SAGA Working Paper
April 2007

Understanding the Formation of Social Networks

Paulo Santos
Cornell University

Christopher B. Barrett
Cornell University

Strategies and Analysis for Growth and Access (SAGA) is a project of Cornell and Clark Atlanta Universities, funded by cooperative agreement #HFM-A-00-01-00132-00 with the United States Agency for International Development.

Understanding the formation of social networks*

Paulo Santos† Cornell University

Christopher B. Barrett‡ Cornell University

April 2007

Abstract

This paper reviews the growing literature that uses social networks as a method to analyze social context, paying special attention to how methods of sampling data on relationships affects inference with respect to the formation of social networks. We use original data from southern Ethiopia to demonstrate a new approach to collecting data on relationships, that starts with a random sample of individuals and then randomly samples from the prospective relationships among sample respondents. We show that this method yields estimates of the structure of social relations that are statistically indistinguishable from those generated using more expensive and time-consuming methods that trace respondents' social networks. We then use Monte Carlo

*This work has been made possible by support from the Social Science Research Council's Program in Applied Economics on Risk and Development (through a grant from the John D. and Catherine T. MacArthur Foundation), The Pew Charitable Trusts (through the Christian Scholars Program of the University of Notre Dame), the Fundação para a Ciência e Tecnologia (Portugal), the Graduate School of Cornell University and the United States Agency for International Development (USAID), through grants LAG-A-00-96-90016-00 to the BASIS CRSP, DAN-1328-G-00-0046-00 and PCE-G-98-00036-00 to the Pastoral Risk Management (PARIMA) project of the Global Livestock CRSP, and the Strategies and Analyses for Growth and Access (SAGA) cooperative agreement, number HFM-A-00-01-00132-00. Thanks are due to the International Livestock Research Institute for its hospitality and support and to Action for Development (Yabello, Ethiopia) for logistical support. We thank Getachew Gebru and our field assistants, Ahmed Ibrahim and Mohammed Ibrahim, for their invaluable assistance in data collection, and Larry Blume, Michael Carter, Marcel Fafchamps, Gueorgi Kossinets, Annemie Maertens, Jacqueline Vanderpuye-Orgle and seminar audiences at Cornell University and NEUDC 2006 (at Cornell University) for helpful comments. The views expressed here and any remaining errors are the authors' and do not represent any official agency.

†PhD Candidate, Department of Applied Economics. Email: ps253@cornell.edu.

‡Professor, Department of Applied Economics.

simulation to test the value of this approach and show that introducing this second level of sampling improves the accuracy of the inference on the determinants of network formation.

1 Introduction

A large and heterogeneous literature under the general label of social capital attempts to quantify the value of social embeddedness in terms of welfare improvements for households and individuals.[1] The concept of a social network plays a prominent motivational role, in that it is through the set of interpersonal links between individuals that the net benefits of social interaction are assumed to flow. In the words of Robert Putnam, an influential author in this literature, "My definition is: social capital is networks".[2]

This conceptual emphasis has not been matched by the use of social networks as a method to explore the effects of social context. Social capital has often been measured through the quantification of the density of membership in voluntary associations (sometimes referred to as "Putnam's instrument")[3] while the related literature on social interactions has largely followed a similar path, using easily available information on community or group membership (ethnicity, gender, geographic neighborhood,etc.) to proxy for social networks. Although this has moved the research on the importance of social context from "being a specialty for network sociologists" (Paldam, 2000, pp.636-7) into what Durlauf (2002, p.459) calls "one of the most striking developments in social science over the last decade", the blurring of the distinction did not help solving the inferential problems on the analysis of social interactions initially pointed out by (Manski, 1993). [4].

It was the recognition of these problems and the need to have data on concrete interactions to overcome them (Manski, 2000) that led to the development, within economics, of a much smaller literature where social networks is not only a metaphor but also a method to characterize social context. The focus of this paper is on the development economics literature that aims at understanding the process underlying network formation, either as a question in itself or as a first step towards the quantification of the

[1] The literature on social capital was recently reviewed by Durlauf and Fafchamps (2004).

[2] Paldam (2000, p. 651, footnote 15).

[3] See, for example, Narayan and Pritchett (1999) for an early use of this type of variable in development economics.

[4] See also Brock and Durlauf (2001) and Moffitt (2001). Both Soetevent (2006) and Blume and Durlauf (2005) present recent reviews of this literature.

instrumental value of social connections.

Social networks are a set of individuals and the relationships among them. This joint focus is the source of differences from data collection strategies centered on the characteristics of individuals alone.[5] The relatively small literature that has collected both types of data is, nevertheless, diverse. Development economists have used a variety of sample designs, both for respondents (from census to random sample) and for relationships (from a complete enumeration to the selection of a pre-determined number of relations, from real to potential behavior). As interest in the empirical analysis of social networks grows and more researchers contemplate the possibility of collecting such data, it is important to understand the implications of these methodological choices.[6] That is the purpose of the next section.

Ultimately, however, we want to probe the validity of one new approach that we introduce in Section 3 and label as *random matching*: individuals who are part of a random sample are randomly matched with other individuals from the same sample and asked about their willingness to establish a link with the random match, hence both individuals and relationships are randomly sampled. We do that in two steps. In section 4 we discuss whether the elicitation of the willingness to establish a relation allows us to understand the process underlying the formation of individuals' actual networks. We use data on the social networks of a random sample of individuals collected in two different ways – through direct elicitation and through random matching – and show that they yield results that are statistically indistinguishable. In Section 5 we demonstrate the importance of sampling relationships. Using Monte Carlo simulation, we compare the accuracy of the inference with respect to the determinants of network formation when data on relationships are collected in two different ways: random matching and the more frequent approach of relying on the set of links from a random

[5]This focus implies also that we consider only those studies where the characteristics of relationships were elicited. We leave outside of this analysis studies such as Bandiera and Rasul (2006) or Behrman, Kohler, and Watkins (2002), where the information on networks is limited to the number of contacts of each respondent.

[6]One strategy that seems not to have been used so far in development economics is "snowball" sampling (Goodman, 1961) where, starting with a set of initial respondents (seeds), one increases the sample by including those individuals named by previous respondents. In this case the sampling of relationships and individuals (after the initial ones) is done simultaneously. Although well-suited for the sampling of "hidden populations", the respondents entering the sample after the seeds are not randomly selected which complicates inference about the population. See Heckathorn (2002) for a discussion of the conditions under which this problem can be solved and Heckathorn and Jeffri (2002) for an application to the analysis of jazz musician communities.

sample of individuals as an accurate image of individuals' networks, which we label as *matches within sample*. Our results show that, for different models of network formation, the random matching approach is, in general, more accurate than using all matches within sample. Section 6 concludes the paper.

2 A review of current approaches

The analysis of networks requires data on both individuals *and* relationships. It is useful to review how the sampling of both units can and has been done.[7] As with every other survey, individuals are the source of information and the existing literature employs two strategies to identify them: a census of all individuals (as in DeWeerdt (2004), Dekker (2004) and, in one village, Goldstein and Udry (1999)) or, more commonly, a random sample of individuals from the population of interest. These lead to different network designs, commonly referred as global versus local network designs, respectively.[8] The pros and cons of each strategy are relatively obvious. Random samples are less expensive but they lead to a loss of information on the network structure as the information generated is essentially limited to dyads, leaving potentially interesting questions outside the range of possible analysis.[9]

Having decided how to sample individuals, the second level of sampling is done through the construction of a "name generator", a question that is used to elicit and identify relationships. If "[...] a network is defined by the links as much as the nodes" (Morris, 2004, p.10), this is a step as important as the selection of the individual respondents although perhaps less visible: "it happens in the questionnaire" (Morris, 2004, p.10). Name generators include two parts the relation/behavior and a rule defining how many relations the researcher identifies.

As for the relationships among individuals, most of the studies by development economists look at potential relations, that is, those elicited through

[7] Much of the systematization that follows borrows from the clear exposition in Morris (2004). Several illustrations of the questions that we deal with in this paper can also be found there, but focusing specifically on the use of social networks to understand the epidemiology of HIV/AIDS.

[8] Global and local networks are also known, in the social networks literature, as sociometric and egocentric networks, respectively.

[9] This also means that much of the work developed within the field of social network analysis, directed to the analysis of complete networks (see Wasserman and Faust (1994) for an extensive treatment of such methods) cannot be directly applied to most of the data used by economists.

questions of the type "Who could you rely on to ...?" (DeWeerdt, 2004, Fafchamps and Gubert, 2007, Santos and Barrett, 2006b), while others focused on real relations through questions such as "From whom did you receive gifts?" (Dekker, 2004, Krishnan and Sciubba, 2005, Conley and Udry, 2005, Udry and Conley, 2005).

When looking at the motive for establishing the link, most studies focused on insurance, the exceptions being the analysis of information networks by Conley and Udry (2005) and Santos and Barrett (2006a), and the analysis of the interpersonal relations through which information, credit, labor and land are transacted in Udry and Conley (2005), all building on the data collected and described by Goldstein and Udry (1999). Finally, concerning the "stopping rule", some studies have asked for all the relationships of the respondents (e.g. DeWeerdt, 2004, Goldstein and Udry, 1999) while others established a maximum number of links (e.g. Fafchamps and Gubert, 2007). This methodological diversity, which reflects both the relative novelty of the approach and the diversity of substantive questions for which such data was collected, is summarized in Table 1.

Several points arise. The first, and most obvious, is the extent of missing information, which is an issue regardless of whether we have a census or a random sample of individuals. For example, DeWeerdt (2004) reports that his analysis is limited to approximately two-thirds of the links identified by his respondents, as the remaining 1/3 were formed with individuals outside his census unit. Krishnan and Sciubba (2005, pp. 19-20), whose data on respondents were collected through a random sample, report a similar magnitude of missing information on the dependent variable,[10] while Fafchamps and Gubert (2007) have much higher values for the amount of information that is lost: of 939 network members identified by 206 households, 750 (or 79.9%) are not part of the sample and are disregarded in their analysis. Other studies, such as Udry and Conley (2005), also mention this problem, but less directly.[11]

[10]The authors have data on "more than two-thirds" of the networks under analysis, reflecting the fact that "in most villages, over 30% of the village forms the sample and in some cases, about three-quarters of the village was surveyed" (**?**, p.19).

[11]In commenting on the graphical representation of the data used in their analysis of the determinants of link formation (Udry and Conley, 2005, Table10.4, p.257)these authors remark that "There are individuals in each village for each network who appear isolated in these graphs. That appearance is a misleading consequence of the strategy of constructing these graphs based on "ego-centric" data from a random sample of the population. In fact for each of these functional networks there is virtually no one in any of these villages who has no interactions with anyone. Virtually everyone in our sample has learning contacts, exchanges credit and/or gifts, hires labor, and has obtained land from someone. If none

Table 1: A summary of approaches to the study of network formation

	Goldstein and Udry (1999)	DeWeerdt (2004)	Dekker (2004)	Krishnan and Sciubba (2005)	Fafchamps and Gubert (2007)	Santos and Barrett (2006)
Sampling of individuals	Random sample, census	Census	Census	Random sample	Random sample	Random sample
Sampling of relationships	Matches within sample, Random matching	Matches within sample	Matches within sample	Matches within sample	Matches within sample	Random matching
Link	Potential, real	Potential	Real	Real	Potential, strong	Potential
Instrumental value	Information, others	Insurance	Insurance	Insurance	Insurance	Insurance
Other references	Conley and Udry (2005), Udry and Conley (2005) Santos and Barrett (2006)	Dercon and DeWeerdt (2006)			Fafchamps and Lund (2002)	

An evaluation of the importance of these losses is beyond the scope of this paper as it would require data on complete networks in order to replicate the effects of missing information.[12] Nevertheless, one suspects that they are important, not only due to the extent of missing information but also because there may be non–random qualitative differences between the links that are left out and those that are identified. For example, even with complete networks (that is, when all individuals in a group are being sampled) well still miss the relationships with individuals outside the census unit. Yet these can be especially valuable if, for example, one is interested in the performance of informal insurance (as income shocks across villages are typically less correlated than within villages, increasing the scope for mutual insurance) or information flows (as outside links may provide access to information that is not easily accessed within the village).

If many relationships are not with individuals who also belong to the sampling unit, one way to diminish the importance of missing information would be to collect detailed information on the attributes of the network members for the sampled individuals. This information could then be used to explain observed patterns of network formation. While there is evidence that very specific details about links' activities may not be accurately known, [13] there seems to be no a priori reason to doubt the validity of information on readily observable attributes such as gender, ethnic affiliation, age (at least within some interval or by comparison with the respondent), migrant status, etc..

The second point that merits reference is the nature of the link that is surveyed. When limiting the number of relationships elicited from a respondent, as in Fafchamps and Gubert (2007), one risks eliciting an implicit ranking of the relationships as these authors recognize.[14] The same is true,

---

of those other parties happens to be in our sample, the individual appears isolated in the graphs."(Udry and Conley, 2005, p.250).

[12]The social network literature dealing with this problem (most recently, Kossinets (2006)), although not focusing on dyad formation, reports discouraging results regarding the reliability of the estimates of network statistics when information on nodes or links is missing.

[13]For example, Goldstein and Udry (1999, p.20) report that, contrary to what is assumed in conventional models of social learning (where a group, such as a village, is assumed to be the network), farmers were not able to provide information about farm operations for a random sample of farmers in the villages they studied. This is further reinforced by Hogset and Barrett (2007), where a similar result is obtained when farmers are asked about details on agricultural practices of farmers that respondents indicated were in their information network.

[14]The authors mention that although they ask for a maximum of four relations per respondent, "In practice, respondents listed on average 4.6 individuals, with a minimum

although perhaps attenuated and less obvious, when one asks for a complete list of relationships. One can expect that those "closer" to the respondents will have a higher probability of being remembered and named (Brewer, 2000). In practice, one is leaving out weak ties, that is, those within the respondent's network who are socially more distant (Granovetter, 1974).[15][16]

Whether this emphasis on strong ties is a problem probably depends on the nature of the purpose for which data on networks are being collected (Sobel, 2002, Chwe, 1999). For some questions (for example, informal insurance), the Folk Theorem of repeated games would suggest that it is not a problem. In this case, the network is conceptualized as both a source of transfers and as a disciplining device that keeps the shadow of defection away; this last function requires proximity between everyone involved.[17] In other contexts (for example, information search), there seems to be less room for such an assumption as respondents may perceive those who are "more distant" as valuable sources of new information even if potentially less motivated to provide it (Santos and Barrett, 2006a). In any case, and in general, it seems that relatively little attention has been given to the importance of "weak ties" (Woolcock and Narayan, 2000, Ionnanides and Loury, 2004).

The distinction between potential and real links is potentially important.[18] Which is more appropriate probably depends on the purpose for which data on social interactions are being collected. Potential links may

of 1 and a maximum of 8. This is because in a number of cases respondents *refused to rank individuals they regarded as equivalently close to them.* (Fafchamps and Gubert, 2007, p. 9, footnote 8, emphasis added).

[15]In part, this is just a refinement of the previous point. Focusing on strong ties is on way of saying that information on weak ties is missing. See Kohler (1998) for an analysis of the effects of truncating the size of elicited networks on estimates of network density.

[16]In the original exposition of the hypothesis of the strength of weak ties, Granovetter (1974, p.1361) writes that "most intuitive notions of the "strength" of an interpersonal tie should be satisfied by the following definition: the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." In an early review of studies that tried to test this hypothesis, Granovetter (1982) identifies two major ways of operationalizing the concept of "strength of tie": (i) frequency of contact, used by Granovetter (1974), and (ii) the assumption that ties with different people (e.g., kin, friends, colleagues and acquaintances) have different strength. See Marsden and Campbell (1984) for a discussion.

[17]But see both Udry (1994), on the role for formal enforcers of such contracts, and Fafchamps (2002) for a discussion of the possibility of contracts when there is no evidence "of a single case of an agent being punished by others for dealing with someone who had previously breached a contract" (pp. 2-3).

[18]See Harrison and Rutstrőm (2004) for evidence on concerns about hypothetical bias in nonmarket value elicitation research.

matter most when analyzing forward–looking behavior, as it is the perception that one can rely on a link, regardless of whether it has been previously used, that likely drives present decisions. Studying real links would perhaps be preferable when the objective is to study past behavior, for example to understanding how information networks have affected learning about and dissemination of a new technology.

Clearly, there does not have to be a perfect juxtaposition between the two. The set of real links will probably be a subset of the potential network as it is improbable that all potential relations are mobilized in a specific period. For example, the data collected by Goldstein and Udry (1999) show that, from the set of individuals who *could* be contacted when searching for information, only a small fraction *was* contacted in the past.

Finally, most analysis to date has implicitly assumed that "everyone knows everyone else in village settings". As a consequence, the possibility that some links are not formed because individuals do not know each other has rarely been raised.[19] How to test this assumption is not trivial. One obviously cannot ask a respondent for a list of individuals that she does not know and to ask for a list of those she knows seems both infeasible (due to respondent fatigue) and, ultimately, unconvincing because those not named could have been just momentarily forgotten, possibly just because of less frequent contact.

The approach first used by Goldstein and Udry (1999) - to ask about social acquaintance between two randomly matched individuals belonging to a sample allows us to take a first look at this question.[20]Besides showing that not everyone knows everyone else, their data also show that knowing one's potential partner is a pre–condition for other interactions, providing support for the idea of embeddedness proposed by Granovetter (1985). Purposeful relations are formed from within a web of social relationships that are not necessarily constructed or maintained with a specific (instrumental) objective but that allow individuals to evaluate the costs and benefits of establishing a link with a specific purpose. The sequential nature of this process has consequences for the econometric model to be estimated (Maddala, 1983) as the analysis of the determinants of an instrumental network

[19]Santos and Barrett (2006a) and Santos and Barrett (2006b) are the exceptions.

[20]Given that, it is not surprising that this approach shares some similarities with previous suggestions in the social networks literature, notably by Granovetter (1973). The main difference is that in the latter, respondents were presented with a roster of all individuals in the group (not a random sample) and asked whether they knew them or not. The results of the application of this approach in a small group are reported in Erickson, Nosanchuck, and Lee (1981) and Erickson and Nosanchuck (1983).

should be done using the subsample of those who know each other and not the full sample.[21]

To summarize, the empirical literature in development economics that has analyzed network formation is small, recent and diverse. The main substantive question about it pertains to the reliability of its conclusions when an important part of the network of interest is missing. In the next section we present an approach, random matching, that largely obviates this problem.

3 Random matching

The approach to the sampling of relationships that we validate was first used by Goldstein and Udry (1999). We label it *random matching*. Starting with a random sample of individuals from a population of interest, one elicits the willingness of each respondent to enter into some specific relation with a match that is randomly selected from the same random sample.[22] Random matching has three major advantages relative to alternative methods. First, it naturally fits into the sampling strategies commonly used to collect micro–level data. Second, by randomly presenting the respondents with different possible matches, one discourages neglect of "weak links". Finally, we know the characteristics of both the respondent and her prospective match, hence no information is lost because one of the nodes is unknown.

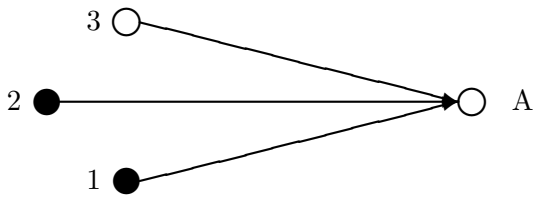
That said, it is important to notice the potential limitations and shortcomings of this approach to the sampling of relationships. In the approaches reviewed in the previous section, information loss occurs because, when free to choose from the population, respondents identified network members who were not in the random sample. With the random matching approach one relaxes the constraint of looking at existing links by imposing a new constraint: respondents must think about forming links with individuals who belong to the random sample. Why can random matching be trusted or even preferable to the matches within sample approach? It is easier to start answering this question by considering an example where random sampling of individuals, that underlies both random matching and matches within sample, *should not* be used.

[21] Or, at least, the interpretation of the results should make clear that their validity also depends on the assumption of generalized inter-knowledge among the sampled individuals.

[22] As we mentioned in the previous section, an important previous step made possible by this approach is to first establish whether the respondent is acquainted with the randomly selected match, allowing for an appreciation of the degree to which instrumental networks are embedded in a wider web of social connections.

Consider patronage relations reviewed by Platteau (1995). In figure 1 we represent an extreme setting where only one individual (labeled A) is a suitable patron for the remaining ones (the clients, labeled by numbers). Clearly if the sample (represented by full circles) is formed only of clients (here, 1 and 2), who do not establish (and are unwilling to establish) links between themselves, both approaches – random matching and matches within sample – would fail in allowing us to understand the process underlying network formation. In the case of random matching, because all individuals, unwilling to establish a link with each other, would (falsely) appear isolated, given that the patron is outside the sample. As for the direct elicitation of (potential or real) links, the absence of survey information on the patron would prevent the use of the matches within sample approach, making it impossible to understand the decision underlying the formation of this link.

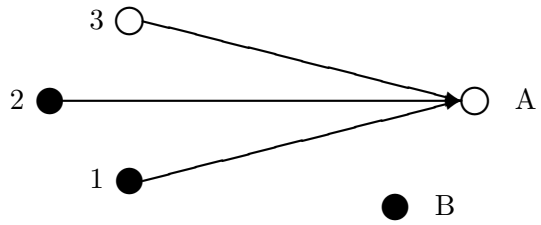
Figure 1: Sampling networks: without a prospective patron in the sample



Imagine now that another patron (labeled B) is available, although all clients still establish their relationship only with A. The picture would be similar (see figure 2) and lets assume that, due to the sampling process, individual B is sampled but A is not. Direct elicitation of links would leave the researcher exactly in the same position as before: all individuals would still appear as isolates. Random matching, on the other hand, has the potential to reveal something about the link formation decision, as it is conceivable that clients would be willing to form a link with B even though, in practice, that link is dominated by that with A.

Patronage as depicted here is an extreme example of a perhaps more general case, as suggested by Cox and Fafchamps (2006): individuals have limits to the number of relations that they can establish and maintain and, as such, social networks are bounded. It is therefore possible that links that were latent(perhaps because others were preferable or just because history

Figure 2: Sampling networks: with a prospective patron in the sample



and inertia led to a particular network configuration) may be "formed" during the questionnaire, allowing for inference that, in some cases (such as the one depicted in figures 1 and 1) would be impossible.

This potential advantage may come at a cost. If the relationships under analysis are the result of a thought experiment during which respondents are (implicitly) asked to reproduce the reasoning underlying the formation of social links but now facing a different set of partners it is not inconceivable, given the artificiality of the experimental setting, that cheap talk (or other noise) might generate connections that are uninformative about the characteristics of specific networks. It is therefore important to probe whether the links elicited following the random matching approach accurately reflect the decision processes underlying actual network formation. We do that in section 4.

Our second concern is that random matching involves the analysis of a subset of the possible relationships between the individuals in the random sample. Is this better than considering all relationships for which data exist, as in the matching within sample approach? In section 5 we show that for several models of network formation the answer is clearly "yes".

4 Can we trust data on hypothetical networks?

Although data on respondents' willingness to form a link has several advantages – it is forward-looking, it can identify feasible and attractive links that have not yet been activated, etc. – economists and other social scientists have a trained reticence to use data on hypothetical behaviors. In this section we ask whether we can trust that the data on hypothetical social network links form the basis for useful inference on the determinants of

network formation. We address this question empirically, using household survey data collected in 2004 from 120 randomly selected pastoralist households in four communities of southern Ethiopia that have been repeatedly interviewed for several years as part of a study that provides rich background data on the respondents.[23]

We collected comparable social networks data from these households using two different approaches. The first is random matching. In each community we randomly matched each respondent with five other respondents that belong to the random sample from the same site. We then asked whether the respondent knew the random match and whether the respondent would ask the match for a gift of one cattle. We subsequently asked our respondents to tell us how many people they could rely on to ask for cattle as a gift and asked for the names of up to five of those individuals, starting with the person they would ask first. These last two questions reveal the size of the respondents' relevant social network and the identities of their stronger links once we remove the random matching constraint that their potential links be with individuals belonging to the random sample.

In one site, we then interviewed as many as possible of the network members identified by the sample respondents, thereby providing a characterization of the respondents' local networks. In this site, one individual was not surveyed during this round. For three of our initial respondents we couldn't find any of the individuals mentioned while seven others were used as starting nodes of a different questionnaire and not subject to this exercise.

In one site, we then proceed to interview as many as possible of these individuals, providing a characterization of the respondents' local networks. In this site, one individual was not surveyed during this round. For three of our initial respondents we couldn't find any of the individuals mentioned while seven others were used as nodes of a different questionnaire and not subject to this exercise.[24] The analysis is therefore limited to the networks of 19 respondents, who named 70 people on who they could rely to ask for cattle as a gift. None of them was in the original sample, hence an analysis of the decision underlying the formation of these networks based on matches within the sample would be impossible. Of these 70 people, we

[23]The original data were collected by the Pastoral Risk Management (PARIMA) project. Barrett et al. (2004) describe the location, survey methods and available variables.

[24]Although we can recover the information on the identity of their network members and most of them were later found and interviewed, the differences in the survey instrument make these data imperfectly comparable. Thus we choose not to use them for this exercise.

could trace and interview 46 (approximately two-thirds).[25] The difficulty we experienced in tracking down the identified network partners underscores the difficulties and costs associated with the characterization of local networks. If random matching generates results that are statistically equivalent to the actual networks, its simplicity would provide a good argument for its use.

Table 2 summarizes the network links established by these two different approaches for the 19 respondent households for whom we have both types of social networks data.[26] On the surface, the resulting network patterns seem quite different. The random matching approach yields 22.6% of the 93 matches as potential providers of a cattle transfer, while our characterization of the respondents' local networks suggests a far lower figure, only 5.7% of the possible matches (where possible matches are defined as the population of individuals named by at least one respondent as someone s/he would approach for a cattle transfer). Nonetheless, it seems hard to extract a conclusion about any behavioral difference from these values, given the differences in the ways that these relationship data were collected.

We therefore test econometrically for the equivalence of the networks generated through random matching and direct elicitation. by estimating the model

$$\text{Prob } (L_{ij} = 1) = \Lambda(\gamma_1 X_{ij}) \tag{1}$$

where the link variable (L_{ij}) is a binary variable that equals one if a link between the respondent (indexed by i) and the match (indexed by j) is formed and is 0 otherwise and X_{ij} is the set of explanatory variables expressed as relative social distance, as in Santos and Barrett (2006a) and Fafchamps and Gubert (2007), that we define and summarize in Table 3. Finally, we

[25]These data represent the respondents' local networks subject to three caveats. First, it is clear that we effectively inquired about the identity of those who are socially closest to the respondent. Second, we assume that those individuals not named by our respondents are not part of their insurance networks. Of course, this may not be true. Perhaps some of them would be the 6^{th} or the 7^{th} person to be contacted in case of need but were omitted by our (arbitrary) rule limiting the insurance network to five individuals. An obvious consequence of this fact is that we are most probably underestimating the density of insurance dyads among this population, although this may not be a serious concern in this case as 10 out of the 19 respondents reported that they could rely on less than 5 individuals. Third, we cannot control for whether our respondents knew all the people named by other respondents and we neglect the possibility that insurance networks are embedded in a wider web of non-instrumental relations of friendship or social acquaintance.

[26]Some individuals were named by more than one of our respondents. We therefore have 50 links elicited among our 19 respondents and the 46 names they generated. Between these two sets of individuals there are 874 possible links, on which we only have direct information on 50. As mentioned in the previous footnote, we must assume that the other 824 links were not formed.

Table 2: Structure of insurance links: two approaches

Link exists?	Yes	No	Total
Random matching[27]	21	72[28]	93
Local network	50	824[29]	874
Total	71	896	967

[27] Data for the 19 respondents for who we found any of the insurance partners belonging to the local network.

[28] Elicited.

[29] Assumed.

assume that the error term, ε_{ij}, follows the logit distribution, where $\Lambda(.)$ is its cumulative distribution function and we further assume that

$$\mathrm{E}(\varepsilon_{\mathrm{ij}}, \varepsilon_{\mathrm{ih}}) \neq 0 \text{ ifj} \neq \mathrm{h} \tag{2}$$

$$\mathrm{E}(\varepsilon_{\mathrm{ih}}, \varepsilon_{\mathrm{jh}}) = 0 \text{ if i} \neq \mathrm{j} \tag{3}$$

Taking advantage of having multiple matches for each respondent, we can then estimate equation 1 using a random effects specification of the logit model. One alternative way of modeling the error term is to assume that,

$$\mathrm{E}(\varepsilon_{\mathrm{ih}}, \varepsilon_{\mathrm{jh}}) \neq 0 \text{ if i} \neq \mathrm{j} \tag{4}$$

that is, to incorporate the effect of matches' unobserved heterogeneity upon the link formation decision. Both (Udry and Conley, 2005) and Fafchamps and Gubert (2007) correct the variance matrix for the possible effect of matches' unobservables, using Conley (1999) estimator but do not find large differences due to this correction.[30]

We follow a different strategy, using a nonparametric permutation test known as Quadratic Assignment Procedure (QAP) (Hubert and Schultz, 1976, Krackhardt, 1987, 1988) to obtain correct p-values. The basic intuition behind this procedure is that the permutation of the data on the dependent variable must maintain its clustered nature. In practice, this means that the same permutation must be applied to respondents and matches. We can then estimate the above model when all correlation between dependent and independent variables is broken through resampling – that is, when the

[30] Although Fafchamps and Gubert (2007) mention that their Monte Carlo simulations support the importance given to this issue, as corrected standard errors can be much larger than uncorrected ones.

Table 3: Variable definitions and descriptive statistics

Variable	Definition	Random matching (1)	Local network (2)
Same clan	Dummy variable, equal to 1 if both i and j belong to the same clan	0.279	0.204
		(0.451)	(0.403)
Same sex	Dummy variable, equal to 1 if both i and j are from the same sex	0.473	0.579
		(0.502)	(0.494)
Bigger family	Absolute value of the difference in family size between i and j if i has a bigger family than j, 0 otherwise	1.882	2.641
		(2.762)	(3.263)
Smaller family	Absolute value of the difference in family size between i and j if i has a smaller family than j, 0 otherwise	2.452	1.960
		(3.070)	(3.362)
More land	Absolute value of the difference in land cropped between i and j if i cultivates more land than j, 0 otherwise	0.339	0.143
		(0.730)	(0.494)
Less land	Absolute value of the difference in land cropped between i and j if i cultivates less land than j, 0 otherwise	1.016	5.989
		(1.263)	(6.472)
More cattle	Absolute value of the difference in cattle owned between i and j if i has more cattle than j, 0 otherwise	2.290	0.815
		(4.608)	(2.885)
Less cattle	Absolute value of the difference in cattle owned between i and j if i has less cattle than j, 0 otherwise	9.720	45.602
		(22.503)	(90.659)
More experience	Absolute value of the difference in experience as herder between i and j if i has more experience than j, 0 otherwise	6.323	6.547
		(11.239)	(11.661)
Less experience	Absolute value of the difference in experience as herder between i and j if i has less experience than j, 0 otherwise	9.839	10.029
		(13.949)	(13.509)
Number of observations		93	874

null hypothesis that all slopes equal zero is known to be true – and compare our first estimates with their empirical distribution obtained through the repetition of this exercise (in our case, 200 times), to generate a sampling distribution for the parameter estimates. Although we present both uncorrected and QAP–corrected p-values, we also find that this added control for unobserved heterogeneity across individuals yields no substantial difference in our results.

Table 4 presents the results of two models. Column (1) reports the parameter estimates when we consider the data obtained through random matching for the 19 respondents for whom we could find any member of her local network. Column (2) presents the analogous regression estimates when we analyze the data on local networks. The qualitative results are quite similar: belonging to the same clan and being of the same sex have a positive effect on the likelihood of a transfer relationship, although there is considerable difference in the precision of these estimates, likely due in large part to the difference in sample size.

To understand if these two approaches produce results that are statistically similar, such that the random-matching approach can guide our understanding of how local networks form just as reliably as direct, unconstrained elicitation of social networks, we pool both sets of observations on links between individuals in this population and estimate the model

$$\text{Prob}\ (l^{ij} = 1) = \Lambda(\gamma_1 X^{ij}, \gamma_2 (X^{ij} \times \text{RM})) \tag{5}$$

under the same assumptions as above. The dummy variable RM takes the value 1 if the observation was obtained through random matching and 0 otherwise. A test of the joint null hypothesis that $H_0: \gamma_2 = 0$ then serves as a test for the statistical equivalence of the two methods at empirically identifying these insurance networks. Failure to reject the null hypothesis indicates that both approaches yield similar information about the structure of social networks.

Table 5 presents the regression coefficient estimates and p–values, as well as the Wald test of the null hypothesis that $\gamma_2 = 0$ for the slope terms (i.e., excluding the intercept, affected by the five name limit we imposed on respondents in reporting their prospective insurance partners). The smallest p–value on a single parameter estimate in γ_2 exceeds 0.2 and the p–value on the joint null hypothesis is 0.858. Turning to the QAP–corrected p–values, we get similar results. In no case can we reject the null hypothesis at the usual levels of statistical significance, although in one case (the interaction between "less land" and the dummy RM) we are clearly at its limit. This

Table 4: Regression results

Variable	Random matching (1)			Local network (2)		
	Coefficient	p–value	QAP p–value	Coefficient	p–value	QAP p–value
same clan	1.170	0.220	0.280	1.231	0.000	0.040
same sex	1.451	0.108	0.220	0.110	0.747	0.440
bigger family	-0.012	0.960	0.530	0.007	0.908	0.450
smaller family	0.065	0.727	0.460	-0.007	0.886	0.560
more land	0.025	0.634	0.470	0.251	0.364	0.300
less land	-0.037	0.304	0.440	-0.002	0.936	0.440
more cattle	-0.223	0.241	0.340	-0.053	0.456	0.360
less cattle	-0.001	0.968	0.510	0.001	0.787	0.330
more experience	0.232	0.815	0.110	0.004	0.788	0.500
less experience	0.417	0.257	0.070	-0.004	0.789	0.460
constant	-3.579	0.033	0.470	-3.263	0.000	0.270
Number of Observations	93			874		
Number of Respondents	19			19		

does not change our conclusion regarding the joint null hypothesis, tested through the statistic

$$\sum \mid \gamma_2(\mathrm{X_{ij}} \times \mathrm{RM}) \mid \tag{6}$$

that generates a measure of how distant the sum of all slopes is from zero. This test statistic equals 1.507 (Table 5) and has a QAP–corrected p–value of 0.985. We clearly cannot reject the null hypothesis that random matching provides a method of identifying the structure of respondents' social networks that is statistically equivalent to direct elicitation following standard methods. Random matching does indeed seem to provide useful inference about the structure of local networks.

One way to overcome, at least partially, the fact that we may be looking at variables that are slightly different is to look at the other piece of information we have about these networks: the number of links that each respondent thinks can be mobilized in case of need, this time without any limit imposed by the interviewer. We have information on this variable for the respondents in the four sites. Does a model such as the one from equation 1 yield predictions of network size that are accurate enough to give us a good idea of the extent of the respondent's network?

To answer this question we re–estimate the model from equation 1 using the data from the four sites.The estimation results are presented in Table 6. [31] We then use these results to predict (out of sample) the probability that each respondent would ask for cattle from any of the 29 potential matches in each village, hence generating a 30 x 30 matrix of predicted values of probability of a link.[32] Assuming that a link is formed if such probability is above some arbitrary threshold (here, 0.5), we can construct a square matrix of links. Finally, summing across the columns of this matrix we can obtain an estimate of the number of individuals that each respondent could ask for a transfer.

How does this estimate correlate with the number of people that could be asked for gifts, as reported by the respondents themselves? Quite highly. The Pearson correlation coefficient equals 0.337 (with a p-value of 0.002).[33] We interpret this result as additional supporting evidence that the random

[31] Because we only use these results to predict out of sample we skip the presentation and discussion of QAP-corrected p-values.

[32] By convention, links with oneself do not exist.

[33] The coefficient of rank correlation may even be a better indicator of the fit between the predictions of the model and the elicited values given that the maximum number of predicted links in each village is constrained to the size of the village sample and no such constraint was imposed when eliciting the size of the network. The Spearman ρ is 0.525 and also statistically significant (p-value=0.000).

matching approach yields data that accurately reflect the behavior underlying the formation of these networks.

These are not necessarily surprising results. An extensive literature on stated choice methods suggests that when properly contextualized, elicitation of hypothetical behaviors can provide an accurate view of actual behaviors (Arrow et al., 1993, Carson and Hanemann, 2005). As a concrete example of this equivalence, Barr (2003) shows that her experimental results, intended to understand how people form insurance networks in villages in Zimbabwe, were mirrored by reality in that the networks of risk pooling contracts constructed during the experiment and the networks existing in real life were significantly correlated.

Table 5: Testing the equivalence between different approaches

Variable	Coefficient	p–value p–value	QAP
same clan	1.228	0.000	0.100
same clan × RM	-0.192	0.783	0.410
same sex	0.130	0.714	0.550
same sex × RM	0.538	0.410	0.320
bigger family	0.010	0.861	0.470
bigger family × RM	-0.091	0.588	0.340
smaller family	-0.006	0.905	0.450
smaller family × RM	0.062	0.605	0.340
more land	0.263	0.352	0.350
more land × RM	-0.104	0.847	0.440
less land	-0.002	0.934	0.520
less land × RM	0.324	0.205	0.050
more cattle	-0.055	0.444	0.450
more cattle × RM	-0.142	0.323	0.280
less cattle	0.001	0.785	0.410
less cattle × RM	-0.010	0.516	0.230
more experience	0.004	0.787	0.420
more experience × RM	-0.025	0.434	0.590
less experience	-0.004	0.763	0.510
less experience × RM	-0.012	0.669	0.450
constant	-3.252	0.000	0.030
constant × RM	1.607	0.069	0.010
H_0: γ_2=0 (not including constant)			
Wald statistic	5.470	0.858	
$\sum \mid \gamma_2(X_{ij} \times RM) \mid$	1.507		0.975
Number observations	967		
Number respondents	19		

Table 6: Asking for gifts

Variable	Coefficient	p–value
same clan	1.947	0.000
same sex	-0.026	0.810
bigger family	0.015	0.821
smaller family	0.007	0.920
more land	-0.054	0.662
less land	0.081	0.505
more cattle	-0.002	0.825
less cattle	0.006	0.505
more experience	0.011	0.538
less experience	-0.016	0.258
village 1	-0.209	0.747
village 2	-0.436	0.479
village 3	1.343	0.008
constant	-2.208	0.000
N	551	

5 Monte Carlo evaluation of different approaches to network sampling

Having shown empirically that randomly matched data on willingness to establish a link can guide the inference on the determinants of network formation, we now turn to our second core question: How reliable are inferences about social network structure based on different approaches to sampling data on individuals and relationships? We answer this question through the use of Monte Carlo simulation so that we can know (by construction) the underlying network formation process and then test which sampling methods generate data that permits accurate inference of that process.

We start by constructing an artificial village of 200 households that mimics, in terms of the distribution of the different variables (clan, gender, cattle ownership, etc.), the data to be used in section 4 (and described in Table **??**, column 1). We then consider three models of link formation. In the first, which we call Random Links, these variables play no role in explaining the relationships between individuals, which originate purely through a random process. Although we do not believe this reflects actual behavior underlying the formation of instrumental networks, it provides a useful benchmark with which to compare the performance of the different sampling strategies, as it helps us establishing whether particular sampling designs might be predisposed to suggest structure where none really exists.

In the second model of link formation, which we call Structured Links, the propensity to form a link is a linear function of the variables included in the characterization of the village, similar to the one that we estimated for the set of all respondents in the previous section (presented in Table **??**). When this propensity is above a certain threshold (here, 0) a link is formed. Our third and final model is a minor variation on the Structured Links model, in which we limit the number of links an individual may form. We call this process Limited Links. Again, a threshold in the propensity to form a link has to be crossed for a link to be formed (the threshold remains 0) but an individual cannot form more than a limited number of links. For those who would surpass the limit, links are randomly deleted down to the imposed (and common, within the village) limit. We obviate this admittedly mechanical way of capping the number of links in a network by considering the effect of different limits (10, 20 and 30 links).

After specifying the structural process of social link generation, we then estimate, in the population, the same logit model from equation 1 (repeated here for convenience),

$$\text{Prob}(\text{L}_{\text{ij}} = 1) = \Lambda(\gamma_1 \text{X}_{\text{ij}})$$

where the variables have the same meaning as above: L_{ij} is a binary variable that is equal to one if a link between i and j is formed, X_{ij} is the set of explanatory variables expressed as relative social distance and $\Lambda(.)$is the logit cumulative distribution function. In table 7 we present the population estimates of this model, the "true" relation between the links and the explanatory variables for each of the three network formation models under consideration.

Table 7: Logit estimates of the link formation decision

	Random Links	Structured Links	Limited Links		
			10	20	30
Same clan	0.0338	2.2467	0.3478	0.4939	0.6817
Same sex	0.0182	0.4027	0.0074	0.4230	0.6005
More experience	-0.0006	0.5565	-0.1211	-0.0271	0.0581
Less experience	0.0003	-0.5605	-0.1528	-0.2428	-0.1174
More land	0.0582	1.4182	1.3666	-0.4339	-1.2254
Less land	0.0136	-1.2401	-1.3031	0.4746	0.0010
More cattle	-0.0002	-0.6689	-0.0485	-0.0401	-0.0422
Less cattle	0.0000	-0.0847	-0.0065	-0.0235	-0.0263
Bigger household	-0.0110	-1.7549	-0.0089	0.1164	0.0586
Smaller household	-0.0065	0.3423	0.3200	0.3446	0.0593
Constant	0.3256	4.5544	-2.0324	-1.8256	-1.8109

In the remainder of this section we analyze how well one can recover the underlying structure of network formation through the use of two different sampling strategies. The first randomly samples individuals and then considers all the links among these individuals, the commonplace *matches within sample* approach. The second is the *random matching* approach, which, as explained above, randomly samples relations among randomly sampled individuals. While the first approach is perhaps easy to understand (we sample individuals and consider *all* the links between them), the second involves a second level of random sampling, as we just consider *some* of the possible links formed by the randomly selected individuals.

Given that we're interested in understanding which approach gives us a more accurate representation of the link formation process in the population (known by construction), we mainly focus in tests of the hypothesis

$$H_0 : \gamma^{\text{sample}} = \gamma^{\text{population}} \tag{7}$$

where $\gamma^{population}$ represents the parameter vector for each underlying model of network formation and is given in Table 7. For each sampling method – matches within sample and random matching, the latter with 5, 10 or 15 random matches – and for each of four different sampling ratios (0.33, 0.50, 0.66 and 0.90) we generate 100 samples and estimate the logit equation 1 each time. Table 8 reports the frequency with which we fail to reject null hypothesis (equation 7), i.e., the frequency with which the resulting sample generates inferences consistent with the true underlying data generating procedure. The Stata code used to generate the village characteristics, the links between individuals, the sampling procedures and how we evaluate their consequences is presented in the Appendix.

This Monte Carlo analysis yields four main results. First, inference based on matches within sample, the most commonly used approach for analyzing local networks, seems valid only when links are formed randomly, an unlikely and uninteresting case, as it would signal that no intentional behavior is present. For other models of network formation, matches within sample seem to perform well only when the sampling ratio is quite high. Under the "structured links" and different "limited links" models, the matches within sample approach is virtually incapable of revealing the structure of link formation for sampling ratios as high as 2/3. This calls into question the reliability of inference about social network formation patterns based on data collected using the matches within sample method.

Second, as a rule, the random matching approach beats the matches within sample approach. Especially in the "limited links" models, the performance of the random matching model is far better than that of the matches within sample approach, albeit still imperfect. Indeed, this is not to say that random matching is adequate under all circumstances. In particular, if social links are formed according to what we termed "structured links", i.e., without limits to the size of networks, then this approach can still perform quite poorly, even if it remains clearly superior to the "matches within sample" approach under standard sampling ratios (i.e., below 90%).

Third, our capacity to accurately describe the link formation decision decreases as we increase the number of relations sampled, emphasizing the importance of sampling relations after sampling individuals, reflecting the double nature of social networks. Given that in the limit, when each respondent in a sample is presented with all possible matches, the two procedures are identical this is a plain consequence of the already discussed superiority of the random matching approach when compared to the matches within sample. This is especially evident in the more interesting models, when links are not randomly formed, and for sampling ratios below 90%.

Table 8: Monte Carlo evaluation of two sampling approaches: Matches within sample vs. Random matching

Sampling ratio (individuals)	*33*	*50*	*66*	*90*
Random Links				
Matches within sample	92	99	100	100
Random matching: 5 relations	96	96	96	94
Random matching: 10 relations	98	94	95	99
Random matching: 15 relations	96	100	95	95
Structured Links				
Matches within sample	0	0	0	92
Random matching: 5 relations	25	29	63	69
Random matching: 10 relations	11	26	47	73
Random matching: 15 relations	1	15	48	78
Limited Links (10)				
Matches within sample	4	2	4	60
Random matching: 5 relations	73	83	91	93
Random matching: 10 relations	68	70	86	93
Random matching: 15 relations	58	57	82	92
Limited Links (20)				
Matches within sample	2	1	4	44
Random matching: 5 relations	74	79	91	95
Random matching: 10 relations	52	70	79	96
Random matching: 15 relations	38	58	74	97
Limited Links (30)				
Matches within sample	0	1	3	30
Random matching: 5 relations	74	84	92	94
Random matching: 10 relations	51	68	77	91
Random matching: 15 relations	38	57	66	93

Finally, we notice that the results regarding the adequacy of the random matching approach under the Limited Links model does not change much with the maximum number of links allowed (and, consequently, with the density of links in the population). Random matching appears slightly more accurate the lower the limit on the number of links formed in the population. But what really seems to matter most is the existence of such a limit.

6 Conclusions

This paper makes a methodological contribution to the growing literature that aims at understanding how social networks are formed, typically as a first step toward analysis of social networks' role in explaining individual behavior and outcomes. We validate a new approach to the collection of data on network structure – which we label "random matching" – where individuals from a random sample are allowed to form links with randomly matched individuals from the same sample. The central advantages of this approach are two: the ease with which it can be integrated into the surveys that economists commonly conduct and use and the fact that both respondent and match are part of the sample.

We compare the determinants of individuals' decision to link or not to link with a random match with the determinants of directly elicited local networks and conclude that these two data collection processes generate statistically identical results with respect to the correlates of social network structure. Furthermore, the size of the predicted network generated by the random matching data is highly correlated with the size of the local network directly elicited from survey respondents. Finally, we demonstrate, via Monte Carlo methods, the superiority of this random matching approach relative to the more conventional method of using all the links between individuals in a random sample.

The way in which we established the relation between the elicited size of the respondent's network and its predicted size at the end of section 4 also suggests how we believe researchers might usefully employ the random matching approach to sampling social networks. In addition to providing a statistically valid means of eliciting data for analysis of social network structure, which may be interesting in its own right, one can also use the resulting parameter estimates to predict respondents' networks and subsequently perform analyzes based on those predicted networks. This is similar, in spirit, to the analysis by Woittiez and Kapteyn (1998), who estimate a latent variable model to infer the unobserved reference groups of respon-

dents, after which the means of behaviors within such groups (in their case, hours of work and labor force participation rate) are used as explanatory variables for individual decisions (in their case, the labor market behaviors of Dutch women). In doing this, one must recognize, however, that we start from simple local rules and aim at the complete structure. Although some evidence exists on the utility of such approach for some questions,[34] more work is probably needed before the validity of these generated variables is reasonably established.

This paper by no means resolves questions of how to identify the structure of social networks of all sorts and under all conditions. Our results reflect only data from insurance networks in just one location, and it is also obvious that the utility of asking questions about potential links is limited in some cases.[35] But, if the validity of the random matching approach to collecting data on social networks is confirmed in other settings, it could help establish a statistically valid and cost-effective method for generating data for social networks analysis to respond to burgeoning questions about the role and importance of social connectivity in processes of economic development, free of some of the key inferential problems that presently plague this literature.

References

Arrow, K.J., R. Solow, P.R. Portney, E.E. Leamer, R. Radner, and H. Schuman. 1993. "Report of the NOAA Panel on contigent valuation." Washington, D.C., National Oceanic and Atmospheric Administration.

Bandiera, O., and I. Rasul. 2006. "Social networks and the adoption of new technology in northern Mozambique." *Economic Journal* 116:869–902.

Barr, A. 2003. "Risk pooling, commitment and information: an experimental test of two fundamental assumptions." Unpublished, University of Oxford, CSAE working paper.

Barrett, C.B., G. Gebru, J.G. McPeak, A.G. Mude, J. Vanderpluye-Orgle, and A.T. Yirbecho. 2004. "Codebook for data collected under the im-

[34] See the discussion in Morris (2003) and an application to the epidemiology of HIV in Kretzschmar and Morris (1997)

[35] It is hard to imagine, for example, how and why one would ask a respondent about his/her willingness to form a sexual network, even if cultural norms made such questions permissible.Yet understanding such networks is essential to study the epidemiology of sexually transmitted diseases.

proving pastoral risk management on East Africa rangelands (PARIMA) project." Unpublished, Cornell University.

Behrman, J., H.P. Kohler, and S. Watkins. 2002. "Social networks, family planning and worrying about AIDS: what are the network effects if network partners are not determined randomly?" Unpublished, University of Pennsylvania, PIER working paper.

Blume, L.E., and S.N. Durlauf. 2005. "Identifying social interactions: a review." Unpublished, Cornell University working paper.

Brewer, D.D. 2000. "Forgetting in the recall-based elicitation of personal and social networks." *Social Networks* 22:29–43.

Brock, W., and S. Durlauf. 2001. "Interactions-based models." In J. Heckman and E. Leamer, eds. *Handbook of Econometrics*. Amsterdam: Elsevier, vol. 5.

Carson, R.T., and W.M. Hanemann. 2005. "Contigent valuation." In K.-G. Maler and J. R. Vincent, eds. *Handbook of Environmental Economics*. Amsterdam: Elsevier, chap. 17.

Chwe, M.S.Y. 1999. "Structure and strategy in collective action." *American Journal of Sociology* 105:128–156.

Conley, T., and C. Udry. 2005. "Learning about a technology: pineapple in Ghana." Unpublished, Yale University, working paper.

Conley, T.G. 1999. "GMM estimation with cross–sectional dependence." *Journal of Econometrics* 92:1–45.

Cox, D., and M. Fafchamps. 2006. "Extended families and kinship networks: economic insights and evolutionary directions." Unpublished, forthcoming in Handbook of Development Economics, vol. 4.

Dekker, M. 2004. "Risk sharing in rural Zimbabwe: an empirical analysis of endogenous network formation." Unpublished, paper presented at the CSAE Conference on growth, poverty reduction and human development.

DeWeerdt, J. 2004. "Risk sharing and endogenous network formation." In S. Dercon, ed. *Insurance against poverty*. Oxford: Oxford University Press.

Durlauf, S. 2002. "On the empirics of Social Capital." *Economic Journal* 112:F459–F479.

Durlauf, S., and M. Fafchamps. 2004. "Social capital." In P. Aghion and S. Durlauf, eds. *Handbook Economic Growth*. Amsterdam: Elsevier.

Erickson, B.H., and T.A. Nosanchuck. 1983. "Applied network sampling." *Social Networks* 5:367–382.

Erickson, B.H., T.A. Nosanchuck, and E. Lee. 1981. "Network sampling in practice: some second steps." *Social Networks* 3:127–136.

Fafchamps, M. 2002. "Spontaneous market emergence." *Topics in Theoretical Economics* 2.

Fafchamps, M., and F. Gubert. 2007. "The formation of risk sharing networks." *Journal of Development Economics* 83:326–350.

Goldstein, M., and C. Udry. 1999. "Agricultural innovation and risk management in Ghana." Unpublished, final report to IFPRI.

Goodman, L. 1961. "Snowball sampling." *Annals of Mathematical Statistics* 32:148–170.

Granovetter, M. 1985. "Economic action and social structure: the problem of embeddedness." *American Journal of Sociology* 91.

—. 1973. "Network sampling: some first steps." *American Journal of Sociology* 81:1287–1302.

—. 1974. "The strength of weak ties." *American Journal of Sociology* 78.

—. 1982. "The strength of weak ties: a network theory revisited." In P. Marsden and N. Lin, eds. *Social structure and network analysis*. Thousand Oaks, CA: Sage Publications, pp. 105–130.

Harrison, G.W., and E.E. Rutstrőm. 2004. "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods." In C. Plott and V. L. Smith, eds. *Handbook of Experimental Economics*. New York: Elsevier Science, vol. 1.

Heckathorn, D. 2002. "Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations." *Social Problems* 49:11–34.

Heckathorn, D., and J. Jeffri. 2002. "Jazz networks: using respondent-driven sampling to study stratification in two jazz musician communities." Unpublished, paper presented at the Annual Meeting of the American Sociological Association, Atlanta, GA.

Hogset, H., and C.B. Barrett. 2007. "Imperfect Social Learning Among Kenyan Smallholders." Unpublished, Cornell University working paper.

Hubert, L., and J. Schultz. 1976. "Predicting with networks: Nonparametric multiple regression analysis of dyadic data." *British Journal of Mathematical and Statistical Psychology* 29:190–241.

Ionnanides, Y., and L.D. Loury. 2004. "Job information networks, neighborhood effects, and inequality." *Journal of Economic Literature* 42:1056–1093.

Kohler, H.P. 1998. "Bias in the estimation of density on the basis of egocentric networks with truncated size." Unpublished, Max Planck Institute for Demographic Research, Rostock, Germany.

Kossinets, G. 2006. "Effects of missing data in social networks." *Social Networks* 28:247–268.

Krackhardt, D. 1988. "Predicting with networks: Nonparametric multiple regression analysis of dyadic data." *Social Networks* 10:359–381.

—. 1987. "QAP partialling as a test of spuriousness." *Social Networks* 9:171–186.

Kretzschmar, M., and M. Morris. 1997. "Concurrent partnerships and the spread of HIV." *AIDS* 11:641–648.

Krishnan, P., and E. Sciubba. 2005. "Links and architecture in village networks." Unpublished, University of Cambridge, working paper CWPE 0462.

Maddala, G.S. 1983. *Limited-dependent and qualitative variables in econometrics*. Econometric Society Monographs, Cambridge University Press.

Manski, C.F. 2000. "Economic analysis of social interactions." *Journal of Economic Perspectives* 14(3):115–136.

—. 1993. "Identification of endogenous social effects: the reflection problem." *Review of Economic Studies*, pp. 531–542.

Marsden, P.V., and K.E. Campbell. 1984. "Measuring tie strength." *Social Forces* 63:482–501.

Moffitt, R. 2001. "Policy interventions, low-level equilibria and social interactions." In S. Durlauf and H. P. Young, eds. *Social Dynamics*. Cambridge, MA: The MIT Press, chap. 3.

Morris, M. 2003. "Local rules and global properties: modeling the emergence of network structure." In R. Breiger, K. Carley, and P. Pattison, eds. *Dynamic social network modeling and analysis: workshop summary and papers*. National Academies Press for the National Research Council, pp. 174–186.

—. 2004. "Overview of network survey designs." In M. Morris, ed. *Network epidemiology: a handbook of survey design and data collection*. Oxford: Oxford University Press, chap. 1.

Narayan, D., and L. Pritchett. 1999. "Cents and sociability: household income and social capital in rural Tanzania." *Economic Development and Cultural Change* 47:871–898.

Paldam, M. 2000. "Social capital: one or many? Definition and measurement." *Journal of Economic Surveys* 14:629–653.

Platteau, J.P. 1995. "A framework for the analysis of evolving patron-client ties in agrarian economies." *World Development* 23:767–786.

Santos, P., and C.B. Barrett. 2006a. "Choosing one's herd: identity and network formation in rural Ghana." Unpublished, Cornell University.

—. 2006b. "Informal insurance in the presence of poverty traps. Evidence from Southern Ethiopia." Unpublished, Cornell University.

Sobel, J. 2002. "Can we trust social capital?" *Journal of Economic Literature* 40:139–154.

Soetevent, A.R. 2006. "Empirics of the identification of social interactions: an evaluation of the approaches and their results." *Journal of Economic Surveys* 20:193–228.

Udry, C. 1994. "Risk and insurance in a rural credit market: an emprirical investigation in Northern Nigeria." *Review of Economic Studies* 61:495–526.

Udry, C., and T. Conley. 2005. "Social networks in Ghana." In C. B. Barrett, ed. *The social economics of poverty: identities, groups, communities and networks*. London: Routledge, chap. 10.

Wasserman, S., and K. Faust. 1994. *Social network analysis. Methods and applications*. Cambridge: Cambridge University Press.

Woittiez, I., and A. Kapteyn. 1998. "Social interactions and habit formation in a model of female labor supply." *Journal of Public Economics* 70:185–205.

Woolcock, M., and D. Narayan. 2000. "Social capital: implications for development theory, research, and policy." *World Bank Research Observer* 15:225–249.

A Monte Carlo simulation code

This is the main structure of the Stata code used to generate the results presented in Table 8. Its use requires small adaptations and extensions (to get different sampling ratios, to allow for other models of network formation, etc) that are duly signaled.

```
*START CODE
drop _all
set obs 200
set seed 12345
gen clan=uniform()
replace clan=1 if clan≤0.20
replace clan=2 if clan≤0.2333
replace clan=3 if clan≤0.30
replace clan=4 if clan≤0.40
replace clan=5 if clan≤0.7667
replace clan=6 if clan≤0.90
replace clan=7 if clan≤0.9667
replace clan=8 if clan≤1.00
set seed 12345
gensex=uniform()
replace sex=1 if sex≤0.633
replace sex=0 if sex>0.633 & sex!=1
set seed 12345
gen hhsize=invnorm(uniform())
replace hhsize=(hhsize*3.59)+7.5
replace hhsize=int(hhsize)
replace hhsize=1 if hhsize≤0
set seed 12345
genexp=invnorm(uniform())
replace exp=(exp*14.94) + 23.2
replace exp=int(exp)
replace exp=0 if exp<0
set seed 12345
gen land=invnorm(uniform())
scalar a=1.48
scalar b=1.37
replace land=ln(a)+sqrt(ln(b))*land
replace land=exp(land)
```

```
set seed 12345
gen ind=uniform()
set seed 12345
gen cat1=invnorm(uniform())
scalar a=5.444
scalar b=4.255
replace cat1=ln(a) + sqrt(ln(b))*cat1 if ind≤0.90
replace cat1=0 if ind>0.90
set seed 12345
gen cat2=invnorm(uniform())
replace cat2=67.333+37.647*cat2 if ind>0.90
replace cat2=0 if ind≤0.90
gen cattle=cat1 + cat2
replace cattle=0 if cattle<0
replace cattle=int(cattle)
drop ind cat1 cat2
gen name=[_n]
tempfile namev1
save "'namev1'"
foreach var in clan sex hhsize exp land cattle {
        ren 'var' 'var'1
}
ren name match
tempfile matchv1
save "'matchv1'"
sort match
save, replace
use "'namev1'"
sort name
expand 200
sort name
gen match=.
replace match=[_n] if [_n]≤200
forvalues x = 2 (1) 200{
       quietly replace match=match[_n-200] if _n>('x'-1)*200 & _n≤'x'*200
}
save, replace
sort match
merge match using "'matchv1'"
drop _merge
```

```
gen sclan=(clan==clan1)
gen ssex=(sex==sex1)
foreach var in exp land cattle hhsize {
        gen m‘var’=‘var’-‘var’1
        replace m‘var’=0 if ‘var’<‘var’1
        gen l‘var’=abs(‘var’-‘var’1)
        replace l‘var’=0 if ‘var’>‘var’1
}
drop clan* sex* hhsize* exp* land* cattle*
save ...\village.dta”, replace
* RANDOM LINKS
sort name match
set seed 123456
gen link=uniform()
replace link=0 if name==match
centile link, c(58.4375)
scalar cut=r(c_1)
replace link=(link<cut)
logit link sclan ssex mexp lexp mland lland mcattle lcattle mhhsize lhhsize
save “...\villageRL.dta”, replace
* STRUCTURED LINKS
use “...\village.dta”, clear
gen link=1.206*sclan + .071*ssex - .029*msize +.007*lsize +.335*mland
        - .024*lland - .071*mcattle -.001*lcattle - .001*mexp -.008*lexp
replace link=0 if name==match
replace link=(link>0)
logit link sclan ssex mexp lexp mland lland mcattle lcattle mhhsize lhhsize
save “...\villageS.dta”, replace
* LIMITED LINKS
use “...\villageS.dta”, clear
sort name match
by name, sort: gen slink=sum(link)
replace link=0 if slink>10
logit link sclan ssex mexp lexp land lland mcattle lcattle mhhsize lhhsize
save “...\villageSL.dta”, replace
/* Simulating the MATCHES WITHIN SAMPLE approach when links are
randomly formed*/
program define networkstructure,rclass
        version 8.0
        drop _all
```

```
set obs 200
gen u=uniform()
centile u, c(33)
scalar r=r(c_1)
replace u=(u≤r)
gen name=_n
sort name
tempfile name
save "`name'", replace
ren name match
tempfile match
sort match
save "`match'", replace
use "...\villageR.dta", clear
sort name
merge name using "`name'"
drop _merge
ren u sample1
sort match
merge match using "`match'"
drop _merge
ren u sample2
keep if sample1==1
keep if sample2==1
scalar bsclan=.0338991
scalar bssex=.0182271
scalar bmexp=-.0006444
scalar blexp=.0003125
scalar bmland=.0582165
scalar blland=.0135889
scalar bmcattle=-.0002283
scalar blcattle=.0000456
scalar bmsize=-.0110378
scalar blsize=-.0065319
scalar bcons=.3256091
logit link sclan ssex mhhsize lhhsize mland lland mcattle lcattle mexp
    lexp
testnl_b[sclan]-bsclan==_b[ssex]-bssex==_b[mhhsize]-bmhhsize==
    _b[lhhsize]-blhhsize==_b[mland]-bmland==_b[lland]-blland==
    _b[mcattle]-bmcattle==_b[lcattle]-blcattle==_b[mexp]-bmexp==
```

```
        _b[lexp]-blexp==_b[_cons]-bcons==0
    return scalar test=r(p)
end
set seed 23456
tempfile structure_R33RSI
simulate "networkstructure" testRRSI33=r(test), reps(100) saving("'structure_RRSI33'")
program drop networkstructure
gen N=_n
sort N
save, replace
/*this program has to be repeated for the remaining sampling ratios (50%,
66%, 90%) and for the remaining models of network formation*/
merge N using structure_R33RSI'
drop _merge
sort N
save, replace
merge N using 'structure_R50RSI'
drop _merge
sort N
save, replace
merge N using 'structure_R66RSI'
drop _merge
save, replace
foreach var in testR33RSI testR50RSI testR66RSI testR90RSI {
    count if 'var'>.05 & 'var'!=.
}
/* Simulating the RANDOM MATCHING approach when links are ran-
domly formed*/
program define networkstructure, class
    version 8.0
    drop _all
    set obs 200
    gen u=uniform()
    centile u, c(33)
    scalar r=r(c_1)
    replace u=(u≤r)
    gen name=_n
    sort name
    tempfile name
    save "'name'", replace
```

```
        ren name match
        tempfile match
        sort match
        save "`match'", replace
        use"...\villageR.dta",clear
        sort name
        merge name using "`name'"
        drop _merge
        ren u sample1
        sort match
        merge match using "`match'"
        drop _merge
        ren u sample2
        keep if sample1==1
        keep if sample2==1
        gen sample3=uniform()
        sort name sample3
        replace sample3=1
        by name, sort: gen sum3=sum(sample3)
        keep if sum3≤5
        scalar bsclan=.0338991
        scalar bsamesex=.0182271
        scalar bmexp=-.0006444
        scalar blexp=.0003125
        scalar bmland=.0582165
        scalar blland=.0135889
        scalar bmcattle=-.0002283
        scalar blcattle=.0000456
        scalar bmsize=-.0110378
        scalar blsize=-.0065319
        scalar bcons=.3256091
        logit link sclan ssex mhhsize lhhsize mland lland mcattle lcattle mexp
            lexp
        testnl _b[sclan]-bsclan==_b[ssex]-bssex==_b[mhhsize]-bmhhsize==
            _b[lhhsize]-blhhsize==_b[mland]-bmland==_b[lland]-blland==
            _b[mcattle]-bmcattle==_b[lcattle]-blcattle==_b[mexp]-bmexp==
            _b[lexp]-blexp==_b[_cons]-bcons==0
        return scalar test=r(p)
end
set seed 23456
```

```
tempfile structure_R33RSR5
simulate "networkstructure" testR33RSR5=r(test), reps(100) saving
        ("'structure_R33RSR5'")
program drop networkstructure
/* this simulation has to be repeated for the remaining sampling ratios, different models of network formation and number of relations to be sampled (10 and 15)*/
```