2} q_n^{1/2}) = o_p(1/\sqrt{nq_n})$. *The variance of the first term is given by*

$$\frac{1}{n^4 q_n^4} E\left[\sum_{i_1}\sum_{i_2}\sum_{j_1}\sum_{j_2} y_{i_1} y_{i_2} K'\left(\frac{F_X(d_X(X_{i_1}, x))}{q_n}\right) K'\left(\frac{F_X(d_X(X_{i_2}, x))}{q_n}\right)\right.$$
$$\left.\times\left[\mathbb{1}\{d_X(X_{j_1}, x) \le d_X(X_{i_1}, x)\} - F_X(d_X(X_{i_1}, x))\right]\left[\mathbb{1}\{d_X(X_{j_2}, x) \le d_X(X_{i_2}, x)\} - F_X(d_X(X_{i_2}, x))\right]\right]$$
$$= \frac{1}{nq_n^4} E\left[y_{i_1} y_{i_2} K'\left(\frac{F_X(d_X(X_{i_1}, x))}{q_n}\right) K'\left(\frac{F_X(d_X(X_{i_2}, x))}{q_n}\right)\right.$$
$$\left.\times\left[F_X(d_X(X_{i_1}, x) \wedge d_X(X_{i_2}, x)) - F_X(d_X(X_{i_1}, x)) F_X(d_X(X_{i_2}, x))\right]\right] + O_p\left(1/\left(n^2 q_n^4\right)\right)$$
$$= V_{1,n} + O_p(1/n) + O_p\left(1/\left(n^2 q_n^4\right)\right)$$

The variance of the second term is given by $(nq_n^2)^{-1} E\left[y_i^2 K\left(\frac{F_X(d_X(X_i,x)}{q_n}\right)^2\right] + o_p((nq_n)^{-1}) =$

$V_{2,n} + o_p((nq_n)^{-1})$. The covariance between the two terms is given by

$$
\begin{aligned}
&E\left[\frac{1}{n^3q_n^3}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j=1}^{n} y_{i_1}K'\left(\frac{F_X(d_X(X_{i_1},x))}{q_n}\right)[\mathbb{1}\{d_X(X_j,x)\le d_X(X_{i_1},x)\} - F_X(d_X(X_{i_1},x))]\right.\\
&\qquad\qquad \left.\times\left(y_{i_2}K\left(\frac{F_X(d_X(X_{i_2},x)}{q_n}\right) - E\left[y_iK\left(\frac{F_X\left(d_X(X_i,x)\right)}{q_n}\right)\right]\right)\right]\\
&= \frac{1}{nq_n^3}E\left[y_iK'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)[\mathbb{1}\{d_X(X_j,x)\le d_X(X_i,x)\} - F_X(d_X(X_i,x))]\right.\\
&\qquad\qquad \left.\times\left(y_jK\left(\frac{F_X(d_X(X_j,x)}{q_n}\right) - E\left[y_iK\left(\frac{F_X\left(d_X(X_i,x)\right)}{q_n}\right)\right]\right)\right] + O_p(1/n^2q_n^3)\\
&= \frac{1}{nq_n^3}E\left[y_iK'\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\left[E\left[\mathbb{1}\{d_X(X_j,x)\le d_X(X_i,x)\}y_jK\left(\frac{F_X(d_X(X_j,x))}{q_n}\right)|X_i\right]\right.\right.\\
&\qquad\qquad \left.\left.-F_X(d_X(X_i,x))E\left[y_iK\left(\frac{F_X(d_X(X_i,x))}{q_n}\right)\right]\right]\right] + O_p(1/n^2q_n^3)\\
&= C_n + O_p(1/n^2q_n^3)
\end{aligned}
$$

since y_i, K, and K' are bounded, the Lemma follows from Serfling (2009) Chapter 5.5, Theorem A. The result $\sqrt{nq_n/V_n}I_4 \to_d \mathcal{N}(0,1)$ is then a consequence of the continuous mapping theorem.

Chapter 3

Nonparametric Estimation of a Link Formation Model with Unobserved Heterogeneity

3.1 Introduction

This chapter demonstrates how the codegree matching strategy from chapter one can be used to estimate the infinite-dimensional component of a semiparametric model of link formation. In the considered model, each agent is associated with an unobserved individual heterogeneity term and each agent-pair is associated with an idiosyncratic error and a vector of covariates. The existence of a link between a pair of agents depends on the agents' individual heterogeneity term, their idiosyncratic error, and covariates. The main contribution of this chapter over the literature on graphon estimation (see for example, Bickel and Chen (2009); Bickel et al. (2011)) is the incorporation of link covariates. The semiparametric structure of the model is also novel, but its implications will be explored in more detail in future work.

The chapter proceeds in three steps. In the first step, it defines the conditional link distribution function. The conditional link distribution function is the conditional probability that two agents are linked given their individual heterogeneity terms and their covariates lie in some fixed hyperplane, as indexed by the parameter θ. In the second step, it defines an analog of network distance to a collection of networks defined over the θ-space, and an empirical analog: conditional empirical codegree distance. In the third step, the conditional codegree distance is used to construct an estimator for the conditional link distribution function via local averaging. Both the conditional links distribution function and the conditional empirical codegree distance are shown to converge uniformly over the θ-space and the collection of agent-pairs.

In future work, I plan to show how the conditional link distribution function can be used as a plug-in estimator in various network econometrics problems. In particular, I plan to use the conditional link distribution function to construct a semiparametric maximum likelihood estimator for the parametric component of the link formation model along the lines of Klein and Spady (1993).

The chapter proceeds as follows. The second section describes the model, the third section introduces the conditional codegree distance, and the fourth section introduces the conditional link distribution function.

3.2 The model

For every pair of agents (i, j) in a random sample of n agents, $D_{ij} = 1$ if i and j are linked and $D_{ij} = 0$ otherwise. Each pair of agents is also assigned a vector of link covariates Z_{ij} with compact support $\mathcal{Z} \subset \mathbb{R}^p$. The network links and link covariates are described by the following model

$$\begin{aligned} D_{ij} &= f\left(w_i, w_j, Z_{ij}\beta_0, \varepsilon_{ij}\right) \\ Z_{ij} &= z\left(x_i, x_j, \xi_{ij}\right) \end{aligned} \tag{3.1}$$

in which the individual characteristics $\{w_i, x_i\}_{i=1}^n$ and idiosyncratic errors $\{\xi_{ij}, \varepsilon_{ij}\}_{i \neq j}$ are mutually independent iid random variables, β_0 is an unknown p-dimensional parameter, and f and z are unknown functions. The joint distribution of w_i and x_i is not restricted so that model (3.1) is essentially a semiparametric analog to Graham (2017).

The natural parameter of interest in (3.1) is β_0. This chapter is instead focuses on the conditional link distribution function

$$g\left(\alpha_i, \alpha_j; \beta, \tau\right) = P\left(D_{ij} = 1 | Z_{ij}\beta \leq \tau, \alpha_i, \alpha_j\right)$$

where $\alpha_i = (w_i, x_i)$, $\eta_{ij} = (\varepsilon_{ij}, \xi_{ij})$. Intutitvely, g returns the probability that a pair of agents with individual characteristics α_i and α_j and link covariates that satisfy the restriction $Z_{ij}\beta \leq \tau$ form a link. Here β is not assumed to be equal to β_0. In future work I plan to show how the conditional link distribution function can be used to estimate β_0.

Another object of potential interest is the conditional link density function

$$f\left(\alpha_i, \alpha_j; z\right) = P\left(D_{ij} = 1 | \alpha_i, \alpha_j, Z_{ij}\beta_0 = z\right)$$

which is often interpreted as the marginal transferable utility agents i and j with link covariates z recieve from forming al link. If g is almost everywhere differentiable in its final argument, the former can be used to construct the latter when β_0 is known

$$f\left(\alpha_i, \alpha_j; z\right) = \lim_{h \to 0} \frac{g\left(\alpha_i, \alpha_j; \beta_0, z + h\right) - g\left(\alpha_i, \alpha_j; \beta_0, z - h\right)}{2h}$$

In this way an estimator for the conditional link distirbution function can be used to construct an estimator for the conditional link density function.

Assumptions

Throughout the chapter I will assume that the link density function f is almost everywhere Lipschitz continuous in its arguments. The density functions for the random variabels α_i and η_{ij} are also assumed to be smooth in $\mathbb{R}^2$, with the components of η_{ij} further asumed to have uniform marginals.

Notation

D and Z are $n \times n$ matricies with the ijth entry of D given by D_{ij}, $D_{ij\tau} = D_{ij}\mathbb{1}\{Z_{ij}\delta \leq \tau\}$, $g_{ij\tau} = g(\alpha_i, \alpha_j, \delta, \tau) = E\left[D_{ijz}|\alpha_i, \alpha_j\right]$, $\sum_{t<s} = \sum_{t=1}^{n-1}\sum_{s=t+1}^{n}$, $\sum_{t<s<r} = \sum_{t=1}^{n-2}\sum_{s=t+1}^{n-1}\sum_{r=s+1}^{n}$, $p_{ij\tau} = p\left(\alpha_i, \alpha_j, \tau\right) = E\left[g_{it\tau}g_{jt\tau}|\alpha_i, \alpha_j\right]$, $r_{\tau,n} = r(\tau, h_n) = P\left(\delta\left(\alpha_i, \alpha_j, \tau\right) \leq h_n\right)$ for a fixed bandwidth sequence h_n and δ to be defined in the next section, $r_n = \inf_{\tau\in\mathbb{R}} r_{\tau,n}$, $f_\alpha(\cdot)$ is the (multivariate) density function of α_i, $f_\eta(\cdot)$ is the density function of η_{ij}, $\theta = (\beta, \tau)$ where β is an element of the p-dimensional hypersphere and $\tau \in \mathbb{R}$. Θ refers to the support of the parameter vector θ.

3.3 Conditional codegree distance

I first define a notion of distance on the space of individual characteristics using a pseudo-metric from Auerbach (2017) based on similarities in linking behavior that I call "network distance." Network distance is defined over each θ.

$$\begin{aligned} d\left(\alpha_i, \alpha_j, \theta\right) &= \left(\int \left(g\left(\alpha_i, t, \theta\right) - g\left(\alpha_i, t, \theta\right)\right)^2 f_\alpha(t)dt\right)^{1/2} \\ &= \left(E\left[E\left[\left(D_{it\theta} - D_{jt\theta}\right)|\alpha_i, \alpha_j, \alpha_t\right]^2 |\alpha_i, \alpha_j\right]\right)^{1/2} \end{aligned}$$

The density of α_i does not depend on θ.

For each network in the process indexed by θ, two agents have similar social characteristics if they have the same probability of forming a link with all other agents (as indexed by their individual characteristics) in the economy.

Network distance is difficult to directly estimate using only a random sample of D and Z. Auerbach (2017) proposes a feasible alternative, "codegree distance." As with conditional network distance, conditional codegree distance is defined over each θ

$$\begin{aligned} \delta\left(\alpha_i, \alpha_j, \theta\right) &= \left(\int \left(p\left(\alpha_i, t, \theta\right) - p\left(\alpha_i, t, \theta\right)\right)^2 f_\alpha(t)dt\right)^{1/2} \\ &= \left(E\left[E\left[D_{ts\theta}\left(D_{is\theta} - D_{js\theta}\right)|\alpha_i, \alpha_j, \alpha_t\right]^2 |\alpha_i, \alpha_j\right]\right)^{1/2} \end{aligned}$$

An application of Jensen's inequality reveals that $\delta(\alpha_i, \alpha_j, \theta) \leq d(\alpha_i, \alpha_j, \theta)$ everywhere, so that codegree distance is a weakly coarser notion of distance than network distance. The two notions of distance are in fact weakly equivalent under Lipschitz continuity of g and f_α so that

Lemma 1: Under the stated assumptions

$$d(\alpha_i, \alpha_j, \theta) \leq C\delta(\alpha_i, \alpha_j, \theta)^{1/4}$$

for some universal $C > 0$ and exactly every $\theta \in \Theta$

The proof of Lemma 1 follows directly from the proof of Lemma 2 in the first chapter of this thesis. Conditional codegree distance has a natural empirical analog, "empirical conditional codegree distance"

$$\hat{\delta}^2_{ij\tau} = \binom{n-2}{3}^{-1} \sum_{t<s_1<s_2} D_{ts_1\tau} D_{ts_2\tau} \left(D_{is_1\tau} - D_{js_1\tau}\right)\left(D_{is_2\tau} - D_{js_2\tau}\right) \mathbb{1}\{t, s_1, s_2 \neq i, j\}$$

For a fixed (i, j)-pair, empirical codegree distance behaves similarly to a third-order U-process (in the sense of Nolan and Pollard 1988) indexed by θ so that

Lemma 2: Under the stated assumptions

$$\sup_{\theta\in\Theta} \max_{i\neq j} \left(\hat{\delta}^2_{ijz} - \delta(\alpha_i, \alpha_j, \tau)^2\right) = o_p\left(\sqrt{\frac{\log(n)}{n}}\right)$$

Theorem 1 essentially follows from an extension of Hoeffding's inequality to dependency graphs due to Janson (2004).

Proof of lemma 2: Fix a pair of agents (i, j) and a parameter $\theta \in \Theta$. Theorem 2.1 of Janson (2004) implies

$$P\left(\left|\hat{\delta}^2_{ij\theta} - \delta(\alpha_i, \alpha_j, \theta)^2\right| > t\right) \leq 2\exp\left(-t^2(n-2)/2\right)$$

Boole's inequality gives

$$P\left(\sup_{\theta\in\Theta} \max_{i\neq j} \left|\hat{\delta}^2_{ij\theta} - \delta(\alpha_i, \alpha_j, \theta)^2\right| > t\right) \leq 2C_n \exp\left(-t^2(n-2)/2\right)$$

where C_n is the cardinality of the support of $\hat{\delta}_{ij\theta}$. Familiar caluclations reveal C_n is $o\left(n^{2(p+2)}\right)$ since the linear classifier $\{Z_{ij}\beta - \tau \leq 0\}$ has VC-dimension less than $p+1$ and g is assumed to be Lipschitz continuous in its third argument. The claim follows from the choice of $t^2 = \frac{2(p+4)\log(n)}{n-2}$.
□

One can also demonstrate that the conditional empirical codegree distance also converges in distribution to a Gaussian process for a finite collection of pairs of agents over Θ at the $\sqrt{n}$-rate (or that the process converges in distribution at that rate for an average of such statistics). However, the process does not converge weakly uniformly over the $\binom{n}{2}$ pairs of agents at the $\sqrt{n}$-rate, because the statistic fails stochastic equicontinuity. This is because the idiosyncratic errors ε_{is} and $\varepsilon_{i's}$ are not similar even when α_i is close to $\alpha_{i'}$.

3.4 Conditional link distribution function

Under the assumption that g is continuous in its arguments, the function $g(w_i, w_j, \theta)$ can be estimated nonparametrically by a local average of the network links using the empirical codegree distance. Let

$$\hat{g}_{it\theta} = \sum_{s=1}^{n} D_{is\theta} k\left(\frac{\hat{\delta}_{ts\theta}}{h_n}\right) / \sum_{s=1}^{n} k\left(\frac{\hat{\delta}_{ts\theta}}{h_n}\right)$$

where k is a continuously differentiable kernel density function and h_n is a bandwidth sequence assumed to be $O\left(n^{-3/7}\right)$. Then

Lemma 3: Under the stated assumptions

$$\sup_{\theta\in\Theta}\max_{i\neq t}\left(\hat{g}_{itz} - g(w_i, w_t, z)\right) = o_p\left(\sqrt{\frac{\log(n)}{nh_n^2}}\right)$$

The proof of lemma 3 essentially follows from lemmas 1 and 2 and the arguments in their proofs.

Proof of lemma 3: Lemma 2 and the smoothness assumption on k implies

$$\sup_{\theta\in\Theta}\max_{i\neq t}\left(\frac{1}{n}\sum_{s=1}^{n} D_{is\theta}^{\iota} k\left(\frac{\hat{\delta}_{ts\theta}}{h_n}\right) - \frac{1}{n}\sum_{s=1}^{n} D_{is\theta}^{\iota} k\left(\frac{\delta_{ts\theta}}{h_n}\right)\right) = o_p\left(\sqrt{\frac{\log(n)}{nh_n^2}}\right)$$

for $\iota \in \{0, 1\}$. Hoeffding's inequality implies that

$$\sup_{\theta\in\Theta}\max_{i\neq t}\left(\frac{1}{n}\sum_{s=1}^{n} D_{is\theta}^{\iota} k\left(\frac{\delta_{ts\theta}}{h_n}\right) - \frac{1}{n}\sum_{s=1}^{n} g_{is\theta}^{\iota} k\left(\frac{\delta_{ts\theta}}{h_n}\right)\right) = o_p\left(\sqrt{\frac{\log(n)}{n}}\right)$$

since $\frac{1}{n}\sum_{s=1}^{n} k\left(\frac{\delta_{ts\theta}}{h_n}\right) = r_n + o_p\left(r_n\right)$. Lemma 1 and continuity on g implies that

$$\sup_{\theta\in\Theta}\max_{i\neq t}\left(\frac{1}{n}\sum_{s=1}^{n} g_{is\theta} k\left(\frac{\delta_{ts\theta}}{h_n}\right) - \frac{1}{n}\sum_{s=1}^{n} g_{it\theta} k\left(\frac{\delta_{ts\theta}}{h_n}\right)\right) = o_p\left(r_n h_n^{1/6}\right).$$

which is $o_p\left(\sqrt{\frac{\log(n)}{nh_n^2}}\right)$ by choice of the bandwidth sequence. The result follows from the continuous mapping theorem. □

Lemma 3 is consistent both with the classic literature on nonparametric regression and recent results on graphon estimators without link covariates (for example, Chan and Airoldi 2014 or Zhang, Levina and Zhao (2016)). The first part demonstrates that the estimator is uniformly consistent over the covariate space and agent-pairs at a rate slightly slower than $(nr_n)^{-1/2}$. Here r_n is the "small deviation" probability that characterizes the probability that two agents have similar linking behavior. More detail on this object is provided in the first chapter.

Bibliography

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Hyungtaik Ahn. Semiparametric estimation of a single-index model with nonparametrically generated regressors. *Econometric Theory*, 13(01):3–31, 1997.

Hyungtaik Ahn and James L Powell. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1-2):3–29, 1993.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

David Aldous and J Michael Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, pages 1–72. Springer, 2004.

Joshua D Angrist. The perils of peer effects. *Labour Economics*, 30:98–108, 2014.

Tiziano Arduini, Eleonora Patacchini, and Edoardo Rainone. Parametric and semiparametric iv estimation of network models with selectivity. Technical report, Einaudi Institute for Economics and Finance (EIEF), 2015.

Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. Who's who in networks. wanted: the key player. *Econometrica*, 74(5):1403–1417, 2006.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.

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

Peter J Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. A distributional framework for matched employer employee data. Technical report, Technical Report, 2015.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

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

Scott E Carrell, Bruce I Sacerdote, and James E West. From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882, 2013.

Gary Chamberlain. Notes on semiparametric regression. *manuscript, Department of Economics, University of Wisconsin-Madison*, 1986.

Arun G Chandrasekhar and Matthew O Jackson. Tractable and consistent random graph models. 2014.

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003.

Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

Lorenzo Ductor, Marcel Fafchamps, Sanjeev Goyal, and Marco J van der Leij. Social networks and research output. *Review of Economics and Statistics*, 96(5):936–948, 2014.

Richard M Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

Marijtje AJ Duijn, Tom AB Snijders, and Bonne JH Zijlstra. p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2):234–254, 2004.

Andreas Dzemski. An empirical model of dyadic link formation in a network with unobserved heterogeneity. Technical report, University of Manheim Working Paper, 2014.

Matthew Elliott, Benjamin Golub, and Matthew O Jackson. Financial networks and contagion. *The American economic review*, 104(10):3115–3153, 2014.

Robert F Engle, Clive WJ Granger, John Rice, and Andrew Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394):310–320, 1986.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

Paul Goldsmith-Pinkham and Guido W Imbens. Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264, 2013.

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

Bryan S Graham. An empirical model of network formation: detecting homophily when agents are heterogenous. 2014.

Bryan S Graham. Methods of identification in social networks. *Annu. Rev. Econ.*, 7(1): 465–485, 2015.

Bryan S Graham and Jinyong Hahn. Identification and estimation of the linear-in-means model of social interactions. *Economics Letters*, 88(1):1–6, 2005.

Jinyong Hahn and Hyungsik Roger Moon. Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(03):863–881, 2010.

Jinyong Hahn and Geert Ridder. Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, 81(1):315–340, 2013.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

Paul W Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50, 1981.

Seok Young Hong and Oliver B Linton. Asymptotic properties of a nadaraya-watson type estimator for regression functions of infinite order. *Available at SSRN 2766822*, 2016.

Bo E Honoré and James Powell. *Pairwise difference estimators for nonlinear models*. na, 1997.

Chih-Sheng Hsieh and Lung Fei Lee. A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 2014.

Matthew O Jackson. Networks in the understanding of economic behaviors. *The Journal of Economic Perspectives*, 28(4):3–22, 2014.

IDA Johnsson and HYUNGSIK ROGER Moon. Estimation of peer effects in endogenous social networks: Control function approach. 2015.

Olav Kallenberg. *Probabilistic symmetries and invariance principles*. Springer Science & Business Media, 2006.

Pavel N Krivitsky, Mark S Handcock, Adrian E Raftery, and Peter D Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks*, 31(3):204–213, 2009.

Michael P Leung. Two-step estimation of network-formation models with incomplete information. *Journal of Econometrics*, 188(1):182–195, 2015.

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

László Lovász and Balázs Szegedy. Szemerédis lemma for the analyst. *GAFA Geometric And Functional Analysis*, 17(1):252–270, 2007.

László Lovász and Balázs Szegedy. Regularity partitions and the topology of graphons. In *An irregular mind*, pages 415–446. Springer, 2010.

Charles F Manski. Semiparametric analysis of random effects linear models from binary panel data. *Econometrica: Journal of the Econometric Society*, pages 357–362, 1987.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Elias Masry. Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications*, 115(1):155–177, 2005.

Konrad Menzel. Strategic networkformation with many agents. 2015.

Carl Nadler. Networked inequality: Evidence from freelancers. 2016.

Whitney K Newey. Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics*, 38(3):301–339, 1988.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

James L Powell. *Semiparametric estimation of bivariate latent variable models*. University of Wisconsin–Madison, Social Systems Research Institute, 1987.

James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

Geert Ridder and Shuyang Sheng. Estimation of large network formation games. 2015.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–54, 1988.

Ian M Schmutte. Free to move? a network analytic approach for learning the limits to job mobility. *Labour Economics*, 29:49–61, 2014.

Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40 (2):211–239, 2011.

Winfried Stute. Asymptotic normality of nearest neighbor regression function estimates. *The Annals of Statistics*, 12(3):917–926, 1984.

Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.

Appendix A

Proofs of Various Lemmas and Theorems

This section contains proofs of the various Lemmas and Theorems from Section 3. Auxiliary lemmas that are not formally stated in the paper are labelled Lemma A1, Lemma A2, et cetera.

Lemmas and Theorems in Section 3.2

Theorem 1: Suppose Assumptions 1-3 hold. Then β is the unique minimizer of $E\left[((y_i - y_j) - (x_i - x_j)b)^2 \mid ||f_{w_i} - f_{w_j}||_2 = 0\right]$ over $b \in \mathbb{R}^k$.

Proof of Theorem 1:

$$\begin{aligned}
&E\left[((y_i - y_j) - (x_i - x_j)b)^2 \mid ||f_{w_i} - f_{w_j}||_2 = 0\right] = E\left[((x_i - x_j)(\beta - b) + (u_i - u_j))^2 \mid ||f_{w_i} - f_{w_j}||_2 = 0\right] \\
&= (\beta - b)' E[(x_i - x_j)'(x_i - x_j)] \mid ||f_{w_i} - f_{w_j}||_2 = 0](\beta - b) + E[(u_i - u_j)^2 \mid ||f_{w_i} - f_{w_j}||_2 = 0] \\
&\qquad - 2(\beta - b)' E[(x_i - x_j)'(u_i - u_j)] \mid ||f_{w_i} - f_{w_j}||_2 = 0]
\end{aligned}$$

The first summand is unique minimized at $b = \beta$ by Assumption 3. The second summand does not depend on b. The third summand is equal to 0 by Assumption 2. Assumptions 2 and 3 are also necessary: if either assumption fails the sum of the first and third terms may be minimized at a b that is not equal β. $\square$

Lemmas and Theorems in Section 3.3.1

Lemma 1: Suppose Assumptions 1 and 5 hold. Then

$$\max_{(i \neq j)} \left| \hat{\delta}_{ij} - ||p_{w_i} - p_{w_j}||_2 \right| = o_{a.s.}(n^{-\gamma/4} h_n)$$

Proof of Lemma 1: The lemma is proved in four steps. Set $h'_n = n^{-\gamma/4}h_n$ and recall $h'_n n^{(1-\gamma)/2} \to \infty$, $p_{w_i w_j} = \int f_{w_i}(\tau) f_{w_j}(\tau) d\tau$ and $\hat{p}_{ij} = \frac{1}{n-2}\sum_{t \neq i,j} D_{it}D_{jt}$. I first show that $\max_{(i\neq j)} h_n'^{-1}|\hat{p}_{w_i w_j} - p_{w_i w_j}| \to_{a.s.} 0$. By Bernstein's Inequality, for any $\epsilon > 0$

$$P\left(h_n'^{-1}|\hat{p}_{w_i w_j} - p_{w_i w_j}| > \epsilon\right) = P\left(h_n'^{-1}\left|(n-2)^{-1}\sum_{t\neq i,j}\left(D_{it}D_{jt} - p_{w_i w_j}\right)\right| > \epsilon\right) \leq 2\exp\left(\frac{-(n-2)(h'_n\epsilon)^2}{2 + 2h'_n\epsilon/3}\right)$$

and so by the union bound

$$P\left(\max_{(i\neq j)} h_n'^{-1}|\hat{p}_{w_i w_j} - p_{w_i w_j}| > \epsilon\right) \leq 2n(n-1)\exp\left(\frac{-(n-2)(h'_n\epsilon)^2}{2 + 2h'_n\epsilon/3}\right)$$

Since $(n-2)^{1-\gamma/2}h_n'^2 \to \infty$ by the assumed choice of bandwidth sequence and $\sum_{n=3}^{\infty} n(n-1)\exp\left(\frac{-(n-2)(h'_n\epsilon)^2}{1+2h'_n\epsilon/3}\right) < \infty$ by the ratio test, $P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}|\hat{p}_{w_i w_j} - p_{w_i w_j}| > \epsilon\right) = 0$ follows from the first Borel-Cantelli Lemma. To see the summability claim, note that $(n-2)h_n'^2 > (n-2)^{\gamma}$ and $2h'_n\epsilon < 1$ eventually, so that $\sum_{n=3}^{\infty} n(n-1)\exp\left(\frac{-(n-2)(h'_n\epsilon)^2}{1+2h'_n\epsilon/3}\right)$ is finite if $\sum_{n=3}^{\infty} n(n-1)\exp\left(\frac{-(n-2)^{\gamma}\epsilon^2}{2}\right)$ is. Letting $m(n) = (n-2)^{1/\gamma}$, the latter sum is eventually less than $\sum_{m=1}^{\infty} 2m^{2/\gamma}\exp\left(\frac{-m\epsilon^2}{2}\right) \times |\{n \in \{\mathbb{N}+2\} : n^{\gamma/2} \in (m-1, m]\}| \leq \sum_{m=1}^{\infty} 2m^{4/\gamma}\exp\left(\frac{-m\epsilon^2}{2}\right)$. This final sum is absolutely convergent by the ratio test, for any $\gamma > 0$.

Second, let $||\hat{p}_{w_i} - p_{w_i}||_{2,n,j} = \left((n-2)^{-1}\sum_{s\neq i,j}\left(\hat{p}_{w_i w_s} - p_{w_i w_s}\right)^2\right)^{1/2}$. Then $\max_{(i\neq j)} h_n'^{-1}|\hat{p}_{w_i w_j} - p_{w_i w_j}| \to_{a.s.} 0$ implies $\max_{(i\neq j)} h_n'^{-1}||\hat{p}_{w_i} - p_{w_i}||_{2,n,j} \to_{a.s.} 0$, since

$$\begin{aligned}
&P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}||\hat{p}_{w_i} - p_{w_i}||_{2,n,j} > \epsilon\right)\\
&= P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left((n-2)^{-1}\sum_{s\neq i,j}\left(\hat{p}_{w_i w_s} - p_{w_i w_s}\right)^2\right)^{1/2} > \epsilon\right)\\
&\leq P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left|\hat{p}_{w_i w_j} - p_{w_i w_j}\right| > \epsilon\right)
\end{aligned}$$

because $h_n'^{-1}\left((n-2)^{-1}\sum_{s\neq i,j}\left(\hat{p}_{w_i w_s} - p_{w_i w_s}\right)^2\right)^{1/2} > \epsilon$ only if $h_n'^{-1}\left|\hat{p}_{w_i w_t} - p_{w_i w_t}\right| > \epsilon$ for some $t \neq i, j$.

Third, for $||p_{w_i} - p_{w_j}||_{2,n} = \left((n-2)^{-1}\sum_{s\neq i,j}(p_{w_i w_s} - p_{w_i w_s})^2\right)^{1/2}$, $\max_{(i\neq j)} h_n'^{-1}\left|||p_{w_i} - p_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_2\right| \to_{a.s.} 0$ since

$$\begin{aligned}
&P\left(h_n'^{-1}\left|||p_{w_i} - p_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_2\right| > \epsilon\right)\\
&= P\left(h_n'^{-1}\left|\left((n-2)^{-1}\sum_{s\neq i,j}\left(p_{w_i w_s} - p_{w_j w_s}\right)^2\right)^{1/2} - \left(\int\left(p_{w_i}(s) - p_{w_j}(s)\right)^2 ds\right)^{1/2}\right| > \epsilon\right)\\
&\leq P\left(h_n'^{-1}\left|(n-2)^{-1}\sum_{s\neq i,j}\left(\left(p_{w_i w_s} - p_{w_j w_s}\right)^2 - \int\left(p_{w_i}(s) - p_{w_j}(s)\right)^2 ds\right)\right|^{1/2} > \epsilon\right)\\
&\leq 2\exp\left(\frac{-(n-2)h_n'\epsilon}{2 + 2\sqrt{h_n'\epsilon}/3}\right)
\end{aligned}$$

with the last line again by Bernstein. So

$$P\left(\max_{(i\neq j)} h_n'^{-1}\left|||p_{w_i} - p_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_2\right| > \epsilon\right) \leq 2(n-2)^2\exp\left(\frac{-(n-2)h_n'\epsilon}{2 + 2\sqrt{h_n'\epsilon}/3}\right)$$

which is again absolutely summable for the assumed choice of h_n', since it is eventually bounded above by the summable sequence considered in the first part of this proof.

Finally, the second and third parts of this proof and a few applications of the triangle inequality yield

$$\begin{aligned}
&P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left|||\hat{p}_{w_i} - \hat{p}_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_2\right| > \epsilon\right)\\
&= P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left|||\hat{p}_{w_i} - \hat{p}_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_{2,n} + ||p_{w_i} - p_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_2\right| > \epsilon\right)\\
&\leq P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left|||\hat{p}_{w_i} - \hat{p}_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_{2,n}\right| > \epsilon/2\right)\\
&\qquad\qquad + P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left|||p_{w_i} - p_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_2\right| > \epsilon/2\right)\\
&= P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left|||\hat{p}_{w_i} - \hat{p}_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_{2,n}\right| > \epsilon/2\right)\\
&\leq P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}||(\hat{p}_{w_i} - \hat{p}_{w_j}) - (p_{w_i} - p_{w_j})||_{2,n} > \epsilon/2\right)\\
&\leq 2P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}||(\hat{p}_{w_i} - p_{w_i})||_{2,n,j} > \epsilon/4\right) = 0
\end{aligned}$$

where $P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}\left|||p_{w_i} - p_{w_j}||_{2,n} - ||p_{w_i} - p_{w_j}||_2\right| > \epsilon/2\right)$ in the second equality follows from the third part of the proof, and

$P\left(\limsup_{n\to\infty}\max_{(i\neq j)} h_n'^{-1}||(\hat{p}_{w_i}-p_{w_i})||_{2,n,j}>\epsilon/4\right)$ in the final inequality from the second part of the proof. Since $h_n' = n^{-\gamma/4}h_n$, this completes the argument. $\square$

Lemma 2: Suppose Assumption 1 holds. Then for every (w_i, w_j) pair

$$||p_{w_i} - p_{w_j}||_2 \leq ||f_{w_i} - f_{w_j}||_2$$

Furthermore, for every $\epsilon > 0$ there exists a $\delta > 0$ such that with probability at least $1-\epsilon^2/4$

$$||p_{w_i} - p_{w_j}||_2 \leq \delta \implies ||f_{w_i} - f_{w_j}||_2 \leq \epsilon$$

Proof of Lemma 2: To see the first part, observe that for every (w_i, w_j) pair

$$\begin{aligned}
||p_{w_i} - p_{w_j}||_2^2 &= \int\left(\int f(\tau,s)\left(f(w_i,s)-f(w_j,s)\right)ds\right)^2 d\tau \\
&\leq \int\int\left(f(\tau,s)\left(f(w_i,s)-f(w_j,s)\right)\right)^2 dsd\tau \\
&\leq \int\left(f(w_i,\tau)-f(w_j,\tau)\right)^2 d\tau = ||f_{w_i}-f_{w_j}||_2^2
\end{aligned}$$

where the first inequality is due to Jensen and the second is due to the fact that $|f(\tau,s)|\leq 1$ for every $(\tau,s)\in[0,1]^2$.

The proof of the second part is more complicated. I first note that since f is Lebesgue measurable, Lusin's theorem (Dudley (2002), Theorem 7.5.2) implies that it is almost everywhere equivalent to a uniformly continuous function. That is, for any $\eta' > 0$, f is uniformly continuous when restricted to a closed subset A of $[0,1]^2$ with measure at least $1-\eta'$.

It follows that for any $\eta > 0$ there must also exist B, a closed subset of $[0,1]$ with measure of at least $1-\eta$ such that for any $b\in B$, there exists another closed subset $C(b)$ of $[0,1]$ with measure of at least $1-\eta$, such that for any $c \in C(b)$, f is uniformly continuous when restricted to the set $A' = \{(b,c)\in[0,1]^2 : b\in B, c\in C(b)\}$.

Second, I show that for all $\epsilon' > 0$ there exists a $\delta(\epsilon',\eta) > 0$ such that $||p_{w_i}-p_{w_j}||_2 \leq \delta(\epsilon',\eta)$ implies $|\int f_{w_i}(s)(f_{w_i}(s)-f_{w_j}(s))ds| < \epsilon'$ with probability at least $1-\epsilon'/4$, so long as $\eta \leq \epsilon'/16$.

I prove the contrapositive. Suppose $|\int f_{w_i}(s)(f_{w_i}(s)-f_{w_j}(s))ds| \geq \epsilon'$. Then by the negative triangle inequality $|\int f_\tau(s)(f_{w_i}(s)-f_{w_j}(s))ds| > \epsilon'/2$ for any $\tau\in[0,1]$ chosen such that $|\int(f_\tau(s)-f_{w_i}(s))(f_{w_i}(s)-f_{w_j}(s))ds| < \epsilon'/4$. Since $||f_{w_i}-f_{w_j}||_2 \leq 1$ for every (w_i,w_j) pair, it follows by Cauchy-Schwartz that $||f_{w_i}-f_\tau||_2\leq\epsilon'/4$ implies

$|\int f_\tau(s)(f_{w_i}(s) - f_{w_j}(s))ds| > \epsilon'/2$.

Since f is uniformly continuous when restricted to A', there exists a universal $\omega(\epsilon', \eta) > 0$ such that $|\tau - w_i| < \omega(\epsilon', \eta)$ implies that $||f_\tau - f_{w_i}||_2 < \epsilon'/8 + 2\eta$ so long as $w_i, \tau \in B$. Taking $\eta \leq \epsilon'/16$ gives $|\tau - w_i| < \omega(\epsilon', \eta)$ implies that $||f_\tau - f_{w_i}||_2 < \epsilon'/4$ so long as $w_i, \tau \in B$. It follows that choosing τ such that $|\tau - w_i| < \omega(\epsilon', \eta)$ implies $|\int f_\tau(s)(f_{w_i}(s) - f_{w_j}(s))ds| > \epsilon'/2$

It is without loss to further restrict $\omega(\epsilon', \eta) < \epsilon'/16$. Since w_i is uniformly distributed on $[0, 1]$, the probability that w_i is in the $\epsilon'/16$ interior of B (that is, the interval $(w_i - \epsilon'/16, w_i + \epsilon'16)$ is contained in B) is greater than $1 - \eta - 2\omega(\epsilon', \eta) \geq 1 - \epsilon'/4$. This implies that $|\int f_\tau(s)(f_{w_i}(s) - f_{w_j}(s))ds| > \epsilon'/2$ on a subset of $[0, 1]$ of measure at least $2\omega(\epsilon', \eta)$ with probability at least $1 - \epsilon'/4$.

Thus $|\int f_{w_i}(s)(f_{w_i}(s) - f_{w_j}(s))ds| \geq \epsilon'$ implies

$$\int \left(\int f_\tau(s)(f_{w_i}(s) - f_{w_j}(s))ds \right)^2 d\tau > (\epsilon'/2)^2 \times 2\omega(\epsilon', \eta)$$

with probability at least $1 - \epsilon'/4$

Since the left hand side is just $||p_i - p_j||_2^2$, it follows that $||p_i - p_j||_2 > (\epsilon'/2) \times (2\omega(\epsilon', \eta))^{1/2}$ with probability at least $1 - \epsilon'/4$, which proves this second part. Taking the contrapositive yields $||p_i - p_j||_2 \leq \delta(\epsilon', \eta)$ implies that $|\int f_{w_i}(s)(f_{w_i}(s) - f_{w_j}(s))ds| < \epsilon'$ with probability at least $1 - \epsilon'/4$, where $\delta(\epsilon', \eta) = (\epsilon'/2) \times (2\omega(\epsilon', \eta))^{1/2}$.

To finish the proof, note that $\left|\int f_{w_i}(s)(f_{w_i}(s) - f_{w_j}(s))ds\right| < \epsilon'$ and $\left|\int f_{w_j}(s)(f_{w_i}(s) - f_{w_j}(s))ds\right| < \epsilon'$ also imply that $\left|\int (f_{w_i}(s) - f_{w_j}(s))(f_{w_i}(s) - f_{w_j}(s))ds\right| < 2\epsilon'$ by the triangle inequality, so that $||p_i - p_j||_2 \leq (\epsilon'/2) \times (2\omega(\epsilon', \eta))^{1/2}$ implies $||f_{w_i} - f_{w_j}||_2 < \sqrt{2\epsilon'}$ with probability at least $1 - \epsilon'/2$. Thus $||p_i - p_j||_2 \leq \delta(\epsilon, \eta)$ implies $||f_{w_i} - f_{w_j}||_2 < \epsilon$ with probability at least $> 1 - \epsilon^2/4$ as claimed, where $\delta(\epsilon, \eta) = (\epsilon^2/4) \times (2\omega(\epsilon^2/2, \eta))^{1/2}$. $\square$

Notice ϵ depends on η for a given δ through the choice of $\omega(\epsilon, \eta)$, so that η cannot be chosen to be arbitrarily small for a fixed δ. Doing so requires a decoupling of the link function approximation error (due to the fact that f might not be smooth off of the set A') from the codegree approximation error (due to the fact that p induces a strictly coarser topology on $[0, 1]$ than f). Lemma 3 accomplishes this by replacing the measurability of f with a stronger continuity assumption, which essentially implies that the former error does not exist.

The proof of Theorem 2 also relies on the auxiliary Lemma A1.

Lemma A1: Suppose Assumption 1 holds. Then for any $\epsilon > 0$, $P\left(||f_{w_i} - f_{w_j}||_2 \leq \epsilon\right) > 0$.

Proof of Lemma A1: As in the proof of the second part of Lemma 2, I begin with an appeal to Lusin's theorem (Dudley (2002), Theorem 7.5.2): for any $\eta > 0$ there must exist B, a closed subset of $[0,1]$ with measure of at least $1-\eta$ such that for any $b \in B$, there exists another closed subset $C(b)$ of $[0,1]$ with measure of at least $1-\eta$, such that for any $c \in C(b)$, f is uniformly continuous when restricted to the set $A' = \{(b,c) \in [0,1]^2 : b \in B, c \in C(b)\}$. That is, for all $\epsilon' > 0$ and $u, v \in B$ there exists a $\omega(\epsilon', \eta) > 0$ such that $|u - v| \leq \omega(\epsilon', \eta)$ implies that $|f(u,t) - f(v,t)| \leq \epsilon'$ for $t \in C(u) \cap C(v)$, a set with Lebesgue measure at least $1 - 2\eta$.

So $|u - v| \leq \omega(\epsilon', \eta)$ and $u, v \in B$ imply that $||f_u - f_v||_2 \leq \left(\epsilon'^2(1-2\eta) + 2\eta\right)^{1/2} \leq \epsilon' + \sqrt{2\eta}$. Since w_i, w_j are independent with standard uniform marginals, this means that $||f_{w_i} - f_{w_j}||_2 \leq \epsilon' + \sqrt{2\eta}$ with probability at least $(1-2\eta)\,\omega(\epsilon', \eta)$. Now just choose $\epsilon' < \epsilon/2$ and $\eta < \epsilon'^2/2$ to get $P\left(||f_{w_i} - f_{w_j}||_2 \leq \epsilon\right) \geq (1 - \epsilon^2/8)\,\omega(\epsilon/2, \epsilon^2/8) > 0$. $\square$

A direct implication of the first part of Lemma 2 and Lemma A1 is that for any $\epsilon > 0$, $P\left(||p_{w_i} - p_{w_j}||_2 \leq \epsilon\right) > 0$.

Theorem 2: Suppose Assumptions 1-5 hold. Then $\hat{\beta} \to_p \beta$.

Proof of Theorem 2: Write

$$\hat{\beta} = \beta + \left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(x_i - x_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)\right)^{-1}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(u_i - u_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)\right)$$

I show $\left(\binom{n}{2}r_n\right)^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(x_i - x_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right) \to_p 2\Gamma_0$, which is positive definite under Assumption 3. Similar arguments yield $\left(\binom{n}{2}r_n\right)^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(u_i - u_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right) \to_p 0$, so that the claim follows from Slutsky and the continuous mapping theorem. Since $r_n > 0$ with high probability from Lemma A1, both statistics are eventually well-defined.

Let $D_n = \left(\binom{n}{2}E\left[K\left(\frac{\delta_{ij}}{h_n}\right)\right]\right)^{-1}\sum_i\sum_j(x_i - x_j)'(x_i - x_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)$ then by the mean value theorem $D_n = \left(\binom{n}{2}E\left[K\left(\frac{\delta_{ij}}{h_n}\right)\right]\right)^{-1}\sum_i\sum_j(x_i - x_j)'(x_i - x_j)\left[K\left(\frac{\delta_{ij}}{h_n}\right) + K'\left(\frac{\iota_{ij}}{h_n}\right)\left(\frac{\hat{\delta}_{ij} - \delta_{ij}}{h_n}\right)\right]$ where $\{\iota_{ij}\}_{i \neq j}$ is the collection of intermediate values implied by that theorem. By Lemma 1 $\max_{i \neq j}\frac{\hat{\delta}_{ij} - \delta_{ij}}{h_n} = o_p\left(n^{-\gamma/4}\right)$ and by Markov's inequality $K'\left(\frac{\iota_{ij}}{h_n}\right) = o_p(r_n n^{\gamma/2})$, since $P\left(K'\left(\frac{\iota_{ij}}{h_n}\right) \geq r_n n^{\gamma/2}\right) \leq \frac{E\left[K'\left(\frac{\iota_{ij}}{h_n}\right)\right]}{r_n n^{\gamma/4}} = o(1)$ by choice of kernel density function in

Assumption 5. It follows that

$$\begin{aligned} D_n &= \left(\binom{n}{2} E\left[K\left(\frac{\delta_{ij}}{h_n}\right)\right]\right)^{-1} \sum_i \sum_j (x_i - x_j)'(x_i - x_j) K\left(\frac{\delta_{ij}}{h_n}\right) + o_p(1) \\ &= D_n' + o_p(1) \end{aligned}$$

since x_i has finite second moments and $K'(u)$ is bounded.

Recall that $\delta_{ij} = \delta(w_i, w_j)$ so that D_n' is a second order U-statistic with kernel depending on n, in the sense of Ahn and Powell (1993). In particular, their Lemma A.3 implies

$$D_n' = \left(E\left[K\left(\frac{\delta_{ij}}{h_n}\right)\right]\right)^{-1} E\left[(x_i - x_j)'(x_i - x_j) K\left(\frac{\delta_{ij}}{h_n}\right)\right] + o_p(1)$$

since $nr_n \to \infty$. Additionally, measurability of f and Assumption 4 imply

$$\begin{aligned} & E\left[(x_i - x_j)'(x_i - x_j) K\left(\frac{\delta_{ij}}{h_n}\right)\right] = \int E\left[(x_i - x_j)'(x_i - x_j) | \delta_{ij} = u\right] K\left(\frac{u}{h_n}\right) dP(\delta_{ij} = u) \\ &= \int \left(\Gamma_0 + o_p(1)\right) K\left(\frac{u}{h_n}\right) dP(\delta_{ij} = u) = \Gamma_0 r_n + o_p\left(r_n\right) \end{aligned}$$

with the second equality is due to $E\left[(x_i - x_j)'(x_i - x_j) | \delta_{ij} \leq u\right] = \Gamma_0 + o_p(1)$ by Lemma 2 and Assumptions 3 and 4. So $D_n = \Gamma_0 + o_p(1)$

A nearly identical argument gives

$$U_n = \left(\binom{n}{2} r_n\right)^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - x_j)'(u_i - u_j) K\left(\frac{\hat{\delta}_{ij}}{h_n}\right) = o_p(1)$$

since $E\left[(x_i - x_j)'(u_i - u_j) | d(w_i, w_j) = h_n\right] = o_p(1)$ by Assumptions 2 and 4. $D_n^{-1} U_n = o_p(1)$ then follows from Slutsky and the continuous mapping theorem. □

Lemmas and Theorems in Section 3.3.2

The proof of Theorem 3 relies on using discreteness of the network types to strengthen Lemma 1 to auxiliary Lemma A2.

Lemma A2: Suppose Assumption 5 holds and f_{w_i} has finite support. Then

$$\max_{(i \neq j)} ||\hat{p}_{w_i} - \hat{p}_{w_j}||_{2,n} \times \mathbb{1}\{||\hat{p}_{w_i} - \hat{p}_{w_j}||_{2,n} \leq \epsilon/2\} = o_{a.s.}(n^{-1/2} h_n)$$

Proof of Lemma A2: The assumption that f_{w_i} has finite support implies $\delta_{ij}1\{\delta_{ij} \le \epsilon\} = 0$ and $m_{ijt}1\{\delta_{ij} \le \epsilon/2\} := \left(p_{w_i w_t} - p_{w_j w_t}\right) \times 1\{\delta_{ij} \le \epsilon/2\} = 0$ both with probability one. Consider the decomposition of $\hat{\delta}_{ij}1\{\hat{\delta}_{ij} \le \epsilon/2\}$ into

$$\hat{\delta}_{ij}\left(1\{\hat{\delta}_{ij} \le \epsilon/2\} - 1\{\delta_{ij} \le \epsilon/2\}\right) + \hat{\delta}_{ij}1\{\delta_{ij} \le \epsilon/2\}$$

I first show $\max_{i \neq j} \sqrt{n} h_n^{-1} \hat{\delta}_{ij} 1\{\delta_{ij} \le \epsilon/2\} = o_{a.s.}(1)$. As in the proof of Lemma 1, Bernstein's inequality gives

$$P\left((n-3)^{-1}\left|\sum_{s \neq i,j,t} D_{ts}(D_{is} - D_{js})1\{\delta_{ij} \le \epsilon/2\}\right| \ge \eta\right) \le 2\exp\left(\frac{-(n-3)\eta^2}{3}\right)$$

so that by the union bound

$$P\left(\sup_{i,j,t}\left[(n-3)^{-1}\sum_{s \neq i,j,t} D_{ts}(D_{is} - D_{js})\right]^2 1\{\delta_{ij} \le \epsilon/2\} \ge \eta\right) \le 2n(n-1)(n-2)\exp\left(\frac{-(n-3)\eta}{3}\right)$$

Averaging over t implies

$$P\left(\max_{i,j} \sqrt{n} h_n^{-1} \hat{\delta}_{ij} 1\{\delta_{ij} \le \epsilon/2\} \ge \eta\right) \le 16(n-3)^3 \exp\left(\frac{-(n-3)\eta h_n}{3\sqrt{n}}\right)$$

so long as $n \ge 6$. Since the right hand side is absolutely summable by arguments made in the proof of Lemma 1, $\max_{i \neq j} \sqrt{n} h_n^{-1} \hat{\delta}_{ij} 1\{\delta_{ij} \le \epsilon/2\} = o_{a.s.}(1)$.

I now show $\max_{i \neq j} \sqrt{n} h_n^{-1} \hat{\delta}_{ij}\left(1\{\hat{\delta}_{ij} \le \epsilon/2\} - 1\{\delta_{ij} \le \epsilon/2\}\right) = o_{a.s.}(1)$. First,

$$\sqrt{n} h_n^{-1} |\hat{\delta}_{ij}\left(1\{\hat{\delta}_{ij} \le \epsilon/2\} - 1\{\delta_{ij} \le \epsilon/2\}\right)| \le 2\sqrt{n} h_n^{-1} \times 1\{|\hat{\delta}_{ij} - \delta_{ij}| > |\epsilon/2 - \delta_{ij}|\}$$

Since $\delta_{ij}1\{\delta_{ij} \le \epsilon\} = 0$ with probability one, $\delta_{ij} \in (\epsilon/4, 3\epsilon/4)$ is a probability zero event, and so it is sufficient to show

$$\max_{i \neq j} \sqrt{n} h_n^{-1} 1\{|\hat{\delta}_{ij} - \delta_{ij}| > \epsilon/4\} = o_{a.s.}(1)$$

Using the inequality from before, the left hand side is nonzero on a set of probability at most $16(n-3)^3 \exp\left(\frac{-(n-3)\epsilon h_n}{12\sqrt{n}}\right)$. Since this is again absolutely summable, $\sup_{i \neq j} \sqrt{n} h_n^{-1} \hat{\delta}_{ij}\left(1\{\hat{\delta}_{ij} \le \epsilon/2\} - 1\{\delta_{ij} \le \epsilon/2\}\right) = o_{a.s.}(1)$ follows.

Taken together, the two arguments demonstrate that $\max_{i \neq j} \sqrt{n} h_n^{-1} \hat{\delta}_{ij} 1\{\hat{\delta}_{ij} \le \epsilon\} = o_{a.s.}(1)$, as claimed. □

Theorem 3: Suppose Assumptions 1-5 hold and the support of f_{w_i} is finite. Then

$$V_3^{-1/2}\left(\hat{\beta}-\beta\right)\rightarrow_d \mathcal{N}\left(0,I_k\right)$$

where $V_3=\Gamma_0^{-1}\Omega_0\Gamma_0^{-1}\times s/n$, Γ_0 is as defined in Assumption 3, I_k is the $k\times k$ identity matrix, and

$$\begin{aligned} s &= P(||p_i-p_j||_2=0,||p_i-p_k||_2=0)/P(||p_i-p_j||_2=0)^2 \\ \Omega_0 &= E\left[(x_i-x_j)'(x_i-x_k)(u_i-u_j)(u_i-u_k)|\,||p_i-p_j||_2=0,||p_i-p_k||_2=0\right] \end{aligned}$$

Proof of Theorem 3: In the proof of Theorem 2, I demonstrate that Assumptions 1-5 are sufficient for

$$\frac{1}{m}\sum_i\sum_{j>i}(x_i-x_j)'(x_i-x_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)\rightarrow_p 2\Gamma_0 E\left[K\left(\frac{\delta_{ij}}{h_n}\right)\right]$$

where $m=n(n-1)/2$. Since the support of f_{w_i} is finite, $E\left[K\left(\frac{\delta_{ij}}{h_n}\right)\right]$
$=K(0)P(||f_{w_i}-f_{w_j}||_2=0)>0$ eventually (for $h_n\leq\epsilon$) since $P(\delta_{ij}=0)>0$.

As for the numerator, I follow the proof of Theorem 2 to write

$$\begin{aligned} U_n &= \frac{1}{m}\sum_i\sum_{j>i}\left((x_i-x_j)'\left(u_i-u_j\right)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)\right) \\ &= \frac{1}{m}\sum_i\sum_{j>i}\left((x_i-x_j)'\left(u_i-u_j\right)\left[K\left(\frac{\delta_{ij}}{h_n}\right)+K'\left(\frac{\iota_{ij}}{h_n}\right)\left(\frac{\hat{\delta}_{ij}-\delta_{ij}}{h_n}\right)1\{\hat{\delta}_{ij}\leq h_n\}\right]\right) \end{aligned}$$

where ι_{ij} is a mean value between δ_{ij} and $\hat{\delta}_{ij}$. I first show
$\frac{1}{m}\sum_i\sum_{j>i}\left(\left(x_{il}-x_{jl}\right)\left(u_i-u_j\right)K'\left(\frac{\iota_{ij}}{h_n}\right)\left(\frac{\hat{\delta}_{ij}-\delta_{ij}}{h_n}\right)1\{\hat{\delta}_{ij}\leq h_n\}\right)=o_p(n^{-1/2})$ for any positive integer $l\leq k$. By Cauchy-Schwartz

$$\begin{aligned} &\frac{1}{m}\left|\sum_i\sum_{j>i}\left(\left(x_{il}-x_{jl}\right)\left(u_i-u_j\right)K'\left(\frac{\iota_{ij}}{h_n}\right)\left(\frac{\hat{\delta}_{ij}-\delta_{ij}}{h_n}\right)\right)\right| \\ &\quad\leq\frac{\bar{K}'}{m}\left(\sum_i\sum_{j>i}\left(\left(x_{il}-x_{jl}\right)\left(u_i-u_j\right)\right)^2\right)^{1/2}\times\left(\sum_i\sum_{j>i}\left(\left(\frac{\hat{\delta}_{ij}-\delta_{ij}}{h_n}\right)1\{\hat{\delta}_{ij}\leq h_n\}\right)^2\right)^{1/2} \end{aligned}$$

where $\bar{K}'=\sup_{u\in[0,1]}K'(u)<\infty$, $\sum_i\sum_{j>i}\left(\left(x_{il}-x_{jl}\right)\left(u_i-u_j\right)\right)^2=O_p(m)$ since x_i and u_i have finite fourth moments, and $\max_{i\neq j}\left(\frac{\hat{\delta}_{ij}-\delta_{ij}}{h_n}\right)1\{\hat{\delta}_{ij}\leq h_n\}=o_{a.s.}\left(n^{-1/2}\right)$ by Lemma A2.

It follows that

$$
\begin{aligned}
U_n &= \frac{1}{m}\sum_i \sum_{j>i}\left((x_i - x_j)'(u_i - u_j)\, K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)\right) \\
&= \frac{1}{m}\sum_i \sum_{j>i}\left((x_i - x_j)'(u_i - u_j)\, K\left(\frac{\delta_{ij}}{h_n}\right)\right) + o_p(n^{-1/2})
\end{aligned}
$$

The first summand is a second order U-statistic with symmetric L^2-integrable kernel, so by Lemma A.3 of Ahn and Powell (1993)

$$\sqrt{n}\,(U_n - U) \to \mathcal{N}(0, V)$$

where $U = E\left[(x_i - x_j))'(u_i - u_j)\, K\left(\frac{\delta_{ij}}{h_n}\right)\right]$ and for $Z_i = (x_i, \nu_i, w_i)$

$$
\begin{aligned}
V &= \lim_{h\to 0} 4E\left[E\left[(x_i - x_j)'(u_i - u_j)K\left(\frac{\delta_{ij}}{h}\right) \mid Z_i\right] E\left[(x_i - x_j)(u_i - u_j)K\left(\frac{\delta_{ij}}{h}\right) \mid Z_i\right]\right] \\
&= \lim_{h\to 0} 4E\left[(x_i - x_j)'(x_i - x_k)(u_i - u_j)(u_i - u_k)K\left(\frac{\delta_{ij}}{h}\right) K\left(\frac{\delta_{ik}}{h}\right)\right]
\end{aligned}
$$

Since f_{w_i} has finite support, $E[\delta_{ij}|\delta_{ij} \le \epsilon] = 0$ for some $\epsilon > 0$, and so $U = E\left[(x_i - x_j))'(u_i - u_j)\, K(0)\, 1\{\delta_{ij} = 0\}\right]$ for n sufficiently large such that $h_n \le \epsilon$. By Lemma 2, $1\{\delta_{ij} = 0\} = 1\{d_{ij} = 0\}$ with probability one, so Assumption 5 implies that $U = 0$ for any choice of $h_n \le \epsilon$ (i.e. $U = 0$ eventually). Similarly $V = 4\Omega_0 K(0)^2 P\left(||f_{w_i} - f_{w_j}||_2 = 0, ||f_{w_i} - f_k||_2 = 0\right)$ so long as $h_n \le \epsilon$. So by Slutsky's Theorem,

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d \mathcal{N}(0, V_3)$$

where $V_3 = \Gamma_0^{-1}\Omega_0\Gamma_0^{-1} \times s$ as claimed. □

Lemma 3: Suppose Assumptions 1 and 6 hold. Then for almost every (w_i, w_j) pair

$$||p_{w_i} - p_{w_j}||_2 \le ||f_{w_i} - f_{w_j}||_2 \le 32\, C_6^{\frac{1}{2+4\alpha}} \left(||p_{w_i} - p_{w_j}||_2\right)^{\frac{\alpha}{1+2\alpha}}$$

so long as $||p_{w_i} - p_{w_j}||_2 < \sqrt{8C_6}K^{-\alpha}$.

Proof of Lemma 3: The first inequality follows from the first part of Lemma 2 holding exactly for every (w_i, w_j) pair. The proof of the second inequality essentially mirrors the second part of Lemma 2, and so only a quick sketch is provded here. I first demonstrate that $||p_{w_i} - p_{w_j}||_2 \le \left(4(4C_6)^{1/\alpha}\right)^{-1} \epsilon'^{\frac{4\alpha+2}{\alpha}}$ and $\left(\frac{\epsilon'}{4C_6}\right)^{\frac{1}{\alpha}} < K^{-1}$ imply that

$||f_{w_i} - f_{w_j}||_2 \leq \sqrt{2\epsilon'}$ with probability one.

Suppose $\left|\int f_{w_i}(s)\left(f_{w_i}(s) - f_{w_j}(s)\right) ds\right| > \epsilon'$. Then $\left|\int f_\tau(s)\left(f_{w_i}(s) - f_{w_j}(s)\right) ds\right| > \epsilon'/2$ for $\tau \in [0,1]$ so long as τ and w_i are in the same block of the partition of $[0,1]$ and $C_6|w_i - \tau|^\alpha < \epsilon'/4$. If $\left(\frac{\epsilon'}{4C_6}\right)^{\frac{1}{\alpha}} < K^{-1}$, then the measure of τ in $[0,1]$ that satisfty these conditions is at least $\left(\frac{\epsilon'}{4C_6}\right)^{\frac{1}{\alpha}}$. It follows that so long as $\left(\frac{\epsilon'}{4C_6}\right)^{\frac{1}{\alpha}} < K^{-1}$

$$\int \left(\int f_\tau(s)\left(f_{w_i}(s) - f_{w_j}(s)\right) ds\right)^2 d\tau > \left(\frac{\epsilon'}{2}\right)^2 \left(\frac{\epsilon'}{4C_6}\right)^{\frac{1}{\alpha}}$$

with probability one.

Following the logic of Lemma 2, I conclude that $||p_i - p_j||_2 \leq \left(4(4C_6)^{1/\alpha}\right)^{-1} \epsilon'^{\frac{4\alpha+2}{\alpha}}$ implies that $||f_{w_i} - f_{w_j}||_2 \leq \sqrt{2\epsilon'}$ with probability one so long as $\left(\frac{\epsilon'}{4C_6}\right)^{\frac{1}{\alpha}} < K^{-1}$. Replacing ϵ' with $\epsilon^2/2$ yields

$$2^{\frac{2\alpha}{4\alpha+2}} 4^{\frac{4}{4\alpha+2}} (4C_6)^{\frac{1}{4\alpha+2}} ||p_i - p_j||_2^{\frac{2\alpha}{4\alpha+2}} \leq \epsilon \text{ implies that } ||f_{w_i} - f_{w_j}||_2 \leq \epsilon$$

with probability one if $\left(\frac{\epsilon^2}{8C_6}\right)^{\frac{1}{\alpha}} < K^{-1}$.

It follows that for almost every w_i and w_j, $2^{\frac{2\alpha+10}{4\alpha+2}} C_6^{\frac{1}{4\alpha+2}} ||p_i - p_j||_2^{\frac{2\alpha}{4\alpha+2}} = \epsilon$ implies that $||f_{w_i} - f_{w_j}||_2 \leq \epsilon$, so long as $\epsilon < \sqrt{8C_6} K^{-\alpha/2}$. The statement of the lemma follows by noting that $2^{\frac{2\alpha+10}{4\alpha+2}}$ is bounded below 32 when $\alpha > 0$. $\square$.

The proof of Theorem 4 relies on the following strengthening of auxiliary Lemma A1 to auxiliary Lemma A3.

Lemma A3: Suppose Assumptions 1 and 6 hold. Then $P\left(||f_{w_i} - f_{w_j}||_2 \leq \epsilon\right) > C_6^{-1/\alpha}\epsilon^{1/\alpha}$, so long as $\epsilon \leq C_6 K^{-\alpha}$

Proof of Lemma A3: The proof of Lemma A3 essentially mirrors that of Lemma A1, except Assumption 6 allows for the replacement of $\omega(\epsilon, \eta)$ with $\left(\frac{\epsilon}{C_6}\right)^{1/\alpha}$. Notice that that so long as $K \leq \left(\frac{\epsilon}{C_6}\right)^{-\frac{1}{\alpha}}$ the probability that w_i and w_j are in the same partition of $[0,1]$ and that $|w_i - w_j| \leq \left(\frac{\epsilon}{C_6}\right)^{1/\alpha}$ is bounded from below by $\left(\frac{\epsilon}{C_6}\right)^{1/\alpha}$. So $P\left(||f_{w_i} - f_{w_j}||_2 \leq \epsilon\right) > \frac{1}{C_6^{1/\alpha}}\epsilon^{1/\alpha}$ as claimed. $\square$

Theorem 4: Suppose Assumptions 1-3 and 6-8 hold and $\alpha \times \zeta > 1/2$. Then

$$V_{4,n}^{-1/2}\left(\hat{\beta} - \beta_{h_n}\right) \to_d \mathcal{N}\left(0, I_k\right)$$

where $V_{4,n} = \Gamma_0^{-1}\Omega_n\Gamma_0^{-1}/n$, Γ_0 is as defined in Assumption 3, r_n is as defined in Assumption 5, and I_k is the $k \times k$ identity matrix, and

$$\beta_{h_n} = \beta + (\Gamma_0)^{-1} E\left[(x_i - x_j)'(u_i - u_j)K\left(\frac{||p_i - p_j||_2}{h_n}\right)\right] / \left(2r_n\right)$$

$$\Omega_n = E\left[(x_i - x_j)'(x_i - x_k)(u_i - u_j)(u_i - u_k)K\left(\frac{||p_i - p_j||_2}{h_n}\right)K\left(\frac{||p_i - p_k||_2}{h_n}\right)\right] / \left(r_n^2\right)$$

Proof of Theorem 4: The proof is simplified by squaring the empirical codegree differences so that

$$\left(\hat{\beta} - \beta\right) = \left(\frac{1}{\binom{n}{2}r_n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(x_i - x_j)K_{1/2}\left(\frac{\hat{\delta}_{ij}^2}{h_n^2}\right)\right)^{-1}$$
$$\frac{1}{\binom{n}{2}r_n}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)'(u_i - u_j)K_{1/2}\left(\frac{\hat{\delta}_{ij}^2}{h_n^2}\right)\right)$$

where $r_n = E\left[K_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right)\right]$ and $K_{1/2}(u) = K(\sqrt{u})$ is supported, positive, and twice differentiable on $[0, 1)$ by Assumption 8. Recall $r_n > 0$ by Lemma A1.

The proof of Theorem 2 demonstrates that Assumptions 1-5 are sufficient for the denominator to converge in probability to $2\Gamma_0$, which is eventually invertible by Assumption 3. As for the numerator,

$$\begin{aligned} U_n &= \frac{1}{\binom{n}{2}r_n}\sum_i\sum_{j>i}\left((x_i - x_j)'(u_i - u_j)K_{1/2}\left(\frac{\hat{\delta}_{ij}^2}{h_n^2}\right)\right) \\ &= \frac{1}{\binom{n}{2}r_n}\sum_i\sum_{j>i}\left((x_i - x_j)'(u_i - u_j)\left[K_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right) + K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(\frac{\hat{\delta}_{ij}^2 - \delta_{ij}^2}{h_n^2}\right)\right.\right. \\ &\qquad \left.\left. + K_{1/2}''\left(\frac{\iota_{ij}}{h_n^2}\right)\left(\frac{\hat{\delta}_{ij}^2 - \delta_{ij}^2}{h_n^2}\right)^2\right]\right) \end{aligned}$$

where ι_{ij} is the intermediate value between $\hat{\delta}_{ij}^2$ and δ_{ij}^2 suggested by the mean value theorem and Taylor's theorem. I consider each of the summands individually. I first show that

$$\frac{1}{\binom{n}{2}r_n}\sum_i\sum_{j>i}\left((x_i - x_j)'(u_i - u_j)K_{1/2}''\left(\frac{\iota_{ij}}{h_n^2}\right)\left(\frac{\hat{\delta}_{ij}^2 - \delta_{ij}^2}{h_n^2}\right)^2\right) = o_p\left(n^{-1/2}\right)$$

Let $s_n = n^{-1/2} h_n^4 r_n$. Since $\delta_{ij} \leq C|w_i - w_j|^\alpha$ by the first part of Lemma 2 and Assumption 6, $r_n \geq KC^{-1/\alpha} h_n^{1/\alpha}$ for $K = \liminf_{h\to 0} E\left[K\left(\frac{\delta_{ij}}{h}\right) | \delta_{ij} \leq h\right] > 0$ by Lemma A2. Since $n^{1/2-\gamma} h_n^{4+1/\alpha} \to \infty$ for some $\gamma > 0$ by Assumption 9, $n^{1-\gamma} s_n \to \infty$, and so Lemma 1 implies that $\sup_{i\neq j}\left(\frac{\hat{\delta}_{ij}^2 - \delta_{ij}^2}{\sqrt{s}}\right)^2 = o_{a.s.}(1)$ or $\sup_{i\neq j}\left(\frac{\hat{\delta}_{ij}^2 - \delta_{ij}^2}{h_n^2\sqrt{r_n}}\right)^2 = o_{a.s.}\left(n^{-1/2}\right)$. It follows that

$$
\begin{aligned}
&\frac{1}{\binom{n}{2} r_n} \sum_i \sum_{j>i} \left((x_i - x_j)' (u_i - u_j) K''_{1/2}\left(\frac{\iota_{ij}}{h_n}\right) \left(\frac{\hat{\delta}_{ij}^2 - \delta_{ij}^2}{h_n^2}\right)^2 \right) \\
&\qquad \leq \frac{\bar{K}''_{1/2}}{\binom{n}{2}} \sum_i \sum_{j>i} \left((x_i - x_j)' (u_i - u_j) \right) \times o_{a.s.}\left(n^{-1/2}\right)
\end{aligned}
$$

where $\bar{K}''_{1/2} = \sup_{u\in[0,1]} K''_{1/2}(u)$ and the last line is $o_p\left(n^{-1/2}\right)$ because x_i and u_i are assumed to have finite fourth moments by Assumption 2. Thus

$$
U_n = \frac{1}{\binom{n}{2} r_n} \sum_i \sum_{j>i} \left((x_i - x_j)' (u_i - u_j) \left[K_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right) + K'_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right) \left(\frac{\hat{\delta}_{ij}^2 - \delta_{ij}^2}{h_n^2}\right) \right] \right) + o_p\left(n^{-1/2}\right)
$$

Now let

$$
\begin{aligned}
\tilde{\delta}_{ij} = \tilde{\delta}(w_i, w_j)^2 = \frac{1}{n} \sum_{t=1}^n &\left(\frac{1}{n} \sum_{s_1=1}^n f(w_t, w_{s_1}) \left(f(w_i, w_{s_1}) - f(w_j, w_{s_1}) \right) \right) \\
&\times \left(\frac{1}{n} \sum_{s_2=1}^n f(w_t, w_{s_2}) \left(f(w_i, w_{s_2}) - f(w_j, w_{s_2}) \right) \right)
\end{aligned}
$$

and rewrite the numerator as

$$
\begin{aligned}
U_n = &\frac{1}{\binom{n}{2} r_n} \sum_i \sum_{j>i} \left((x_i - x_j)' (u_i - u_j) \left[K_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right) + K'_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right) \left(\frac{\tilde{\delta}_{ij}^2 - \delta_{ij}^2}{h_n^2}\right) \right] \right) \\
&+ \frac{1}{\binom{n}{2} r_n} \sum_i \sum_{j>i} \left((x_i - x_j)' (u_i - u_j) \left[K'_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right) \left(\frac{\hat{\delta}_{ij}^2 - \tilde{\delta}_{ij}^2}{h_n^2}\right) \right] \right) + o_p\left(n^{-1/2}\right)
\end{aligned}
$$

In the remainder of this proof, I show that the second summand is $o_p(n^{-1/2})$, while the first part is a fifth-order U-statistic. First,

$$
\frac{1}{\binom{n}{2} r_n} \sum_i \sum_{j>i} \left((x_i - x_j)' (u_i - u_j) \left[K'_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right) \left(\frac{\hat{\delta}_{ij}^2 - \tilde{\delta}_{ij}^2}{h_n^2}\right) \right] \right) = o_p\left(n^{-1/2}\right)
$$

by Chebyshev's inequality, since $E\left[\left(\frac{\hat{\delta}_{ij}^2-\tilde{\delta}_{ij}^2}{h^2}\right)|x_i,x_j,u_i,u_j,w_i,w_j\right]=0$ implies $\frac{1}{\binom{n}{2}r_n}\sum_i\sum_{j>i}\left((x_i-x_j)'(u_i-u_j)\left[K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(\frac{\hat{\delta}_{ij}^2-\tilde{\delta}_{ij}^2}{h_n^2}\right)\right]\right)$ is mean zero and

$$E\left[\left(\frac{1}{\binom{n}{2}r_n}\sum_i\sum_{j>i}\left((x_i-x_j)'(u_i-u_j)\left[K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(\frac{\hat{\delta}_{ij}^2-\tilde{\delta}_{ij}^2}{h_n^2}\right)\right]\right)\right)^2\right]$$
$$=\frac{1}{\binom{n}{2}^2n^6r_n^2h_n^4}E\Bigg[\sum_{i_1}\sum_{i_2}\sum_{j_1}\sum_{j_2}\sum_{t_1}\sum_{t_2}\sum_{s_{11}}\sum_{s_{12}}\sum_{s_{21}}\sum_{s_{22}}$$
$$(x_{i_1}-x_{j_1})'(x_{i_2}-x_{j_2})(u_{i_1}-u_{j_1})(u_{i_2}-u_{j_2})K_{1/2}'\left(\frac{\delta_{i_1j_1}^2}{h_n^2}\right)K_{1/2}'\left(\frac{\delta_{i_2j_2}^2}{h_n^2}\right)$$
$$\times\left[D_{t_1s_{11}}D_{t_1s_{12}}(D_{i_1s_{11}}-D_{j_1s_{11}})(D_{i_1s_{12}}-D_{j_1s_{12}})-f_{t_1s_{11}}f_{t_1s_{12}}(f_{i_1s_{11}}-f_{j_1s_{11}})(f_{i_1s_{12}}-f_{j_1s_{12}})\right]$$
$$\times\left[D_{t_2s_{21}}D_{t_2s_{22}}(D_{i_2s_{21}}-D_{j_2s_{21}})(D_{i_2s_{22}}-D_{j_2s_{22}})-f_{t_2s_{21}}f_{t_2s_{22}}(f_{i_2s_{21}}-f_{j_2s_{21}})(f_{i_2s_{22}}-f_{j_2s_{22}})\right]\Bigg]$$

is $o\left(n^{-1}\right)$. To see this, note that unless two elements from the set $\{i_1,j_1,t_1,s_{11},s_{12}\}$ equal two in $\{i_2,j_2,t_2,s_{21},s_{22}\}$, $\{\eta_{t_1s_{11}},\eta_{t_1s_{12}},\eta_{i_1s_{11}},\eta_{j_1s_{11}},\eta_{i_1s_{12}},\eta_{j_1s_{12}}\}$ is independent of $\{\eta_{t_2s_{21}},\eta_{t_2s_{22}},\eta_{i_2s_{21}},\eta_{j_2s_{21}},\eta_{i_2s_{22}},\eta_{j_2s_{22}}\}$ and so

$$E\Bigg[\left[D_{t_1s_{11}}D_{t_1s_{12}}(D_{i_1s_{11}}-D_{j_1s_{11}})(D_{i_1s_{12}}-D_{j_1s_{12}})-f_{t_1s_{11}}f_{t_1s_{12}}(f_{i_1s_{11}}-f_{j_1s_{11}})(f_{i_1s_{12}}-f_{j_1s_{12}})\right]$$
$$\times\left[D_{t_2s_{21}}D_{t_2s_{22}}(D_{i_2s_{21}}-D_{j_2s_{21}})(D_{i_2s_{22}}-D_{j_2s_{22}})-f_{t_2s_{21}}f_{t_2s_{22}}(f_{i_2s_{21}}-f_{j_2s_{21}})(f_{i_2s_{22}}-f_{j_2s_{22}})\right]$$
$$|Z_{i_1},Z_{i_2},Z_{j_1},Z_{j_2},Z_{t_1},Z_{t_2},Z_{s_{11}},Z_{s_{12}},Z_{s_{21}},Z_{s_{22}}\Bigg]=0$$

where $Z_i=\{x_i,w_i,\nu_i\}$. Since $K_{1/2}'\left(\frac{\delta_{i_1j_1}^2}{h^2}\right)$ is $O_p(r_n)$ by Assumption 8 (see the proof of Theorem 2 for the formal argument), the desired term is $o\left(n^{-1}\right)$ since $nh_n^4\to\infty$.

It follows that

$$U_n=\frac{1}{\binom{n}{2}r_n}\sum_i\sum_{j>i}\left((x_i-x_j)'(u_i-u_j)\left[K_{1/2}\left(\frac{\delta_{ij}^2}{h^2}\right)+K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(\frac{\tilde{\delta}_{ij}^2-\delta_{ij}^2}{h_n^2}\right)\right]\right)+o_p\left(n^{-1/2}\right)$$
$$=\frac{1}{\binom{n}{5}^2r_n}\sum_i\sum_{j>i}\sum_{t>j}\sum_{s_1>t}\sum_{s_2>s_1}(x_i-x_j)'(u_i-u_j)\Bigg[K_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right)$$
$$+h_n^{-2}K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(f_{ts_1}f_{ts_2}(f_{is_1}-f_{js_1})(f_{is_2}-f_{js_2})-\delta_{ij}^2\right)\Bigg]+o_p\left(n^{-1/2}\right)$$

so that U_n is equivalent to a 5th order U-statistic up to a $o_p(1/\sqrt{n})$ error. As in Theorem 3, I apply Lemma 3.2 from Powell et al. (1989) to rewrite this statistic as the sum of first order projections.

$$\begin{aligned}
U_n = E[U_n] &+ \frac{2}{nr_n}\sum_{\tau=1}^{n}\left(E\left[(x_\tau - x_j)'(u_\tau - u_j)K_{1/2}\left(\frac{\delta_{\tau j}^2}{h_n^2}\right)|Z_\tau\right] - E[U_n]\right) \\
&+ \frac{1}{nr_nh_n^2}\sum_{\tau=1}^{n}E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(f_{\tau s_1}f_{\tau s_2}(f_{is_1} - f_{js_1})(f_{is_2} - f_{js_2}) - \delta_{ij}^2\right)|Z_\tau\right] \\
&+ \frac{2}{nr_nh_n^2}\sum_{\tau=1}^{n}E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(f_{t\tau}f_{ts_2}(f_{i\tau} - f_{j\tau})(f_{is_2} - f_{js_2}) - \delta_{ij}^2\right)|Z_\tau\right] \\
&+ o_p\left(n^{-1/2}\right)
\end{aligned}$$

where $E[U_n] = E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right)\right]$ and $Z_\tau = \{x_\tau, w_\tau, \nu_\tau\}$.

When $\alpha \times \zeta > 1/2$ the second and third terms are both $o_p\left(n^{-1/2}\right)$. For the second term, I show this by fixing some $\epsilon > 0$ and writing

$$\begin{aligned}
&P\left(E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(f_{\tau s_1}f_{\tau s_2}(f_{is_1} - f_{js_1})(f_{is_2} - f_{js_2}) - \delta_{ij}^2\right)|Z_\tau\right] \geq r_nh_n^2\epsilon\right) \\
&= P\left(E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(E[f_{\tau s}(f_{is} - f_{js})|Z_i, Z_j, Z_\tau]^2 - \delta_{ij}^2\right)|Z_\tau\right] \geq r_nh_n^2\epsilon\right) \\
&\leq E\left[\left|E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)|Z_\tau\right]\right| \times \left(E[f_{\tau s}(f_{is} - f_{js})|Z_i, Z_j, Z_\tau]^2 + \delta_{ij}^2\right)|Z_\tau\right]/r_nh_n^2\epsilon
\end{aligned}$$

with the last line by Markov's inequality and the triangle inequality. Since $\left|E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)|Z_\tau\right]\right| = o_p(r_n)$ and both $E[f_{\tau s}(f_{is} - f_{js})|Z_i, Z_j, Z_\tau]^2$ and δ_{ij}^2 are $O_p(h_n^2)$, the term is $o_p(1)$. So the second summand

$$\frac{1}{nr_nh_n^2}\sum_{\tau=1}^{n}E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(f_{\tau s_1}f_{\tau s_2}(f_{is_1} - f_{js_1})(f_{is_2} - f_{js_2}) - \delta_{ij}^2\right)|Z_\tau\right]$$

is an average of n independent random variables with finite third moments (since x_i and u_i have finite sixth moments) that are each $o_p(1)$, and so must be $o_p\left(n^{-1/2}\right)$ by the Lindeberg-Levy central limit theorem.

Bounding the third term is a bit more complicated. Again fix some $\epsilon > 0$ and write

$$P\left(E\left[(x_i - x_j)'(u_i - u_j)K_{1/2}'\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(f_{t\tau}f_{ts_2}(f_{i\tau} - f_{j\tau})(f_{is_2} - f_{js_2}) - \delta_{ij}^2\right)|Z_\tau\right] \geq r_nh_n^2\epsilon\right)$$

However, this time Markov's inequality only provides the upper bound

$$E\left[\left|E\left[(x_i-x_j)'\,(u_i-u_j)\,K'_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right)|Z_\tau\right]\right|\right.$$
$$\left.\times\left(E[f_{t\tau}(f_{i\tau}-f_{j\tau})|Z_i,Z_j,Z_\tau]E[f_{ts}(f_{is}-f_{js})|Z_i,Z_j]+\delta_{ij}^2\right)|Z_\tau\right]/r_n h_n^2\epsilon$$

Here δ_{ij}^2 is $O_p(h_n^2)$ and $E[f_{ts}(f_{is}-f_{js})|Z_i,Z_j]$ is $O_p(h_n)$ by Jensen's inequality, but it is only possible to demonstrate that $E[f_{t\tau}(f_{i\tau}-f_{j\tau})|Z_i,Z_j,Z_\tau]\leq||f_{w_i}-f_{w_j}||_2=O_p(h_n^{2\alpha/(1+2\alpha)})$ by Lemma 3. This is where I use the $\zeta\times\alpha>1/2$ condition so that $\left|E\left[(x_i-x_j)'\,(u_i-u_j)\,K'_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right)|Z_\tau\right]\right|$ is not just $o_p(r_n)$ but $o_p(h_n^{2\alpha\zeta/(1+2\alpha)}r_n)$. Together, these rates imply that the term is $o_p(1)$, and that the third summand

$$\frac{1}{nr_nh_n^2}\sum_{\tau=1}^{n}E\left[(x_i-x_j)'\,(u_i-u_j)\,K'_{1/2}\left(\frac{\delta_{ij}^2}{h_n^2}\right)\left(f_{t\tau}f_{ts_2}(f_{i\tau}-f_{j\tau})(f_{is_2}-f_{js_2})-\delta_{ij}^2\right)|Z_\tau\right]$$

is $o_p\left(n^{-1/2}\right)$ by previous arguments.

It follows from these two arguments that

$$U_n=E[U_n]+\frac{2}{nr_n}\sum_{\tau=1}^{n}\left(E\left[(x_\tau-x_j)'(u_\tau-u_j)K_{1/2}\left(\frac{\delta_{\tau j}^2}{h_n^2}\right)|Z_\tau\right]-E[U_n]\right)+o_p\left(n^{-1/2}\right)$$

U_n is simply an iid sum of random variables with bounded third moments, so by the Lindeberg-Levy central limit theorem

$$V_n''^{-1/2}\left(U_n-E[U_n]\right)\to_d\mathcal{N}\left(0,I_k\right)$$

where

$$\begin{aligned}V_n''&=E\left[\left(\frac{4}{r_n^2}\left(E\left[(x_\tau-x_j)'(u_\tau-u_j)K_{1/2}\left(\frac{\delta_{\tau j}^2}{h_n^2}\right)|Z_\tau\right]-E[U_n]\right)\right)\right.\\&\qquad\left.\times\left(E\left[(x_\tau-x_j)(u_\tau-u_j)K_{1/2}\left(\frac{\delta_{\tau j}^2}{h_n^2}\right)|Z_\tau\right]-E[U_n]\right)\right]\\&=\frac{4}{r_n^2}E\left[(x_\tau-x_j)'(x_\tau-x_k)(u_\tau-u_j)(u_\tau-u_k)K_{1/2}\left(\frac{\delta_{\tau j}^2}{h_n^2}\right)K_{1/2}\left(\frac{\delta_{\tau k}^2}{h_n^2}\right)\right]\end{aligned}$$

because $E[U_n]\to_p 0$ by Theorem 2. It follows from Slutsky's Theorem that

$$V_{4,n}^{-1/2}\left(\hat{\beta}-\beta-(2\Gamma_0)^{-1}E\left[U_n\right]\right)\to_d\mathcal{N}\left(0,I_k\right)$$

where $E[U_n]=r_n^{-1}E\left[(x_i-x_j)'(u_i-u_j)K\left(\frac{\delta(w_i,w_j)}{h_n}\right)\right]$ as claimed. $\square$

Theorems in Sections 3.3.3 and 3.3.4

Theorem 5: Suppose Assumptions 1-3 and 6-9 hold, and $L > \zeta\alpha/(2\theta(1+2\alpha))$. Then

$$V_{5,n}^{-1/2}\left(\bar{\beta}_L - \beta\right) \to_d \mathcal{N}\left(0, I_k\right)$$

where $V_{5,n} = \sum_{l_1=1}^{L}\sum_{l_2=1}^{L} a_{l_1}a_{l_2}\Gamma_0^{-1}\Omega_{n,l_1l_2}\Gamma_0^{-1}/n$, Γ_0 is as defined in Assumption 3, r_n is as defined in Assumption 5, I_k is the $k \times k$ identity matrix, and

$$\Omega_{n,l_1l_2} = E\left[(x_i - x_j)'(x_i - x_k)(u_i - u_j)(u_i - u_k)K\left(\frac{||p_i - p_j||_2}{h_n}\right)K\left(\frac{||p_i - p_k||_2}{h_n}\right)\right]/\left(r_n^2\right)$$

Proof of Theorem 5: Since $\bar{\beta}_L = \sum_{l=1}^{L} a_l\hat{\beta}_{C_lh_n}$, the logic of Theorem 4 and the continuous mapping theorem imply

$$\sqrt{n}\left(\bar{\beta}_L - \bar{\beta}_{L,h_n}\right) = \sum_{l=1}^{L} a_l\sqrt{n}\left(\hat{\beta}_{C_lh_n} - \beta_{C_lh_n}\right) \to_d \mathcal{N}\left(0, \sum_{l_1=1}^{n}\sum_{l_2=1}^{n}\Gamma_0^{-1}\Omega_{l_1l_2,h_n}\Gamma_0^{-1}\sigma_{l_1,l_2,h_n}\right)$$

where $\bar{\beta}_{L,h} = \sum_{l=1}^{L} a_l\beta_{C_lh}$ and
$\Omega_{n,l_1l_2} = E\left[(x_i - x_j)'(x_i - x_k)(u_i - u_j)(u_i - u_k)K\left(\frac{||p_i-p_j||_2}{h_n}\right)K\left(\frac{||p_i-p_k||_2}{h_n}\right)\right]/\left(r_n^2\right)$. By Assumption 10 and the definition of $\{a_1, ..., a_L\}$, $\bar{\beta}_{L,h}$ can be written as

$$\begin{aligned}\bar{\beta}_{L,h} &= \beta + \sum_{l_1=1}^{L}\sum_{l_2=1}^{L} a_{l_1}\left(2\Gamma_0\right)^{-1}C_{l_2}\left(c_{l_1}h\right)^{l_2/\theta} + o_p\left(n^{-1/2}\right) \\ &= \beta + \left(2\Gamma_0\right)^{-1}\sum_{l_2} C_{l_2}\left[\sum_{l_1} a_{l_1}c_{l_1}^{l_2/\theta}\right]h^{l_2/\theta} + o_p\left(n^{-1/2}\right)\end{aligned}$$

since $\sum_{l_2} a_{l_2} = 1$ by choice of $\{a_1, ..., a_L\}$. Furthermore, $\{a_1, ..., a_L\}$ also satisfies $\left[\sum_{l_1} a_{l_1}c_{l_1}^{l_2/\theta}\right] = 0$ for all $l_2 \in \{1, ..., L\}$, so the second summand is 0 and $\bar{\beta}_{L,h} = \beta + o_p\left(n^{-1/2}\right)$. The claim follows. □

Theorem 6: Suppose Assumptions 1-5 hold. Then $\hat{\Gamma}_{h_n}^{-1}\hat{\Omega}_{h_n,h_n}\hat{\Gamma}_h^{-1}/\sqrt{n} \to_p V_{4,n}$ and $\sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\hat{\Gamma}_{c_{l_1}h_n}^{-1}\hat{\Omega}_{c_{l_1}h_n,c_{l_2}h_n}\hat{\Gamma}_{c_{l_2}h_n}^{-1}/\sqrt{n} \to_p V_{5,n}$

Proof of Theorem 6 It is sufficient to prove the second result, which nests the first as a special case. In the proof of Theorem 2 I demonstrate that Assumptions 1-5 are sufficient

for $\left(E\left[K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)\right]\right)^{-1}\hat{\Gamma}_{ch_n} = 2\Gamma_0 + o_p(1)$ for any constant $c > 0$. It remains to be shown that $\left(E\left[K\left(\frac{\delta(w_i,w_j)}{c_1h_n}\right)\right]\right)^{-1}\left(E\left[K\left(\frac{\delta(w_i,w_j)}{c_2h_n}\right)\right]\right)^{-1}\hat{\Omega}_{c_1h_n,c_2h_n}$ converges to $\Omega_{nc_1c_2}$.
I first fix agent i and $Z_i = \{x_i, w_i, \nu_i\}$ and study the average
$\left(E\left[K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)\right]\right)^{-1}(n-2)^{-1}\sum_{j>i}(x_i-x_j)'(\hat{u}_i-\hat{u}_j)K\left(\frac{\hat{\delta}_{ij}}{ch_n}\right)$ for some fixed $c > 0$. Since $\hat{u}_i = u_i + x_i(\hat{\beta}-\beta)$ this average can be rewritten

$$\left(E\left[K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)\right]\right)^{-1}(n-2)^{-1}\sum_{j>i}(x_i-x_j)'\left[(u_i-u_j)-(x_i-x_j)(\hat{\beta}-\beta)\right]K\left(\frac{\hat{\delta}_{ij}}{ch_n}\right)$$
$$=\left(E\left[K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)\right]\right)^{-1}(n-2)^{-1}\sum_{j>i}(x_i-x_j)'(u_i-u_j)K\left(\frac{\hat{\delta}_{ij}}{ch_n}\right)$$
$$-\left(E\left[K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)\right]\right)^{-1}(n-2)^{-1}\sum_{j>i}(x_i-x_j)'(x_i-x_j)K\left(\frac{\hat{\delta}_{ij}}{ch_n}\right)(\hat{\beta}-\beta)$$

The first summand converges to
$\left(E\left[K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)\right]\right)^{-1}E\left[(x_i-x_j)'(u_i-u_j)K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)|Z_i\right]$ following from arguments made in Theorem 3. The first part of the second summand
$\left(E\left[K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)\right]\right)^{-1}(n-2)^{-1}\sum_{j>i}(x_i-x_j)'(x_i-x_j)K\left(\frac{\delta(w_i,w_j)}{ch_n}\right)$ is bounded following arguments made in Theorem 2, and so the second summand converges to 0 in probability since $(\hat{\beta}-\beta) = o_p(1)$ by Theorem 2. As a result,
$\left(E\left[K\left(\frac{\delta(w_i,w_j)}{c_1h_n}\right)\right]\right)^{-1}\left(E\left[K\left(\frac{\delta(w_i,w_j)}{c_2h_n}\right)\right]\right)^{-1}\hat{\Omega}_{c_1h_n,c_2h_n}$ can be written as

$$(n-2)^{-1}4\sum_{i=1}^{n-1}E\left[(x_i-x_j)'(u_i-u_j)K\left(\frac{\delta(w_i,w_j)}{c_1h_n}\right)|Z_i\right]E\left[(x_i-x_j)(u_i-u_j)K\left(\frac{\delta(w_i,w_j)}{c_2h_n}\right)|Z_i\right]$$
$$\times\left(E\left[K\left(\frac{\delta(w_i,w_j)}{c_1h_n}\right)\right]E\left[K\left(\frac{\delta(w_i,w_j)}{c_2h_n}\right)\right]\right)^{-1}$$
$$=(n-2)^{-1}4\sum_{i=1}^{n-1}E\left[(x_i-x_j)'(x_i-x_k)(u_i-u_j)(u_i-u_k)K\left(\frac{\delta(w_i,w_j)}{c_1h_n}\right)K\left(\frac{\delta(w_i,w_k)}{c_2h_n}\right)|Z_i\right]$$
$$\times\left(E\left[K\left(\frac{\delta(w_i,w_j)}{c_1h_n}\right)\right]E\left[K\left(\frac{\delta(w_i,w_j)}{c_2h_n}\right)\right]\right)^{-1}$$

Together, the two results imply that $\hat{\Gamma}^{-1}_{c_{l_1}h_n}\hat{\Omega}_{c_{l_1}h_n,c_{l_2}h_n}\hat{\Gamma}^{-1}_{c_{l_2}h_n} \to_p \Gamma_0^{-1}\Omega_{c_1h_n,c_2h_n}\Gamma_0^{-1}$, and the claim follows from the continuous mapping theorem. □