Estimating Network Externalities in Undirected Link Formation Games [*]

Margherita Comola[†] and Amit Dekel[‡]

December 26, 2023

Abstract

This paper explores the existence of externalities from network architecture (so-called network externalities) in link formation games of incomplete information. It extends the structural estimation method by Leung (2015) to games where links are undirected and proposals are only partially observable. We provide an econometric characterization of the proposed two-step estimator, and we document its performance through a simulation exercise. When the estimation method is applied to data on risk-sharing arrangements in a Tanzanian village, results indicate that indirect connections matter. Assuming that link formation follows a bilateral process, the estimated probability of proposing a link to a potential partner increases by 9% for any additional indirect connection provided.

Keywords: Undirected Networks; Network Externalities; Incomplete Information; Risk-sharing

JEL codes: C45; D85; O12

[*] Earlier versions of this paper have circulated under the name "Estimating local network externalities", "Network Externalities in Bilateral Link Formation: What's the Value of Friends of Friends?". We thank Antonio Cabrales, Marcel Fafchamps, Rasa Karapandza, Philipp Ketz, Angelo Mele, Dotan Persitz, Angelo Secchi and several anonymous referees for their helpful comments, and Joachim De Weerdt for making the data available. Margherita Comola acknowledges the financial support of the EUR grants ANR-17-EURE-0001 and ANR-21-CE26-0002-01.

[†] University Paris-Saclay (RITM) and Paris School of Economics: margherita.comola@psemail.eu

[‡] Paris School of Economics: amit.dekel@psemail.eu

1 Introduction

From its very first steps network theory has claimed that the formation of links may depend strategically on the entire graph (Jackson and Wolinsky, 1996; Bala and Goyal, 2000). However, evidence based on experimental and observational data still lags behind, and empirical questions about the value of indirect connections remain largely unexplored.[1] Building on Leung (2015), we design an estimation protocol for network formation games where links are undirected and proposals are only partially observable. This procedure accommodates undirected links formed by bilateral or unilateral link formation rules. In our setting, agents play a simultaneous game of incomplete information where they form undirected links on the basis of their beliefs about the emerging network architecture. Assuming that these beliefs satisfy a number of regularity conditions (discussed in Section 2), the estimation strategy boils down to a two-step procedure where the first stage consistently estimates agents' beliefs, and the second stage estimates the role of network externalities.[2] We provide existence, consistency and asymptotic normality results for the two-step estimator, and we conduct a set of simulation exercises to investigate its performance as sample size grows.

We illustrate the procedure using data on risk-sharing arrangements from the Tanzanian village of Nyakatoke. Lacking access to formal insurance, most households in developing countries rely on informal risk-sharing arrangements in face of shocks such as health-related expenses, injuries, funerals and job losses. These arrangements have long captured the attention of economists, for several reasons. On the one hand, the prevalence of the phenomenon makes it of paramount importance for economic development.[3] On the other hand, most arrangements do not take place at the level of the entire community but among pairs of house-

[1] Most of the available evidence relates to specific settings. For instance, the study of cross-firm collaborative networks suggests that information flows are insignificant for indirect neighbors (Breschi and Lissoni, 2005; Singh, 2005). On the other hand, experimental evidence with dictator games shows that further-away connections are relevant and decay with the inverse of distance (Goeree et al., 2010). Graham and Pelican (2019) provide a test for interdependencies in link-formation preferences and conclude for the presence of externalities in the same data we use here.

[2] A two-step approach is also taken by König et al. (2019).

[3] Coate and Ravallion (1993), Townsend (1994), Udry (1994), Fafchamps and Lund (2003).

holds, which makes risk-sharing a compelling application of network theory for economists.[4]

We use the self-declared information in Nyakatoke data to draw the undirected village network and to investigate the role of network architecture. Specifically, we test whether agents choose between risk-sharing partners on the basis of their individual characteristics only or whether indirect connections also play a role in these decisions. Much of the economic literature assumes that informal risk-sharing arrangements require the consent of the two parties involved, which implies that link formation follows a bilateral process.[5] Following this literature, in the empirical illustration of Section 5 we assume that links are formed bilaterally. Appendix B shows how our procedure also accommodates undirected networks issued by unilateral link formation rules. Results from Section 5 indicate that Nyakatoke villagers evaluate potential partners' connections in a positive manner. Our estimates suggest that for a given pair of potential partners ij, the probability that i proposes a link to j increases on average by 0.016 for any additional indirect connection j provides. This increase is sizeable, as it corresponds to approximately 9% of the average fitted probability of link proposal.

From an econometric standpoint, testing whether network architecture predicts link formation has proved to be a complex task. Our paper deals with the case where the researcher observes one single network at one single period and wants to include network covariates in the objective function of agents. In this scenario the structural equation can have multiple solutions (Bjorn and Vuong, 1984; Bresnahan and Reiss, 1991; Tamer, 2003), and the calculation may become intractable due to the combinatorial complexity of networks. One solution is provided by the exponential random graph models where a dynamic meeting protocol acts as an equilibrium selection mechanism (Hsieh and Lee, 2016; König, 2016; Mele,

[4] Risk-sharing networks have been studied from multiple angles, including the efficiency and sustainability of the resulting arrangements, the determinants of link formation and the structural properties of the network architecture (Genicot and Ray, 2003; Bramoullé and Kranton, 2007; Bloch et al., 2008; Jackson et al., 2012; Banerjee et al., 2013; Ambrus et al., 2014; Ambrus and Elliott, 2020).

[5] Most models of risk sharing and favor exchange assume that agents can refuse transactions that are against their self-interest (Kimball, 1988; Coate and Ravallion, 1993; Kocherlakota, 1996; Bloch et al., 2008; Jackson et al., 2012).

2017; Badev, 2020). Another solution is to condition on models that replicate some observed topological patterns or to limit the degree to which other players can affect one's utility.[6] Alternatively, one can simplify the estimation procedure by relying on incomplete information to induce symmetry and independence in agents strategies (Leung, 2015; De Paula and Tang, 2012), which is the approach we take here.

This paper's main contribution is methodological: it develops a protocol to estimate network externalities in undirected link formation games of incomplete information. This builds on Leung (2015) who also relies on incomplete information to estimate a simultaneous game of link formation. Our paper differs in one substantive aspect, however: Leung (2015)'s procedure requires data on directed links, which are interpreted as observed proposals in a game of unilateral link formation. Our protocol is designed for undirected link data, which we interpret as the equilibrium outcome of a link formation process where proposals are only partially observed. This opens to the possibility of coordination failures, which we preclude by restricting to *admissible* Bayesian Nash equilibria banning weakly dominated strategies (Section 2.2). The resulting log-likelihood function generalizes the partial observability bivariate model of Comola and Fafchamps (2014) to include network covariates in the objective function of agents. This method is naturally suited for bilateral as well as unilateral link formation rules, as long as the econometrician only observes the undirected link outcome. In Appendix B we revisit the empirical illustration from Leung (2015), showing that directed and undirected models of link formation can yield different results when applied to the same data. Our work also relates to Ridder and Sheng (2020), who generalize Leung (2015) by relaxing the separability assumption to include additional non-linear net-

[6] One can identify structural parameters by aggregating individuals into 'types' and assuming that agents have preferences only over the type of their partners (De Paula et al., 2018), or by the rate at which various sub-graphs are observed in the overall network (Chandrasekhar and Jackson, 2016). Along similar lines, Boucher and Mourifie (2017) study a setting where individual preferences display weak homophily.

work externalities.[7,8] As an additional contribution, our paper also advances the knowledge of risk-sharing arrangements in developing countries by providing first-hand evidence that indirect connections affect linking choices, while previous literature has focused mostly on documenting the number and characteristics of risk-sharing partners.[9]

Network formation models have proved difficult to estimate in presence of externalities. Most of the existing tools were developed for directed networks and expect two distinct reports per each dyad (Leung, 2015; Mele, 2017; Badev, 2020). On the other hand, the available models of undirected network formation rely on complete information and achieve set identification (Miyauchi, 2016; Sheng, 2020; De Paula et al., 2018). The procedure we propose is computationally parsimonious, providing a convenient alternative to complete-information models. As such it can prove useful in a variety of applications where links are undirected for conceptual and/or practical reasons. From the conceptual viewpoint, in many instances it is legitimate to assume that link formation requires the consent of both parties. For example, link formation is 'naturally' interpreted as bilateral when data represent risk-sharing, trade deals, co-authorship amongst researchers, communication flows, and industrial executive linkages (Banerjee et al., 2013; Buchel et al., 2020; Lalanne and Seabright, 2022). In these cases the practitioner may want to draw undirected links on the basis of multiple (possibly discordant) survey reports (Section 5.1). Also, practically speaking, many data sources contain no information on linking intentions and one single link outcome per pair. This is mainly the case when data originate from administrative sources (rather than individual surveys): for example, communication records retrieved from digital social networks, exchange data from online marketplaces, import-export shipment registries,

[7] The methodology in Ridder and Sheng (2020), which is designed for directed networks, also extends to a scenario where (agents form directed links but) the observed links are undirected because of data collection and reporting issues. This is conceptually different from our setting where undirected links are formed based on players' proposals which are not fully observed.

[8] For the estimation of social interaction models with incomplete information, see also Gilleskie and Zhang (2009) and Hoshino (2019).

[9] An exception is Krishnan and Sciubba (2009), who identify the common features of all equilibrium configurations in a model with negative network externalities and test these predictions against data on labor exchange arrangements in Ethiopia.

scientific publication records and patent repositories only report ‘successful’ matches (Gaulier and Zignago, 2010; Hitsch et al., 2010; Ductor et al., 2014; Bailey et al., 2018). In all these situations, the toolbox developed for directed networks is inadequate and our estimator provides a valid alternative.

The paper is organized as follows. Section 2 introduces the theoretical setting. Section 3 presents the estimation method. Section 4 describes a simulation exercise. Section 5 applies the estimation method to risk-sharing data from rural Tanzania. Section 6 concludes. Appendix A discusses the inclusion of continuous attributes and the smoothing of discrete variables. Appendix B revisits the data illustration from Leung (2015) to compare different models of link formation. All proofs are relegated to Appendix C.

2 The Model

2.1 The game

Let $N = \{1, 2, ..., n\}$ be a set of agents who play to form an undirected network. For agent i, let $X_i = [X_{i,1}, ..., X_{i,q}]$ be a vector of individual attributes of dimension $[1 \times q]$ and $X = \{X_1, ..., X_n\}$ denote the set of these vectors. For ease of exposition in this section we assume that X is composed of discrete attributes only (this assumption is relaxed in Appendix A).

Assumption 1 (Discrete X)**.** *For every $i \in N$, X_i has finite support and for any x in the support $Pr(X_i = x)$ is bounded away from zero.*

Let $\epsilon_i = [\epsilon_{i,1}, ..., \epsilon_{i,i-1}, 0, \epsilon_{i,i+1}, ..., \epsilon_{i,n}]$ be a $[1 \times n]$ vector of shocks of agent i with all other agents (ϵ_{ij} does not necessarily equal ϵ_{ji}), which are stochastically independent from X. ϵ denotes the collection ϵ_i over all $i \in N$.

Assumption 2 (i.i.d. Shocks)**.** $\{\epsilon_{ij} \mid i, j \in N, i \neq j\}$ *are independently drawn from the* *standard normal distribution.*[10]

Thus, shocks are assumed to be uncorrelated across and within individuals.[11] The set of attributes vectors X is common knowledge, while the shocks are private information, i.e. only i knows ϵ_i.

Agents play a simultaneous-move game of link formation, where everyone announces independently the links they wish to form. The action of agent i is represented by a binary vector of length n, where the jth entry ($j \neq i$) equals 1 if i proposes j to form a link and 0 otherwise.[12] The actions of all agents stacked on top of each other, denoted S, can be interpreted as an adjacency matrix of link proposals:

$$S = \begin{bmatrix} 0 & S_{1,2} & \dots & S_{1,n} \\ S_{2,1} & 0 & \dots & S_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,1} & \dots & S_{n,n-1} & 0 \end{bmatrix} \tag{1}$$

These link proposals give rise to a network G. We consider two alternative rules by which G is formed. In the "bilateral rule" an undirected link is formed if and only if both sides propose to one another. Formally, $G_{ij} = G_{ji} = S_{ij} \cdot S_{ji}$. The interpretation is that pairs of agents need bilateral consent in order to form an undirected link between them. In the "unilateral rule" an undirected link is formed if and only if at least one side proposes to the other. Formally, $G_{ij} = G_{ji} = S_{ij} + S_{ji} - S_{ij} \cdot S_{ji}$. The interpretation is that agents may unilaterally form undirected links with others. Note that the issue of transforming proposals into links only arises when links are undirected, as when links are directed it is straightforward to set $G_{ij} = S_{ij}$ and $G_{ji} = S_{ji}$. For concreteness, we assume a bilateral rule

[10] The standard normal distribution is chosen here for the sake of convenience, but our results hold for other full-support distributions.

[11] While independence across individuals is essential to our estimation strategy, independence within individuals (i.e. between ϵ_{ik} and ϵ_{il}) is imposed for simplicity and could be relaxed by adding an agent-level unobserved effect as in Graham (2017).

[12] Since an agent cannot form a link with herself, the ith entry always equals 0.

throughout the theoretical discussion (the unilateral rule is explored in Appendix B).

Given network G, agent i's utility is given by:

$$u_i(X, G; \theta_0) = \sum_{j \neq i} G_{ij} \cdot (v_{ij}(X, G_{-i}; \theta_0) + \epsilon_{ij}) \tag{2}$$

where G_{-i} indicates G with the i^{th} row and column deleted, and $\theta_0 \in \Theta$ is a $[p \times 1]$ vector of parameters from a compact set Θ. Estimating the parameters in θ_0 is the goal of the procedure described in Section 3.

Assumption 3 (Linearity, Separability and Anonymity)**.** *The $v_{ij}(\cdot)$ function: (i) is linear in θ_0 i.e. can be written in the form $Z_{ij}\theta_0$ with Z_{ij} satisfying $\|Z_{ij}\| < \bar{Z} < \infty$ for all $i, j \in N$; (ii) depends on G only through G_{-i}; (iii) is insensitive to permutations of the agents' labels.*

The separability condition, borrowed from Leung (2015), requires that the i's marginal utility from a link with j is independent from other links she may have.[13] In Section 2.4 below we discuss which types of externalities from indirect connections this assumption is compatible with.

2.2 Equilibrium

Let i's (pure) strategy be a function from commonly observed attributes and privately observed shocks to an action: $S_i : (X, \epsilon_i) \to \{0, 1\}^n$ (henceforth we omit the dependency on X). A Bayes Nash Equilibrium (BNE) is a strategy profile $[S_i(\epsilon_i), S_{-i}(\epsilon_{-i})]$ such that for all $i \in N$ and for all $S'_i(\epsilon_i)$:

$$\mathbb{E}_{\epsilon_{-i}} \left[u_i(X, G[S_i(\epsilon_i), S_{-i}(\epsilon_{-i})]; \theta_0) \right] \geq \mathbb{E}_{\epsilon_{-i}} \left[u_i(X, G[S'_i(\epsilon_i), S_{-i}(\epsilon_{-i})]; \theta_0) \right] \tag{3}$$

Due to the separability assumption, in any BNE agents consider proposal decisions separately. Hence, we can write $S_i(X, \epsilon_i) = [S_{ij}(X, \epsilon_{ij})]_{j \in N}$, where $S_{ij} : (X, \epsilon_{ij}) \to \{0, 1\}$. In

[13] Separability is relaxed by Ridder and Sheng (2020) in the context of directed network formation.

addition, in any BNE, S_{ij} must prescribe i to propose to j whenever it strictly increases her expected utility and not to propose whenever it strictly reduces it. Formally:

$$S_{ij}(\epsilon_{ij}) = \begin{cases} 1 & \text{if } \mathbb{E}_{\epsilon_{ji}}[S_{ji}(\epsilon_{ji})] \cdot \left(\mathbb{E}_{\epsilon_{-i}}[v_{ij}(X, G_{-i}[S_{-i}(\epsilon_{-i})]; \theta_0)] + \epsilon_{ij}\right) > 0 \\ 0 & \text{if } \mathbb{E}_{\epsilon_{ji}}[S_{ji}(\epsilon_{ji})] \cdot \left(\mathbb{E}_{\epsilon_{-i}}[v_{ij}(X, G_{-i}[S_{-i}(\epsilon_{-i})]; \theta_0)] + \epsilon_{ij}\right) < 0 \end{cases} \tag{4}$$

Whenever proposing to j does not change i's expected utility, proposing and not proposing are both best-replies. This shows that Bayes Nash equilibria do not exclude coordination failures. For instance, a pair $S_{ij}(\epsilon_{ij})$ and $S_{ji}(\epsilon_{ji})$ that prescribed i and j (respectively) not to propose for any ϵ_{ij} and ϵ_{ji} (respectively) may well be part of a BNE profile, *even if* both i and j stand to gain (in expectation) from forming a link. Since we are interested in modeling bilateral network formation, where pairs of agents are free to coordinate their actions, we wish to rule out such equilibria. We do so by restricting attention to *admissible* Bayes Nash equilibria, i.e. equilibria where no player uses a (weakly) dominated strategy. In any *admissible* BNE, S_{ij} must prescribe i to propose to j whenever, *assuming j proposes to i*, her expected utility from proposing is strictly positive, and not to propose if it is strictly negative. Formally:

$$S_{ij}(\epsilon_{ij}) = \begin{cases} 1 & \text{if } \mathbb{E}_{\epsilon_{-i}}[v_{ij}(X, G_{-i}[S_{-i}(\epsilon_{-i})]; \theta_0)] + \epsilon_{ij} > 0 \\ 0 & \text{if } \mathbb{E}_{\epsilon_{-i}}[v_{ij}(X, G_{-i}[S_{-i}(\epsilon_{-i})]; \theta_0)] + \epsilon_{ij} < 0 \end{cases} \tag{5}$$

Given this decision rule, one may reformulate the equilibrium condition in terms of beliefs over proposal probabilities. To that end, let $\sigma^{S_{-i}}$ be a $[(n-1) \times n]$ matrix representing i's beliefs about the probabilities that each agent $j \neq i$ proposes to another agent $k \neq j$ (including i herself). Given the decision rule in Equation (5), and letting Φ denote the CDF of the standard normal distribution, the ex-ante probability that i proposes to j is:

$$Pr(S_{ij} = 1 | X, \sigma^{S_{-i}}) = Pr\left(\mathbb{E}[v_{ij}(X, G_{-i}; \theta_0) | X, \sigma^{S_{-i}})] + \epsilon_{ij} > 0\right) \tag{6}$$

$$= \Phi\left(\mathbb{E}[v_{ij}(X, G_{-i}; \theta_0)|X, \sigma^{S_{-i}})]\right) \tag{7}$$

Note that since ϵ_{ij} is drawn from a continuous distribution, is makes no difference whether i's strategy prescribes to propose or not when $\mathbb{E}_{\epsilon_{-i}}[v_{ij}(X, G_{-i}[S_{-i}(\epsilon_{-i})]; \theta_0)] + \epsilon_{ij}$ is exactly zero. A belief matrix σ^S corresponds to an admissible BNE if and only if it satisfies the following equality for all i and j:

$$\sigma_{ij}^{S_{-i}} = Pr(S_{ij} = 1|X, \sigma^{S_{-i}}) \tag{8}$$

The fact that $v_{ij}(\cdot)$ depends on G_{-i}, rather than S_{-i} allows conditioning its expected value on beliefs over *linking* probabilities rather than proposal probabilities. In addition, due to Assumption 2, the probability that a link exists is simply the product of the proposal probabilities of the two parties involved. This allows reformulating the equilibrium condition in terms of beliefs over linking probabilities. To that end, we let σ^G denote a $[n \times n]$ matrix representing agents' common beliefs about linking probabilities among all pairs of agents, and $\sigma^{G_{-i}}$ denote the same matrix but with its i^{th} row and column deleted. A belief matrix σ^G corresponds to an admissible BNE if and only if it satisfies the condition below for all i and j. We call such σ^G an "equilibrium belief".

$$\sigma_{ij}^G = \underbrace{\Phi\left(\mathbb{E}[v_{ij}(X, G_{-i}; \theta_0)|X, \sigma^{G_{-i}}]\right)}_{Pr(i \text{ proposes to } j)} \underbrace{\Phi\left(\mathbb{E}[v_{ji}(X, G_{-j}; \theta_0)|X, \sigma^{G_{-j}}]\right)}_{Pr(j \text{ proposes to } i)} \tag{9}$$

Given an equilibrium belief σ^G, a network G is said to be an "equilibrium" if the following holds for all i and j:

$$G_{ij} = \underbrace{\mathbb{1}\left\{\mathbb{E}[v_{ij}(X, G_{-i}; \theta_0)|X, \sigma^{G_{-i}}] + \epsilon_{ij} > 0\right\}}_{i \text{ proposes to } j} \underbrace{\mathbb{1}\left\{\mathbb{E}[v_{ji}(X, G_{-j}; \theta_0)|X, \sigma^{G_{-j}}] + \epsilon_{ji} > 0\right\}}_{j \text{ proposes to } i} \tag{10}$$

Note that due to admissibility, an equilibrium network G is one that satisfies the pairwise stability conditions *in expectation*: *(i)* if i and j are linked in G then the marginal expected utilities this link provides each player is positive; *(ii)* if i and j are not linked in G then then the marginal expected utility this link provides is negative for at least one of them. Hence, even though the solution concept we deploy is non-cooperative, in equilibrium no pair of players fail to coordinate on forming a link.

Following Leung (2015), from here on we restrict attention to symmetric equilibria. A symmetric equilibrium is an equilibrium in which all pairs of agents that are observationally equivalent have the same linking probabilities. Formally, an equilibrium σ^G is symmetric if for all $i, j \neq k, l \in N$:

$$(X_i = X_k \ \textit{and} \ X_j = X_l) \ \textit{or} \ (X_i = X_l \ \textit{and} \ X_j = X_k) \implies \sigma^G_{ij} = \sigma^G_{kl} \tag{11}$$

Figure 1 illustrates this definition. Agents in this network have a single binary attribute – being either black or white – depicted by the colors of the nodes. Beliefs are depicted by weights on edges and their values by their color (i.e. all red beliefs equal each other, and all blue beliefs equal each other). All pairs consisting of two black agents have the same σ^G value (red), and the same holds for pairs of white and black agents (blue) and pairs of two white agents (green). The described beliefs are therefore symmetric.

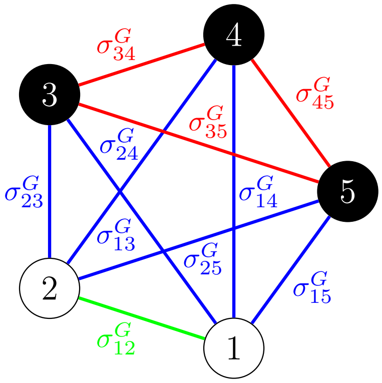


Figure 1: Example of a symmetric belief matrix

For given X and θ_0, we let $\omega(X, \theta_0)$ denote the set of admissible and symmetric BNE. Proposition 1 establishes that $\omega(X, \theta_0)$ is non-empty. Assumption 4 states that the observed data is generated by some equilibrium within $\omega(X, \theta_0)$.[14]

Proposition 1 (Existence)**.** *Under assumptions 1-3, there exists an admissible and symmetric BNE, i.e.* $\omega(X, \theta_0) \neq \emptyset$*.*

Assumption 4 (Admissible and Symmetric BNE)**.** *The observed network is generated according to Equation (10) where* $\sigma^G \in \omega(X, \theta_0)$*.*

2.3 Example

Consider the case where 3 agents have one binary attribute X_i, and their utility function is as follows:

$$v_{ij}(X, G_{-i}; \theta_0) = \theta_1 + \theta_2 X_i + \theta_3 |X_i - X_j| + \theta_4 \frac{1}{n-1} \sum_{k \neq i} G_{jk} \tag{12}$$

with $\theta_0 = [-1, 1, -0.5, 1]'$. The term $|X_i - X_j|$ represents a measure of similarity between i and j. It thus accounts for homophily. The term $\frac{1}{n-1}\sum_{k\neq i} G_{jk}$ represents the average number of indirect connections (i.e. paths of length 2) that i gains by forming a link with j. It thus accounts for externalities from the network topology.

Columns 1 and 2 in Table 1 present all possible ordered pairs in the 3-agent network. Columns 3 and 4 report the binary attributes of agents i and j respectively. Column 5 reports $|X_i - X_j|$. The third term in the utility function $\frac{1}{n-1}\sum_{k\neq i} G_{jk}$ depends on the network structure G. Its expected value therefore depends on the beliefs about the network structure σ^G.

Let us consider a given set of beliefs which are reported in column 6. Column 7 uses these beliefs to compute $\frac{1}{n-1}\sum_{k\neq i} \sigma_{jk}^{G_{-i}}$. Using columns 3, 5 and 7 and the functional form

[14] Note that Assumption 4 does not impose any restrictions on the probability that a given equilibrium is selected. This stands in contrast to the "many markets asymptotics" setting where the econometrician observes many repetitions of the game and assumes that the probability distribution over (not necessarily symmetric) equilibria is degenerate. As a result, the equilibrium being played in all repetitions of the game is guaranteed to be the same one. Following Leung (2015), we are able to avoid this assumption and achieve point identification with one large network ("large market asymptotics") by allowing only symmetric equilibria to be selected.

we can now compute the expected value of v_{ij} for all pairs of agents. This is reported in column 8. Now, given that the ϵ_{ij} values are drawn independently from the standard normal distribution, the probability that i would propose to j (that is, that $\mathbb{E}[v_{ij}] + \epsilon_{ij} \geq 0$) is $\Phi(\mathbb{E}[v_{ij}])$. This is reported in column 9. Finally, the probability that a link exists in G is the product of the proposal probabilities of the two agents involved. This is reported in column 10.

1	2	3	4	5	6	7	8	9	10
i	j	X_i	X_j	$\lvert X_i - X_j \rvert$	σ^G	$\frac{1}{n-1}\sum_{k\neq i}\sigma_{jk}^{G_{-i}}$	$\mathbb{E}[v_{ij}]$	$\Phi(\mathbb{E}[v_{ij}])$	$\Phi(\mathbb{E}[v_{ij}]) \cdot \Phi(\mathbb{E}[v_{ji}])$
1	2	0	1	1	0.027	$0.5 \cdot 0.255$	-1.3725	0.0850	0.027
1	3	0	1	1	0.027	$0.5 \cdot 0.255$	-1.3725	0.0850	0.027
2	1	1	0	1	0.027	$0.5 \cdot 0.027$	-0.4865	0.3133	0.027
2	3	1	1	0	0.255	$0.5 \cdot 0.027$	0.0135	0.5054	0.255
3	1	1	0	1	0.027	$0.5 \cdot 0.027$	-0.4865	0.3133	0.027
3	2	1	1	0	0.255	$0.5 \cdot 0.027$	0.0135	0.5054	0.255

Table 1: Example

Note that in this example $\sigma_{ij}^G = \Phi(\mathbb{E}[v_{ij}])\Phi(\mathbb{E}[v_{ji}])$ for all i and $j \neq i$. This means that the beliefs σ^G in column 6 are equilibrium beliefs. Also note that all pairs of agents which are observationally equivalent have the same linking probabilities, e.g. the pairs $\{1, 2\}$ and $\{1, 3\}$ have the same linking probability under σ^G. This means that the beliefs σ^G are symmetric.

2.4 Separability and Externalities

The utility agents gain from the network might be related to different measures of their centrality in it. The assumptions we take on the form of agents' utility function, however, limits the type of centrality measures whose effect on proposal decision can be estimated. This subsection discusses what centrality measures are compatible with our assumptions.

Let $c_i(G)$ denote a generic centrality measure of player i in network G. The separability assumption requires that it can be written in the form $c_i(G) = \sum_{j\neq i} G_{ij} \cdot f(G_{-i})$ for some function f. This condition can alternatively be written as $c_i(G + ij) - c_i(G - ij) = f(G_{-i})$, where $G + ij$ (respectively, $G - ij$) denote the network G with the link between i and j added (respectively, removed). Hence, our model allows for the marginal contribution of a link ij to i's centrality to be a function of all walks in G besides those that pass through i. Centrality measures compatible with this condition include "information centrality" (Stephenson and Zelen, 1989) and "targeting centrality" (Bramoullé and Genicot, 2023).[15]

Information centrality assign weights for every path emanating from i and sum those weights up. Since paths are sequences of agents and links in which no agent appears twice, this measure is compatible with the separability assumption. While information centrality defines a specific weighting scheme, one could generalize it by leaving the weighting scheme open. The externality we use in the empirical illustration of Section 5 corresponds to the special case where the weights on every paths of length larger then three are set to zero.

To give the intuition behind targeting centrality, consider a dynamic process of information diffusion that takes place in discrete time. At time period $l = 0$ an agent i passes a message to each of her friends with some fixed probability p. At every subsequent period $l > 1$ any agent that received the message at period $l - 1$ passes it to each of her friends with probability p. Now suppose that the message is targeted towards a specific agent j. j's targeting centrality measures the expected number of times she receives messages from others assuming she does not participate in the diffusion process. The idea that the transmission process stops at the target node makes this centrality measure compatible with our separability assumption.

While the discussion above presents centrality measures that are compatible with the separability assumption, some are clearly not. The following equation provides a generic way to construct a separable counterpart for *any* centrality measure $c_i(G)$: $\tilde{c}_i(G) = \sum_{j \neq i} G_{ij} \cdot c_i(G_{-i} + ij)$. As an illustration, suppose $c_i(G)$ denote i's diffusion centrality (Banerjee et al., 2013), which is based on the same information diffusion process described above. The interpretation of $\tilde{c}_i(G)$ is that i diffuses the message in period 1 and then never re-transmits it again.

[15] Brandes and Fleischer (2005) show that information centrality is equivalent to current-flow closeness centrality.

3 Estimation

Imagine we observe a single network G and agents' attributes X.[16] Let us assume that G is formed according to the model specified above, that is, the network results from all agents behaving optimally given the symmetric equilibrium belief σ^G and their realization of the error terms ϵ_i that we do not observe. Our goal is to estimate and conduct inference on the true parameter vector θ_0. In what follows we describe the building blocks of our procedure.

3.1 Log-likelihood function

Let us denote by δ_{ij} a function that takes X_i, X_j and returns a vector of covariates of dimension $[1 \times (p-k)]$ (e.g. i's attributes and the distance between i and j's attributes, in the example above). Denote by γ_{ij} a function that takes i's beliefs about the emerging network (possibly together with X) and returns a vector of covariates of dimension $[1 \times k]$ (e.g. the number of length-two paths i gains from linking with j, in the example above). To facilitate an intercept, assume that δ_{ij} always returns 1 as a first element. We call the first type of covariates 'exogenous' as they do not depend on the network structure, and the second type 'endogenous', as they do. Using this terminology, while $\gamma_{ij}(X, G_{-i})$ represents the endogenous covariates associated with i's linking with j, $\gamma_{ij}(X, \sigma^{G_{-i}})$ represents their expected value. By Assumption 3 $v_{ij}(\cdot)$ is a linear function of the exogenous and endogenous covariates:

$$v_{ij}(X, G_{-i}; \theta_0) = [\delta_{ij}(X_i, X_j), \gamma_{ij}(X, G_{-i})] \cdot \theta_0 \tag{13}$$

The expected value of v_{ij} conditional on X and the event that σ^G is the selected equilibrium

[16] Measurement error in the network topology is an important, yet largely unexplored issue that goes beyond the scope of this paper (De Paula, 2017; Advani and Malde, 2014; Bramoullé et al., 2020). Our estimator relies on the assumption that the network in measured in an accurate and complete manner, like other methods do (Leung, 2015; De Paula et al., 2018). In particular, the beliefs estimates (Subsection 3.2) may not be consistent in presence of link measurement error of general form.

is therefore:

$$\mathbb{E}[v_{ij}(X, G_{-i}; \theta_0)|X, \sigma^{G_{-i}}] = [\delta_{ij}(X_i, X_j), \gamma_{ij}(X, \sigma^{G_{-i}})] \cdot \theta_0 \tag{14}$$

Suppressing some of the input arguments, we can now rewrite Equation (9) as:

$$P(G_{ij} = 1|X, \sigma^G) = \Phi([\delta_{ij}, \gamma_{ij}(\sigma^{G_{-i}})]\theta_0) \cdot \Phi([\delta_{ji}, \gamma_{ji}(\sigma^{G_{-j}})]\theta_0) \tag{15}$$

Since $\{\epsilon_{ij}|i, j \in N, i \neq j\}$ are drawn independently from one another, conditional on X and the event that σ^G is selected, the likelihood of observing a network G is:

$$\begin{aligned} L(\theta, \sigma^G) = \prod_{i,j>i}^{n} \Bigg[&\bigg(\Phi\big([\delta_{ij}, \gamma_{ij}(\sigma^{G_{-i}})]\theta\big) \cdot \Phi\big([\delta_{ji}, \gamma_{ji}(\sigma^{G_{-j}})]\theta\big) \bigg)^{G_{ij}} \\ &\times \bigg(1 - \Phi\big([\delta_{ij}, \gamma_{ij}(\sigma^{G_{-i}})]\theta\big) \cdot \Phi\big([\delta_{ji}, \gamma_{ji}(\sigma^{G_{-j}})]\theta\big) \bigg)^{1-G_{ij}} \Bigg] \end{aligned} \tag{16}$$

By taking the log of this expression and dividing by the number of observations we obtain the following log-likelihood function:

$$\begin{aligned} l(\theta, \sigma^G) = \frac{2}{n(n-1)} \sum_{i,j>i}^{n} \Bigg[&\bigg(G_{ij} \cdot \log \Big(\Phi\big([\delta_{ij}, \gamma_{ij}(\sigma^{G_{-i}})]\theta\big) \cdot \Phi\big([\delta_{ji}, \gamma_{ji}(\sigma^{G_{-j}})]\theta\big) \Big) \bigg) \\ + &\bigg(\Big(1 - G_{ij}\Big) \cdot \log \Big(1 - \Phi\big([\delta_{ij}, \gamma_{ij}(\sigma^{G_{-i}})]\theta\big) \cdot \Phi\big([\delta_{ji}, \gamma_{ji}(\sigma^{G_{-j}})]\theta\big) \Big) \bigg) \Bigg] \end{aligned} \tag{17}$$

This function depends on the unobserved beliefs σ^G. We therefore cannot directly proceed to maximize it with respect to θ. Instead, we follow a two-step procedure, where in the first stage we consistently estimate the symmetric equilibrium beliefs (Subsection 3.2), and in the second stage we plug the estimated beliefs into the log-likelihood function to recover the estimands (Subsection 3.3).

Two comments about the log-likelihood function are in place. First, note that if we rule out endogenous covariates from the marginal utility the model boils down to a bivariate probit with partial observability (Poirier, 1980). Partial observability occurs when a positive outcome for one response variable is only observed if the other response variable is also positive.[17,18] This model has been used to model undirected network formation in the absence of externalities by Comola and Fafchamps (2014). Second, note that under uniqueness of equilibria, resorting to recovering σ^G from the data is not strictly necessary. Instead, we could analytically calculate the unique equilibrium beliefs for any candidate θ that is being considered by the optimization algorithm and evaluate the log-likelihood function at these beliefs.[19]

3.2 Estimating Beliefs

Under the assumption that beliefs satisfy the symmetric equilibrium condition, producing a consistent estimate of the beliefs $\hat{\sigma}^G$ is straightforward. Consider a set of observationally equivalent pairs of agents. In a symmetric equilibrium, the belief that any of these pairs are linked is identical (due to symmetry) and correct (since it is an equilibrium). Thus, the proportion of pairs within this set that are linked in the observed network is a consistent estimator for the belief that any of the pairs in the set are linked. In the case of discrete attributes, the estimator for the belief that i and j are linked $\hat{\sigma}^G_{ij}$ is defined as:

$$\hat{\sigma}^G_{ij} \equiv \frac{\sum_{l,k>l} \Big[G_{kl} \cdot \mathbb{1}\big\{ (X_i = X_k \wedge X_j = X_l) \vee (X_i = X_l \wedge X_j = X_k) \big\} \Big]}{\sum_{l,k>l} \Big[\mathbb{1}\big\{ (X_i = X_k \wedge X_j = X_l) \vee (X_i = X_l \wedge X_j = X_k) \big\} \Big]} \tag{18}$$

Proposition 2. *Under assumptions 1 and 4, $\hat{\sigma}^G_{ij}$ is consistent for σ^G_{ij} for all $i, j \in N$ such*

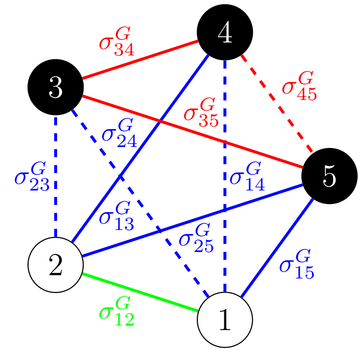
[17] In our context the proposals of the two agents can be interpreted as two partially-observed latent response variables, where the θs are by construction the same across the two equations. For a discussion of how identification depends on the functional form of the payoff function, see Poirier (1980).

[18] Note that in our setting the two latent response variables are partially observed, but the equilibrium link is observed accurately. This stands in contrast with situations where links are measured with error (Chandrasekhar and Lewis, 2012; Candelaria and Ura, 2018; Thirkettle, 2019).

[19] Under multiplicity, one could in principle calculate all equilibria for a candidate θ and compare their likelihood value. However, this approach could be difficult to implement (Aguirregabiria and Mira, 2007).

that $i \neq j$.

Figure 2 provides an example of how this estimator is calculated. As in Figure 1, the colors of the agents depict their one-dimensional binary attribute (being either black or white) and the colors of the edges and weights illustrate which pairs of agents have identical ex-ante linking probabilities (due to symmetry). The type of the edges illustrate which links are realized in the observed network – full lines describe realized links and dashed lines describe unrealized ones. The $\hat{\sigma}^G$ matrix presents the estimated beliefs. Concentrating on the black pairs, for instance, since two out of the three potential links between this type of pairs are realized we estimate the belief that these pairs are linked to be $\frac{2}{3}$.



$$\hat{\sigma}^G = \begin{bmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & \frac{2}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & \frac{2}{3} & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & \frac{2}{3} & \frac{2}{3} & 0 \end{bmatrix}$$

Figure 2: Example of beliefs estimation

To get a better understanding of the advantages of this estimation method, it is useful to contrast this "large-market" framework with an alternative "many-markets" framework (Graham and De Paola, 2020). Assume we were to observe many repetitions of the game over a constant set of agents ("many-markets"). The same pairs of agents are expected to have the same ex-ante linking probabilities across games, regardless of anonymity of preferences or symmetry of beliefs. As mentioned in Subsection 2.2, this only holds when agents are guaranteed to play the same equilibrium across games, which can be obtained by assuming a degenerate equilibrium selection mechanism. Thus, the proportion of games in which a given pair is linked gives a consistent estimate for the belief that this pair would be linked as the number of games increases to infinity. In our context of "large-market" framework we can relax the assumption that the equilibrium selection mechanism is degenerate and estimate symmetric beliefs from one single network realization. This broadens the applicability of our estimator, since many network datasets depict a single network (Goyal et al., 2006; Mele, 2017).[20]

Two additional points are worth mentioning. First, since the denominator sums up pairs that are exactly identical, it is only applicable to cases where all attributes in X are discrete. Second, since the estimator divides the set of observations into bins of identical pairs of agents, we risk not having enough observations within each bin when the sample size is small, the number of attributes is high, and their support is large. Both of these concerns are formally addressed in Appendix A. Subsection A.1 allows for the inclusion of continuous attributes, thereby resolving the first concern. Subsection A.2 discusses smoothing of discrete variables, which addresses the second one.

3.3 Estimating Preferences

Once $\hat{\sigma}^G$ is computed, plugging it into Equation (17) and maximizing with respect to θ yields our estimates $\hat{\theta}$ of θ_0. Since $\hat{\sigma}^G$ is consistent $\hat{\theta}$ is also consistent under standard regularity conditions. Below we state the consistency and asymptotic normality results for the second-stage estimator.

Proposition 3 (Consistency)**.** *Under assumptions 1-4 and standard regularity conditions, $\hat{\theta}$ is consistent for θ_0.*

Since the endogenous covariates are computed based on the estimated beliefs rather than the true ones, standard errors should be adjusted. Proposition 4 shows how to do so provided that the aggregate values of the true endogenous covariates and the estimated ones are identical.

[20] Our estimation procedure also carries over to the case of multiple networks (Appendix B).

Proposition 4 (Asymptotic Normality)**.** *Assume the endogenous covariates satisfy*

$$\sum_{i,j\neq i}\gamma_{ij}(X,G_{-i})=\sum_{i,j\neq i}\gamma_{ij}(X,\hat{\sigma}^{G_{-i}}). \tag{19}$$

Let γ_{ij}^0 denote the output of $\gamma_{ij}(X,\sigma^G)$ and γ_0 denote the set of γ_{ij}^0 for all i,j. Then, under assumptions 1-4:

$$\sqrt{\frac{1}{2}n(n-1)}(\hat{\theta}-\theta_0)\xrightarrow{d}N(0,[V(\gamma_0,\theta_0)]^{-1}\Psi(\gamma_0,\theta_0,G)[V(\gamma_0,\theta_0)]^{-1}) \tag{20}$$

where V and Ψ are defined as in Equations 73 and 93 in the Appendix.

As mentioned above, proposition 4 relies on the endogenous covariates satisfying condition (19).[21] Lemma 1 proves this property for endogenous covariates of the form $\frac{1}{n-1}\sum_{k\neq i}G_{jk}\cdot\mu(X_k)$, where $\mu(X_k)$ represents some weighting function of agent k's attributes, assuming the beliefs are estimated according to (18). $\mu(\cdot)$ captures any sort of observed attributes that agents might care about in their indirect contacts. For instance, when deciding to form a link with someone, they may care not only about the number of this potential partner's friends but also about their wealth. The illustration of Section 5 makes use of covariates of this form.

Lemma 1. *Let $\gamma_{ij}(X,G_{-i})\equiv\frac{1}{n-1}\sum_{k\neq i}G_{jk}\cdot\mu(X_k)$, where $\mu(X_k)$ is some weighting function of the attributes of agent k and $\hat{\sigma}^{G_{-i}}$ be defined as in (18), then, for any G_{-i}, condition (19) holds.*

4 Simulations

We now describe the simulation exercise we designed to evaluate the asymptotic performance of the estimator in networks of increasing size (from $n=100$ to $n=500$). First we describe

[21] If condition (19) does not hold, one could still compute standard errors with an appropriately-designed bootstrap test.

the data generating process, then the estimation results.

4.1 Data Generating Process

For a given number of agents n with a two-dimensional attribute vector X_i, we posit a data generating process of the form:

$$X_{i,1} \sim U\{0,1\} \tag{21}$$

$$X_{i,2} \sim U\{0,1,2,3,4\} \tag{22}$$

$$\epsilon_{ij} \sim N(0,1) \tag{23}$$

$$v_{ij} = \theta_1 + \theta_2 X_{i,1} + \theta_3 X_{i,2} + \theta_4 \mathbb{1}\{X_{i,1} = X_{j,1}\} + \theta_5 |X_{i,2} - X_{j,2}| + \theta_6 \frac{1}{n-1} \sum_{k \neq i} G_{jk} \tag{24}$$

$$\theta_0 = [-2.8, 1, 0.5, 1, -0.1, 1]' \tag{25}$$

where $\frac{1}{n-1}\sum_{k\neq i} G_{jk}$ represents the average number of indirect friends that j grants access to, as in the example of Section 2.3. θ_0 is set so that the utility function is not dominated by its deterministic component, i.e. so that proposal decisions are sensitive to ϵ_{ij}.

The data generating process consists of three steps: first we draw the attribute $X_{i,1}$ and $X_{i,2}$ for all i. Second we find a corresponding symmetric equilibrium σ^G. We use an algorithm that starts from a randomly drawn belief matrix, computes the corresponding linking probabilities, and updates beliefs accordingly until convergence is achieved. Algorithm 1 describes the process in more detail.[22]

[22] For further details on its convergence behaviour see Rabinovich et al. (2013).

Algorithm 1: Search Algorithm

```
1 Generate a random belief matrix $\sigma^G$
2 Calculate the matrix of linking probabilities $L$, given $\sigma^G$, $X$ and $\theta_0$:
3     $L_{ij} = L_{ji} = \Phi(\mathbb{E}[v_{ij}(X, \sigma^{G_{-i}}, \theta_0)]) \cdot \Phi(\mathbb{E}[v_{ji}(X, \sigma^{G_{-j}}, \theta_0)])$
4 If $\sigma^G \not\approx L$:
5     Re-assign $\sigma^G = L$ and go back to line 2
6 Else:
7     Return $\sigma^G$
```

As a third step we draw the ϵ_{ij} values and construct a network realization G according to the following rule: a link in G exists if and only if the realization of ϵ_{ij} and ϵ_{ji} are such that $v_{ij}(X, \sigma^{G_{-i}}, \theta_0) + \epsilon_{ij} \geq 0$ and $v_{ji}(X, \sigma^{G_{-j}}, \theta_0) + \epsilon_{ji} \geq 0$.

For each $n \in \{100, 250, 500\}$ we generate 500 networks according to the procedure above. The networks that result from this process exhibit many commonly observed characteristics of real-world networks: the average geodesic distance between connected agents is low (≈ 2.2); the clustering coefficient is high compared to the linking probability of a comparable Poisson random network (≈ 0.27 vs. ≈ 0.1); and the degree distribution is positively skewed. The average degrees are approximately 10.9, 27.6 and 55.6 for $n \in \{100, 250, 500\}$ respectively.

4.2 Simulation Results

In the estimation step, for each simulation draw we use the realized network G and the agents attributes X (but not the error terms and beliefs) to estimate σ^G (as explained in Section 3.2). Then we maximize Equation (17) by replacing σ^G with $\hat{\sigma}^G$ to obtain $\hat{\theta}$.

Table 2 presents histograms for the exogenous coefficients capturing homophily $\hat{\theta}_4$ and $\hat{\theta}_5$ and for the endogenous coefficient $\hat{\theta}_6$. The values of the true coefficients are depicted by the vertical lines at the center of each sub-figure. As n increases the distributions of the estimated values become increasingly tight around the true values. This illustrates consistency.

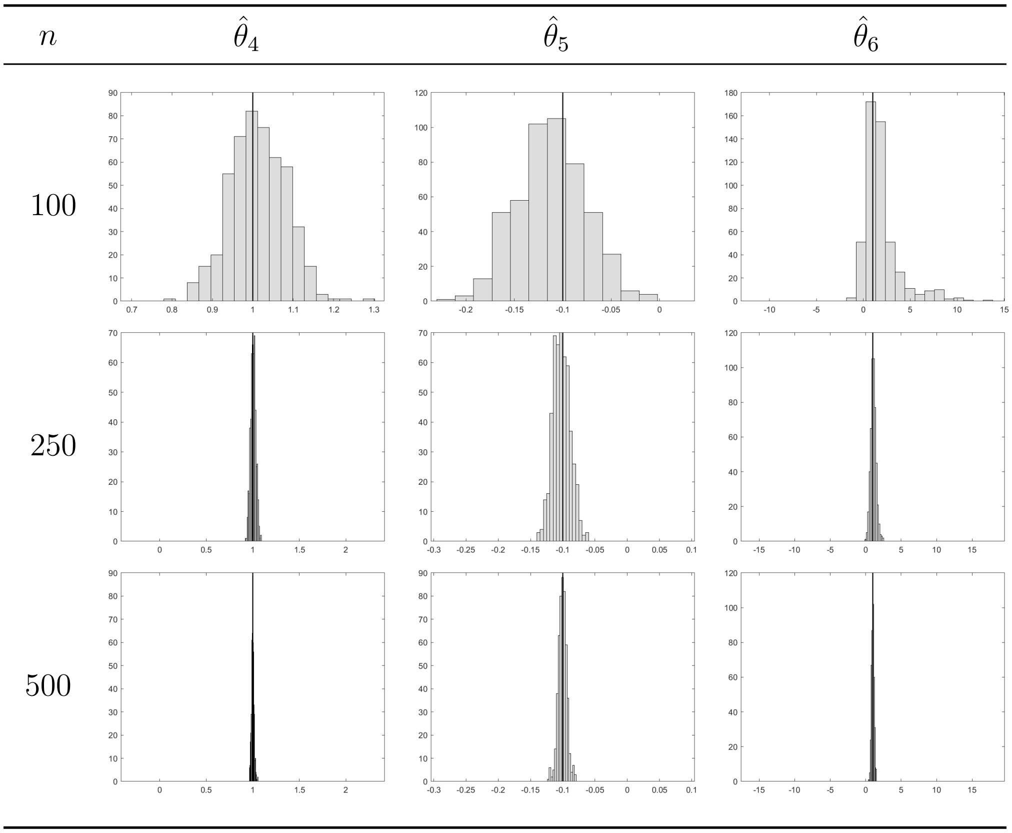


Table 2: Consistency

Note: The table reports histograms of estimated coefficients. The true values of the coefficients are depicted by the vertical line at the center of each sub-figure.

Table 3 presents the fitted Kernel distributions of $\sqrt{\frac{1}{2}n(n-1)}(\hat{\theta}-\theta_0)$ over all 500 iterations (in dashed lines) as well as true normal distributions with mean zero and variance $V^{-1}\Psi V^{-1}$ (in full lines). As n increases, the dashed lines converge to the full lines. This illustrates asymptotic normality.

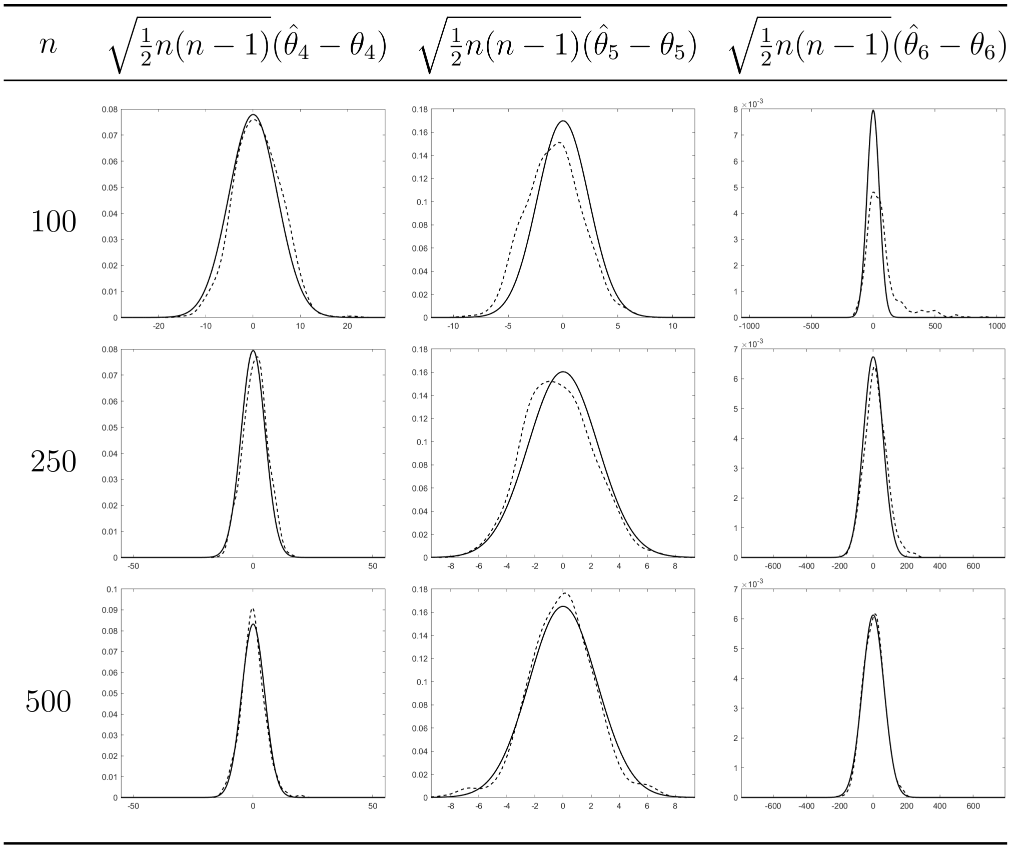


Table 3: Asymptotic normality

Note: The dashed lines depict the fitted Kernel distributions of $\sqrt{\frac{1}{2}n(n-1)}(\hat{\theta}-\theta_0)$. The full lines depict true normal distributions with mean 0 and variance $V^{-1}\Psi V^{-1}$.

5 Empirical Illustration

5.1 Data Description

We use data on the risk sharing network of Nyakatoke, a small village in the Buboka rural district of Tanzania.[23] Rural villages are an appropriate setting to study network formation, because the population can be entirely surveyed and several confounding effects (such as spatial and informational barriers) can be reasonably ruled out. The village of Nyakatoke consists of 119 households which have been interviewed in five regular intervals from February to December 2000. The data contains information on households' demographics, wealth,

[23] These data have been the object of numerous articles (De Weerdt and Dercon, 2006; De Weerdt and Fafchamps, 2011; Vandenbossche and Demuynck, 2013; Comola and Fafchamps, 2014).

income sources and income shocks, transfers and risk-sharing links. At the time of the study, the village of Nyakatoke is isolated (the few unpaved roads leading to the village are hardly passable during rains), densely inhabited (90% of households live within a distance of 1 kilometer from each other) and relatively poor (consumption for adult equivalent is less than 2 US$ a week, and average food share in consumption is 77%). Households earn most of their income from agricultural activities, especially the cultivation of coffee and banana; other sources of income are rare and off-farming activities are mostly considered supplementary to farming.

During the first survey round all respondents were asked *'Can you give a list of people from inside or outside of Nyakatoke, who you can personally rely on for help and/or that can rely on you for help in cash, kind or labour?'*.[24] The phrasing of this survey question was intended to capture undirected links of mutual assistance, and qualitative interviews and pilot tests suggested that respondents have understood it that way.[25] Our empirical exercise assumes that these survey responses represent undirected bilateral agreements of mutual help which could be activated if one of the partners is struck by an income shock. This is in line both with the survey design and with theoretical work on the voluntary nature of risk-sharing arrangements (Bloch et al., 2008; Jackson et al., 2012).[26] In Appendix B we revisit the trust network data analysed by Leung (2015), and show that our method could accommodate different network generation processes and yield different conclusions.

The resulting risk-sharing network of Nyakatoke consists of 490 links among $(119 \cdot 118)/2 = 7021$ household dyads. This network displays a mean geodesic distance of 2.5

[24] Respondents could list as many names as they wanted. They could also mention partners who live outside the village (this occurs in 34% of all declared partners). Since we have no information on the attributes of households outside the village we omit them from the analysis.

[25] This phrasing was first piloted in the Philippines (Fafchamps and Lund, 2003) and subsequently adopted in the Nyakatoke survey, because respondents understand it and are willing to answer. Other survey questions on directed flows were tried during the pilots, for instance drawing a distinction between people which respondents would help and people which respondents would seek help from. But respondents were confused by this distinction, which they perceived as non-existent, and complained they are asked the same question twice. See also Comola and Fafchamps (2014).

[26] In case of discordant reports, we assume that an undirected link exists whenever it is declared by at least one of the households involved. This is the most common stand in the empirical literature on risk-sharing links, and it is equivalent to assuming that all observed discordances are due to under-reporting.

steps and an average degree of 8.2. No household is isolated, and the network exhibits all the empirical regularities of large social networks.[27]

5.2 Main Results

We now illustrate the estimation procedure described in Section 3 using the Nyakatoke data. We take the household as a unit of observation ($n = 119$) and we include as covariates: a constant, the geographical distance between households (in meters), the wealth of j,[28] three types of homophily regressors, and two types of endogenous regressors. The homophily regressors are binary variables that take the value 1 if i and j belong to the same family,[29] same clan[30] or same religion[31] respectively. These exogenous covariates (i.e., distance, wealth and dummies for same family, clan and religion respectively) were identified by the previous literature as strong predictors of risk-sharing link formation in developing countries. The endogenous regressors are the number of j's friends ($\sum_{k \neq i} G_{jk}$) and the total wealth of j's friends ($\sum_{k \neq i} G_{jk} \cdot \text{Wealth}_k$).[32]

We run the first stage using the individual attributes that are used in the second stage (Wealth_j), as well as those implied by the relational attributes in the second stage (Family_i, Clan_i, Religion_i). Since the relational attribute "Distance_{ij}" does not imply a unique individual geographic location, we treat the entire vector of distances between i and the rest

[27] The Nyakatoke network has a unique component covering the entire population, the diameter is in the order of $\log(n)$ and the clustering coefficient is 7 times larger than in a randomly generated Poisson network with similar characteristics.

[28] The wealth of a household is defined as the total monetary value of its land and livestock assets (1 unit = $100,000$ Tanzanian shillings). Data on land were originally in acres and were transformed in monetary equivalent with a conversion rate of 300,000 *tzs* for 1 acre which reflects average local prices in 2000. For international comparisons, the exchange rate in 2000 was 1 US dollar for 800 *tzs*. Since land and livestock are publicly observable with a good degree of precision, we argue that the regressor satisfies the common-knowledge assumption (Section 2.1).

[29] Two households i and j are said to belong to the same family if there is some blood relation between at least one of the members of i and at least one of the members of j.

[30] There are 26 clans in Nyakatoke. 10 of them have only one household.

[31] There are three religions in Nyakatoke: Roman Catholic (49 households), Lutheran (46 households) and Muslim (24 households).

[32] For presentation purposes we do not re-scale these variables in the results of Table 4. In fact, the normalization is only needed to facilitate the asymptotic case where n approaches infinity.

of the households as i's individual attribute.[33] The categorical variables (family, clan, religion) and continuous variables (distance, wealth) are combined as described in Appendix A (in particular, Equation (31)), with $\lambda = 0.1$ and h set according to the "normal reference rule-of-thumb") and a normal kernel function. Figure 3 presents a histogram of the resulting estimated beliefs.

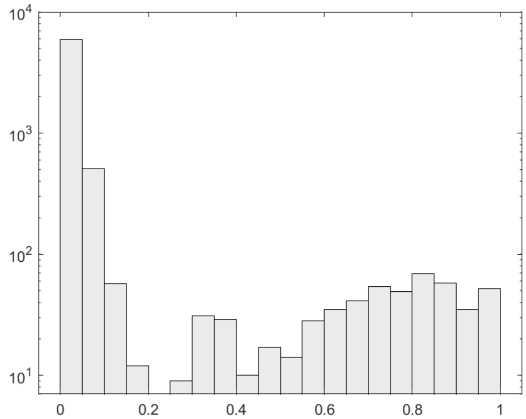


Figure 3: Histogram of the estimated beliefs in the Nyakatoke network

Note: the y-axis is on a logarithmic scale.

The results of the second stage are reported in Table 4. Column 1 presents a specification without endogenous regressors, for reference. Column 2 includes the endogenous regressors (number of j's friends only, total wealth of j's friends only, both). Column 3 presents the marginal effects that correspond to the specification of column 2. Standard errors are computed according the expression given in Proposition 4 with the true parameters replaced by their estimates.

As for the endogenous regressors, the coefficient of the number of j's friends can be positive or negative depending on whether households prefer potential partners to have many or few other partners. In principle, both types of externalities are conceivable in the context of risk-sharing arrangements: if j has many friends she may have a rather limited

[33] Consider a three-agent network in which agents 1 and 2 have the same geographic distances from (2,3) and (1,3), respectively. These distance profiles can be obtained by assuming various individual locations for agents 1 and 2, e.g. all location configurations in which all agents are located on a line and agents 1 and 2 are located symmetrically around agent 3.

amount of resources to devote to i, implying a negative coefficient. If j has many friends she is likely to be well-positioned to provide i with financial support in case of need, and is also less likely to rely heavily on i in case she herself is in need, implying a positive coefficient. The sum of wealth of j's friends is expected to be positive, as this grants j access to greater wealth which may indirectly benefit i.

Results in Table 4 provides evidence for the existence of network externalities. The positive sign of the coefficient of the number of j's friends suggests that the benefits from having a partner with many other partners (greater financial resilience) outweigh the costs (dilution of attention and/or resources). For the average pair i and j, an increment of one unit in the expected number of j's friends ($\approx 12\%$ of the average expected number of j's friends) is associated with an increase of roughly 0.016 in the probability of a proposal ($\approx 9\%$ of the average predicted proposal probability).

The signs of the other coefficients conform to our expectations. The constant appears negative, reflecting the idea that maintaining links is costly. The coefficient of the geographical distance between households is also negative, as distance is likely to render links harder to maintain. The coefficient of wealth is positive, as the wealthier a potential partner is the more helpful she could be in case of a negative income shock. The coefficients of the homophily regressors are all positive, in line with the large evidence that similarity between agents makes them more desirable to each other.

In Appendix B we present estimates obtained under different hypotheses about misreporting and the data generation process. The scope of the exercise is to illustrate the use of our estimation protocol in the context of self-reported network data. In particular, we modify our estimator to accommodate for a unilateral link formation rule, and we show that it yields different results from the directed unilateral estimator by Leung (2015).

	coefficients		mfx
	(1)	(2)	(3)
Same family	0.8436***	0.8493***	0.2934***
	(0.0627)	(0.0644)	(0.0256)
Same clan	0.1661***	0.1485**	0.0415**
	(0.0579)	(0.0602)	(0.0177)
Same religion	0.1649***	0.1751***	0.0495***
	(0.0401)	(0.041)	(0.0118)
Distance$_{ij}$	-0.0009***	-0.0009***	-0.0002***
	(0.0001)	(0.0001)	(0.0000)
Wealth$_j$	0.0586***	0.0376**	0.0098**
	(0.0069)	(0.0155)	(0.004)
Number of j's friends		0.0607***	0.0159***
		(0.0113)	(0.003)
Wealth of j's friends		-0.0002	0.000
		(0.0013)	(0.0003)
Constant	-0.6482***	-1.0967***	
	(0.0563)	(0.1063)	
# observations	7021	7021	

Notes: Column 3 reports the marginal effects for the specification of column 2. Standard errors in parentheses. Significance level based on false discovery rate q-values (Benjamini and Hochberg, 1995): *q<10%, **q<5%, and ***q<1%.

Table 4: Estimated coefficients.

6 Concluding remarks

Data on network interactions were previously scarce but are now becoming more available to economists. The current enthusiasm for network data from digital interaction platforms (Vosoughi et al., 2018; Blumenstock, 2018) has refueled the research interest about how non-digital links are formed, and how they respond to strategic incentives. Models of link formation with network externalities are at the frontier of the econometric research, facing difficulties related to dimensionality and equilibria multiplicity (Graham, 2015; Chandrasekhar, 2016; De Paula, 2017). Our paper fills a void in the literature by proposing a versatile method to estimate network externalities in a simultaneous-move game of undirected link formation. This method is naturally suited for bilateral link formation models, but it could also be applied to unilateral models where only the undirected link outcome (rather than the proposals) is observable. We provide existence, consistency and asymptotic normality results for the proposed estimator, and we test its asymptotic performance through a simulation exercise. In the context of bilateral link formation, this procedure provides a simpler alternative to methods exploiting pairwise stability under complete information (De Paula et al., 2018; Sheng, 2020). Importantly, it allows to make inference about various aspects of agents' preferences over network topology when data on a single (and possibly large) network are available. For instance, our method could be paired with data issued from a randomized experiment, allowing the researcher to disentangle endogenous network externalities from other exogenous factors (e.g., agents randomly allocated treatment status).[34]

We illustrate the method using data on risk-sharing in a Tanzanian village named Nyakatoke. Risk-sharing links are commonly assumed to be mutually agreed upon and provide an intriguing case for the role of externalities from indirect connections. Results confirm that the network architecture has an explanatory value: households seem to take into consideration the number of indirect friends they stand to gain when making linking decisions. Our estimates suggest that an additional two-steps-away connection is associated with an average increase of roughly 9% in the predicted proposal probability.

References

Advani, A. and Malde, B. (2014). Empirical Methods for Networks Data: Social Effects, Network Formation and Measurement Error. IFS Working Paper, page 91.

Aguirregabiria, V. and Mira, P. (2007). Sequential Estimation of Dynamic Discrete Games. Econometrica, 75(1):1 – 53.

Ambrus, A. and Elliott, M. (2020). Investments in social ties, risk sharing and inequality. Review of Economics Studies (forthcoming).

Ambrus, A., Mobius, M., and Szeidl, A. (2014). Consumption risk-sharing in social networks. American Economic Review, 104(1):149–82.

[34] The assumption that the attributes of others are observable suits well the case of a medium-sized village community where randomization is implemented through a public lottery.

Badev, A. (2020). Nash equilibria on (un)stable networks. Econometrica (forthcoming).

Bailey, M., Cao, R., Kuchler, T., and Stroebel, J. (2018). The economic effects of social networks: Evidence from the housing market. Journal of Political Economy, 126(6):2224–2276.

Bala, V. and Goyal, S. (2000). A noncooperative model of network formation. Econometrica, 68:1181–1230.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57(1):289–300.

Bjorn, P. A. and Vuong, Q. H. (1984). Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation. Technical report, WP, California Institute of Technology.

Bloch, F., Genicot, G., and Ray, D. (2008). Informal insurance in social networks. Journal of Economic Theory, 143(1):36–58.

Blumenstock, J. (2018). Dont forget people in the use of big data for development. Nature, 561:170–172.

Boucher, V. and Mourifie, I. (2017). My friend far, far away: a random field approach to exponential random graph models. The Econometrics Journal, 20(3):S14–S46.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2020). Peer effects in networks: a survey. Annual Review of Economics, 12:603–629.

Bramoullé, Y. and Genicot, G. (2023). Diffusion and targeting centrality.

Bramoullé, Y. and Kranton, R. (2007). Risk-sharing networks. Journal of Economic Behavior & Organization, 64(3-4):275–294.

Brandes, U. and Fleischer, D. (2005). Centrality measures based on current flow. In Annual symposium on theoretical aspects of computer science, pages 533–544. Springer.

Breschi, S. and Lissoni, F. (2005). " cross-firm" inventors and social networks: Localized knowledge spillovers revisited. Annales d'Economie et de Statistique, pages 189–209.

Bresnahan, T. F. and Reiss, P. C. (1991). Empirical models of discrete games. Journal of Econometrics, 48(1-2):57–81.

Buchel, K., Ehrlich, M., Puga, D., and Viladecans-Marsal, E. (2020). Calling from the outside: The role of networks in residential mobility. Journal of Urban Economics, C(119).

Candelaria, L. and Ura, T. (2018). Identification and inference of network formation games with misclassified links. working paper arXiv:1804.10118.

Chandrasekhar, A. (2016). Econometrics of network formation. The Oxford Handbook of the Economics of Networks, pages 303–357.

Chandrasekhar, A. G. and Jackson, M. O. (2016). A network formation model based on subgraphs. Available at SSRN 2660381.

Chandrasekhar, A. G. and Lewis, R. (2012). Econometrics of sampled networks.

Charness, G. and Jackson, M. O. (2007). Group play in games and the role of consent in network formation. Journal of Economic Theory, 136(1):417–445.

Coate, S. and Ravallion, M. (1993). Reciprocity without commitment: Characterization and performance of informal insurance arrangements. Journal of development Economics, 40(1):1–24.

Comola, M. and Fafchamps, M. (2014). Testing Unilateral and Bilateral Link Formation. Economic Journal, 124(579):954–976.

De Paula, A. (2017). Econometrics of network models. In Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress, pages 268–323. Cambridge University Press Cambridge.

De Paula, A., Richards-Shubik, S., and Tamer, E. T. (2018). Identifying preferences in networks with bounded degree. Econometrica, 86:263 – 288.

De Paula, A. and Tang, X. (2012). Inference of signs of interactions effects in simultaneous games with incomplete information. Econometrica, 80(1):143–172.

De Weerdt, J. and Dercon, S. (2006). Risk-sharing networks and insurance against illness. Journal of development Economics, 81(2):337–356.

De Weerdt, J. and Fafchamps, M. (2011). Social identity and the formation of health insurance networks. Journal of Development Studies, 47(8):1152–1177.

Ductor, L., Fafchamps, M., Goyal, S., and Van der Leij, M. (2014). Social networks and research output. Review of Economics and Statistics, 96(5):936948.

Fafchamps, M. and Lund, S. (2003). Risk-sharing networks in rural philippines. Journal of development Economics, 71(2):261–287.

Gaulier, G. and Zignago, S. (2010). Baci: International trade database at the product-level. CEPII Working Paper 2010-23.

Genicot, G. and Ray, D. (2003). Group formation in risk-sharing arrangements. The Review of Economic Studies, 70(1):87–113.

Gilleskie, D. and Zhang, Y. (2009). Friendship formation and smoking initiation among teens. unpublished.

Goeree, J. K., McConnell, M. A., Mitchell, T., Tromp, T., and Yariv, L. (2010). The 1/d law of giving. American Economic Journal: Microeconomics, 2(1):183–203.

Goyal, S., Van der Leij, M., and MoragaGonzlez, J. L. (2006). Economics: An emerging small world. Journal of Political Economy, 114(2):403–412.

Graham, B. (2017). An econometric model of network formation with degree heterogeneity. Econometrica, 85(4):1033 – 1063.

Graham, B. and De Paola, A. (2020). The Econometric Analysis of Network Data. Elsevier Academic Press.

Graham, B. and Pelican, A. (2019). Testing for externalities in network formation using simulation. In The Econometric Analysis of Network Data (B. Graham and A. de Paula, Eds.), pages 63 – 82.

Graham, B. S. (2015). Methods of identification in social networks. Annual Review of Economics, 7(1):465–485.

Hitsch, G., Hortasu, A., and Ariely, D. (2010). Matching and sorting in online dating. American Economic Review, 100(1):130–163.

Hoshino, T. (2019). Two-step estimation of incomplete information social interaction models with sample selection. Journal of Business Economic Statistics, 37(4):598–612.

Hsieh, C.-S. and Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. Journal of Applied Econometrics, 31(2):301–319.

Jackson, M. O., Rodriguez-Barraquer, T., and Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. American Economic Review, 102(5):1857–97.

Jackson, M. O. and Wolinsky, A. (1996). A strategic model of social and economic networks. Journal of economic theory, 71(1):44–74.

Kimball, M. S. (1988). Farmers' cooperatives as behavior toward risk. The American Economic Review, 78(1):224–232.

Kocherlakota, N. R. (1996). Implications of efficient risk sharing without commitment. The Review of Economic Studies, 63(4):595–609.

König, M., Liu, X., and Zenou, Y. (2019). R&d networks: Theory, empirics and policy implications. Review of Economics and Statistics, 101(3):476–491.

König, M. D. (2016). The formation of networks with local spillovers and limited observability. Theoretical Economics, 11(3):813–863.

Krishnan, P. and Sciubba, E. (2009). Links and architecture in village networks. The Economic Journal, 119(537):917–949.

Lalanne, M. and Seabright, P. (2022). The old boy network: are the professional networks of female executives less effective than men's for advancing their careers? Journal of Institutional Economics, pages 1–20.

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

Li, Q. and Racine, J. S. (2007). Nonparametric econometrics: theory and practice. Princeton University Press.

Mele, A. (2017). A structural model of dense network formation. Econometrica, 85(3):825–850.

Miyauchi, Y. (2016). Structural estimation of pairwise stable networks with nonnegative externality. Journal of Econometrics, 195(2):224–235.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of econometrics, 4:2111–2245.

Poirier, D. J. (1980). Partial observability in bivariate probit models. Journal of Econometrics, 12(2):209–217.

Rabinovich, Z., Naroditskiy, V., Gerding, E. H., and Jennings, N. R. (2013). Computing pure bayesian-nash equilibria in games with finite actions and continuous types. Artificial Intelligence, 195:106–139.

Ready, E. and Power, E. A. (2021). Measuring reciprocity: Double sampling, concordance, and network construction. Network Science, 9(4):387402.

Ridder, G. and Sheng, S. (2020). Estimation of large network formation games. working paper arXiv:2001.03838.

Sheng, S. (2020). A structural econometric analysis of network formation games through subnetworks. Econometrica, 88(5):1829–1858.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. Management science, 51(5):756–770.

Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. Social networks, 11(1):1–37.

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

Thirkettle, M. (2019). Identification and estimation of network statistics with missing link data. working paper.

Townsend, R. M. (1994). Risk and Insurance in Village India. Econometrica, 62(3):539–591.

Udry, C. (1994). Risk and insurance in a rural credit market: An empirical investigation in northern nigeria. The Review of Economic Studies, 61(3):495–526.

Vandenbossche, J. and Demuynck, T. (2013). Network formation with heterogeneous agents and absolute friction. Computational Economics, 42(1):23–45.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. Science, 359(6380):1146–1151.