Testing Unilateral and Bilateral Link Formation *

SHORT TITLE: Testing Link Formation

Margherita Comola † Marcel Fafchamps‡

May 13, 2013

Abstract

Empirical analysis of social networks is often based on self-reported links from survey data. How we interpret such data is crucial for drawing correct inference on network effects. We propose a method to test whether survey responses can safely be interpreted as a link and, if so, whether links are generated by a unilateral or bilateral link formation process. We present two empirical illustrations of the test on risk-sharing links in Tanzania and on communication among Indian farmers respectively, demonstrating the ability of the methodology to discriminate between competing data generating processes.

JEL codes: C12; C52; D85
Keywords: pairwise stability; self-reported link; non-nested test

---

*We thank Michael Wooldridge for his useful suggestions and Joachim De Weerdt for making the data available. We have benefitted from useful comments from Yann Bramoullé and from seminar participants at the Paris School of Economics, Oxford University, University of Nottingham, Stanford and Yale. All remaining errors are our own.

†Paris School of Economics (Université Paris 1 Pantheon-Sorbonne), Maison des Sciences Économiques, 106-112 Boulevard de l'Hôpital, 75647 Paris Cedex 13, France. Email: `margherita.comola@univ-paris1.fr`

‡Department of Economics, Oxford University, Manor Road, Oxford OX1 3UQ. Email: `marcel.fafchamps@economics.ox.ac.uk`

1. Introduction

A growing literature documents how social networks influence economic outcomes (Jackson, 2008, 2011), with a special focus on phenomena that fall in the category of favor exchange: risk sharing, mutual assistance, and information exchange (*e.g.*, about products and technologies, jobs, contract compliance, and market opportunities). The econometric analysis of social networks is relatively new, however, and there often is a lack of clarity on the implicit assumptions necessary for estimating models of this kind. In particular, many empirical studies on networks rely on survey questions to elicit social links between individuals: examples of these types of questions include 'who is your friend' (*e.g.*, Ballester *et al.*, 2006; Calvó-Armengol *et al.*, 2009); 'who do you exchange favors with' (*e.g.*, Jackson *et al.*, 2012); 'who do you turn to in times of trouble' (*e.g.*, Fafchamps and Lund, 2003; De Weerdt and Dercon, 2006); 'with whom do you exchange information about employment opportunities' (*e.g.*, Granovetter, 1995; Topa, 2001); 'do you discuss agricultural practices with individual X' (Conley and Udry, 2010). It is, however, not entirely clear what is the exact nature of the information that has been collected.

Two major issues are at hand. First, even when researchers make the survey question as precise and factual as possible, responses contain a residual subjective element because social relationships are subjective by nature.[1] Evidence of this is most appar-

[1] It is rarely possible to eliminate all ambiguity by asking more precise questions, if only because long convoluted questions are least well understood by the average survey respondent. Complex survey questions are particularly unfit for developing countries, where many respondents are illiterate or interviewed in a language they do not master fully. Semantic issues may also arise: for instance, many respondents answer the question 'who do you turn to in times of trouble' the same as 'who would you turn to in times of trouble' (this is particularly true if their language does not have a conditional tense, but even in English many respondents will confuse the two questions).

ent when two individuals are asked about the link between them: it is very common for their responses to be discordant, i.e., i cites j but j does not cite i (Fafchamps and Lund, 2003; De Weerdt, 2004; De Weerdt and Fafchamps, 2011; Liu *et al.,* 2011; Banerjee *et al.*, 2012). This feature raises concerns about the nature of self-reported link data: if a link truly exists between two individuals, they should both be aware of it.[2] The usual approach to discordant data is to assume mis-reporting, *i.e.,* people forget to mention some of their links. While this is reasonable in some cases, there are other possibilities (Marsden, 1990). One of them is that respondents list links that they wish to form but do not yet exist – in the sense that there is no arrangement with the prospective partner yet. Testing whether discordant survey responses can safely be interpreted as existing links is our first goal.

The second issue relates to whether the underlying link formation process is bilateral or unilateral, *i.e.*, whether links are mutually agreed or not. Bilateral link formation is central to much of economics: producer and consumer theory, general equilibrium and game theory, and intra-firm and intra-household bargaining all assume that agents can refuse transactions that are against their self-interest. Many economic models of link formation make similar assumptions, including most models of risk sharing and favor exchange (*e.g.*, Kimball, 1988; Coate and Ravallion, 1993; Kocherlakota, 1996; Bloch *et al.*, 2008; Jackson *et al.*, 2012). The same is true for the link formation models of Jackson and Wolinsky (1996), Boucher (2012), and Christakis *et al.* (2010),

[2] All link questions listed above aim at eliciting social relationships which result in some favor exchange arrangement. Depending on the context the link may be based on implicit rights or explicit contractual obligations, and may be in the interest of one or both parts, but both parts are expected to be aware of it. Note that there is a distinction between links defined on expected future exchanges *versus* past exchanges: although past exchanges often predict future exchanges, the fact that two people have exchanged favors in the past does not necessarily imply the arrangement is still on.

among others. Yet there are circumstances in which some form of pressure – external or internal – makes people form a link that is against their self-interest. This is for instance the case when transfers to others are compulsory by law *(e.g.*, alimony or child support payments). It is also conceivable that social norms make it difficult to refuse flows to and from others (*e.g.*, it goes against social norms not to contribute to a colleague's parting gift, or to refuse a parting gift).[3] If this is the case, link formation is best seen as unilateral.

The interpretation of self-declared links is not just a semantic distinction, it is essential for drawing inference about network effect. To illustrate with an example, suppose we ask students '*who do you go to when you have a question about econometrics*' and we find that students who mention the econometrics lecturer perform better at the exam. What can we conclude about network effects?

In most universities, students are entitled to ask the lecturer questions about taught material. Link formation is thus unilateral: the lecturer cannot refuse to see a student. If all students have equal access to the lecturer, better performance cannot, by design, be 'caused' by social links.[4] Correlation between social links and performance may nevertheless suggest that access to the lecturer helps those students who take advantage of it, and thus is a valuable information channel.

If the lecturer can agree to see some students and refuse others, link formation is

[3]Social norms can be enforced through social pressure (*e.g.*, the threat of ostracism) or internalized during upbringing. Platteau (1996) argues that many agrarian societies, especially in sub-Saharan Africa, cultivate egalitarian norms. The same point has repeatedly been made by anthropologists and other social scientists (*e.g.*, Scott, 1976). Barr and Stein (2008) provide some recent evidence.

[4]Some students may have reduced access to advice because office hours clash with their lecture schedule, or because of their poor English. If this is the case, unequal access may thus cause unequal performance correlated with social links. The study of network effects therefore demands a proper understanding of how the link formation process determines the differences in links across individuals.

bilateral. Here it is conceivable that some students perform better because of a social link: had the lecturer refused to see them, they would have performed worse, and *vice versa*. Here the link has a potentially causal effect on performance in the sense usually associated with the phrase 'network effects', *i.e.*, better networks cause better outcomes. Whether link formation is unilateral or bilateral thus affects inference about network effects.

It is also conceivable that the question does not measure social links. To see this, imagine that no student has seen the lecturer but, when answering the question, some confident students listed the lecturer because they knew that, if they had a question, they would go to him or her. If these students are also better on average, a correlation will result between listing the lecturer and exam performance even though there were no social link – and hence there cannot be network effects. Researchers who use answers to questions of this type to draw inference about network effects would thus like to know whether answers can safely be regarded as measuring social links.

The contribution of this paper is to propose a testing methodology to shed light on the interpretation of self-reported link data in relation to the two issues outlined above. Operationally, the test assumes that the researcher has proxies for the values U_i and U_j that i and j respectively assign to a link g_{ij} between them. The contribution is articulated in two parts: first, if one observes discordant responses from i and j, our test evaluates whether these responses are more likely to represent existing but misreported links, or a desire to link. The logic of this test is that, if mis-reporting is random, i and j are equally likely to report the link. But if respondents report mostly links they care about, whether or not these links exist, then i is more likely to report the link than j if U_i is larger than U_j. The second contribution is to propose a test of whether existing links are best understood as the result of a unilateral or bilateral link formation process. The intuition behind this test is that, if a link is bilateral, the link must be in the self-interest of both i and j. In contrast, if link formation is unilateral, the link may be in the interest of only i or j.[5]

We provide two illustrations of our methodology using observational data. The first illustration relies on self-reported risk-sharing links in a Tanzanian village. Risk sharing has received considerable attention, especially in the development literature (Scott, 1976; Altonji *et al.*, 1992; Townsend, 1994). Since the survey question is intended to capture mutual assistance, it should in principle be answered in the same way by i and j irrespective of whether flows between them are one-sided or reciprocated.[6] In practice, however, discordant responses are very frequent. One possible interpretation is that respondents give the names of households from whom they intend to seek assistance in case of need; another interpretation is that they provide information on existing risk-sharing links but their responses differ because of mis-reporting. When we peg the models against each other, we find that the desire-to-link model provides the best fit to these data. Our result is important because we know from research undertaken on the same – or similar – data (*e.g.*, Fafchamps and Lund, 2003; De Weerdt and Dercon, 2006; Fafchamps and Gubert, 2007; De Weerdt and Fafchamps, 2011) that when i lists

[5]Reporting a desire to link is *a priori* more compatible with unilateral link formation: if link formation is unilateral, i is more likely to anticipate future interaction with j irrespective of whether j has been made aware of this. In contrast, if link formation is bilateral, responses are more likely to capture situations in which i has previously verified j's willingness to interact and thus can reasonably anticipate future favor exchanges.

[6]Our test does *not* rely on whether favors are reciprocated, and we do not look at directed flows of favors. As illustrated for instance in Goyal (2007), the two concepts are distinct. On the one hand, it is possible for a one-way transfer of favors to require the assent of both parties – *e.g.*, someone who has been awarded a grant has the right to refuse it, for instance because it comes from a disreputable source. On the other hand, a two-way transfer of favors may be impossible to refuse – *e.g.*, someone offers you a gift that you cannot turn down but feel obliged to reciprocate.

j as a partner, a subsequent transfer between i and j is more likely to be observed. This has until now been interpreted as evidence that access to favors is limited by pre-existing social relationships. But if survey responses represent desire to link, we cannot rule out that people receive help from those they wish to receive help from, irrespective of whether these flows are *ex-ante* expected by the other side involved.

We further illustrate our methodology using data on information sharing among farmers in the Indian province of Maharashtra. In rural economies, social learning through friends and relatives is particularly important in agriculture, given the prevalence of information asymmetries and high search costs (Foster and Rosenzweig, 1995; Munshi, 2004; Bandiera and Rasul, 2006; Conley and Udry, 2010). Our data contains detailed information on a sample of farming households growing a diverse range of commercial crops for many villages. For each other sampled farmer in the village, respondents were asked *'how often do you discuss agricultural issues with this person or member of his household?'* Responses to this question constitute a self-reported link, which may as well be discordant. Given the question – it is hard to imagine that respondents report conversations which did not take place – our prior is that discordant responses correspond to misreported links. This is indeed what we find: the desire-to-link model is outperformed by both bilateral and unilateral link formation models, and overall we find that the unilateral model wins. This suggests that respondents are able to initiate conversations about farming practices with knowledgeable farmers who are themselves unlikely to learn anything from the conversation. Given that information about farming practices is non-rival, unilateral link formation may improve efficiency by circulating information more widely.[7]

[7]Abstracting from the strategic substitute effects that arise in learning about new technology

The paper is organized as follows. In Section 2 we provide a conceptual framework and describe our empirical strategy. The first illustration of our methodology on risk-sharing data from Tanzania is described in Section 3. The application on Maharashtra farmers' communication data is discussed in Section 4. Section 5 concludes. The estimation of hybrid models is discussed in Appendix A. The Online Appendix B investigates our test's performance in presence of clustering.

2. Conceptual framework and testing strategy

2.1. Overview and notation

In this section we present our estimation and testing strategy. Our tests build on the work of Comola (2012) and take pairwise stability as starting point for the estimation process. First introduced by Jackson and Wolinsky (1996), pairwise stability has established itself as a cornerstone equilibrium condition in the study of bilateral link formation processes (Goyal, 2007). Comola (2012) has shown that the restrictions imposed by pairwise stability take the form of a bivariate probit model with partial observability (Poirier, 1980). We extend this approach by noting that, under unilateral link formation, the absence of a link is formally equivalent to a pairwise stable decision by *both* agents (nodes) *not* to form a link. In contrast, if responses only represent desire to link, the relevant regression model is a simple probit.

Building on these insights, we propose a method to test whether information collected from survey respondents is most consistent with desire to link, bilateral, or unilateral link formation. This is achieved using the non-nested likelihood ratio test

(Foster and Rosenzweig, 1995).

proposed by Vuong (1989), that we correct for network dependence across residuals. Using simulations, we show that our test is able to select the correct model, even in the presence of measurement error in the regressors. As an extension, in the Appendix A we discuss how our methodology can be modified to incorporate elements of self-censoring (Hitsch *et al.*, 2005).

Throughout the paper, we define the link g_{ij} as a favor exchange arrangement between nodes i and j.[8] Links are symmetric, *i.e.*, $g_{ij} = g_{ji}$. Formally, for each pair of nodes (dyad) ij, define $g^i_{ij} = 1$ if i reported a link with j, and 0 otherwise. Similarly define $g^j_{ji} = 1$ if j reported a link with i. Variables g^i_{ij} and g^j_{ji} provide a representation of the data. Their interpretation varies depending on what the data generation process is. In subsection (2.2) we assume that g^i_{ij} and g^j_{ji} capture desire to link and we specify the corresponding data generation process. In subsections (2.3) and (2.4) we regard g^i_{ij} and g^j_{ji} as two different measurements of the same link g_{ij}. Subsection (2.3) specifies the data generation process if link formation is bilateral while subsection (2.4) focuses on the unilateral scenario.

2.2. Desire to link

In this subsection the variable g^i_{ij} is interpreted as i's interest in forming a favor exchange link with j – and similarly for g^j_{ij}. We focus on a single link ij at a time, keeping the rest of the network $g = [g_{mn \neq ij}]$ constant. By a standard abuse of notation, let g_{-ij} denote the network g without the link g_{ij}, that is, with $g_{ij} = 0$. Similarly, let g_{+ij}

[8] What this means depends on whether link formation is bilateral or unilateral. If it is bilateral, $g_{ij} = 1$ means that i and j have agreed to exchange favors because it is in their mutual self-interest. If it is unilateral, $g_{ij} = 1$ means that both i and j expect favor exchanges, even though it may only be in the interest of i or j.

denote the network with the link g_{ij}, that is, with $g_{ij} = 1$. The utility that node i derives from network g is written $U_i(g)$. Thus, the gain to node i of forming the link g_{ij} is written $U_i(g_{+ij}) - U_i(g_{-ij})$. We approximate this gain by a linear function of a vector of observables X_{ij} and a zero-mean residual ε_{ij}:

$$
\begin{aligned}
U_i(g_{+ij}) - U_i(g_{-ij}) &= \alpha + X'_{ij}\beta - \varepsilon_{ij} &\quad (1)\\
U_j(g_{+ji}) - U_j(g_{-ji}) &= \alpha + X'_{ji}\beta - \varepsilon_{ji} &\quad (2)
\end{aligned}
$$

The key maintained assumption of our testing strategy is that X_{ij} contains a suitable predictor of the gain that i would obtain from a link with j, and that $X_{ji} \neq X_{ij}$ for at least some dyad – for a discussion on identification see subsection (2.7). Since the order in which i and j appear in the data is arbitrary, equations (1) and (2) must be interchangeable, which implies that the coefficient vector β must be the same in the two equations. Assuming that $(\varepsilon_{ij}, \varepsilon_{ji})$ are jointly normal, equations (1) and (2) can be estimated as a standard probit by stacking observations g^i_{ij} and g^j_{ji}:

$$
\begin{aligned}
\Pr(g^i_{ij} &= 1) = \Pr\left(U_i(g_{+ij}) \geq U_i(g_{-ij})\right) = \Pr\left(\varepsilon_{ij} \leq \alpha + X'_{ij}\beta\right)\\
\Pr(g^j_{ji} &= 1) = \Pr\left(U_j(g_{+ji}) \geq U_j(g_{-ji})\right) = \Pr\left(\varepsilon_{ji} \leq \alpha + X'_{ji}\beta\right) \quad (3)
\end{aligned}
$$

In what follows we refer to equation (3) as to the desire-to-link model.[9] Since interdependencies are likely we allow for possible correlation between ε_{ij} and ε_{ji}, as discussed in subsection (2.5).

[9] Reporting desire to link is more compatible with a scenario where link formation is unilateral, because respondents may list links that have not yet been activated. In this case, we cannot rule out that respondents have also listed some links they would not be able to form if they tried, perhaps because the perspective partner is not sufficiently close socially and/or geographically.

2.3. Bilateral link formation

Let us now interpret g_{ij}^{i} and g_{ji}^{j} as two separate measurements of the same existing link g_{ij}. This implies that discrepancies in survey responses must be imputed to misreporting, as we discuss in what follows. We first consider bilateral link formation. To specify the bilateral data generation process, we impose the local equilibrium conditions implied by pairwise stability: a link between i and j exists if both i and j wish to form it, and it is severed if any of them wishes so (*e.g.*, Jackson and Wolinsky, 1996). Formally, a network is pairwise stable if and only if:

$$\begin{aligned}\forall g_{ij} &= 1, U_i(g_{+ij}) \geq U_i(g_{-ij}) \text{ and } U_j(g_{+ij}) \geq U_j(g_{-ij}) \\ \forall g_{ij} &= 0, \text{ if } U_i(g_{-ij}) < U_i(g_{+ij}) \text{ then } U_j(g_{-ij}) > U_j(g_{+ij})\end{aligned}$$

This set of conditions implies that:

$$\Pr(g_{ij} = 1) = \Pr\left(U_i(g_{+ij}) \geq U_i(g_{-ij}) \text{ and } U_j(g_{+ij}) \geq U_j(g_{-ij})\right) \tag{4}$$

Using (1) and (2), equation (4) is equivalent to:

$$\Pr(g_{ij} = 1) = \Pr\left(\varepsilon_{ij} \leq \alpha + X_{ij}'\beta \text{ and } \varepsilon_{ji} \leq \alpha + X_{ji}'\beta\right) \tag{5}$$

where $(\varepsilon_{ij}, \varepsilon_{ji})$ are jointly normal.[10] Model (5) has a single dependent variable but two regressing equations. Such model, proposed by Poirier (1980) and first used by Comola

[10] Fox (2008) proposes an alternative estimation strategy based on pairwise stability. This strategy, however, only applies to link formation processes that satisfy transferable utility, and therefore cannot be used here.

(2012) to model network formation, is known as a partial observability bivariate probit. The link g_{ij} can be understood as the product of two distinct and unobservable binary events: i's to form the link ij, and j's to form that same link. Let us define these unobservable variables as w^i_{ij} and w^j_{ji}, with $w^i_{ij} = 1$ if $\varepsilon_{ij} \leq \alpha + X'_{ij}\beta$ and similarly for w^j_{ji}. Under pairwise stability, a link is formed if and only if both i and j wish to form it, *i.e.*, $g_{ij} = 1$ if and only if $w^i_{ij} = 1$ and $w^j_{ji} = 1$ or, more succinctly, $w^i_{ij}w^j_{ji} = 1$. The term 'partial observability' comes from the fact that we only observe the product $w^i_{ij}w^j_{ji}$, not each of them separately. That is, whenever $g_{ij} = 0$ we can not observe whether one or both nodes are not willing to form the link.

In practice, in our data we have two measurements g^i_{ij} and g^j_{ji} of g_{ij}. Since we have no reason to believe one measurement more than the other when they are discordant, we take the most neutral stand on mis-reporting and give each measurement equal weight.[11] With this assumption the estimated model becomes:

$$\begin{aligned}
\Pr(g^i_{ij} &= 1) = \Pr\left(\varepsilon_{ij} \leq \alpha + X'_{ij}\beta \text{ and } \varepsilon_{ji} \leq \alpha + X'_{ji}\beta\right) \\
\Pr(g^j_{ji} &= 1) = \Pr\left(\varepsilon_{ji} \leq \alpha + X'_{ji}\beta \text{ and } \varepsilon_{ij} \leq \alpha + X'_{ij}\beta\right) \qquad (6)
\end{aligned}$$

Estimating β under the assumption of bilateral link formation boils down to maximizing

[11] Giving discordant responses different weights is equivalent to making an assumption on whether links are more likely to be over-reported or under-reported. Assuming no over-reporting, one should give zero weight to $g^i_{ij} = 0$ if $g^j_{ji} = 1$. Conversely, assuming no under-reporting, one should give zero weight to $g^i_{ij} = 1$ if $g^j_{ji} = 0$. Giving g^i_{ij} and g^j_{ji} equal weight is equivalent to assuming that over-reporting and under-reporting are equally likely, which is a necessary condition for the Vuong test to be applicable in the current setting. In fact if we estimate the model assuming that discordant pairs are due to over-reporting or under-reporting only, *de facto* we transform the dependent variable, so that the Vuong test cannot be used to compare these models to model (3). In a different estimation framework that is not applicable here, Comola and Fafchamps (2010) propose a structural approach for dealing with either over-reporting or under-reporting – but not with both simultaneously.

the likelihood function defined by (6).

In contrast to the recent literature on the estimation of network formation models (*e.g.*, Christakis *et al.*, 2010; Mele, 2011; Chandrasekhar and Jackson, 2012; Sheng, 2012), we sidestep the thorny issue of multiple equilibria by using pairwise stability purely as a local equilibrium condition. To convey the intuition behind this approach, imagine having data on the actions and payoffs of players in a game with multiple equilibria. The network formation papers cited above seek to test whether the combined actions of all players are consistent with the game being in a particular type of equilibrium or configuration (*e.g.*, with a certain proportion of triads and dyads). The multiplicity of equilibria is a serious concern here: the researcher must evaluate the likelihood of the data under alternative configurations, and compare this to the likelihood of the data under the configuration of interest. When the number of equilibria is very large, the evaluation of the first likelihood is extremely difficult. Our approach is different in that we test whether the actions of each *pair* of players are compatible with pairwise stability, keeping the actions of other players constant. The fact that the game may have multiple equilibria is now irrelevant, because our focus is local: as long as the conditional payoffs of each players' pair are identified, it is possible to estimate the likelihood that an observed link is compatible with pairwise stability using (6). Interdependence among players' actions is dealt with by allowing for possible correlation not only between ε_{ij} and ε_{ji} in (6), but also across different dyads, as explained in subsection (2.5).

2.4. Unilateral link formation

A network may also result from a process of unilateral link formation, which corresponds to a situation where only one side's desire to link is sufficient for a link to be formed (Goyal, 2007). As in the bilateral case, let w^i_{ij} and w^j_{ji} represent the nodes' unobserved desire to form link g_{ij}. Under unilateral link formation, $g_{ij} = 1$ whenever either of the two nodes wishes to form a link. It follows that $g_{ij} = 0$ only when both nodes do not wish to form the link. This simple observation forms the basis of our estimation strategy because it implies that, using a change of variable, the unilateral link formation model can also be estimated as a partial observability model. To see how this is possible, we begin by noting that:

$$\begin{aligned} \Pr(g_{ij} = 0) &= \Pr\left(U_i(g_{+ij}) < U_i(g_{-ij}) \text{ and } U_j(g_{+ij}) < U_j(g_{-ij})\right) \\ &= \Pr\left(\varepsilon_{ij} > \alpha + X'_{ij}\beta \text{ and } \varepsilon_{ji} > \alpha + X'_{ji}\beta\right) \end{aligned} \tag{7}$$

Let $h_{ij} \equiv 1 - g_{ij}$. We have $h_{ij} = 1$ if and only if $w^i_{ij} = 0$ and $w^j_{ji} = 0$ or, more succinctly, $(1 - w^i_{ij})(1 - w^j_{ji}) = 1$. Estimation can proceed by applying a partial observability bivariate probit to the transformed model:

$$\Pr(h_{ij} = 1) = \Pr\left(-\varepsilon_{ij} \leq -\alpha - X'_{ij}\beta \text{ and } -\varepsilon_{ji} \leq -\alpha - X'_{ji}\beta\right) \tag{8}$$

The dependent variable is still binary, and the partial observability feature ensures that the absence of a link is interpreted as implying that both nodes do not wish to form that link. As it is clear from (8), estimated coefficients have the reverse sign compared to (5). This is because we are estimating nodes' desire *not* to form a link.

Once again, we have two measurements h_{ij}^i and h_{ji}^j of h_{ij}, and we take a neutral stand on mis-reporting. The estimated model is thus:

$$\begin{aligned}\Pr(h_{ij}^i &= 1) = \Pr\left(-\varepsilon_{ij} \leq -\alpha - X_{ij}'\beta \text{ and } -\varepsilon_{ji} \leq -\alpha - X_{ji}'\beta\right) \\ \Pr(h_{ji}^j &= 1) = \Pr\left(-\varepsilon_{ji} \leq -\alpha - X_{ji}'\beta \text{ and } -\varepsilon_{ij} \leq -\alpha - X_{ij}'\beta\right) \end{aligned} \tag{9}$$

Estimating β under the assumption of unilateral link formation thus boils down to maximizing the likelihood function implicitly defined by (9). Note that unilateral link formation does not imply that any individual i can form a link with any j if he so wishes: individuals may be restricted in the links they can form unilaterally, *e.g.*, some links may be off limit. This is captured by including in vector X_{ij} regressors that proxy for geographical and social distance between i and j.

2.5. Standard errors

Decisions to link are not independent of each other. Model prediction errors are therefore correlated, sometimes negatively, across observations. This is a common problem in dyadic data which can seldom if ever be regarded as made of independent observations. This invalidates inference unless standard errors are corrected to account for non-independence.

For dyadic data, the most pressing concern is the correlation in the residual for observation g_{ij}^i with those pertaining to all observations involving nodes i and j. This is because i's decision to form a link with j potentially affects his or her decision to form a link with any other node. Extending Conley (1999), Fafchamps and Gubert

(2007) have proposed a correction for dyadic standard errors dependence of the form:

$$AVar(\widehat{\beta}) = \frac{1}{N-K}(X'X)^{-1}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{m_{ijkl}}{2N}X_{ij}u_{ij}u'_{kl}X_{kl}\right)(X'X)^{-1} \qquad (10)$$

where β denotes the vector of coefficients, N is the number of dyadic observations, K is the number of regressors, X is the matrix of all regressors, X_{ij} is the vector of regressors for dyadic observation ij, and $m_{ijkl} = 1$ if $i = k, j = l, i = l$ or $j = k$, and 0 otherwise.[12] The only structure imposed is that $E[u_{ij}, u_{ik}] \neq 0$, $E[u_{ij}, u_{kj}] \neq 0$, $E[u_{ij}, u_{jk}] \neq 0$ and $E[u_{ij}, u_{ki}] \neq 0$ for all k but that $E[u_{ij}, u_{km}] = 0$ otherwise.

In this paper we use formula (10) to correct the standard errors when our data belong to a single population (Section 3).[13] When we have data from several unlinked populations (Section 4) we cluster standard errors at the level of each village (Arcand and Fafchamps, 2012; Barr *et al.*, 2012). This latter solution is more flexible as it allows for arbitrary cross-observation dependence (*i.e.*, $E[u_{ij}, u_{km}] \neq 0$ for $i \neq k, m$ and $j \neq k, m$).

2.6. Non-nested tests

Our aim is to test which one of the models presented above best accounts for the data. To this effect we proceed by pairwise comparisons. Vuong (1989) has proposed a framework for hypothesis testing in non-nested models. Say we want to test which of two alternative, non-nested models k and m fit the data best. Let $M = N(N-1)$ be

[12] Formula (10) was developed for linear regressions where u_{ij} denotes the residual from observation ij. To apply it to maximum likelihood estimation, simply replace u_{ij} by the corresponding log likelihood contribution (score) l_{ij}.

[13] Bester *et al.* (2011) have suggested an alternative approach to eliminate bias in standard errors by dividing the data into large blocks and clustering within blocks. Unfortunately this approach requires a large sample, which is not our case.

the total number of dyadic observations. The original form of the Vuong test statistic is

$$V = \frac{M^{-1/2} LR(k,m)}{\hat{\omega}} \xrightarrow{d} N(0,1)$$

where $LR(k,m) \equiv L^k - L^m$ is the log of the likelihood ratio statistic and:

$$\hat{\omega}^2 = \frac{1}{M} \sum_{ij=1}^{M} \left[\log \frac{l_{ij}^k}{l_{ij}^m} \right]^2 - \left[\frac{1}{M} \sum_{ij=1}^{M} \log \frac{l_{ij}^k}{l_{ij}^m} \right]^2$$

where l_{ij}^k and l_{ij}^m are the observation-specific scores for each model k and m. This test can be implemented more simply by regressing the difference between scores on a constant:[14]

$$l_{ij}^k - l_{ij}^m = \alpha_{km} + v_{ij}^{km}$$

The t-value on the constant α_{km} is the Vuong statistic that tests whether model k outperforms model m. For inference to be valid, we correct the standard error of the constant $\widehat{\alpha}_{km}$ for cross-dependence across observations using formula (10) when we have data from one fully-connected population (Section 3), and we cluster standard errors at the village level when we have data from several unlinked populations (Section 4).

2.7. Identification

Identification of the model that best fits the data relies on two main data features. The first one is critical for all tests, the second one is essential to test the desire-desire-to-link

[14]The Vuong test requires that the models have the same dependent variable. This condition is satisfied by construction for models (3) and (6). In spite of the change of variable from g_{ij}^i to $h_{ij}^i = 1 - g_{ij}^i$, it is also satisfied for model (9) because the scores are the same except for a sign change, which we correct for.

model against the other two.

The first requirement is to have good *a priori* predictors X_{ij} of each node i's self-interest in forming the link with j, and *vice versa*. Furthermore these predictors must satisfy $X_{ij} \neq X_{ji}$ for at least some X, or else identification fails. $X_{ij} = X_{ji}$ arises generically for predictors aimed at capturing homophyly – *e.g.*, same gender, kinship, geographical distance. $X_{ij} \neq X_{ji}$ generically arises for predictors that capture common preferences – *e.g.*, everyone prefers to link to the richest or most knowledgeable nodes. If the researcher has proxies for these common preferences, and $X_{ij} \neq X_{ji}$ is satisfied for at least some X, it is possible to test unilateral versus bilateral link formation. That identification is achieved can be seen by noting that, in the bilateral case, it is unlikely to observe a link g_{ij} when either i or j strongly wishes not to link. In contrast, in the unilateral case, it is unlikely *not* to observe a link when one of the nodes strongly wishes to link. It is this difference between the predictions made by the two models when $X_{ij} \neq X_{ji}$ that makes identification possible.

To test the desire-to-link model (3) against either of the models (6) or (9), an additional requirement must be satisfied: we must have, for each pair of nodes, separate statements g_{ij}^i and g_{ji}^j from i and j regarding link g_{ij}. Since identification is achieved from the pattern of discordant responses when $g_{ij}^i \neq g_{ji}^j$, the test does not work if link information is collected only from i or j but not both.[15] If each respondent has listed links he or she wishes to form, differences in responses $g_{ij}^i \neq g_{ji}^j$ should match differences in self-interest $X_{ij} \neq X_{ji}$, as assumed in (3). If this is the case, it follows

[15] The focus of the paper is on those link questions which contain a subjective element. Obviously this test does not apply to other types of network data from administrative sources (*e.g.,* data recording 'objective' transactions, for instance phone conversations) – although the bilateral *vs.* unilateral test would still apply.

that model (3) is the correct data generating process. To see this, assume instead that each respondent reports only existing links, that is, situations in which both i and j are aware that $g_{ij} = 1$. If i and j correctly report all their links, $g_{ij}^i = g_{ij}^i = g_{ij}$. If $g_{ij}^i \neq g_{ji}^j$, respondents must have misreported some of their links. If mis-reporting is uncorrelated with X_{ij} and X_{ji}, then both g_{ij}^i and g_{ji}^j should be equally likely, and models (6) and (9) are the correct data generating process, under the maintained assumptions common to all three models.

As a last caveat, it is also possible that respondents list links that have a high value for them, *i.e.*, links with a high $X_{ij}'\beta$. If mis-reporting takes this specific form, g_{ij}^i is better predicted by X_{ij} while g_{ji}^j is better predicted by X_{ji}, as in model (3), thus we have no way of distinguishing whether survey responses correspond to desire to link or to those existing links respondents most care about.

2.8. Simulations

Our methodology relies on proven maximum likelihood methods and non-nested tests. The reader may nevertheless wonder whether it works in practice. To investigate this issue, we present results from a simulation analysis along three lines.

First, we present a simulation exercise which compares the three baseline models, to reassure the reader that our test is able to select the right data generation process under standard conditions. For each process described in Section 2 we generate many artificial network draws with 100 nodes (*i.e.*, 9900 dyads) and a single predictor $X_{ij} \neq X_{ji}$. We then estimate and compare the three models (3), (6) and (9) using the Vuong test described above, and replicate this procedure for each artificially generated sample. To make the simulation results as comparable as possible across models, we impose link probabilities that are similar for the three processes.[16] Table 1 reports the simulation results, consisting in nine Vuong tests (three pairwise comparisons for each of the three true data generating processes).[17] The mean of the test statistic α_{km} across replications is reported, together with its standard deviation, minimum, and maximum. We see that the test nearly always picks the correct model and, in most cases, the power of the test is good, with few replications where the absolute value of the test statistic falls below 1.96, the critical value for a 5% level of significance. Critical to this good performance is the correction of the standard error of α_{km} using formula (10) to correct for non-independence across observations.

Table 1: Simulation results

True model	Test	Mean	St.d.	Min	Max
desire to link (D)	D vs. U	6.01	0.76	3.93	8.00
desire to link (D)	D vs. B	5.83	0.80	3.41	8.08
desire to link (D)	B vs. U	-3.67	0.79	-5.32	-1.63
unilateral (U)	D vs. U	-5.48	1.13	-12.75	-3.24
unilateral (U)	D vs. B	-5.03	1.17	-9.32	-2.33
unilateral (U)	B vs. U	-2.91	0.97	-8.44	-0.26
bilateral (B)	D vs. U	-4.31	0.91	-7.06	-2.33
bilateral (B)	D vs. B	-5.92	1.09	-10.36	-3.31
bilateral (B)	B vs. U	4.36	0.84	2.54	7.84

Second, we investigate how sensitive the testing strategy is to measurement error, proceeding as follows. First, for each of the three models (3), (6) and (9) we generate

[16] Data are generated in such a way that the average of the dyadic dependent variable is the same whether the true process is desire to link, bilateral or unilateral link formation. This proportion of $g^i_{ij} = 1$ is chosen to match the observational data of Section 3.

[17] Partial observability models tend to encounter convergence difficulties. Using a simplex algorithms for non-concave regions of the likelihood function alleviates the problem most of the time, but occasionally convergence cannot be achieved. 250 samples were generated for each of the nine test pairs, but for an handful of samples a partial observability model failed to converge. For this reason, the number of usable replications can be less than 250.

250 networks based on one uniformly distributed regressor X_j.[18] When we estimate the Vuong tests over these 250×3 observations. The tests select the correct model in all cases. We then repeat the experiment replacing regressor X_j with a mis-measured $Z_j \equiv X_j + \varepsilon_2$ with $\varepsilon_2 \sim N(0, 0.3)$.[19] The test still selects the correct model in all 250 replications for the bilateral and desire-to-link models, and in 245 out of 250 cases for the unilateral link formation model. From this we conclude that moderate measurement error does not constitute a serious threat to identification as long we have a sufficiently strong set of predictors for the desire to link.

Finally, we investigate whether our test is able to select the right data generation process in presence of strong inter-relatedness in linking decisions of the form theoretically proposed by Jackson, Rodriguez-Barraquer and Tan (2012). The simulation exercise is explained in details in Online Appendix B. The results confirm that, with the correction for cross-dependence across observations that we discussed earlier, such inter-relatedness does not invalidate our testing strategy.

3. Risk sharing in rural Tanzania

3.1. The data

We illustrate our testing strategy using two datasets: one from Tanzania and one from India. The Tanzanian dataset comes from a village community named Nyakatoke in the Buboka Rural District of Tanzania, at the west of Lake Victoria. The dataset is a census

[18] The parameter vectors are specified as follows: we first generate a uniformly distributed variable $X_i \in (-1, 1]$ over 100 nodes and we set $\alpha + X'_{ij}\beta = k + 0.4X_j + \varepsilon_1$ with $\varepsilon_1 \sim N(0, 0.3)$. The constant term k is chosen such that the average percentage of links is roughly the same (20%) under all data generating processes.

[19] The R^2 of the regression of X_j on Z_j is 0.78.

covering all 119 households in the village and includes information on households' demographics, wealth and assets, income sources and income shocks, transfers and interpersonal relations. The Nyakatoke data have been the object of numerous articles (*e.g.* De Weerdt and Dercon, 2006; De Weerdt and Fafchamps, 2011; Vandenbossche and Demuynck 2012).

During the first survey round, each adult in the village was 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 labor?'.[20] Aggregated at the level of each household, the responses to this question constitute variables g^i_{ij} and g^j_{ji}. In other words, $g^i_{ij} = 1$ if an adult member of household i mentions an adult member of household j in their response to the above question.[21] In the dataset there are 119 households, which make $119(118) = 14042$ dyadic observations. The proportion of pairs for which g^i_{ij} or $g^j_{ji} = 1$ is 7%, and the proportion of discordant responses is very large. If we interpret all responses are capturing a link, the village forms a single giant component involving 117 of the 119 households. The network is sparse, with no evidence of quilt structure.

Given the cultural context, it is not obvious how to interpret villagers' responses to the link question. One possible interpretation is that responses represent the desire

[20]Note that we do not look at directed flows of favors: this survey question is intended to capture mutual assistance, and field work has indeed suggested that this is how respondents perceive it. The question was first piloted in the Philippines (Fafchamps and Lund, 2003) and subsequently adopted in the Tanzania survey, because respondents understand it and are willing to answer it. Other survey questions on directed flows were tried, for instance drawing a distinction between those the respondent would help and those the respondent would seek help from. But respondents were confused by the distinction which they perceived as non-existent, and complained they are asked the same question twice.

[21]34% of the mentioned partners live out of the village. They are omitted from the analysis since we have no information on the partner and hence we cannot apply our testing methodology.

to establish a link of mutual assistance that does not exist yet. This interpretation is particularly appealing when the responses are discordant, that is, when $g_{ij}^i \neq g_{ji}^j$. It is nevertheless possible that discordant responses as due to measurement error and that the data describe, albeit with some error, existing links between villagers; if this is the case the link formation process can be bilateral or unilateral.

Much of the economic literature on informal risk sharing in developing countries has assumed that households willingly enter in such arrangements (Kimball 1988; Coate and Ravallion, 1993). In our context, this approach implies bilateral link formation. In contrast, much of the anthropological literature has emphasized the difficulty for individuals to abstract themselves from the moral and social obligation to assist others in need (Scott, 1976; Platteau, 1996). This point has been made by a number of economists as well.[22] This line of reasoning implies unilateral link formation, possibly limited to households that are sufficiently close geographically and/or socially.

The covariates X_{ij} used in the regression analysis are illustrative of the type of variables included in an analysis of this kind. Two regressors capture the risk-sharing attractiveness of the potential partner j. The first regressor $w_j - w_i$ is the simple difference in total wealth between i and j.[23] The second regressor is the number of times that the members of household j are mentioned as risk-sharing partners by other Nyakatoke villagers: this variable, that we call $popularity_j$, is meant to proxy for

[22] Lucas and Stark (1985) and Azam and Gubert (2006) discuss social obligations in the context of remittance flows. Anderson and Baland (2002) provide evidence that individuals living in Kenyan slums put money in rotating savings and credit associations (ROSCAs) to avoid claims on their resources by spouse and relatives. Ambec (1998) and Banerjee and Mullainathan (2007) take these observations as starting point to model the saving behavior of poor households.

[23] Total wealth is computed as the sum of land and livestock assets. Data on land was collected in acres, but transformed in monetary equivalent using a conversion rate of 300000 Tanzanian shillings for 1 acre, which reflects the average local price in 2000.

various unobservable characteristics – *e.g.*, sociability, generosity, moral sense – that make j an attractive partner for many villagers.[24]

A second set of regressors seeks to control for homophyly, that is, the desire to link with similar or proximate households. The literature has shown that social ties depend to a large extent on social and spatial proximity, which reduces transaction costs and facilitate monitoring (Fafchamps and Gubert, 2007; De Weerdt and Fafchamps, 2011). To control for geographical proximity, we introduce a dummy that takes value one if i and j are neighbors, that is, live less than 400 meters apart. Blood ties are controlled for using a kinship dummy that takes value one if i and j – or members of their household – are strictly related (parents, children and siblings). We also include an education dummy taking value one if i and j have the same educational level.[25] As De Weerdt and Fafchamps (2011) show, informal transfers in Nyakatoke respond to health shocks. Since they pool labor resources, larger households should find it easier to deal with health shocks than smaller ones – and hence are less in need of forming mutual insurance links with other villagers (Binswanger and McIntire, 1987). Thus, we also control for the total number of adult members in households i and j respectively. Descriptive statistics are reported in Table 2.

[24]While computing $popularity_j$ we omit the reports from household i, to rule out spurious correlation. Note that in this context we are not interpreting $popularity_j$ causally.

[25]The educational level in Nyakatoke is rather low, with 23% of households where no member has completed primary education. The dummy takes value one if both households have or both households do not have at least one member who completed primary education, and zero otherwise.

Table 2: *Nyakatoke Survey – descriptive statistics*

dichotomous variable	definition			frequency
g^i_{ij}, g^j_{ij}	$g^i_{ij} = g^j_{ij} = 1$			280 (2%)
	$g^i_{ij} \neq g^j_{ij}$			700 (5%)
	$g^i_{ij} = g^j_{ij} = 0$			13062 (93%)
neighbors	distance ij <400 m			5550 (40%)
same family	ij have strict blood ties			218 (2%)
same education	ij have same education			9074 (65%)
continuous variable	mean	min	max	s.d.
$popularity_j$ (*)	0.52	0	2.30	0.45
$w_j - w_i$ (**)	0	-27.97	27.97	6.84
n. adult members of i	2.55	1	9	1.31
n. adult members of j	2.55	1	9	1.31

Notes: computed on the estimation sample of 14042 dyads. (*) 1 unit corresponds to 10 reports; it excludes the ij link. (**) 1 unit corresponds to 100000 Tanzanian Shillings.

3.2. Results

In Table 3 we report estimate of models (3), (6) and (9). Standard errors are corrected using formula (10). In column (1) we report the estimation results obtained when we assume that responses to the risk-sharing question capture desire to link, as explained in subsection (2.2). Estimates suggest that respondents prefer to link with popular households who live nearby and are related, while other regressors are not significant.

In column (2) we estimate the bilateral link formation model of equation (6). Coefficient estimates are similar to those reported in column (1) ($popularity_j$, neighbors and blood link dummies remain significant), and again are suggestive of homophyly. Additionally, household sizes of i and j appear significant with the expected sign: large households seek less links, and are seen as better partners.

In column (3) we present the results assuming that the data were generated by the unilateral link formation model (9). As explained in subsection (2.4), we transform household responses g_{ij}^i and g_{ij}^j into the equation-level dependent variables $h_{ij}^i \equiv 1 - g_{ij}^i$ and $h_{ij}^j \equiv 1 - g_{ij}^j$. To facilitate comparison with columns (1) and (2), we report estimated coefficients $\widehat{\beta}$ directly, which means inverting the sign of the coefficient estimates obtained from estimating (9) with partial observability bivariate probit. In terms of coefficient estimates, results are similar to those reported in column (2), except for i's household size which becomes (marginally) positive. Additionally, the coefficient of $w_j - w_i$ is small in magnitude but significantly negative, suggesting that under the hypothesis of unilateral link formation respondents tend not to link with much wealthier households.

Table 3: Nyakatoke survey – regression coefficients

	(1) desire to link	(2) bilateral	(3) unilateral
$w_j - w_i$	-0.002	0.004	-0.023***
	(0.009)	(0.010)	(0.007)
$popularity_j$	0.451***	0.152**	0.405***
	(0.102)	(0.065)	(0.065)
neighbors	0.500***	0.126***	0.463***
	(0.118)	(0.040)	(0.054)
same family	1.526***	0.602***	1.395***
	(0.199)	(0.099)	(0.120)
same education	0.005	-0.002	-0.003
	(0.116)	(0.012)	(0.047)
n. adult members of i	0.049	-0.055*	0.043*
	(0.059)	(0.029)	(0.026)
n. adult members of j	0.049	0.083***	0.043**
	(0.033)	(0.029)	(0.021)
constant	-2.576***	-0.249*	-2.777***
	(0.160)	(0.147)	(0.064)
arc tan(ρ)		-1.871***	-0.435**
		(0.497)	(0.219)

Note: dyadic standard errors in parenthesis.

We now turn to the main object of the paper, which is to compare the performance of the different models in accounting for the data. As explained in Section 2, we proceed by pairwise comparisons, adapting the non-nested Vuong test to the dyadic structure of the data. To compare two models k and m we calculate, for each observation ij, the scores under the two models and we regress the difference $l_{ij}^k - l_{ij}^m$ on a constant, correcting the standard errors using formula (10). The t-value of the constant is the Vuong test corrected for dyadic dependence. Since the distribution of the Vuong test is asymptotically normal, the relevant critical value for a 5% level of significance is 1.96. Note that the test works in two directions: if $t > 1.96$ model k is to be preferred to model m; if $t < -1.96$ model m is to be preferred to model k. For values of t between

-1.96 and 1.96 the test is inconclusive – both models fit the data equally well.

Table 4 reports the result of the pairwise comparisons between the desire-to-link, bilateral, and unilateral models. When the bilateral and unilateral models are compared to each other, the bilateral model is found superior. But the results unambiguously shows that the desire-to-link model fits the data best.[26]

Table 4: Nyakatoke Survey – Vuong tests

Model k	Model m	Vuong test	Best fit
bilateral	unilateral	2.44**	bilateral
desire to link	bilateral	1.98**	desire to link
desire to link	unilateral	3.22***	desire to link

In the Appendix A we also extend out testing strategy to incorporate an element of self-censoring. It is indeed possible that respondents may refrain from reporting a desire to form certain links if they anticipate rejection (Hitsch *et al.*, 2005; Belot and Francesconi, 2007; Fisman *et al.*, 2008). To capture this idea, we incorporate an additional parameter representing self-censoring into the model. Results are reassuring: although we find some evidence of self-censoring in survey responses, the findings of Table 4 remain valid. Overall, results suggest that desire to link, tempered somewhat by self-censoring, is the most appropriate model to interpret the links reported by Nyakatoke villagers.

These findings are important because they suggest that more caution should be taken when interpreting self-reported risk-sharing links. From the work of De Weerdt and Dercon (2006) on the same data, we know that cash and in-kind transfers are

[26] To test the robustness of our findings, we re-estimated all models using different regressors sets. Vuong test results are generally consistent across specifications but are less conclusive when regressors have little predictive power. This is a common feature of non-nested tests.

much more likely between households that listed each other as source of help. If listed households are interpreted as existing links, this implies that pre-existing social networks shape future transfers – with possible implications regarding efficiency (mutual gains from risk sharing are limited by the extent of the pre-existing network) and equity (some households have fewer links or links to less helpful people, and consequently are less well insured). If survey responses are interpreted as desire to link, however, obstacles to efficiency and equity result from individual linking preferences.[27]

4. Communication among Indian farmers

4.1. The data

The second illustration of our methodology uses data from Indian farmers. In rural economies farmers are often faced with incomplete information about the use of a technology, such as chemical fertilizer and high-yielding seeds varieties. In such circumstances, learning from peers is a major determinant of adoption (Munshi, 2004; Bandiera and Rasul, 2006; Conley and Udry, 2010). The data in use were collected from May to July 2009 in 100 villages of the state of Maharashtra, as a part of a project funded by the International Food Policy Research Institute and the World Bank. Within each of the sampled villages, 10 farmers growing a diverse range of crops were randomly selected, and detailed information was collected on their cropping practices, land holdings, trading activities, information utilization, and network-based

[27] The distinction is subtle, but relevant: if risk sharing is imperfect because social links are limited, the policy implication is to create more links; if risk sharing is imperfect because households are unwilling to form certain links, the policy implication is to shape linking preferences, *e.g.*, social integration – which may be much harder.

learning.[28]

Each respondent was asked about each other farmer in his village sample "How often do you discuss agricultural issues with this person or members of his household?' Our dependent variable $g_{ij}^i = 1$ if farmer i states he discusses agricultural issues with farmer j at least once a week. Communication among surveyed farmers appears quite high: on average farmers declare to discuss agricultural issues at least once a week with 67% of their village sample.

Survey respondents were also asked to assess their own knowledge about agriculture ('Would you describe yourself as a knowledgeable, well informed farmer?') and to evaluate the knowledge of the partner ('How knowledgeable about farming is this person?') in a scale from 1 to 5.[29] We use this piece of information to build the first proxy for the attractiveness of the partner in terms of agricultural knowledge: the difference $knowdledge_j - knowdledge_i$ measures how much more knowledgeable j is relative to i (*i.e.,* it takes value 1 if i consider j one point more knowledgeable than he is, and value -1 if i considers himself 1 point more knowledgeable than j).[30] The

[28]The survey was originally designed to have 9000 dyads in the data: in each of the 100 sampled village, 10 farmers were randomly selected, and each farmer was interviewed about the 9 other selected farmers. However, because the selected names were pre-printed on the questionnaire, 27% of listed network partners could not be matched with a survey respondent, either because the names in the questionnaire did not uniquely identify villagers, or because of the sample attrition (*i.e.,* if a pre-selected farmer could not be interviewed, another household in the village was randomly selected). After having dropped all dyads where at least one respondent was not in the original sample, we are left with 5080 dyadic observations.

[29]Where 1 is defined as 'well below average', 2 as 'slightly below average', 3 as 'average', 4 as 'slightly above average', and 5 as 'well above average'.

[30]Note that i's opinion about j's farming knowledge is only poorly correlated either with j's assessment of his own farming knowledge, or with objective proxies for j's farming knowledge (*e.g.,* various measures of experience, agricultural yields and good practices). This may be because what i regards as knowledgeable about j is not the absolute level of knowledge that j has, but rather the knowledge that j has and i does not have. Since this is the most relevant measure of desire to link, it is the one we use here.

second proxy for partner's attractiveness is a dummy $responsibility_j$ which equals one if farmer j occupies a position of responsibility within the village (*e.g.* elected official, village leader, or board member).

Our list of covariates also includes: the sum and absolute difference in land owned (in acres), the sum and absolute difference in labor intensity (defined as total man days of work per acre of land owned), the age of respondent, the absolute difference in age between the partners, and a same gender dummy. Descriptive statistics are reported in Table 5.

Table 5: Maharashtra data – descriptive statistics

dichotomous variable	definition			frequency
g^i_{ij}, g^j_{ij}	$g^i_{ij} = g^j_{ij} = 1$			2468 (48%)
	$g^i_{ij} \neq g^j_{ij}$			1872 (37%)
	$g^i_{ij} = g^j_{ij} = 0$			740 (15%)
same gender	both females or both males			4880 (96%)
$responsibility_j$	j has a responsibility position			714 (14%)
continuous variable	mean	min	max	sd
$land_i + land_j$	21.17	2.5	155	16.07
$\lvert land_i - land_j \rvert$	8.27	0	76	9.87
$intensity_i + intensity_j$	147.54	0	6133	286.27
$\lvert intensity_i - intensity_j \rvert$	60.30	0	5867	263.95
age_i	51.04	21	90	12.87
$\lvert age_i - age_j \rvert$	13.90	0	66	10.66
$knowdledge_j - knowdledge_i$	0.10	-4	3	0.83

Note: computed on the estimation sample of 5080 dyads.

4.2. Results

In Table 6 we estimate models (3), (6) and (9), clustering the standard errors at the village level. In column (1) we report the estimation results obtained when we assume desire to link. In columns (2) and (3) we estimate the bilateral and unilateral link formation models respectively. As before, estimated coefficients of column (3) are presented with the corrected sign to facilitate comparison.

Table 6: Maharashtra data – regression coefficients

	(1) desire to link	(2) bilateral	(3) unilateral
$land_i + land_j$	-0.0022	-0.0015	-0.0008
	(0.003)	(0.002)	(0.001)
$\|land_i - land_j\|$	0.0021	0.0018	0.0009
	(0.004)	(0.003)	(0.002)
age_i	-0.0035	-0.0073**	-0.0052***
	(0.003)	(0.003)	(0.002)
$\|age_i - age_j\|$	-0.0076***	-0.0055***	-0.0039***
	(0.002)	(0.002)	(0.001)
$knowdledge_j - knowdledge_i$	0.1536***	0.1811***	0.1382***
	(0.037)	(0.041)	(0.030)
same gender	0.3508**	0.2602**	0.1879*
	(0.169)	(0.130)	(0.102)
$intensity_i + intensity_j$	0.0002	0.0001	0.0001
	(0.000)	(0.000)	(0.000)
$\|intensity_i - intensity_j\|$	-0.0003	-0.0002	-0.0001
	(0.000)	(0.000)	(0.000)
$responsibility_j$	0.0553	0.1867**	0.1243**
	(0.068)	(0.089)	(0.048)
constant	0.3952	1.1340***	-0.3158**
	(0.253)	(0.227)	(0.143)
$arc\ tan(\rho)$		-0.3115*	-4.0625***
		(0.177)	(0.978)

Note: standard errors clustered at the village level are in parenthesis.

Results are similar across the three columns. Large age differences are associated with less communication between farmers, and younger farmers seem more keen to communicate – which is consistent with their limited experience of farming. The difference $knowdledge_j - knowdledge_i$, which proxies for the partner's attractiveness in terms of agricultural knowledge, is positive and significant: *ceteris paribus* respondents are more willing to engage in a conversation if they believe that the partner is more knowledgeable than they are. The dummy for same gender is strongly significant and positive across the three columns, suggesting homophyly by gender. The sum and absolute difference in land owned and in labor intensity do not seem to play a role in explaining link formation. Partners who hold a position of responsibility in the community appear more attractive in the bilateral and unilateral link formation models, but the variable is not significant in the desire-to-link model.

Table 7 reports the results of pairwise comparisons between models. We find that the unilateral model dominates the bilateral model, which in turn unambiguously dominates the desire-to-link model.

Table 7: Maharashtra data – Vuong tests

Model k	Model m	Vuong test	Best fit
bilateral	unilateral	-1.71*	unilateral
desire to link	bilateral	-1.98**	bilateral
desire to link	unilateral	-2.14**	unilateral

We also estimate the hybrid model discussed in Appendix A. We find that self-censoring does not affect the rankings in Table 7. To summarize, the results suggest that survey responses represent existing links, confirming that the methodology can distinguish between different data generating processes. Results further suggest that links among Maharashtra farmers are unilateral, *i.e.*, an exchange of agricultural information between two farmers can take place even if only one of the two expects to benefit from the conversation. This reassuringly suggests that information about agricultural technology is likely to flow from more knowledgeable to less knowledgeable farmers, and it opens interesting opportunities from the policy intervention perspective. Unilateral link formation does not imply that any link can be formed, however: as shown in our analysis, respondents are much more likely to list people with similar age and same gender as those with whom to communicate. Obstacles to information sharing subsist even though link formation is unilateral.

5. Conclusions

We have proposed a methodology that enables a researcher to investigate the nature of self-reported links. Its purpose is to help draw more accurate inference about network effects. We illustrate our methodology with two separate observational datasets.

The first illustration focuses on informal risk sharing in a Tanzanian village, where respondents were asked to enumerate all their risk-sharing partners. The literature is uncertain as to whether risk-sharing links should be seen as implicit contracts grounded in mutual self-interest, or whether social norms impose an element of moral or social pressure making it difficult for households to refuse helping (and being helped by) others. We find that the desire-to-link model best fits the data: respondents list households with whom it is in their objective interest to link. This finding takes all its meaning when compared to the work of Comola and Fafchamps (2010) who analyze cash and in-kind transfers in the same community. They conclude that the pattern of gifts and loans is more consistent with unilateral link formation, suggesting that villagers find it difficult to refuse assisting others in need. Given this finding, it is not surprising if, when asked to list those to whom they would turn in times of need, respondents simply list those who can help them – irrespective of whether a link already exists or not.

The second illustration uses survey data from farmers in the Indian region of Maharashtra. We find that survey responses are best interpreted as existing links and that the unilateral link formation model fits the data better than the bilateral one. This suggests that, in the communities studied, less experienced farmers can secure information about farming practices from more knowledgeable neighbors even though the latter have little to gain objectively from agricultural information exchange. This opens interesting opportunities from a policy intervention perspective, *e.g.*, targeting extension services to more knowledgeable farmers who are better able to absorb and subsequently disseminate information about new technology.

Two final caveats are in order. First, to use our methodology to confirm whether survey responses can safely be interpreted as existing links, the researcher must have reports from the two end-points of the same link. This is because identification is achieved from the pattern of discordance between the two responses. Secondly, to distinguish between unilateral and bilateral link formation, the researcher must be able to proxy for the interest of individual i in forming a link with j in a way that is distinct from the interest of j in forming a link with i. This does not, however, preclude the inclusion of variables to proxy for homophyly, transaction costs, or limits to the reach of social norms. The formation of favor exchange links is typically embedded in a pre-existing social context that may limit the set of potential links, and this is also what we find in our data.

SUBMITTED ON: January 5th, 2012

CONDITIONALLY ACCEPTED ON: May 4th, 2013

References

[1] Altonji, J.G., Hayashi, F. and Kotlikoff, L.J. (1992). 'Is the extended family altruistically linked? direct tests using micro data', *American Economic Review*, vol. 82(5), pp. 1177-98.

[2] Ambec, S. (1998). 'A theory of African's income-sharing norm with enlarged families', unpublished.

[3] Anderson, S. and Baland, J.M. (2002). 'The economics of Roscas and intrahousehold resource allocation', *Quarterly Journal of Economics*, vol. 117(3), pp. 963-95.

[4] Arcand, J.L. and Fafchamps, M. (2012). 'Matching in community-based organizations', *Journal of Development Economics*, vol. 98(2), pp. 203-19.

[5] Azam, J.P. and Gubert, F. (2006). 'Migrants' remittances and the household in Africa: a review of evidence', *Journal of African Economies*, vol. 15(2), pp. 426-62.

[6] Ballester, C., Calvó-Armengol, A. and Zenou, Y. (2006). 'Who's who in networks. Wanted: the key player', *Econometrica*, vol. 74, pp. 1403-17.

[7] Bandiera, O. and Rasul, I. (2006). 'Social networks and technology adoption in northern Mozambique', *ECONOMIC JOURNAL*, vol. 116, pp. 862-902.

[8] Banerjee, A., Chandrasekhar, A., Duflo E. and Jackson, M.O. (2012). 'The diffusion of microfinance', NBER Working Paper No. 17743.

[9] Banerjee, A. and Mullainathan, S. (2007). 'Climbing out of poverty: long term decisions under income stress', unpublished.

[10] Barr, A., Dekker, M. and Fafchamps, M. (2012). 'Who shares risk with whom under different enforcement mechanisms?', *Economic Development and Cultural Change*, vol. 60(4), pp. 677-706.

[11] Barr, A. and Stein, M. (2008). 'Status and egalitarianism in traditional communities: an analysis of funeral attendance in six Zimbabwean villages', CSAE Working paper No. 2008-26.

[12] Belot, M. and Francesconi, M. (2007). 'Can anyone be "The" one? field evidence on dating behavior', ISER Working Paper Series No. 2007-17.

[13] Bester, C.A., Conley, T.G. and Hansen, C.B. (2011). 'Inference with dependent data using cluster covariance estimators', *Journal of Econometrics*, vol. 165(2), pp. 137-51.

[14] Binswanger, H.P. and McIntire, J. (1987). 'Behavioral and material determinants of production relations in land-abundant tropical agriculture', *Economic Development and Cultural Change*, vol. 36(1), pp. 73-99.

[15] Bloch, F., Genicot, G. and Ray, D. (2008). 'Informal insurance in social networks', *Journal of Economic Theory*, vol. 143(1), pp. 36-58.

[16] Boucher, V. (2012). 'Structural homophily', Working Papers HAL-00720825.

[17] Calvó-Armengol, A., Patacchini, E. and Zenou, Y. (2009). 'Peer effects and social networks in education', *Review of Economic Studies*, vol. 76, pp. 1239-67.

[18] Chandrasekhar, A. and Jackson M.O. (2012). 'Tractable and consistent random graph models', unpublished.

[19] Christakis, N.A., Fowler, J.H., Imbens, G.W., and Kalyanaraman, K. (2010). ‘An empirical model for strategic network formation’, NBER Working Paper No. 16039.

[20] Coate, S. and Ravallion, M. (1993). ‘Reciprocity without commitment: characterization and performance of informal insurance arrangements’, *Journal of Development Economics*, vol. 40, pp. 1-24.

[21] Comola, M. (2012). ‘Estimating local network externalities’, SSRN WP No. 946093.

[22] Comola, M. and Fafchamps, M. (2010). ‘Are gifts and loans between households voluntary?’, Paris School of Economics Working Paper No. 2010-19.

[23] Conley, T. (1999). ‘GMM estimation with cross-sectional dependence’, *Journal of Econometrics*, vol. 92(1), pp. 1-45.

[24] Conley, T. and Udry, C. (2010). ‘Learning about a new technology: pineapple in Ghana,’ *American Economic Review,* vol. 100(1), pp. 35-69.

[25] De Weerdt, J. (2004). ‘Risk sharing and endogenous network formation’, in (S. Dercon, ed.), *Insurance Against Poverty*, pp: 197-216, Oxford: Oxford University Press.

[26] De Weerdt, J. and Dercon, S. (2006). ‘Risk-sharing networks and insurance against illness’, *Journal of Development Economics*, vol. 81(2), pp. 337-56.

[27] De Weerdt, J. and Fafchamps, M. (2011). 'Social identity and the formation of health insurance networks', *Journal of Development Studies*, vol. 47(8), pp. 1152-77.

[28] Fafchamps, M. and Gubert, F. (2007). 'The formation of risk-sharing networks', *Journal of Development Economics*, vol. 83(2), pp. 326-50.

[29] Fafchamps, M. and Lund, S. (2003). 'Risk-sharing networks in rural Philippines', *Journal of Development Economics*, vol. 71, pp. 261-87.

[30] Fisman, R., Iyengar, S., Kamenica, E. and Simonson, I. (2008). 'Racial preferences in dating', *Review of Economic Studies*, vol. 75, pp. 117-32.

[31] Foster, A.D. and Rosenzweig, M.R. (1995). 'Learning by doing and learning from others: human capital and technical change in agriculture', *Journal of Political Economy*, vol. 103(6), pp. 1176-1209.

[32] Fox, J.T. (2008). 'Estimating matching games with transfers', NBER Working Paper No. 14382.

[33] Goyal, S. (2007). *Connections: An Introduction to the Economics of Networks,* Princeton and Oxford: Princeton University Press.

[34] Granovetter, M.S. (1995). *Getting a Job: A Study of Contacts and Carreers*, Chicago: University of Chicago Press.

[35] Hitsch, G.J., Hortacsu, A. and Ariely, D. (2005). 'What makes you click: an empirical analysis of online dating', Society for Economic Dynamics, 2005 Meeting Papers 01/2005.

[36] Jackson, M.O. (2008). *Social and Economic Networks*, Princeton: Princeton University Press.

[37] Jackson, M.O. (2011). 'An overview of social networks and economic applications', in (J. Benhabib, A. Bisin, and M.O. Jackson, eds.) *Handbook of Social Economics*, pp. 511-586. Amsterdam: North Holland Press.

[38] Jackson, M.O. and Wolinsky, A. (1996). 'A strategic model of social and economic networks', *Journal of Economic Theory*, vol. 71(1), pp. 44-74.

[39] Jackson, M.O., T. Rodriguez-Barraquer, and X. Tan (2012). 'Social capital and social quilts: network patterns of favor exchange', *American Economic Review*, vol. 102(5), pp. 1857-97.

[40] Kimball, M.S. (1988). 'Farmers' cooperatives as behavior toward risk', *American Economic Review*, vol. 78(1), pp. 224-232.

[41] Kocherlakota, N.R. (1996). 'Implications of efficient risk-sharing without commitment', *Review of Economic Studies*, vol. 63(4), pp. 595-609.

[42] Marsden, P.V. (1990). 'Network data and measurement', *Annual Review of Sociology*, vol. 16, pp. 435-63.

[43] Mele, A. (2011). 'A Structural model of segregation in social networks', unpublished.

[44] Munshi, K. (2004). 'Social learning in a heterogeneous population: technology diffusion in the Indian green revolution', *Journal of Development Economics*, vol. 73(1), pp. 185-213.

[45] Liu, X., Patacchini, E., Zenou, Y. and Lee, L. (2011). 'Criminal networks: who is the key player?', Stockholm University's Research Papers in Economics No. 2011/7.

[46] Lucas, R.E. and Stark, O. (1985). 'Motivations to remit: evidence from Botswana', *Journal of Political Economy*, vol. 93(5), pp. 901-918.

[47] Platteau, J.P. (1996). 'Traditional sharing norms as an obstacle to economic growth in tribal societies', Technical report, CRED, Facultés Universitaires Note-Dame de la Paix, Namur, Belgium. Cahiers No. 173.

[48] Poirier, D. (1980). 'Partial observability in bivariate probit models,' *Journal of Econometrics*, vol. 12, pp. 209-17.

[49] Scott, J.C. (1976). *The Moral Economy of Peasant: Rebellion and Subsistence in South-East Asia*, New Haven: Yale University Press.

[50] Sheng, S. (2012). 'Identification and estimation of network formation games', unpublished.

[51] Topa, G. (2001). 'Social interactions, local spillovers, and unemployment', *Review of Economic Studies*, vol. 68(2), pp. 261-95.

[52] Townsend, R. M. (1994). 'Risk and insurance in village India,' *Econometrica*, vol. 62(3), pp. 539-591.

[53] Vandenbossche, J. and Demuynck, T. (2012). 'Network formation with heterogeneous agents and absolute friction', *Computational Economics*, vol. 42(1), pp. 23-45.

[54] Vuong, Q.H. (1989). 'Likelihood ratio tests for model selection and non-nested hypotheses', *Econometrica*, vol. 57(2), pp. 307-33.

6. Appendix A: the hybrid model

Self-censoring of reported desire to link has long plagued the study of link formation (Hitsch, Hortacsu and Ariely, 2005; Belot and Francesconi, 2007; Fisman *et al.*, 2008).[31] In our context, a respondent i may refrain from reporting a desire to link to j if he anticipates rejection, *i.e.*, if j is unlikely to desire a link with i. The more respondent i internalizes a possible rejection by j, the more i's reported desire to link resembles a mutually agreed link. To capture this idea, an additional parameter δ can be introduced in (6) to represent the extent to which each respondent internalizes the other's desire to link in his survey response. The corresponding data generating process is a hybrid model written as:

$$\begin{aligned} \Pr(g^i_{ij} &= 1) = \Pr\left(\varepsilon_{ij} \leq \alpha_1 + X'_{ij}\beta \text{ and } \varepsilon_{ji} \leq \alpha_2 + \delta X'_{ji}\beta\right) \\ \Pr(g^j_{ji} &= 1) = \Pr\left(\varepsilon_{ji} \leq \alpha_1 + X'_{ji}\beta \text{ and } \varepsilon_{ij} \leq \alpha_2 + \delta X'_{ij}\beta\right) \end{aligned} \tag{11}$$

In this setting g^i_{ij} incorporates not only i's desire to link with j, but also i's expectation of whether the link would be accepted by j. If $\delta = 1$, this boils down to the bilateral link formation model (5). In contrast, if $\delta = 0$, the second term in the right-hand becomes a constant and the regression boils down to the desire-to-link model

[31] In their study of internet dating sites, Hitsch, Hortacsu and Ariely (2005) note that the emails participants send to each other to initiate interaction may not reflect their true desire to link if they refrain from making openings they know will be rejected. Belot and Francesconi (2007) make similar observations in their study of speed dating. Self-censoring has also been discussed in the context of matching models in which individuals can only rank a subset of their possible choices (*e.g.*, the University Centralised Application System in the UK where students can only list 5 universities of their choice). In such contexts, it is optimal for low ranked individuals not to 'waste' limited slots on options they are unlikely to get.

(3).[32] If $\delta > 1$, this means that i puts more weight on j's desire to link with him than to his own desire to link with j. To estimate (11), we perform a grid search on δ with values ranging from 0.05 to 20 and, for each, we maximize the corresponding likelihood function. The point estimate of δ is the value that yields the highest likelihood value. Identification is achieved via a functional form assumption, which some may regard as unconvincing. Stronger identification would require a regressor that predicts censoring but does not affect the utility players derive from the link. Future work is needed in this area.

We estimate the hybrid model described above using both datasets at hand, starting from the risk sharing data from the Tanzania village. The best likelihood value is obtained for $\delta = 0.2$, which suggests that survey responses g^i_{ij} capture mostly desire to link, but with some self-censoring when j is unlikely to want to link. Coefficients estimates are similar to those of Table 3 and therefore are not reported to save space. Table A1 is the continuation of Table 4, where we add the hybrid model to the comparison. The Vuong test shows that the hybrid model proves significantly superior to the standard unilateral and bilateral models – it also seems to perform better than the standard desire-to-link model, but the difference is not significant. This is perhaps not surprising, it incorporates a strong desire-to-link component. Overall, results reconfirm that desire to link, tempered somewhat by self-censoring, is the most appropriate model to interpret the self-reported risk-sharing links by Nyakatoke villagers: respon-

[32]Note that we allow intercepts α_1 and α_2 to differ. To see why this is necessary, imagine we did not and the correct model is (3). In this case we want estimation of (11) to yield $\delta = 0$. If we force $\alpha_2 = \alpha_1 = \alpha$, however, model (11) implicitly requires that *at the same time* $\varepsilon_{ji} \leq \alpha$ (from the second term in the first equation of 11) and $\varepsilon_{ji} \leq \alpha + X'_{ji}\beta$ (from the first term in the second equation of 11). Since this is not possible unless $\beta = 0$, imposing $\alpha_2 = \alpha_1$ biases β and thus δ. By the same reasoning, when $\delta = 1$ estimation automatically yields $\alpha_1 = \alpha_2$ since the two inequalities now coincide.

dents take partially into account whether others are willing to link with them before listing them.

Table A1: Nyakatoke Survey – hybrid model

Model k	Model m	Vuong test	Best fit
desire to link	hybrid model	-1.61	undefined
bilateral	hybrid model	-3.17***	hybrid model
unilateral	hybrid model	-3.82***	hybrid model

As for the Maharashtra farmers data, the optimal δ found though the grid search is 1.1, which basically boils down to bilateral link formation – except for a slight over-weighting of the partner's desire to link. Again, coefficients estimates are not reported here to save space. In Table A.2 we continue Table 7 by comparing the first three models to the hybrid model. The Vuong test shows that the hybrid model fits the data as well as bilateral (which is not surprising since, with $\delta = 1.1$, the two models are nearly identical) and other rankings are preserved. These results suggest that information sharing among Maharashtra farmers is most likely to be initiated unilaterally, that is, even if the selected partner does not expect to gain from the conversation.

Table A.2: Maharashtra data – hybrid model

Model k	Model m	Vuong test	Best fit
desire to link	hybrid model	-2.05**	hybrid model
bilateral	hybrid model	-0.40	undefined
unilateral	hybrid model	1.69*	unilateral

Online Appendix B: Simulations with clustering

This appendix describes the simulation exercise we run to investigate how the non-nested test performs in presence of an element of inter-relatedness in linking decisions (clustering). Observational evidence and theoretical contributions suggest that clustering is a pervasive feature of social networks. Our exercise is motivated in particular by Jackson, Rodriguez-Barraquer and Tan (2012), who propose a game theoretic foundation for informal favor exchange motivated by threats of ostracism or loss of multiple relationships. In their setting link formation is bilateral, and in equilibrium all links must be supported (*i.e.* any two individuals exchanging favors must have a common partner). Although there is little evidence of the quilt network structure they describe in the two datasets we investigate, we wish to verify that our methodology would not be affected by the presence of such pattern.

For this mis-specification test, we draw 250 bilateral networks with 100 nodes (*i.e.* 9900 dyads) each where all links are supported by triads. Put differently, a link between i and j can only exist if i and j have at least one partner in common (*i.e.,* there is a path of length two between i and j). In order to generate each network we proceed in two steps: first, we generate a network under bilateral link formation, and then we drop all links not embedded in a triad. The bilateral data generation process takes the form

$$U_i(g_{+ij}) - U_i(g_{-ij}) = -0.4 + 0.8X_{1j} - |X_{2i} - X_{2j}| - \varepsilon_{ij}$$

where X_1 and X_2 are uniformly distributed and $\varepsilon \sim N(0, 0.3)$. Under these parameter values the proportion of non-supported links is important (18% of all bilateral links) and the final proportion of linked dyads (6.5% of all dyads) is comparable to the observational data of Section 3. We find that the test is able to correctly select bilateral link formation in all 250 instances. This reconfirm that our testing strategy is not invalidated by the presence of triadic clustering of the type hypothesized by Jackson, Rodriguez-Barraquer and Tan (2012).