Two-Step Estimation of Network-Formation Models with Incomplete Information*

Michael Leung†

April 19, 2013

Abstract. We characterize network formation as a game of incomplete information in which agents simultaneously form links. The model allows linking decisions to depend on the structure of the network as well as the characteristics of agents, thereby generalizing dyadic regression. When the data is rationalized by an *anonymous* equilibrium, meaning observationally identical agents choose the same *ex-ante* strategies, the model can be estimated using a computationally simple two-step procedure. Estimation is possible with only a single network because the frequency with which agents with differing attributes link provides identifying variation. Under a sequence of experiments that sends the number of agents in the network n to infinity, our estimator is consistent at the parametric rate n when attributes are discrete and at a slower nonparametric rate when attributes are continuous. We apply the model to study the formation of risk-sharing networks in rural Indian villages.

JEL Codes: C13, C31, D85
Keywords: social networks, network formation, multiple equilibria, large-market asymptotics, discrete games of incomplete information, conditional inference, risk-sharing networks

*I thank Han Hong for guidance and many helpful discussions and close readings. I also thank Trevor Davis, Caroline Hoxby, Guido Imbens, Matt Jackson, and participants at the Stanford structural I.O. lunch for feedback and criticism.

†Department of Economics, Stanford University. E-mail: mpleung@stanford.edu

1 Introduction

Social networks influence a broad range of economic behavior and phenomena, including workplace productivity (Bandiera, Barankay and Rasul, 2009), student achievement (Calvó-Armengol, Patacchini and Zenou, 2009), social learning (Conley and Udry, 2010), welfare participation (Bertrand, Luttmer and Mullainathan, 2010), obesity (Christakis and Fowler, 2007), and product demand (Farrell and Klemperer, 2007), among others. Because networks are a crucial medium through which economic agents interact, understanding the incentives responsible for the formation of networks is valuable. Toward this end, this paper develops a method to estimate a general class of models of network formation.

Such models are useful for the empirical study of networks for several reasons. First, the determinants of network formation have intrinsic economic significance. For instance, it has long been recognized that homophily, the principle that "similarity breeds connection," is pervasive in social networks (McPherson, Smith-Lovin and Cook, 2001). That is, an agent i is more likely to link with another agent j if the two share similar attributes. However, it is also possible that i wishes to link with j because she anticipates that many other agents will also link with j. Disentangling such endogenous determinants of network structure, so called because they depend on the linking decisions of other agents, from exogenous determinants that depend solely on agent characteristics, such as homophily, cannot be accomplished using experimental variation and requires a model-based approach, as proposed here. Second, to the extent that networks matter for economic outcomes, it is important to understand how policy interventions might reshape networks in order to improve outcomes. A key advantage of model-based approaches is that they permit researchers to simulate the impact of counterfactual policies. Third, many studies of the effect of networks on economic outcomes treat networks as exogenous when in reality agents often choose to form networks in anticipation of their future economic benefits. Formally modeling the incentives that give rise to networks is a step toward credibly controlling for network endogeneity.

This paper develops a computationally simple two-step estimation strategy to estimate strategic models of network formation under incomplete information. We assume that agents simultaneously form links to maximize utility, given beliefs about the anticipated state of the network. For a given application, the task of the researcher is to specify a form for the utility function that governs the incentives agents face when deciding with whom to link. We propose the following estimation approach: agents' beliefs about the state of the network are nonparametrically estimated in the first step, and structural parameters are estimated in the second step using maximum likelihood, replacing the unknown beliefs with their estimates (other methods, such as GMM, are possible). A key assumption that makes the first step possible is anonymity of beliefs, meaning that in equilibrium, identical agents choose identical linking strategies *ex-ante*, prior to their draws of unobserved heterogeneity. This assumption solves a curse of dimensionality problem. As we discuss below, network-formation models can induce likelihoods that are computationally intractable, precluding the possibility of maximum likelihood. In our model, the likelihood is tractable due to separability restrictions on the utility function analogous to assumptions made in dyadic regression models. Our estimation strategy effectively utilizes each of the $O(n^2)$ linking decisions as individual observations. Consequently, the estimator converges at the rate n for the case of discrete attributes. When attributes are continuous, convergence occurs at a slower nonparametric rate $nh^{d+c/2}$, where h is the kernel smoothing parameter, and c and d are constants related to the number of possible agent characteristics.

We apply our approach to study the formation of risk-sharing networks in villages in southern India. In these networks, a link from i to j exists if i trusts j enough to lend her a substantial amount of money. We find that the key determinants of whether or not a person i trusts j are whether or not j also trusts i (reciprocal trust), the number of individuals who trust both i and j (supported trust), and whether or not i and j are relatives. In contrast, the popular dyadic regression approach *a priori* rules out the possibility of the first two determinants.

Related Literature. Estimation of network formation models has been a recent topic of interest in statistics and economics. The challenge of developing network formation models is that the space of possible networks is enormous, with $2^{n(n-1)/2}$ possible undirected networks on n nodes.[1] For $n = 30$ the number of possible networks already exceeds the number of elementary particles in the universe. This fact often creates a curse of dimensionality problem for estimating network-formation models. Perhaps the most common approach in applied economics is dyadic or pairwise regression, which is a discrete choice model with network links as the dependent variable (see e.g. Bramoullé and Fortin, 2010; Fafchamps and Gubert, 2007). Such models assume that an agent i's decision to form a link with an agent j only depends on the characteristics of i and j and not, for instance, on which agents choose to link with j. This avoids the curse of dimensionality by ruling out strategic interactions. Most of the statistics and econometrics literature on network formation are concerned with estimation methods when strategic interactions are permitted.

[1] A network is undirected if an agent i is linked to an agent j if and only if agent j is linked to agent i. Otherwise it is directed.

The leading class of models in the statistics literature is the class of exponential random graph models (Snijders, 2002). Such models directly specify a probability distribution over the space of networks based on network statistics deemed relevant by the researcher, such as the prevalence of dyads.[2] See Robbins, Pattison, Kalish and Lusher (2007) for a review of exponential random graph models. The computational complexity of the likelihood in such models has led to the adoption of Markov Chain Monte Carlo (MCMC) methods, but this approach may not be a panacea. Bhamidi, Bresler and Sly (2011) demonstrate that the rate of convergence for MCMC is fast only when the model is indistinguishable from a Poisson random graph, in which case the problem is trivial since links are i.i.d., and that otherwise the rate is $O(e^{n^2})$. This rate suggests that MCMC may be infeasible even for networks with as few as 30 agents, since one would need an extremely large number of simulation draws on the order of e^{n^2} to achieve convergence.[3] Chandrasekhar and Jackson (2012) provide simulation evidence that this translates to poor performance in practice.

In the econometrics literature, network formation is typically modeled as a game in which agents form links to maximize payoffs that can generally depend on the linking decisions of others. The first pioneering attempts in this literature model the network-formation process as a dynamic game in which a subset of the agents can form or break links each period (Christakis, Fowler, Imbens and Kalyanaraman, 2010; Mele, 2011). In these models, links are formed myopically, so that agents form links without anticipating consequent changes to the network. This guarantees equilibrium uniqueness and simplifies computation. Because these models induce likelihoods that are computationally intractable, estimation is done using MCMC. Unfortunately, while MCMC avoids the computation of intractable likelihoods, the curse of dimensionality can reappear as an $O(e^{n^2})$ rate of convergence (Hsieh and Lee, 2012; Mele, 2011). That is, dynamic models often suffer from the same computational problem faced by exponential random graph models. The essential problem is that when MCMC is used to crawl the space of all possible networks, which contains $O(e^{n^2})$ elements, this requires a prohibitively large number of simulation draws to ensure convergence.[4]

An alternative method is to model network formation as a static game, which is the approach we take. Static network-formation models are often estimable using standard fre-

[2] A dyad consists of two nodes that are linked.

[3] However, recent work by Chandrasekhar and Jackson (2012) outlines a new, computationally feasible, frequentist estimation approach that can consistently estimate a broad class of ERGMs (as well as other random graph models) as the number of nodes goes to infinity.

[4] Christakis *et al.* (2010) appear to avoid the Bhamidi *et al.* (2011) exponential bound because they do not simulate a distribution over the space of networks. The rate of convergence for their algorithm is presently unknown.

quentist techniques. For instance, Boucher and Mourifié (2012) provide a model that can be estimated using maximum likelihood, using a novel approach viewing networks as random fields in attribute space. Sheng (2012) studies partially identified network-formation models, drawing on the set estimation literature for inference. The difference between these two approaches lies in the fact that Boucher and Mourifié (2012) assume there exists a unique equilibrium, while Sheng (2012) permits multiple equilibria and derives identification primarily from the direct implications of pairwise stability. Because network-formation models often admit many equilibria (Sheng, 2012), uniqueness is a strong assumption.

As far as we are aware, all strategic models proposed thus far have assumed complete information. Our estimation strategy differs markedly from these papers because we assume incomplete information, which brings several advantages. For example, unlike Boucher and Mourifié (2012), our model can admit multiple equilibria, and unlike Sheng (2012), the parameters of our model are point-identified, even without strong support assumptions.[5] Moreover, we consider estimation under a sequence of sampling experiments in which the number of agents goes to infinity, while Sheng assumes instead that the number of network observations goes to infinity. We argue below that this is more useful for network data (also see the discussion in Chandrasekhar and Jackson, 2012).

This paper contributes to the literature on estimating games of incomplete information (e.g. Aradillas-Lopez, 2010; Aguirregabiria and Mira, 2007; Bajari, Hong, Kraimer and Nekipelov, 2010) by providing a computationally feasible model for which the action space is large and multidimensional and that can be estimated consistently as the number of players in the game goes to infinity. A multidimensional action space is a difficult case to handle because of the potential enormity of this space, as even with k binary actions, its cardinality is 2^k. Indeed, in our case, the dimension of the action space grows with the network size ($k = n - 1$), which appears to add to the computational complexity. We use this to our advantage by treating each linking decision as an observation. Whereas most papers in this literature assume that the econometrician observes a large number of independent games, we assume she instead observes a small number of large games.[6] In other words, we send

[5] Achieving point identification for games of complete information often requires large-support assumptions on covariates. See e.g. Kline (2012).

[6] Other papers considering large-market asymptotics in game-theoretic settings include Brock and Durlauf (2001), Bisin, Moro and Topa (2011), Fox (2010), Menzel (2012), Shang and Lee (2011), and Song (2012). These papers do not accommodate network-formation games. Chandrasekhar and Jackson (2012) consider large-market asymptotics for exponential random graph models, and Boucher and Mourifié (2012) analyze strategic network-formation models for large markets. Our estimation strategy bears some resemblance to that of Shang and Lee (2011), who estimate a peer effects model that admits multiple equilibria by conditioning on groups to eliminate correlation between agents' actions.

the size of the network to infinity ("large-market" asymptotics), rather than the number of network observations ("multi-market" asymptotics). In network data, researchers typically observe only a small number of networks, but these networks tend to have many agents. Hence, what we would like to say is that having a large number of agents is akin to having a large number of observations. Deriving consistency under large-market asymptotics provides a formal justification for this idea.

In the incomplete-information setting, the multi-market approach requires that the same equilibrium is played in games with identical agents (e.g. Aguirregabiria and Mira, 2007). Our assumption that the equilibrium in the data is anonymous plays an analogous role in the large-market context. Interestingly, when attributes are continuous, we need not impose smoothness assumptions on the equilibrium selection mechanism, unlike the multi-market case (Bajari *et al.*, 2010).

The paper is structured as follows. Section 2 presents an overview of our estimation strategy as it relates to the popular dyadic regression approach. In section 3, we develop the model and derive the likelihood. We outline our estimation strategy in section 4 and derive the asymptotic properties of our estimator. In section 5, we apply our model to study the formation of risk-sharing networks in Indian villages. We provide a method for simulating counterfactuals in section 6. Section 7 extends the model to accommodate undirected networks. Finally, section 8 concludes.

2 Overview

Consider the formation of a friendship network among n students. Assume a student i's vector of attributes X_i is two-dimensional, consisting of her race and her parents' income, so that $X_i = (R_i, M_i)$, where R_i is a race indicator (assume two races for simplicity) and M_i is parental income. Suppose a researcher is interested in whether or not friendships are homophilous in race. A common approach is to estimate a dyadic regression model. This is a binary-choice model in which the dependent variable is a potential link G_{ij} that evaluates to one if i is friends with j and zero otherwise.[7] The right-hand side variables are M_i, M_j, and $|R_i - R_j|$, the latter capturing homophily. The model can be microfounded by assuming that i receives utility

$$u_{ij}(X_i, X_j; \theta) + \varepsilon_{ij} = \theta_0 + M_i\theta_1 + M_j\theta_2 + |R_i - R_j|\theta_3 + \varepsilon_{ij} \tag{1}$$

[7] We assume here that friendships are directed. This is actually consistent with the Add Health data on high school friendship networks in which we see that friendships are not necessarily reciprocated.

from linking with j, so she forms a link if and only if $u_{ij}(X_i, X_j; \theta) + \varepsilon_{ij} \geq 0$. In other words, the model assumes an agent i's decision to form a link with an agent j only depends on the characteristics of agents i and j and not, for instance, on which agents choose to link with j. This assumption of zero network externalities is attractive because it tremendously simplifies the problem by ruling out any strategic considerations. However, it also assumes away potentially crucial incentives in the network-formation process.

Now suppose that student i also wishes to link with popular students. Let G be the network adjacency matrix and G_{-i} be the matrix with row i removed. Then we might model payoffs as

$$u_{ij}(G_{-i}, X_i, X_j; \theta) = \theta_0 + M_i\theta_1 + M_j\theta_2 + |R_i - R_j|\theta_3 + \theta_4 \sum_{k \neq i,j} G_{kj}. \tag{2}$$

Popularity in this model is measured by j's in-degree, i.e. the number of links to j. This new model cannot be estimated using dyadic regression because including popularity creates a simultaneity problem: an agent's linking decisions now depend on other agents' linking decisions. This creates new challenges for estimation because the model may have no reduced form; it may be incomplete or incoherent, meaning there may be multiple equilibria or no equilibria for certain values of X_i, X_j, and ε_{ij} (see Tamer, 2003).

We provide an estimation approach that can handle these complications. In our model, as for e.g. Brock and Durlauf (2001) and Bajari *et al.* (2010), the realization of ε_{ij} is private information for agent i. Hence, agents form links by maximizing expected utility given beliefs about the state of the network, replacing network links G_{kj} in (2) with conditional probabilities $\sigma_{kj}(X) \equiv \mathbb{P}(G_{kj} = 1 \mid X_1, ..., X_n)$. We say a network is in *equilibrium* if the beliefs coincide with the actual linking probabilities. Let $\big((\sigma_{kl}(X)\big)_{k \neq i}$, be the analog of G_{-i}, where we replace each entry G_{kj} in G_{-i} with $\sigma_{kj}(X)$. The likelihood of our model turns out to be

$$\mathbb{P}(G \mid X) = \prod_{ij: i \neq j} \Phi\left(u_{ij}\big((\sigma_{kl}(X))_{k \neq i}, X_i, X_j; \theta\big)\right)^{G_{ij}} \left(1 - \Phi\left(u_{ij}\big((\sigma_{kl}(X))_{k \neq i,l}, X_i, X_j; \theta\big)\right)\right)^{1 - G_{ij}},$$

where Φ is the CDF of ε_{ij}. This likelihood is identical to the likelihood of the dyadic regression model if u_{ij} is given by (1). In general, however, u_{ij} can depend on the network G, as in specification (2), in which case the likelihood is a function of unknown nuisance parameters $\sigma_{kl}(X)$. An equilibrium always exists in our model but is not necessarily unique, so the model is incomplete. Nonetheless, we show that when the size of the network is large, a separability restriction on payoffs and the assumption of anonymous beliefs allow us to consistently estimate the structural parameters θ by standard methods, after replacing the nuisance parameters with nonparametric estimates. Thus, our model generalizes the dyadic regression approach by permitting strategic externalities at the cost of adding an additional estimation step.

Estimating $\sigma_{kl}(X)$ for all $k, l \in \{1, ..., n\}$ is nontrivial when $n \to \infty$, since the number of such functions goes to infinity and their dimensions go to infinity, as well. Nevertheless, we show that this curse of dimensionality problem can be avoided if we assume that equilibrium linking probabilities are the same for pairs of agents with identical characteristics, i.e. $\sigma_{kl}(X) = \sigma_{ij}(X)$ if $X_i = X_k$ and $X_j = X_l$. Estimating the nuisance parameters is then possible using simple frequency or kernel estimators. This is the assumption of equilibrium *anonymity* or *symmetry* theorists commonly impose due to the natural symmetry of the game-theoretic environment. The idea is that if agents are similar and all possess the similar information, as in our setting, then they should act similarly from an *ex-ante* perspective. Notice that it does *not* imply that identical agents form the same links because observationally equivalent agents still possess different draws of ε_{ij}.

3 Model

We model the formation of a directed network as a static game of incomplete information.[8] Agents are endowed with exogenous attributes, which are common knowledge. An agent's payoff from forming a particular link depends on a random utility component that is private information. Given beliefs over the linking decisions of others, agents simultaneously form links. Formally, the model is as follows.

Players. There are n agents, each endowed with an exogenous vector of attributes $X_i \in \mathbf{X}$. Components of X_i can include attributes such as race and income. We assume that $\mathbf{X}$ is a bounded subset of $\mathbb{R}^d$ and let $X = (X_1' \ ... \ X_n')$, which we call the *profile*.

Each pair of agents is endowed with a *pair-specific characteristic* Z_{ij} that lies in some bounded set $\mathbf{Z} \subset \mathbb{R}^c$. We arbitrarily set Z_{ii} equal to the zero vector for all i. In general, Z_{ij} can be a vector, so we collect them in a three-dimensional array Z. For instance, in our application we consider the formation of risk-sharing networks, controlling for family relationships. Such relationships are modeled as a network, so that Z_{ij} denotes whether

[8] We later discuss how to extend our approach to accommodate undirected networks.

or not i and j are relatives. Z_{ij} can also include variables such as the geographic distance between i and j.

Actions. Any directed network on n nodes $G \equiv G_n$ is formally a matrix with ijth component $G_{ij} \equiv G_{n,ij}$, such that $G_{ij} = 1$ if agent i links to agent j and $G_{ij} = 0$ otherwise. Then agents select actions $G_i = (G_{i1}, ..., G_{i,i-1}, G_{i,i+1}, ..., G_{in}) \in \mathbb{R}^{n-1}$. We call G_{ij} a *potential link*. There are no self-links, so $G_{ii} = 0$ for all i. Let G_{-i} be the matrix G with the ith row deleted. Usually we will suppress the dependence of $G_{n,ij}$ on n. We will often do so similarly for preferences $u_{ij} \equiv u_{n,ij}$ and beliefs $\sigma_{ij} \equiv \sigma_{n,ij}$, defined below.

Payoffs. Agent i receives a payoff that can be decomposed into a random component and a deterministic component, $\pi_i(g, X, Z)$. We impose the following substantive restrictions on payoffs.

Assumption 1. *Deterministic preferences are given by*

$$\pi_i(g, X, Z) = \sum_{j \neq i} G_{ij} u_{n,ij}(G_{-i}, X, Z_{ij}),$$

which satisfies the following restrictions.

1. *(Additive Separability) As displayed above, $\pi_i(g, X, Z)$ is additively separable in each G_{ij}, and the link-specific payoff $u_{n,ij}$ does not depend on G_i.*
2. *(Linearity) $u_{ij}(\cdot, X, Z_{ij}) \equiv u_{n,ij}(\cdot, X, Z_{ij})$ is linear in each G_{jk} for $j \neq i$.*
3. *(Anonymity) For any $i \neq j$, u_{ij} is an anonymous function at the realized values of (G_{-i}, X, Z_{ij}), so that for any bijective function $\varphi : \{1, ..., n\} \mapsto \{1, ..., n\}$ (a "permutation" of labels),*

$$u_{\varphi(i)\varphi(j)}\big((G_{kl})_{k \neq i,l}, X_1, ..., X_n, Z_{ij}\big) = u_{ij}\big((G_{\varphi(k)\varphi(l)})_{k \neq i,l}, X_{\varphi(1)}, ..., X_{\varphi(n)}, Z_{\varphi(i)\varphi(j)}\big).$$

We will discuss Assumption 1 in more detail in section 3.2. For now we make two points. First, anonymity simply means that payoffs do not depend on agents' identities or labels, which is natural when the labels given to nodes are arbitrary, as in many applications. However, it rules out models in which different agents have different roles. Anonymity implies that we can write

$$u_{ij}(G_{-i}, X, Z_{ij}) = u(X_i, X_j, X_{-i,-j}, G_j, G_{-i,-j}, Z_{ij}),$$

where $X_{-i,-j}$ is the attribute profile with the attributes of agents i and j removed, and similarly for $G_{-i,-j}$. The function u is invariant with respect to permutations of indices in $X_{-i,j}$ and $G_{-i,-j}$. Second, the linearity assumption is without loss of generality. For example, take a simple case where $u_{ij} = f(g_{ji}, g_{j3})$. This function can be rewritten as $g_{ij}(1-g_{j3})f(1,0)+g_{ji}g_{j3}f(1,1)+(1-g_{ji})g_{j3}f(0,1)+(1-g_{ji})(1-g_{j3})f(0,0)$, which is linear in each link.

We also require that payoffs satisfy certain regularity conditions. In what follows, the derivative $\frac{\partial}{\partial G_{kl}}u_{ij}$ is well defined by the linearity assumption above if we reinterpret each G_{kl} as a continuous variable on $[0,1]$ with support $\{0,1\}$.

Assumption 2. *Link-specific payoffs satisfy the following conditions.*

1. *(Parametrization) The function $u_{n,ij}(G_{-i}, X, Z_{ij})$ is known up to a finite-dimensional parameter $\theta_n^\circ \in \Theta \subset \mathbb{R}^p$, so that $u_{ij}(G_{-i}, X, Z_{ij}) = \tilde{u}_{ij}(G_{-i}, X, Z_{ij}, \theta_n^\circ)$, differentiable in θ_n°.*

2. *(Finiteness) Let $\theta_{n,m}^\circ$ be the mth component of θ_n°. Given a fixed sequence of parameters $\{\theta_n^\circ\}_{n=1}^\infty$, the following random variables are finite almost surely: $\sup_{i,j,n} u_{n,ij}$, $\sup_{i,j,n} \frac{\partial}{\partial\theta_{n,m}^\circ}u_{n,ij}$, $\sup_{i,j,n}\sum_{k\neq l}\frac{\partial}{\partial G_{kl}}\frac{\partial}{\partial\theta_{n,m}^\circ}u_{n,ij}$, and $\sup_{i,j,n}\sum_{q\neq r}\sum_{k\neq l}\frac{\partial^2}{\partial G_{kl}\partial G_{qr}}\frac{\partial}{\partial\theta_{n,m}^\circ}u_{n,ij}$.*

Finiteness ensures that the link-specific payoff function and its derivatives do not become infinite as more agents are added. This condition is needed because we consider a sequence of experiments in which the number of agents goes to infinity.

Overall utility $U_i(g, X)$ is the sum of the deterministic component and a random component, so that

$$U_i(g, X, Z) = \sum_{j\neq i} G_{ij}u_{n,ij}(G_{-i}, X, Z_{ij}) + \sum_{j\neq i} G_{ij}\varepsilon_{ij}.$$

The term ε_{ij} is a link-specific random shock, which captures unobserved factors that influence linking decisions. We call (X_i, ε_i) the *type* of agent i, where $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{i,i-1}, \varepsilon_{i,i+1}, \dots, \varepsilon_{in})$.

Observables. The econometrician observes the network G, attribute profile X, and pair-specific characteristics Z. Link-specific shocks ε_{ij} are unobserved.

Assumption 3 (Distribution)**.** *Shocks ε_{ij} are i.i.d. with full support on $\mathbb{R}$, density ϕ, strictly monotonic CDF Φ, and distribution symmetric about zero and independent of X and Z. Attributes $(X_i)_i$ and pair-specific characteristics $(Z_{ij})_{ij}$ are identically distributed.*

Symmetry of the distribution is used to simplify the exposition and is inessential. It is straightforward to allow ε_{ij} to depend on n and to be correlated with (X_i, X_j, Z_{ij}). However, as with standard discrete choice models, the conditional CDF must then be known to the researcher. The i.i.d. and independence assumptions are common in the econometric literature on network formation (e.g. Boucher and Mourifié, 2012; Christakis *et al.*, 2010; Mele, 2011; Sheng, 2012). Indepenence between shocks can be relaxed to allow arbitrary correlation between ε_{ij} and ε_{ik} for any $j, k \neq i$, in which case the potential links will be sparsely correlated conditional on (X, Z). Then the rate of convergence in the case of discrete attributes should be $\sqrt{n}$ rather than n, since the number of independent observations is reduced to $O(n)$ from $O(n^2)$. Limit theorems for such data exist (see e.g. Lumley and Mayer-Hamblett, 2003), so extending the theorems in Appendix B to allow for sparse correlation is feasible. We leave this to future research.

Notice that the i.i.d. assumption is weaker than the usual requirement in single-agent discrete choice models that the random utility components for each choice are mutually independent. In this setting, an agent's choice is a vector of $n-1$ links, so for two choices $G_i, \tilde{G}_i$, if $G_{ij} = \tilde{G}_{ij} = 1$, then ε_{ij} enters the payoffs of both choices. Hence, the random utility components are not mutually independent across different choices.

Shocks ε_{ij} can capture factors such as search costs, idiosyncratic network shocks, or intangibles such as the disposition of agents when they first "meet" and decide whether or not to link. The following assumption reflects the fact that agents have noisy information about the linking decisions of others due to these shocks, a possibility that is ruled out by complete-information models that predominate in the literature.

Assumption 4 (Information)**.** *The realization of $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{i,i-1}, \varepsilon_{i,i+1}, ..., \varepsilon_{in})'$ is private information for agent i, and all other features of the model are common knowledge.*

The model is therefore a static game of incomplete information, and our solution concept is Bayesian equilibrium, defined below. This assumption follows the econometric literature on discrete games of incomplete information (e.g. Bajari *et al.*, 2010) and social interactions (e.g. Brock and Durlauf, 2001, 2007). We further discuss Assumption 4 in section 3.2.

3.1 Anonymous Equilibrium

Let $\sigma_{ij}(X, Z) = \mathbb{P}(G_{ij} = 1 \mid X, Z)$ be the equilibrium belief that agent i will link with agent j given covariates. In equilibrium, each agent i simultaneously chooses a vector of directed links G_i that maximizes expected utility conditional on private information, commonly known covariates, and beliefs $(\sigma_{ij}(X,Z))_{i\neq j}$. Then agent i's expected utility from network G, given attributes X and pair-specific characteristics Z, is

$$\mathbb{E}_{G_{-i}}[U_i((G_i,G_{-i}),X,Z)\,|\,X,Z,\varepsilon_i] = \sum_{j\neq i} G_{ij}\left[u_{ij}((\sigma_{kl}(X,Z))_{k\neq i,l},X,Z_{ij}) + \varepsilon_{ij}\right],$$

by the linearity restriction in Assumption 1. Thus, an agent i chooses action G_i if and only if $\sum_{j\neq i}(G_{ij}-\tilde{G}_{ij})\left[u_{ij}((\sigma_{ij}(X,Z))_{i\neq j},X,Z_{ij})+\varepsilon_{ij}\right] \geq 0$ for all $\tilde{G}_i$ in i's action space, and therefore,

$$G_{ij} = 1 \text{ if and only if } u_{ij}((\sigma_{kl}(X,Z))_{k\neq i,l},X,Z_{ij}) + \varepsilon_{ij} \geq 0. \tag{3}$$

This implies that agent i has a separate decision rule for each G_{ij}. In other words, the center agent in Figure 1 forms a direct link to a particular subtree if and only if the total expected utility of that link is positive. Hence, the chance that agent i links to agent j is

$$\mathbb{P}(G_{ij}=1\,|\,X,Z_{ij}) = \mathbb{P}\Big(\varepsilon_{ij} \geq -u_{ij}((\sigma_{kl}(X,Z))_{k\neq i,l},X,Z_{ij})\,\Big|\,X,Z_{ij}\Big), \tag{4}$$

The system (4) defines an (*ex-ante*) best-response mapping $\Gamma(\cdot,X,Z) : [0,1]^{n(n-1)} \to [0,1]^{n(n-1)}$, which takes as its argument a vector of beliefs and outputs a vector of conditional linking probabilities. Following Aguirregabiria and Mira (2007), we define a *Bayesian equilibrium* as a vector-valued "belief function" $\sigma_n(\cdot)$ such that for all X and Z, $\sigma_n(X,Z) = \Gamma(\sigma_n(X,Z),X,Z)$. That is, $\sigma_n(X,Z)$ is a fixed point in the best-response mapping Γ. The ijth component of σ_n corresponds to the function $\sigma_{ij}(X,Z)$ defined in the previous subsection. In general, this Γ may have multiple fixed points, each of which corresponds to a different equilibrium. In the following theorem, we demonstrate that for a given (X,Z) the mapping has a fixed point that is differentiable and *anonymous at* (X,Z), meaning that for a permutation of labels φ,

$$\sigma_{\varphi(i)\varphi(j)}(X_1,...,X_n,Z_{12},...,Z_{n,n-1}) = \sigma_{ij}(X_{\varphi(1)},...,X_{\varphi(n)},Z_{\varphi(1)\varphi(2)},...,Z_{\varphi(n)\varphi(n-1)}).[9]$$

That is, an equilibrium is anonymous if agents with identical attributes act identically. If the equilibrium is also differentiable, then anonymity also implies that similar agents act similarly. Just as with anonymous preferences, anonymity of beliefs implies the existence of

[9] Anonymity has also been called symmetry. Our choice of name is motivated by the similarity between the current definition and the definition of anonymous payoffs.

a function ρ such that

$$\sigma_{ij}(X,Z) = \rho(X_i, X_j, Z_{ij}, X_{-i,-j}, Z_{-ij}). \tag{5}$$

Thus, the function does not depend on labels i and j. We discuss the plausibility of anonymity in section 3.2.

Theorem 1 (Existence)**.** *If* $\mathbf{X}$ *and* $\mathbf{Z}$ *are bounded,* u_{ij} *satisfies anonymity (Assumption 1) and is differentiable up to order* s *in argument* (X,Z)*, and the density for shocks* ε_{ij} *is differentiable up to order* s*, then there exists a Bayesian equilibrium that is anonymous at any* (X,Z) *that is* s*-times differentiable.*

Smoothness is only needed for nonparametric estimation of beliefs when attributes are continuous. Note that the theorem is stronger than what is actually needed; it demonstrates the existence of an equilibrium that is *globally* anonymous, meaning anonymous at any (X,Z), whereas we will only require that the selected equilibrium is anonymous at the *realized* (X,Z). Henceforth, we refer to the latter simply as an *anonymous equilibrium* and drop the reference to (X,Z).

Since an anonymous equilibrium exists, we now make the following assumption.

Assumption 5 (Sampling Experiment)**.** *Let* $\{(X_i, \varepsilon_i)\}_{i=1}^{\infty}$ *be a sequence of types, and let* $\{\theta_n^\circ\}_{n=1}^{\infty}$ *be a sequence of parameters. In the* n*th experiment, the linking probabilities in the induced network-formation game with agents* $1,\ldots,n$ *are rationalized by a single anonymous equilibrium under the parameter* θ_n°*.*

Assumption 5 incorporates two equilibrium restrictions. First, when multiple Bayesian equilibria exist at the realized (X,Z), it requires that a particular equilibrium is chosen by a degenerate equilibrium selection mechanism. This is a common assumption in the literature on estimating games of incomplete information, as it helps ensure that beliefs are immediately identified from the data. For large-market asymptotics, this is not sufficient to ensure identification. The second restriction is that the selected equilibrium is anonymous, which then guarantees identification of beliefs. Notice we need not assume that this selection mechanism is known, nor that it is smooth.

Anonymity "typically" imposes no restriction on the data if attributes and pair-specific characteristics contain a continuous component. Notice that because beliefs coincide with *ex-ante* strategies in equilibrium, anonymity implies that the chance that i links to j equals the chance that k links to l, if both i and k share the same attributes and both j and l share

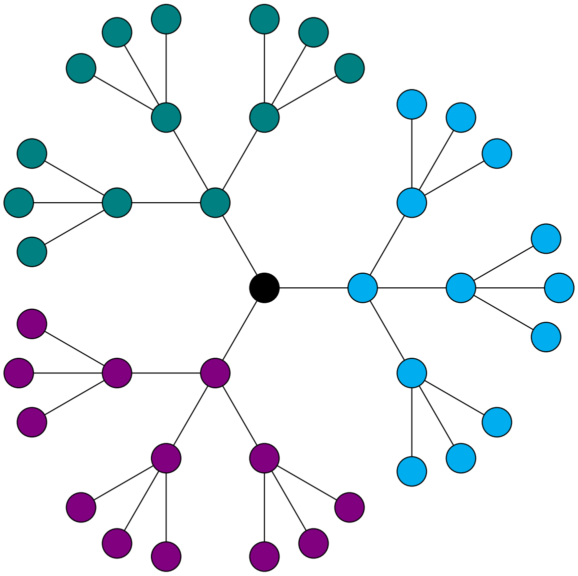


Figure 1: (Additively separable utility.) The utilities from the three trees enter separately into the central agent's utility function. Note that the trees can have nodes in common.

the same attributes. This is the only restriction anonymity imposes on the data. Notice that linking probabilities can be rationalized by an anonymous equilibrium generically if agents possess a continuous attribute, simply because two agents share the same attribute vector if their continuous attributes have the same realization, an event that has zero probability. Therefore, anonymity generically imposes no restriction on the data in the presence of a continuous attribute.

3.2 Discussion of Assumptions

Preferences. The additively separable form of $\pi_i(g, X, Z)$ implies that an agent cares separately about each "tree" subnetwork emanating from each of her direct links. See Figure 1. Separability is assumed in dyadic regression models and is part of the reason why they are computationally attractive. The assumption plays an analogous role in our setting. Anonymity ensures the existence of an anonymous equilibrium, which is crucial for nonparametric estimation of linking probabilities.

Many models of network formation specify utility functions that satisfy Assumption 1, such as the model of Mele (2011). He uses the specification

$$\pi_i((G_i, G_{-i}), X) = \sum_{j=1}^{n} G_{ij} \Bigg[\mu(X_i, X_j) + G_{ji} m(X_i, X_j) + \sum_{\substack{k=1 \\ k \neq i,j}}^{n} G_{jk} v(X_i, X_k) + \sum_{\substack{k=1 \\ k \neq i,j}}^{n} G_{ki} v(X_k, X_j) \Bigg]. \tag{6}$$

Here the first term captures the direct benefit μ of a directed link, the second term the additional benefit m of a reciprocated link, the third term the utility v derived from friends of friends, and the fourth term what Mele calls "popularity." The elements in the brackets correspond to our function u_{ij}. Under our Assumption 1, utility can depend on other statistics, as well, such as the number of agents who link to j ($\sum_{k \neq i,j} G_{kj}$), whether or not there exists a third party linking with both i and j ($\sum_{k \neq i,j} G_{ki} G_{kj}$), or the number of indirect friends who share i's attributes ($\sum_{k \neq i,j} G_{jk} \mathbf{1}\{X_i = X_j\}$).

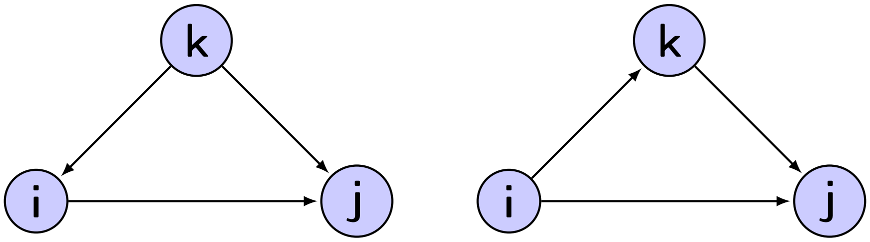
The finiteness assumption simply ensures that link-specific payoffs (and its derivatives) are $O_p(1)$, which is sensible since the shocks ε_{ij} are also $O_p(1)$, and infinite utility is undesirable. Specification (6) does not satisfy finiteness. However, if we assume that μ, m, and v are bounded on their domains, then for most parametrizations (e.g. the common linear-in-parameters specification as in Assumption 6), simply scaling (6) by some $O(\frac{1}{n})$ sequence of constants ensures finiteness. Many utility functions take a form similar to (6) and can be bounded after rescaling by either an $O(\frac{1}{n})$ sequence or an $O(\frac{1}{n^2})$ sequence.[10]

Our assumptions do not permit utility functions of the form in Jackson and Wolinsky (1996) in which agent i derives utility from agent j according to some function of the network distance (length of the shortest path) between the agents.[11] These functions violate separability of the components of G_{ij}, G_{ik} in the utility function. However, we do permit agent i to derive utility separately from *every* path to an agent j, as opposed to only just the shortest path.

Our assumptions do permit utility to depend on the existence of certain subnetwork structures. A common structure of interest is the triangle, which consists of three linked nodes. In our model, the link-specific payoff that i receives from linking with j can depend on type 1 triangles (Figure 2a) but not on type 2 triangles (Figure 2b). In the former case, u_{ij} is a function of $G_{ki} G_{ji}$, while in the latter case, it is a function of $G_{ik} G_{ij}$, violating

[10] In the linear utility specification, these constants are absorbed by the parameters, which are allowed to vary with n, so the data can determine the right scaling.

[11] A path between i and j on the network G is a sequence of links $G_{ia_1}, G_{a_1 a_2}, ..., G_{a_{p-1} a_p}, G_{a_p j}$.



(a) Type 1 triangle (b) Type 2 triangle

Figure 2: To satisfy Assumption 1, agent i's utility can depend on type 1 but not type 2 triangles.

separability. Whether or not this is restrictive depends on the application. In the context of lending networks, in which a directed link from i to j signifies the willingness of i to link to j, it is more sensible to have type 1 triangles enter into the utility function, since the existence of agent k can be a credible signal to agent i regarding the trustworthiness of agent j.

Information. In a game of incomplete information, actions are unobserved, so agents in our model do not observe the network when making their linking decisions. This is a realistic assumption in many applications, as rarely is it the case that agents have complete knowledge of the network. For example when a friendship forms, seldom do the individuals know each others' friends with certainty, let alone their "indirect" friends, those twice or more times removed. People certainly have very little knowledge of the peer groups of individuals to which they are not connected. This supports a model of incomplete information. Furthermore, if i considers j to be a friend, j may not necessarily consider i a friend. This is indeed the case for friendship networks in the Add Health dataset in which friendships can be unidirectional. This may be a reflection of the fact that friendships are intangible and not fully observed, so psychologically their existence may be a matter of probabilistic degree or belief.

Incomplete information allows for *ex-post* regret, meaning agents can make mistakes. Whether or not this is reasonable depends on the context. In friendship networks, the colloquial phenomena of "missed connections" and "third wheels" are indicative of mistakes, and they occur in practice more frequently than not. Of course, if agents eventually receive full information about the network and are not locked into existing relationships, allowing for mistakes may be unreasonable. However, this depends on whether or not the current network is in a long-run state. To the extent that the observed network is only a snapshot of an evolving network, one might expect that some links are not optimally chosen in the data. Moreover, if agents do not fully observe the network at any stage of the game, it is sensible to permit mistakes. In our application, links denote trust, which arguably is never fully observed.

Anonymity. The idea behind anonymity is that if identities are irrelevant for utility by anonymity of preferences, they should also be irrelevant for *ex-ante* equilibrium strategies and therefore beliefs. In the theoretical literature, the focus on anonymous equilibria is usually justified by the natural symmetry of the game-theoretic environment. In our case, agents face the same environment prior to learning their private information, and agents have no intrinsic preference for other particular agents given attributes. This payoff- and informational symmetry should lead to *ex-ante* behavioral symmetry.

Complete informational symmetry is a strong assumption. This is a simplification used extensively in the literature on discrete games of incomplete information, as heterogeneous information is very difficult to analyze tractably. In our model, the largeness of the network lends plausibility to informational symmetry. Consider a school friendship network. If the school is large, as is required by our estimation strategy, most students likely have very little knowledge about who the immediate friends of most students are, let alone their "indirect" friends. In that case, violations of informational symmetry may be localized and minimal, and anonymity is defensible.

4 Estimation

We consider a simple two-step estimation strategy utilized by Aguirregabiria and Mira (2007), Bajari *et al.* (2010), and Brock and Durlauf (2001). In the first step, we estimate beliefs $\sigma_{ij}(X, Z)$ nonparametrically. In the second step, we plug the estimated values from the first step into (10) and then choose the parameters that maximize the resulting pseudo-likelihood. Other estimation methods are also possible for the second step, such as GMM.

4.1 Estimating Beliefs

Nonparametric estimation of $\sigma_{ij}(X, Z)$ is a trivial problem if one observes many independent repetitions of the network-formation game, since the probability that i links to j can be consistently estimated using the empirical frequency with which i links to j (assuming each repetition plays the same equilibrium). However, motivated by the fact that a large number of network observations is rare in practice, we assume that we observe only a single network. Under Assumption 5, we have at our disposal a large number of links G_{ij} from the same market, all of which depend on (X, Z), a vector whose dimension is grows quickly with n. This requires a new estimation approach.

Anonymity of beliefs implies that the chance that i links to j is just the chance that an X_i agent links to an X_j agent, ignoring pair-specific characteristics for simplicity. Moreover if the equilibrium is smooth, if x is "near" X_i and x' is "near" X_j, then the probability that an x agent links to an x' agent is close to the probability that i links to j. The key insight is that we can then approximate the latter probability by the empirical frequency with which x agents link to x' agents using x near X_i and x' near X_j.

Define the probability that an x agent links to an x' agent given pair-specific characteristics z when the set of covariates is $\mathcal{X} = (X, Z)$ as the function

$$\tau_{\mathcal{X}}(x, x', z) := \begin{cases} \sigma_{ij}(X, Z) & \text{if } x = X_i,\ x' = X_j,\ z = Z_{ij} \\ \beta & \text{otherwise} \end{cases} . \tag{7}$$

That is, the function is defined to be equal to the conditional probability that i links to j if agent i has attributes x, agent j has attributes x', and $Z_{ij} = z$. If (x, x', z) corresponds to no observed pair of agents, e.g. there is no i such that $X_i = x$, then the function is arbitrarily defined to be some value β. This is well defined by anonymity of beliefs by (5).

Let $\bar{x} = (x, x', z)$, $\mathbf{x} = (x, x', X_{-i,-j}, z, Z_{ij})$ an unordered random vector,[12] and $\mathcal{X}_{ij} = (X_i, X_j, Z_{ij})$. To motivate our proposed estimators, consistency for the second-stage estimator will eventually require a convergence rate for

$$\sup_{i \neq j} |\hat{\tau}_{\mathcal{X}}(\mathcal{X}_{ij}) - \tau_{\mathcal{X}}(\mathcal{X}_{ij})| \leq \sup_{x, x', z} |\hat{\tau}_{\mathbf{x}}(x, x', z) - \tau_{\mathbf{x}}(x, x', z)|, \tag{8}$$

as usual for two-step estimators with nonparametric nuisance parameters. Note the supremum on the right-hand side is taken with respect to both the arguments in the parentheses and the relevant components of $\mathbf{x}$. If the realization of $\mathcal{X}_{ij}$ is $\bar{x}$, we show that we can find uniformly consistent estimates for the parameter $\tau_{\mathbf{x}}(\bar{x})$. Notice that this is a *random* function, since it depends on $\mathcal{X}$.[13] If the joint support of attributes and pair-specific characteristics is

[12] That is, $\mathbf{x}$ is the equivalence class of all vectors that are component-wise permutations of $(x, x', X_{-i,-j}, z, Z_{ij})$. Technically when we write $\tau_{\mathcal{X}}$, we treat $\mathcal{X}$ as an unordered random vector.

[13] Estimation of random parameters has precedent in the econometrics literature; for instance, Abadie, Imbens and Zheng (2011) study the estimation of conditional best linear predictors, defined by the minimization of an objective that depends on the sample distribution of covariates.

discrete, we show consistency can be achieved with a simple frequency estimator.

Proposition 1. *Let x_i be the realization of X_i and z_{ij} the realization of Z_{ij}, for all i, j. Assume the following.*

(i) The support of (X_i, X_j, Z_{ij}) is finite.

(ii) $\frac{1}{n^2}\sum_{k\neq l}\mathbf{1}\{\mathcal{X}_{kl}=\bar{x}\}\xrightarrow{p}\alpha\in(0,1]$.

Define the frequency estimator for $\tau_{\mathbf{x}}(\bar{x})$ as

$$\hat{\tau}(\bar{x}) := \begin{cases} \frac{\sum_{k\neq l} G_{kl}\mathbf{1}\{X_k=x, X_l=x', Z_{kl}=z\}}{\sum_{k\neq l}\mathbf{1}\{X_k=x, X_l=x', Z_{kl}=z\}} & \text{if } \exists\, i,j : (x_i, x_j, z_{ij}) = \bar{x} \\ \beta & \text{otherwise} \end{cases} . \tag{9}$$[14]

Then $\sup_{\bar{x}}|\hat{\tau}(\bar{x})-\tau_{\mathbf{x}}(\bar{x})|\xrightarrow{p}0$ at rate $\frac{1}{n}$.

That is, to estimate the high-dimensional function $\sigma_{ij}(\mathcal{X})$ at $\mathcal{X}=\bar{x}$, we can use a frequency estimator that averages over all links from X_i agents to X_j agents with pair-specific covariates Z_{ij}.

If covariates have continuous support, we show that beliefs can be nonparametrically estimated using a kernel estimator, provided the equilibrium is smooth.[15]

Proposition 2 (Kernel Rate of Convergence)**.** *Let x be the realization of the attribute profile and z the realization of pair-specific characteristics. Assume the following holds for any i, j.*

(i) The support of X_i is a convex subset of $\mathbb{R}^d$, and the support of Z_{ij} is a convex subset of $\mathbb{R}^c$.

(ii) The density of (X_i, X_j, Z_{ij}) is bounded away from zero on its support.

(iii) $\sigma_{n,ij}(\mathcal{X})$ is s-times differentiable at (x, z), and $\sup_n \sigma^{(s)}_{n,ij}(\mathcal{X})<\infty$ a.s.

(iv) The kernel $K(\cdot)$ has bounded range and satisfies $\int u^r K(u)\,\mathrm{d}u = 0$ for all $r < s$; $\int |u^s| K(u)\,\mathrm{d}u < \infty$; and $\sup_u K(u) < \infty$.

(v) For any $t\in\{0,1,\ldots,s\}$, $\frac{1}{n(n-1)}\sum_{k\neq l}\frac{1}{h^q}K\left(\frac{\bar{x}-\mathcal{X}_{kl}}{h}\right)b^t_{kl}$ converges in probability to its expectation, where b_{kl} is a vector defined in the proof that includes $\mathcal{X}_{kl}$.[16,17]

[14] The choice of β is arbitrary. This definition emphasizes that estimating out of sample is meaningless.

[15] For the case of mixed discrete and continuous attribute components, one can use the approach of Racine and Li (2004).

[16] A sufficient condition for this is $X_i \perp\!\!\!\perp X_j$ if $i\neq j$, and $Z_{ij}\perp\!\!\!\perp Z_{kl}$ if $i\neq k$.

[17] Powers of vectors, e.g. b^t_{kl}, are defined using standard multi-index notation. See the definitions immediately preceding the proof of this proposition in Appendix A.

Define the kernel estimator for $\tau_{\mathbf{x}}(\bar{x})$ *as*

$$\hat{\tau}(\bar{x}) := \begin{cases} \frac{\sum_{k\neq l} K\left(\frac{\bar{x}-\mathcal{X}_{kl}}{h}\right) G_{kl}}{\sum_{k\neq l} K\left(\frac{\bar{x}-\mathcal{X}_{kl}}{h}\right)}. & \textit{if } \exists\, i,j : (x_i, x_j, z_{ij}) = \bar{x} \\ \beta & \textit{otherwise} \end{cases}.$$

Then $|\hat{\tau}(\bar{x}) - \tau_{\mathbf{x}}(\bar{x})| \xrightarrow{p} 0$ *at rate* $h^s + \sqrt{\frac{\log n}{n^2 h^{c+2d}}}$.

Assumption (iii) is an equilibrium restriction, requiring that the sequence of equilibria chosen by the equilibrium selection mechanism satisfy the stated smoothness and finiteness conditions. Recall that the existence of a smooth equilibrium is guaranteed by Theorem 1. This assumption imposes no smoothness conditions on the equilibrium selection mechanism itself, which distinguishes this result from the multi-market setup.

4.2 Consistency and Asymptotic Normality

We next provide conditions under which the second-stage estimator is consistent and asymptotically normal under a sequence of experiments that sends the number of agents to infinity. The form of the likelihood provides the intuition behind these results. By independence of the ε_{ij} shocks and symmetry of the distribution (Assumption 3), using (4), the log-likelihood is

$$\log \mathbb{P}(G \,|\, X, Z) = \sum_{ij:i\neq j} \log \Phi\Big((-1)^{1-G_{ij}} u_{ij}\big((\sigma_{kl}(X,Z))_{k\neq i,l}, X, Z_{ij}\big)\Big) \tag{10}$$

This is a sum over all possible $n(n-1)$ potential links in the network, and computing this sum is tractable. By (3), linearity and additive separability imply that each agent i has a separate decision rule for each potential link G_{ij} she may form, and the rule does not depend on any of the other links G_{ik} for $k \neq j$. This essentially transforms the game between n agents into a game between $n(n-1)$ agents, where each agent is a pair ij (with agent ij distinct from agent ji) that takes a binary action, whether or not to form a directed link. Viewed in this way, the form of the likelihood is reminiscent of the standard discrete choice setting in which $n(n-1)$ agents choose binary actions, the difference being that the index function includes nonparametric nuisance parameters.

The presence of strategic interactions creates additional complications, as (X, Z) appears in each summand of the likelihood and each action G_{ij} is a function of the entire matrix (X, Z). Notice, however, that if (X, Z) are fixed regressors, then the summands are fully independent, since shocks ε_{ij} are independent. By studying the asymptotic behavior of our estimator *conditional on* (X, Z), we effectively fix covariates. Toward this end, we employ and develop extensions of standard limit theory to accommodate triangular arrays with conditionally independent row elements (see Appendix B).

The next theorems will require some notation. Let $\mathcal{X} = (X, Z)$, whose dimension depends on n, and let (x, z) be the realization of $\mathcal{X}$. As defined in (7), $\tau_{\mathcal{X}}$ is the anonymous belief function when covariates are $\mathcal{X}$, and $\hat{\tau}$ is its nonparametric estimate. For convenience, we will suppress the subscript in $\tau_{\mathcal{X}}$. Denote the true parameter for the game with n agents by θ_n°. Further define

- $m_{ij}(G_{ij}, \tau, \theta_n^\circ) \equiv m_{n,ij}(G_{ij}, \tau, \theta_n^\circ) = \nabla_\theta \log \Phi\left((-1)^{1-G_{ij}} \tilde{u}_{ij}\left(\tau, \theta_n^\circ\right)\right)$, the summands of the first derivative of the log-likelihood, abbreviating $\tilde{u}_{ij}\left(\tau, \theta_n^\circ\right) = \tilde{u}_{ij}\left((\tau(\mathcal{X}_{kl}))_{k \neq l}, X, Z_{ij}, \theta_n^\circ\right)$,
- $\bar{m}_n(\tau, \theta_n^\circ) = \frac{1}{n(n-1)} \sum_{i \neq j} m_{ij}(G_{ij}, \tau, \theta_n^\circ)$, the first derivative of the average log-likelihood,
- $M_\theta(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n(n-1)} \sum_{i \neq j} \frac{\phi(\tilde{u}_{ij}(\tau, \theta_n^\circ))^2}{\tau(\mathcal{X}_{ij})(1-\tau(\mathcal{X}_{ij}))} \nabla_\theta \tilde{u}_{ij}(\tau, \theta_n^\circ) \nabla_\theta \tilde{u}_{ij}(\tau, \theta_n^\circ)'$.

The finiteness restriction on preferences ensures that $M_\theta(\mathcal{X})$ is finite almost surely, if this limit exists.

The following theorems require assumptions 1, 2, 3, 4, and 5. The first result is that the second-stage estimator is consistent.

Theorem 2. *Under the following conditions,* $\hat{\theta} - \theta_n^\circ \xrightarrow{p} 0$.

(i) For all $\theta \neq \theta_n^\circ$, $\liminf_n \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}\left\{\tilde{u}_{ij}(\tau, \theta) \neq \tilde{u}_{ij}(\tau, \theta_n^\circ)\right\} > 0$ *for a.s. all* $\mathcal{X}$.

(ii) $\sup_{\bar{x}} |\hat{\tau}(\bar{x}) - \tau(\bar{x})| \xrightarrow{p} 0$.

(iii) $\{\Theta_n\}_{n=1}^\infty$ *is a sequence of compact subsets of* Θ *compact, with* $\theta_n^\circ \in \Theta_n$ *for each* n.

Assumption (ii) merely requires a consistent first-stage estimator. Assumption (i) is an identification condition that assumes sufficient variation in agents' payoffs. As we show in the proof, the assumption implies that the log likelihood is identifiably unique (see Theorem 7 in Appendix B). Proposition 3 provides a sufficient rank condition for this assumption under a linear utility specification.

The following theorem establishes asymptotic normality under the assumption that covariates have finite support and beliefs are estimated using a frequency estimator.

Theorem 3. *Assume the following.*

(i) The assumptions in Proposition 1 hold.

(ii) $\hat{\theta} - \theta_n^\circ \xrightarrow{p} 0$.

(iii) $M_\theta(\mathcal{X})$ *has full rank a.s., and the limits* $M_\theta(\mathcal{X})$ *and* $S(\mathcal{X})$ *(below) exist a.s.*

Then

$$n(\hat{\theta} - \theta_n^\circ) \xrightarrow{\mathcal{X}-d} N\left(0, M_\theta(\mathcal{X})^{-1} + M_\theta(\mathcal{X})^{-1} S(\mathcal{X}) M_\theta(\mathcal{X})^{-1}\right),$$

where

$$\begin{aligned} S(\mathcal{X}) &= \lim_{n\to\infty} S_n(\mathcal{X}) = \lim_{n\to\infty} \Lambda_n(\mathcal{X})\Psi_n(\mathcal{X})\Lambda_n(\mathcal{X})', \\ \Lambda_n(\mathcal{X}) &= \frac{1}{n(n-1)} \sum_{i\neq j} \frac{\phi(\tilde{u}_{ij}(\tau, \theta_n^\circ))^2}{\tau(\mathcal{X}_{ij})(1-\tau(\mathcal{X}_{ij}))} \nabla_\theta \tilde{u}_{ij}(\tau, \theta_n^\circ) \nabla_\tau \tilde{u}_{ij}(\tau, \theta_n^\circ)', \end{aligned}$$

and the $(ij\text{-}kl)$*th element of the* $n(n-1) \times n(n-1)$ *matrix* $\Psi_n(\mathcal{X})$ *is zero if* $\mathcal{X}_{ij} \neq \mathcal{X}_{kl}$ *and otherwise equal to*

$$\frac{\tau(\mathcal{X}_{ij})(1-\tau(\mathcal{X}_{ij}))}{\frac{1}{n(n-1)} \sum_{q\neq r} \mathbf{1}\{\mathcal{X}_{qr} = \mathcal{X}_{ij}\}}.$$

We show in the proof that if the limits in (iii) exist, then they must be finite. By $\xrightarrow{\mathcal{X}-d}$ we mean the convergence in distribution occurs conditional on $\mathcal{X}$ (see Definition 1). The rate of convergence is the parametric rate, which is n rather than $\sqrt{n}$ because each observation is a potential link, not an agent, and there are $O(n^2)$ links in a network of n agents.

The next theorem establishes asymptotic normality under the assumption that covariates have continuous support.

Theorem 4. *Assume the following.*

(i) The assumptions in Proposition 2 hold.

(ii) The kernel smoothing parameter $h = o\left(n^{-\frac{2s}{2s+(c+2d)}}\right)$.

(iii) $\hat{\theta} - \theta_n^\circ \xrightarrow{p} 0$.

(iv) $M_\theta(\mathcal{X})$ *has full rank a.s., and the limits* $M_\theta(\mathcal{X})$ *and* $\Omega(\mathcal{X})$ *(below) exist a.s.*

Then

$$nh^{c+2d}(\hat{\theta} - \theta_n^\circ) \xrightarrow{\mathcal{X}-d} N\left(0, M_\theta(\mathcal{X})^{-1} \Omega(\mathcal{X}) M_\theta(\mathcal{X})^{-1}\right),$$

where

$$\Omega(\mathcal{X}) = \lim_{n\to\infty} \Omega_n(\mathcal{X}) = \lim_{n\to\infty} \Lambda_n(\mathcal{X})\Sigma_n(\mathcal{X})\Lambda_n(\mathcal{X})'$$

and the $(ij\text{-}kl)$*th element of the* $n(n-1) \times n(n-1)$ *matrix* $\Sigma_n(\mathcal{X})$ *is given by*

$$n^2 h^{c+2d} \sum_{q \neq r} \tau(\mathcal{X}_{qr})(1-\tau(\mathcal{X}_{qr})) \frac{K\left(\frac{\mathcal{X}_{ij}-\mathcal{X}_{qr}}{h}\right) K\left(\frac{\mathcal{X}_{kl}-\mathcal{X}_{qr}}{h}\right)}{\left(\sum_{s \neq t} K\left(\frac{\mathcal{X}_{ij}-\mathcal{X}_{st}}{h}\right)\right)^2}.$$

We show in the proof that if the limits in (iv) exist, then they must be finite. Assumption (ii) simply says that we choose h to undersmooth in order to eliminate a bias term. In this theorem, convergence occurs slower than the parametric rate because we condition on covariates. Semiparametric estimators that converge at the parametric rate typically need to average over covariates, meaning the estimators need to be full means, in order to achieve the parametric rate (see Newey, 1994). In our model, because there are $n(n-1)$ link observations but only n independent attribute vectors, kernel estimators for the density of (X_i, X_j, Z_{ij}) converge at a $\sqrt{n}$ rate, rather than $\sqrt{n^2}$. Due to a lack of averaging over covariates, this divergence in rates leads to a violation of Newey's mean-square continuity condition needed to achieve a parametric $\sqrt{n^2}$ rate of convergence.

We lastly provide a primitive condition for assumption (i) of Theorem 2 when the utility function satisfies a commonly used linearity restriction.

Assumption 6. *The function* $\tilde{u}_{ij}(G_{-i}, X, Z_{ij}, \theta_n^\circ)$ *is linear in* $\theta_n^\circ \in \Theta \subset \mathbb{R}^p$*, so that there exists a vector-valued function* $H_{ij}(G_{-i}, X, Z_{ij})$*, with range in* $\mathbb{R}^p$ *such that* $\tilde{u}_{ij}(G_{-i}, X, Z_{ij}, \theta_n^\circ) = H_{ij}(G_{-i}, X, Z_{ij})'\theta_n^\circ$*.*

An example of this is if preferences take the form in (6) with μ, m, and v linear in parameters, i.e. for $w \in \{\mu, m, v\}$, $w_{ij} = \sum_{q=1}^{p} \theta_{wq} H^{wq}(X_i, X_j)$.

Proposition 3 (Identification in the Linear Case)**.** *Under the following conditions, assumption* (i) *of Theorem* 2 *is satisfied.*

(i) Preferences satisfy Assumption 6*.*

(ii) Let H *be a* $p \times n(n-1)$ *matrix with* (q, ij)*th entry* $H_{ij}^q\big((\tau(\mathcal{X}_{kl}))_{k \neq l})_{k \neq i,l}, X, Z_{ij}\big)$ *for* $q \in \{1, ..., p\}$*. For any* n*,* H *has full rank for a.s. all* $\mathcal{X}$*.*

The rank condition also ensures that $M_\theta(\mathcal{X})$ is invertible.

The asymptotic variances can be consistently estimated by replacing θ_n° and τ with their estimators (because the nonparametric estimator is uniformly consistent) and expectations with their sample analogs. Consistency for these estimators follows from arguments made in the proof of Theorem 2.

Finite-Sample Bias. Two-step estimators for games of incomplete information can have large finite-sample bias in the first-step estimator when the attribute space $\mathbf{X}$ is high-dimensional. These concerns can be addressed through the use of various smoothing estimators, as we do in our application (see e.g. Delgado and Mora, 1995; Racine and Li, 2004). The parameter estimates we find are fairly robust across a reasonable range of smoothing parameters, despite the inclusion of many covariates. In fact the frequency estimator without smoothing yields quite similar estimates to the smoothed estimators. Another alternative is the nested pseudo-likelihood method proposed by Aguirregabiria and Mira (2007). This method has been shown to have finite-sample advantages over the two-step estimator in some contexts.

Sampled Networks. Most models of network formation require data on the full network, which is seldom the case in practice. Our estimation strategy can allow for sampled link data but not sampled covariate data. That is, we need to know the characteristics of all agents in the network, but we only need to observe a subset of the links. Consistent estimation is then possible using only the observed subset of links because linking probabilities for pairs of agents whose links are unobserved can be estimated from the first stage using the linking probabilities of similar pairs of agents.

Suppose network links are sampled as follows. A researcher chooses m agents to survey and asks each agent to name her social connections to any agents in the network. In this case, we observe all links G_{ij} such that i is a surveyed agent. Treating the sampled network as the full network creates bias because if G_{ij} is unobserved, then it is coded as $G_{ij} = 0$. Such bias can be avoided by computing the first-stage estimates for all pairs of agents using only the set of observed links and forming the second-stage likelihood only using this observed set.

5 Application: Risk-Sharing Networks

We examine the formation of risk-sharing networks in rural villages in southern India. In particular, we study the extent to which such networks are homophilous in caste, religion, and gender and the extent to which they depend on endogenous signals of trustworthiness such as the number of agents who trust the borrower and the number of agents that the borrower trusts to lend money. We use data on risk-sharing networks from 75 rural villages in India collected in 2006 (Banerjee, Chandrasekhar, Duflo and Jackson, 2012; Jackson, Rodriguez-Barraquer and Tan, 2012). Household characteristics were collected in full village censuses, while individual and network data were collected from random samples of individuals in each village. We assume networks are closed societies and that we observe the full network of lending relations. In our model, direct links are obtained from the survey question, "Whom do you trust enough that if he or she needed to borrow Rs. 50 for a day you would lend it the him or her?"[18] Hence, a direct link from i to j exists if, in the survey, i names j as an individual who i trusts enough to lend a substantial amount of money. A link is therefore a social relation, rather than an indication of an actual monetary transfer.

Villages are mostly homogeneous along linguistic and religious lines with the majority being Hindu, although there are some Muslim and Christian minorities. Villages are quite heterogeneous in caste. Because we are interested in homophily in religion, we use villages that are at least 10 percent non-Hindu to avoid collinearity problems, so we only use nine of the 75 villages. Despite the small sample of networks, our actual sample size is large because there are 492690 link observations. Table 1 presents summary statistics on these nine villages, and Figure 3 displays one such village with nodes colored by caste.

Table 1: Summary statistics

	mean	sd	min	max
# villagers	226	67	98	303
average age	38.5	1.4	35.8	40.6
% female	.56	.02	.55	.59
% Hindu	.79	.11	.58	.92
% OBC	.62	.13	.43	.76
% Scheduled	.30	.09	.21	.44

Note: Scheduled castes are at the bottom of the caste hierarchy, and OBC castes are just above them. All other castes fall into a general category at the top of the hierarchy.

We use a linear utility specification for u_{ij} (Assumption 6) and include controls for characteristics of i and j, including age, gender, religion, caste, education level, spoken languages, and whether or not they are heads of their households. Individuals can either be Hindu or non-Hindu, the latter including Muslims and Christians; there are very few Christians in the sample. Individuals can belong to one of three possible caste categories; see Table 1.

We also assume u_{ij} depends on whether or not i and j share the same religion, gender, spoken languages, or caste and also whether or not they are related; these factors capture homophily. We include the following endogenous determinants of lending: number of people j trusts ($\sum_{k\neq i,j} G_{jk}$), number of people i trusts; whether or not j trusts i (G_{ji}); number

[18] Rs. 50 is roughly a dollar, and per capita income in India is around three dollars per day (Jackson *et al.*, 2012).

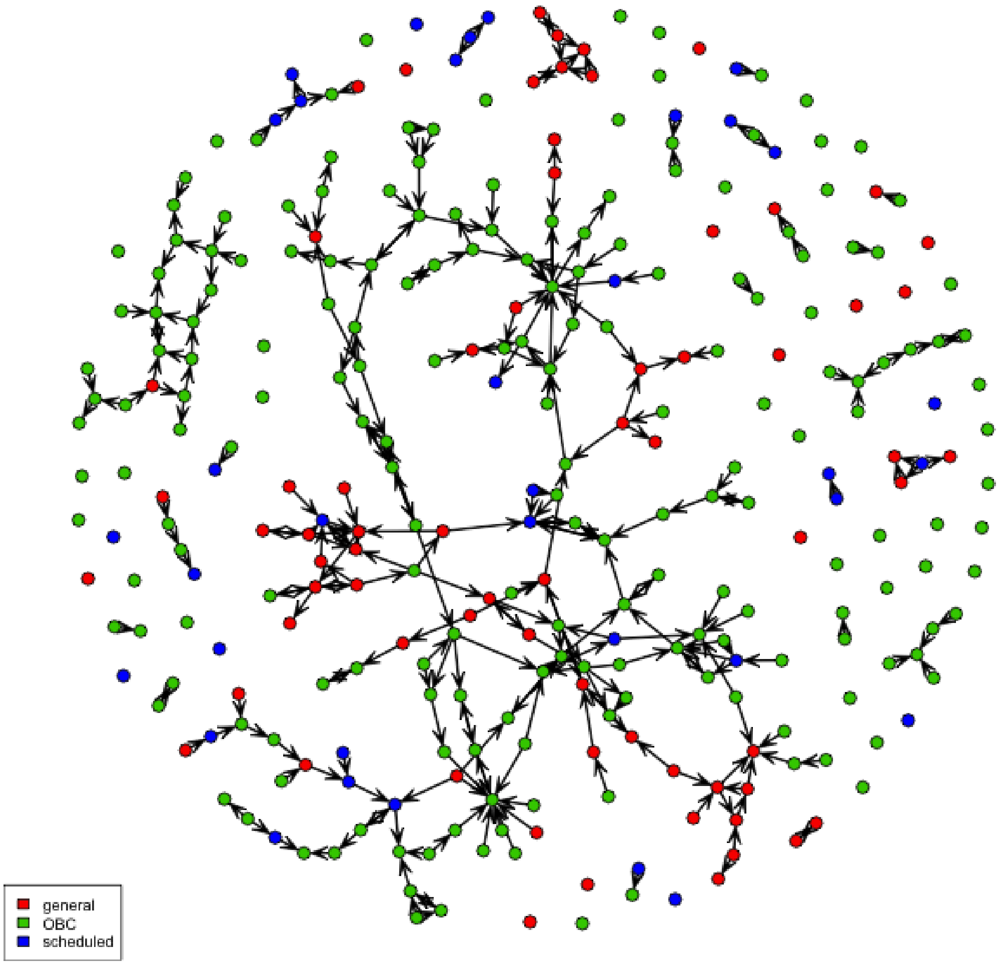


Figure 3: Homophily in caste: scheduled castes are at the bottom of the caste hierarchy, and OBC castes are just above them. All other castes fall into a general category at the top of the hierarchy.

of people who trust *both* i and j ($\sum_{k\neq i,j} G_{ki}G_{kj}$); and the number of people who trust j ($\sum_{k\neq i,j} G_{kj}$). We also allow u_{ij} to depend on the number of people j trusts or who trust j and additionally share i's caste or religion (e.g. $\sum_{k\neq i,j} G_{jk}\mathbf{1}\{C_i = C_k\}$, where C_i is i's caste).

The random utility component is assumed to be normally distributed. To deal with finite-sample bias in the first stage, we use a smoothed frequency estimator proposed by Racine and Li (2004).[19] The amount of smoothing is controlled by a weighting parameter λ.

[19] In the case of purely categorical data, this estimator is given by

$$\hat{\tau}(\bar{x}) = \frac{\sum_{k\neq l} G_{kl}\lambda^{d(\mathcal{X}_{kl},\,\bar{x})}}{\sum_{k\neq l}\lambda^{d(\mathcal{X}_{kl},\,\bar{x})}},$$

The case of no smoothing is $\lambda = 0$, in which case the smoothed frequency estimator coincides with the standard frequency estimator, while the case of $\lambda = 1$ corresponds to placing full weight on all observations: $\frac{1}{n(n-1)} \sum_{i \neq j} G_{ij}$.

Table 2 presents coefficient estimates for the homophily parameters and the constant term across a range of smoothing parameters. The constant is negative and large because networks are sparse; most potential links do not form. It is clear from the table that the estimates are highly robust across a range of smoothing parameters and that homophily in religion, caste, gender, and family are always statistically significant. By far the most important determinant among these four is whether or not i and j are relatives.

Table 2: Estimates for homophily parameters.

	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
cons	-4.0282***	-4.1258***	-4.2342***	-4.366***
	(0.10864)	(0.11399)	(0.12208)	(0.13781)
related	1.3984***	1.3836***	1.3899***	1.415***
	(0.045419)	(0.046387)	(0.047299)	(0.049006)
same caste	0.23882***	0.24055***	0.24336***	0.24842***
	(0.027907)	(0.029973)	(0.033053)	(0.03965)
share language	0.020824	0.020124	0.019208	0.01761
	(0.015689)	(0.01601)	(0.016505)	(0.017675)
same religion	0.44888***	0.44072***	0.4323***	0.42244***
	(0.037424)	(0.040643)	(0.045682)	(0.054691)
same gender	0.69143***	0.69527***	0.70167***	0.71072***
	(0.024763)	(0.024923)	(0.025075)	(0.02529)

Note: Standard errors are in parentheses. (*) denotes significance at the 10% level, (**) the 5% level, and (***) the 1% level.

Table 3 presents coefficient estimates for some endogenous determinants of lending. These point estimates are less robust to different smoothing parameters. However, all coefficients in the table are always highly statistically significant. The largest magnitudes are the coefficients for reciprocal trust (whether or not j trusts i) and what might be called supported trust (the number of individuals willing to lend to both i and j; see figure 4). The latter is likely important because the willingness of k to lend to j is a positive signal for i regarding

where $\lambda \in [0, 1]$ and $d(\mathcal{X}_{kl}, \bar{x})$ is the number of disagreeing components between $\mathcal{X}_{kl}$ and $\bar{x}$. The standard frequency estimator divides the links into bins according to the covariates of the lender and the lendee and averages links only within bins. For instance, $\hat{r}(x, x', z)$ is computed by averaging all links from x agents to x' agents when the pair-specific characteristic is z. In contrast, the smoothed frequency estimator places positive weight on all observations, with more weight placed on links that have covariates similar (i.e. having fewer disagreeing components) to the covariates (X_i, X_j, Z_{ij}) that define the bin.

j's trustworthiness. Agent k's trust in j may be a more credible signal for i than other lenders because she also trusts i. The importance of reciprocal trust is intuitive: if j trusts i, then i is more likely to trust j. The remaining coefficient estimates and standard errors can be found in Appendix C.

Overall we find that the most important determinants of trust-in-lending relations are reciprocal trust, supported trust, and the existence of family ties. A dyadic regression approach would miss the first two determinants because such methods cannot control for endogenous determinants of network structure.[20]

Table 3: Estimates for endogenous determinants.

	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
# lendees of j	-0.20424***	-0.23549***	-0.27069***	-0.32916***
	(0.036231)	(0.042925)	(0.053407)	(0.074983)
j lends to i	1.4085***	1.5732***	1.7686***	2.0752***
	(0.046862)	(0.05674)	(0.071177)	(0.098348)
# lenders to i	-0.04664***	-0.045534***	-0.040353***	-0.032367***
	(0.0081956)	(0.0091383)	(0.010327)	(0.012265)
# lenders to i and j	0.73392***	0.92516***	1.0882***	1.2695***
	(0.069765)	(0.099101)	(0.14493)	(0.24224)
# lenders to j	0.3845***	0.45108***	0.54373***	0.69565***
	(0.017494)	(0.022379)	(0.031274)	(0.049119)

Note: Standard errors are in parentheses. (*) denotes significance at the 10% level, (**) the 5% level, and (***) the 1% level.

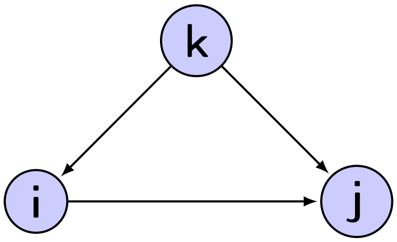


Figure 4: Supported trust: k supports the link G_{ij}.

[20] We have often received the comment that our results invalidate the incomplete-information assumption because the significance of these coefficients demonstrates that villagers do in fact know who trusts (links with) whom. This comment confuses restrictions on information with restrictions on preferences. In any game of incomplete information, actions are necessarily unobserved, which is why agents must form beliefs about the actions of others. This clearly does not preclude the possibility that an agent cares about the actions of other agents, as this possibility is precisely what defines the model as a *game* rather than a single-agent decision problem. Thus, while a villager may not know with certainty who trusts whom, she forms beliefs about the trust network and accordingly decides who she trusts based on these beliefs. See section 3.2 for more discussion of the incomplete-information assumption.

6 Simulating Counterfactuals

When there are multiple equilibria, counterfactual simulation requires some way to choose among them. One method is to select the equilibrium that maximizes the likelihood. In this case, simulation will require finding a solution to a system of equations that maximizes the likelihood, a program that may impose a higher computational burden than estimation. Once the program has been solved, however, simulation is computationally simple, since each link is just a Bernoulli random variable with known success probability. If one wishes to simulate counterfactuals for other equilibria, other methods are needed. See for example Bajari *et al.* (2010).

Recall that in equilibrium, beliefs satisfy

$$\sigma_{n,ij}(X,Z) = \Phi\big(\tilde{u}_{ij}(\sigma_n, X, Z_{ij}, \theta_n^\circ)\big). \tag{11}$$

When simulating counterfactuals, the variables θ_n° (or its estimate), X, and Z are given, so the only unknown is the $n(n-1)\times 1$ dimensional vector $\sigma_n(X,Z)$. In this case, (11) defines a system of $n(n-1)$ equations in $n(n-1)$ unknowns. If these equations have multiple solutions, as they do in general, one possible choice is the equilibrium solution with the highest likelihood. We can compute this solution by solving the following program:

$$\max_{\{\sigma_{ij}\}_{i\neq j}} \sum_{ij:i\neq j} \log\left(\sigma_{ij}\right) \text{ subject to } (11).$$

This is a constrained optimization problem with a smooth, concave objective and smooth, nonlinear constraints, which can be solved by KNITRO (Byrd, Nocedal and Waltz, 2006). We also need to require that the equilibrium is anonymous, which means we need an additional constraint that the belief functions are the same for agents who share the same attributes. With a solution in hand, we can simulate networks by setting $G_{ij} = 1$ with probability $\hat{\sigma}_{ij}$ given by the selected solution.

7 Model for Undirected Networks

Thus far we have assumed that networks are directed. With some modification, the model and estimation strategy can accommodate undirected networks. Let G be an unobserved latent network of *directed* link "proposals." The econometrician does not observe the latent network G. Instead, she observes an undirected network $\tilde{G}$ with ijth entry equal to $G_{ij}G_{ji}$.

Let preferences be given by

$$U_i(g, X, Z) = \sum_{i \neq j} G_{ij} G_{ji} \big(u_{ij}(\tilde{G}_{-i}, X, Z_{ij}) + \varepsilon_{ij} \big).$$

That is, agents choose whether or not to propose links, but a link forms only if the two agents in question propose to each other. Hence, under this model, agents receive utility only under mutual consent; otherwise they receive the outside option of zero. Note that the model is simply an incomplete-information version of Myerson's link announcement model (Myerson, 1977). Equilibria are not necessarily pairwise stable, which is sensible given the incomplete-information environment.

We assume that the function u_{ij} satisfies Assumption 1, with the proviso that the restrictions apply to the observed network $\tilde{G}$ rather than the latent network G.

Likelihood. Let $\tilde{p}_{kl}$ be the belief that k proposes to l and vice versa. Notice $\tilde{p}_{kl} = p_{kl} p_{lk}$ because, as in the original model, link proposals G_{ij} are conditionally independent. Similar to (3), $\tilde{G}_{ij} = 1$ if and only if

$$p_{ji} u_{ij}\big((\tilde{p}_{kl})_{k \neq i,l}, X, Z_{ij} \big) + p_{ji} \varepsilon_{ij} \geq 0 \text{ and } p_{ij} u_{ji}\big((\tilde{p}_{kl})_{k \neq j,l}, X, Z_{ij} \big) + p_{ij} \varepsilon_{ji} \geq 0. \tag{12}$$

As in the directed link setting, (12) defines an *ex-ante* best-response mapping $\Gamma(\cdot, X, Z)$ for beliefs with respect to *link proposals*, and an anonymous Bayesian equilibrium can be defined accordingly and shown to exist. As in the directed model, the finiteness assumption will bound equilibrium beliefs away from zero, so that (12) reduces to

$$u_{ij}\big((\tilde{p}_{kl})_{k \neq i,l}, X, Z_{ij} \big) + \varepsilon_{ij} \geq 0 \text{ and } u_{ji}\big((\tilde{p}_{kl})_{k \neq j,l}, X, Z_{ij} \big) + \varepsilon_{ji} \geq 0. \tag{13}$$

Now the best response functions only depend on beliefs about the undirected network. Thus the log-likelihood is given by

$$\log \mathbb{P}(\tilde{G} \,|\, \mathcal{X}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Bigg[\tilde{G}_{ij} \log \Bigg(\Phi\left(u_{ij}\big((\tilde{p}_{kl})_{k \neq i,l}, X, Z_{ij} \big) \right) \Phi\left(u_{ji}\big((\tilde{p}_{kl})_{k \neq j,l}, X, Z_{ij} \big) \right) \Bigg) \\ + (1 - \tilde{G}_{ij}) \log \Bigg(1 - \Phi\left(u_{ij}\big((\tilde{p}_{kl})_{k \neq i,l}, X, Z_{ij} \big) \right) \Phi\left(u_{ji}\big((\tilde{p}_{kl})_{k \neq j,l}, X, Z_{ij} \big) \right) \Bigg) \Bigg].$$

where the sum contains $\frac{n(n-1)}{2}$ elements. We can estimate equilibrium linking probabilities $\tilde{p}_{kl}$ by the same method detailed in section 4.1 and the structural parameters by the same two-step estimation procedure.

8 Conclusion

This paper develops a strategic model of network formation that allows an agent's linking decisions to depend on the linking decisions of others. We demonstrate that preference restrictions previously utilized in the literature lead to a simple likelihood reminiscent of standard discrete choice models if we model the network-formation process as a static game of incomplete information. The restriction that the observed equilibrium is anonymous allows us to circumvent a curse of dimensionality problem and estimate the model consistently as the number of agents goes to infinity. This is advantageous because network data often feature a small number of large networks. Applying the model to study risk-sharing networks, we demonstrate the importance of endogenous determinants of trust in lending, such as reciprocated trust, which would be missed by a dyadic regression model.

We lastly note that the estimation approach proposed in this paper can be easily applied to a broad class of discrete games with incomplete information, for example models of social interactions with binary actions. If the action space does not grow with n, the required restrictions on the utility function are even weaker, as we no longer need to impose additive separability. This encompasses peer-effects models much more general than the discrete-choice model of Brock and Durlauf (2001), allowing utility to depend on peer actions in more general ways than as group means.

Appendix A: Proofs of Main Theorems

Appendix B: Conditional Asymptotics

Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space and $\mathcal{F}_n \subset \mathcal{G}$ a σ-algebra for all $n \geq 1$. Unless stated otherwise, all random variables in this section are measurable functions from $(\Omega, \mathcal{G})$ to $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra, and have finite expectations. We will sometimes write $\mathbb{E}^{\mathcal{F}} X$ in place of $\mathbb{E}[X \,|\, \mathcal{F}]$ and similarly for conditional probabilities and variances. As a reminder, for a random variable Z, $\mathbb{E}[X \,|\, Z] = \mathbb{E}[X \,|\, \sigma(Z)]$, where $\sigma(Z) = \sigma(\{Z^{-1}(B) : B \in \mathcal{B}\})$, the smallest σ-algebra that contains $\{Z^{-1}(B) : B \in \mathcal{B}\}$. In our application, we will condition on the sequence of σ-algebras with nth element $\mathcal{F}_n = \sigma(X_1, ..., X_n)$.

We now define two notions of conditional convergence. These definitions are related to those of Nowak and Zięba (2005), but we permit the conditioning σ-algebra to depend on n. Note that unlike conditions for martingale limit theorems, $\{\mathcal{F}_n\}_n$ is not required to be a filtration. In this respect, the results differ from those of Menzel (2012).

Definition 1. Let $\{b_n\}$ be a sequence of positive numbers. We denote a random variable $c(\omega)$ on $(\Omega, \mathcal{F})$ by $c(\mathcal{F})$.

- We say X_n *strongly converges in probability to* X *conditionally on* $\{\mathcal{F}_n\}_n$ or sometimes more simply *conditionally converges in probability* if $\mathbb{P}(|X_n| \geq c_n(\mathcal{F}_n)b_n \,|\, \mathcal{F}_n) \xrightarrow{a.s.} 0$ for all sequences $\{c_n(\mathcal{F}_n)\}_n$ such that $\liminf_n c_n(\mathcal{F}_n) > 0$ a.s. We say X_n *weakly converges in probability to* X *conditionally on* $\{\mathcal{F}_n\}_n$ if $\mathbb{P}(|X_n| \geq c_n(\mathcal{F}_n)b_n \,|\, \mathcal{F}_n) \xrightarrow{p} 0$.
- Write $X_n = o_{p|\mathcal{F}_n}(b_n)$ if X_n strongly converges in probability to zero conditionally on $\{\mathcal{F}_n\}_n$.
- A sequence of random variables $\{X_n\}$ $\mathcal{F}_n$*-conditionally converges in distribution* to X if for each continuity point of the distribution of X, we have $\mathbb{P}^{\mathcal{F}_n}(X_n < x) \to \mathbb{P}(X < x)$ a.s. as $n \to \infty$. We denote this by $X_n \xrightarrow{\mathcal{F}_n - d} X$.

Notice that weak or strong conditional convergence in probability implies its unconditional analog by an application of the law of total probability and the dominated convergence theorem. Hence anything that is $o_{p|\mathcal{F}_n}(b_n)$ is also $o_p(b_n)$.

Definition 2. Let $(x_1, ..., x_n) \in \mathbb{R}^n$. A set of random variables $X_1, ..., X_n$ is conditionally independent given $\mathcal{F}$, or $\mathcal{F}$*-independent*, if

$$\mathbb{E}\left[\prod_{i=1}^n \mathbf{1}\{X_i \leq x_i\} \,\middle|\, \mathcal{F}\right] = \prod_{i=1}^n \mathbb{E}\left[\mathbf{1}\{X_i \leq x_i\} \,|\, \mathcal{F}\right] \text{ a.s.}$$

The first result extends the weak law of large numbers to triangular arrays of random vectors with conditionally independent rows.

Theorem 5 (Conditional Weak Law)**.** *Let $X_{n,t}$, $t = 1, ..., v_n$ be row-wise $\mathcal{F}_n$-independent and satisfy $\mathbb{E}^{\mathcal{F}_n}|X_{n,t}| < \infty$ a.s. Define $\tilde{X}_{n,t} = X_{n,t}\mathbf{1}\{|X_{n,t}| \leq v_n\}$, and assume*

(i) $\sum_{t=1}^{v_n} \mathbb{P}^{\mathcal{F}_n}(|X_{n,t}| > v_n) \xrightarrow{a.s.} 0$, and

(ii) $\frac{1}{v_n^2} \sum_{t=1}^{v_n} \mathrm{Var}^{\mathcal{F}_n}(\tilde{X}_{n,t}) \xrightarrow{a.s.} 0$.

Then $\frac{1}{v_n}\left(\sum_{t=1}^{v_n} X_{n,t} - \sum_{t=1}^{v_n} \mathbb{E}^{\mathcal{F}_n}\tilde{X}_{n,t}\right) = o_{p|\mathcal{F}_n}(1)$. If additionally $|X_{n,t}| < M < \infty$ for all n, t, then we can replace $\tilde{X}_{n,t}$ with $X_{n,t}$ in the last expression.

The following theorem extends the Lindeberg CLT.

Theorem 6 (Conditional CLT). *Let $\{X_{n,t}, t = 1, ..., v_n\}$ be $\mathcal{F}_n$-independent random vectors of dimension k with conditional mean zero satisfying $\sum_{t=1}^{v_n} \mathbb{E}^{\mathcal{F}_n} X_{n,t}X'_{n,t} = I_k$, the identity matrix. If the conditional Lindeberg condition holds, i.e. for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sum_{t=1}^{v_n} \mathbb{E}^{\mathcal{F}_n} \left[||X_{n,t}||^2 \mathbf{1}\{||X_{n,t}|| > \varepsilon\} \right] = 0,$$

then $\sum_{t=1}^{v_n} X_{n,t} \overset{\mathcal{F}_n - d}{\longrightarrow} N(0, I_k)$.

The next theorem provides sufficient conditions for the conditional consistency of extremum estimators when the objective is a random variable on $(\Omega, \mathcal{F}_n)$ and both the objective and the underlying parameter depend on n. For fixed $\omega \in \Omega$, define $Q_n^{\mathcal{F}_n} : \Theta \mapsto \mathbb{R}$ for $\Theta \subset \mathbb{R}^p$. For a fixed θ, this is an $\mathcal{F}_n$-measurable function, e.g. an expectation conditional on $\mathcal{F}_n$. Let $\hat{\theta} = \arg\max_\theta \hat{Q}_n(\theta)$, where $\hat{Q}_n(\theta)$ is a function of the data.

Theorem 7 (General Consistency). *Let $\{\Theta_n^\circ\}$ be a sequence of subsets of Θ. Under the following assumptions, $\hat{\theta}_n^\circ - \theta_n^\circ = o_{p|\mathcal{F}_n}(1)$.*

(i) (Compactness) $\theta_n^\circ \in \Theta_n^\circ$ compact.

(ii) (Continuity) For each n, $Q_n^{\mathcal{F}_n}(\cdot)$ is continuous on Θ a.s.

(iii) (Identifiable Uniqueness) For any $\eta > 0$, $\liminf_n \left\{ Q_n^{\mathcal{F}_n}(\theta_n^\circ) - \sup_{\theta:|\theta-\theta_n^\circ|\geq\eta} Q_n^{\mathcal{F}_n}(\theta) \right\} > 0$ a.s.

(iv) (Uniform Convergence) $\sup_\theta |\hat{Q}_n(\theta) - Q_n^{\mathcal{F}_n}(\theta)| = o_{p|\mathcal{F}_n}(1)$.

References

Abadie, A., Imbens, G. and Zheng, F. (2011). Robust inference for misspecified models conditional on covariates. *Working paper, Harvard University*. 18

Aguirregabiria, V. and Mira, P. (2007). Sequential estimation of dynamic discrete games. *Econometrica*, **75** (1), 1–53. 5, 6, 12, 17, 24

Aradillas-Lopez, A. (2010). Semiparametric estimation of a simultaneous game with incomplete information. *Journal of Econometrics*, **157** (2), 409–431. 5

Bajari, P., Hong, H., Kraimer, J. and Nekipelov, D. (2010). Estimating static models of strategic interactions. *Journal of Business and Economic Statistics*, **28**, 469–482. 5, 6, 7, 11, 17, 29

Bandiera, O., Barankay, I. and Rasul, I. (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica*, **77** (4), 1047–1094. 2

Banerjee, A., Chandrasekhar, A., Duflo, E. and Jackson, M. (2012). The diffusion of microfinance. *Working paper, Massachusettes Institute of Technology*. 24

Bertrand, M., Luttmer, E. and Mullainathan, S. (2010). Network effects and welfare cultures. *Quarterly Journal of Economics*, **115** (3), 1019–1055. 2

Bhamidi, S., Bresler, G. and Sly, A. (2011). Mixing time of exponential random graphs. *The Annals of Applied Probability*, **21** (6), 2146–2170. 4

Bisin, A., Moro, A. and Topa, G. (2011). The empirical content of models with multiple equilibria in economies with social interactions. *Working paper, Vanderbilt University*. 5

Boucher, V. and Mourifié, I. (2012). My friend far far away: Asymptotic properties of pairwise stable networks. *Working paper, University of Montreal*. 5, 11

Bramoullé, Y. and Fortin, B. (2010). Social networks: Econometrics. In S. N. Durlauf and L. E. Blume (eds.), *The New Palgrave Dictionary of Economics.*, Basingstoke: Palgrave Macmillan. 3

Brock, W. and Durlauf, S. (2001). Discrete choice with social interactions. *Review of Economic Studies*, **68**, 235–260. 5, 7, 11, 17, 31

— and — (2007). Identification of binary choice models with social interactions. *Journal of Econometrics*, **140** (1), 52–75. 11

Byrd, R., Nocedal, J. and Waltz, R. (2006). Knitro: An integrated package for nonlinear optimization. In G. di Pillo and M. Roma (eds.), *Large-Scale Nonlinear Optimization*, Springer Verlag, pp. 35–59. 29

Calvó-Armengol, T., Patacchini, E. and Zenou, Y. (2009). Peer effects and social networks in education. *Review of Economic Studies*, **76**, 1239–1267. 2

Chandrasekhar, A. and Jackson, M. (2012). Tractable and consistent random graph models. *Working paper, Stanford University*. 4, 5

Christakis, N. and Fowler, J. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, **357** (4), 370–379. 2

—, —, Imbens, G. and Kalyanaraman, K. (2010). An empirical model for strategic network formation. *Working paper, Harvard University*. 4, 11

Conley, T. and Udry, C. (2010). Learning about a new technology: Pineapple in ghana. *American Economic Review*, **100**, 35–69. 2

Delgado, M. and Mora, J. (1995). Nonparametric and semiparametric estimation with discrete regressors. *Econometrica*, **63** (6), 1477–1484. 24

Fafchamps, M. and Gubert, F. (2007). The formation of risk-sharing networks. *Journal of Economic Development*, **83**, 326–350. 3

Farrell, J. and Klemperer, P. (2007). Coordination and lock-in: Competition with switching costs and network effects. In M. Armstrong and M. Porter (eds.), *Handbook of Industrial Organization*, vol. 3, *31*, New York: Elsevier, pp. 1967–2072. 2

Fox, J. (2010). Estimating matching games with transfers. *Working paper, University of Michigan*. 5

Hsieh, C. and Lee, L. (2012). A structural modeling approach for network formation and social interactions—with applications to students' friendship choices and selectivity on activities. *Working paper, Ohio State University*. 4

Jackson, M., Rodriguez-Barraquer, T. and Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, **102** (5), 1857–1897. 24, 25

Jackson, M. O. and Wolinsky, J. (1996). A strategic model of social and economic networks. *Journal of Economic Theory*, **71** (1), 44–74. 15

Kline, B. (2012). Identification of complete information games. *Working paper, University of Texas at Austin*. 5

Lumley, T. and Mayer-Hamblett, N. (2003). Asymptotics for marginal generalized linear models with sparse correlations. *Working Paper, University of Washington.* 11

McPherson, M., Smith-Lovin, L. and Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, **27**, 415–444. 2

Mele, A. (2011). A structural model of segregation in social networks. *Working paper, John Hopkins University.* 4, 11, 15

Menzel, K. (2012). Inference for large games with exchangeable players. *Working paper, New York University.* 5, 32

Myerson, R. (1977). Graphs and cooperation in games. *Mathematics of Operations Research*, **2**, 225–229. 30

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, **72** (6), 1349–1382. 23

Nowak, W. and Zięba, W. (2005). Types of conditional convergence. *Annales Universitatis Mariae Curie-Skłodowska. Sectio A. Mathematica*, **59**, 97–105. 32

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, **119** (1), 99–130. 19, 24, 26

Robbins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, **29**, 173–191. 4

Shang, Q. and Lee, L. (2011). Two-step estimation of endogenous and exogenous group effects. *Econometric Reviews*, **30** (2), 173–207. 5

Sheng, S. (2012). Identification and estimation of network formation games. *Working paper, University of Southern California.* 5, 11

Snijders, T. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**, 2. 4

Song, K. (2012). Econometric inference on a large bayesian game. *Working paper, University of British Columbia.* 5

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies*, **70** (1), 147–165. 7