What do we learn about social networks when we only sample individuals? Not much.*

Paulo Santos[†] University of Sydney

Christopher B. Barrett[‡] Cornell University

May 2008

Abstract

Much of the empirical analysis of social networks is based on a sample of individuals, rather than a sample of matches between pairs of individuals. This paper asks whether that approach is useful when one wants to understand the determinants of variables that are inherently dyadic, such as relationships. After reviewing the shortcomings of the data used in the literature, we use Monte Carlo simulation to show that the answer is positive only when relationships are themselves randomly formed, a very special and uninteresting case. Additional work that supports strategies to collect dyadic data as part of surveys usually used by economists seems to be needed.

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

*We thank Larry Blume, Michael Carter, Marcel Fafchamps, Gueorgi Kossinets, Annemie Maertens, Jacqueline Vanderpuye-Orgle and seminar audiences at Cornell University and NEUDC 2006 for helpful comments. The views expressed here and any remaining errors are the authors' and do not represent any official agency.

[†]Lecturer, Discipline of Agricultural and Resource Economics. Contact address: 107 Watt Building, Science Road, The University of Sydney, Camperdown NSW 2006, Australia. Email: p.santos@usyd.edu.au

[‡]Professor, Department of Applied Economics.

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, "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 (for an influential example, see Narayan and Pritchett (1999)) 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). [3]

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

[1] The literature on social capital was recently reviewed by Durlauf and Fafchamps (2005).

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

[3] Soetevent (2006) and Blume and Durlauf (2005) present recent reviews of this literature.

context. The focus of this paper is on the development economics literature that aims at understanding the process of network formation, either as a question in itself or as a first step towards the quantification of the instrumental value of social connections – as Jackson (2007) argues, the two questions are intimately connected.

Although relatively small, this literature is diverse and development economists have used a variety of sample designs, both for respondents and for relationships. As interest in the empirical analysis of the effects and the structure of social networks grows and more researchers contemplate the possibility of using such data, it is important to understand the implications of these methodological choices. That is the purpose of the next section.

Social networks are a set of individuals and of relationships among them. The ultimate objective of this paper is to understand whether this joint focus needs to be taken seriously and reflected in the data collection strategies used by economists. In Section 3 we use Monte Carlo simulation to compare the accuracy of the inference with respect to the determinants of network formation when data on relationships are collected in two different ways: the frequent approach of relying on the set of all links formed by a random sample of individuals with other individuals also in the sample as an accurate image of individuals' networks, which we label as *matches within sample* and a different approach where relationships are randomly selected, which we label as *random matching*. Our results show that the random matching approach is, in general, more accurate than the matches within sample. Section 4 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.[4] 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. [5] These lead to different network structures, commonly referred as global versus local networks respectively.[6] 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 analysis.[7]

Having decided how to sample individuals, the second level of sampling is done through the construction of a "name generator", a question that

[4] 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.

[5] 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.

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

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

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 questions of the type "Who could you rely on to ...?" (DeWeerdt, 2004, Fafchamps and Gubert, 2007, Santos and Barrett, 2007), while others focused on real relations through questions such as "On whom did you rely to ...?" (Dekker, 2004, Krishnan and Sciubba, 2005, Conley and Udry, 2005, Udry and Conley, 2005). 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 to be identified by the respondents (e.g. Fafchamps and Gubert, 2007). These different approaches are 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 the census unit. Krishnan and Sciubba (2005, pp. 19-20), whose data on respondents were collected through a random sample, report a similar mag-

nitude of missing information on the dependent variable,[8] 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.

One suspects that the importance of these losses is a major drawback of an approach that limits itself to look at the links between randomly selected individuals. The discussion in Udry and Conley (2005) is especially illuminating in this regard: in commenting on the graphical representation of the data used in their analysis of the determinants of link formation, 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. (...) If none of those other parties happens to be in our sample, the individual appears isolated in the graphs." (Udry and Conley, 2005, p.250, emphasis added).

Concerns on the validity of the inference being made raised by the extent of missing information can only be augmented if there are reasons to suspect that there may be non–random qualitative differences between the links that are left out and those that are identified.[9]

[8] 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" (Krishnan and Sciubba, 2005, p.19).

[9] For example, even when all individuals in a group are being sampled we'll 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).

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.[10] The same is true, 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). [11]

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 example, in the analysis of informal insurance, 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, suggesting that focusing on strong ties should not be a problem. In other contexts such as, for example, information search, there seems to be less room for such an assumption as

[10]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 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).

[11]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.

respondents may perceive those who are "more distant" as valuable sources of new information even if potentially less motivated to provide it. 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).

Finally, the distinction between potential and real links may be important. Which is more appropriate probably depends on the purpose for which data on social interactions are being collected. Potential links may 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 understand how information networks have affected learning about and dissemination of a new technology. [12]

To summarize, the empirical literature in development economics that has analyzed network formation is small, recent and diverse. It has mainly grown out of surveys of individuals, where some questions regarding their social networks are asked. Although random sampling of individuals guarantees that the results can be extrapolated to the population of individuals, it is not clear why it should be used to guide inference on the determinants of dyadic variables such as relationships. In the next section we address this question through the use of Monte Carlo simulation.

[12] 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.

3 Should we sample relationships?

How reliable is inference on the determinants of social networks structure based on different approaches to sampling data on individuals and relationships? To be more specific: When looking at the determinants of social networks, do we need to sample relationships or can we just sample individuals?

We address this question through the use of Monte Carlo simulation so that we can know (by construction) the underlying network formation process and then test whether random sampling of individuals is enough to recover the determinants of that process or, on the contrary, we need to go one step further and sample relationships.

We start by constructing an artificial village of 200 households that can be characterized by a set of variables such as clan, gender, cattle ownership, etc.. 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. 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 group) 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).

It should be clear that none of these models has a clear foundation on individual behavior. This reflects both the lack of such a generally accepted model and our relatively limited objectives: we aim only at comparing two approaches to data collection and it is not clear that offering yet another model of network formation would help us on that.

After specifying the process of link generation, we then estimate, in the population, a logit model of the form,

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

where 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 logistic cumulative distribution function. In table 2 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.

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 - an approach that we call *matches within sample* approach. The second strategy randomly samples matches between randomly individuals in a sample (and as such, it randomly samples relationships) and we label it *random matching*. While the first approach is perhaps easy to understand (we sample individuals and consider *all* the links between them) and has been used in the literature, the second involves a second level of random sampling, as we just consider *some* of the possible links formed by the randomly selected individuals. [13]

Given that we are interested in understanding which of the two approaches (matching within sample and random matching) 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

$$H_0 : \gamma^{\text{sample}} = \gamma^{\text{population}} \tag{2}$$

where $\gamma^{\text{population}}$ represents the parameter vector for each underlying model of network formation and is given in Table 2. 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,

[13] This approach was used, for example, by Aggarwal (2007), in the analysis of contract choice in groundwater sales: see Aggarwal (2007, p.479) for an explanation of the sampling procedure. It is also similar to the approach used to collect some of the data on networks described in Goldstein and Udry (1999).

0.50, 0.66 and 0.90) we generate 100 samples and estimate equation 1 each time. Table 3 reports the frequency with which we fail to reject the null hypothesis (equation 2), 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.

Our 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 random matching 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, reflecting the double nature of social networks and the importance of sampling relations after sampling individuals, the capacity to accurately describe the link formation decision decreases as we increase the number of relations sampled. 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%.

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.

4 Conclusions

In a recent review of the economics literature on social networks, Jackson (2007) remarks that interest in the structure of social networks goes together with interest in their importance. But before that relation can be explored further, one needs to be able to understand how such networks are formed. Solely relying on information on specific links formed between individuals

that happen to be both randomly selected from the population of individuals, as done in most of the existing literature, does not seem to be able to tell us much about that decision. In short, understanding dyadic variables, such as relationships, seems to require sampling approaches that cannot be focused on individuals alone - a conclusion also reached in the conclusion of closely related empirical literature on the determinants of contract choice (Ackerberg and Botticini, 2002).

References

Ackerberg, D.A., and M. Botticini. 2002. "Endogenous matching and the empirical determinants of contract form." *Journal of Political Economy* 110:564–591.

Aggarwal, R.M. 2007. "Role of risk-sharing and transaction costs in contract choice: Theory and evidence from groundwater contracts." *Journal of Economic Behavior and Organization* 63:475–496.

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.

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.

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. 2005. "Social capital." In P. Aghion and S. Durlauf, eds. *Handbook Economic Growth*. Amsterdam: Elsevier.

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. 1974. "The strength of weak ties." *American Journal of Sociology* 78:1360–1380.

—. 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.

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

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

Jackson, M. 2007. "The study of social networks in economics." Unpublished, forthcoming in The Missing Link: Formation and Decay of Economics Networks.

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

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

—. 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.

Morris, M. 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.

Santos, P., and C.B. Barrett. 2007. "Persistent poverty and informal credit." 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., 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.

Woolcock, M., and D. Narayan. 2000. "Social capital: implications for de-

velopment theory, research, and policy." *World Bank Research Observer* 15:225–249.

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

	Udry and Conley (2005)	DeWeerdt (2004)	Dekker (2004)	Krishnan and Sciubba (2005)	Fafchamps and Gubert (2007)
Sampling of individuals	Random sample, census	Census	Census	Random sample	Random sample
Sampling of relationships	All matches within sample	All matches within sample	All matches within sample	All matches within sample	All matches within sample
Link	Real	Potential	Real	Real	Potential
Instrumental value	Information, land, labor, credit	Insurance	Insurance	Insurance	Insurance

Table 2: 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

Table 3: 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

A Monte Carlo simulation code

This is the main structure of the Stata code used to generate the results presented in Table 3. 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
*Constructing the fictitious village
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
** Defining the different models of network formation
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)          defining the sample ratio
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
Keeping the matches within sample
keep if sample1==1
keep if sample2==1
Including the population estimates
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
```

```
        Comparing sample estimates with population estimates
        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)
 Defining the number of sampled relationships
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)*/
```