Mis-measurement and Social Networks: Evidence from survey and census data in Ethiopia

Pramila Krishnan
University of Cambridge

Irina Shaorshadze
University of Cambridge

Abstract

We use data on informal networks of relatives and people relied upon for help obtained from both a household survey and a separate census in a single village in Ethiopia to examine the robustness of data on sampled networks. We study three issues: first, we examine the role of measurement error in links; second, we examine the difficulties of ascribing unilateral and bilateral links and finally, we examine the relationship between degree and economic outcomes and contrast the sample data with information from the census.

1 Introduction

We have seen a tremendous increase in the empirical study of informal networks in developing countries [5],[9], . With this growth in the literature, there is new interest in whether the information on networks derived from surveys is reliable enough to afford accurate inference on behaviour and outcomes. In general, data derived from household surveys based on large-scale data collection do allow accurate inference on many aspects of economic interest but empirical work on networks is probably the least well established It is increasingly clear that traditional sampling methods, while inexpensive and easy to implement, are unlikely to result in a 'representative' network of the population network, because the sampling frame is often based on that used for the sampling of households, and consequently, a subset of nodes is sampled and are asked to name connections and (direct) links to other nodes It might thus omit critical nodes such as moneylenders or only capture some nodes and those of their links that can be identified within the sample, thus giving us a sparse and potentially biased view of the network graph ([16], [?],). There is now a small set of papers that investigate these and related difficulties in the empirical study of networks ([3], [6], [?], [16], [14]) and the aim here is to examine similar issues using data on a variety of informal networks.

We use data on networks collected based on a well-established panel survey of households in rural Ethiopia [1]) and a separate census on all informal networks in a singe village of a15 village survey, to examine potential biases in measurement and inference,linked to having access only to a survey-based and not a census based network . In this paper, we describe the difficulties in using data from a sample in comparison to a census, and use as illustration two examples of informal networks. The first is usually described as an informal insurance network and usually refers to the persons who can be relied on to help in times of trouble. ([12], [8]) and the second is simply the set of relatives in the village - but the results we see here extend to the other informal networks observed in this context In brief, we concentrate on informal insurance links and the network of relatives and reserve discussion of other informal networks such as labour sharing, or fully connected groups such as funeral societies and Roscas. We focus on a number of standard issues in empirical studies on networks, such as the measurement of links, whether links are reciprocal and the relationship between network variables and outcomes. We ask whether a sample-based approach results in incorrect inference due to non-random measurement error. In the next section, we introduce the notation and the data. followed by a discussion of summary network characteristics in section 3. In section 4, we set up some of the empirical tests and explore the results. We conclude that measurement error linked to sample-based approaches is likely to be a serious issue.

1.1 Notation and description of data

We represent a network by a graph (N, g), which consists of a set of nodes N $= \{1,2,...n\}$ and an n x n matrix G= [g_{ij} $]_{i,j\in N}$ (adjacency matrix), where g_{ij} $\in \{0,1\}$ represents the potential link of an edge between nodes i and j[2] . Define a directed graph (or digraph) if $g_{ij} = g_{ji}$ and an undirected graph if $g_{ij} = g_{ji}$ for all i , j $\in$ N. Alternatively, we might define a network or a graph as a pair G = (N; E) consisting of a set N of nodes and a set E of edges, where E might be directed or undirected. Let w(G) denote graph-level network statistics for the network G, which include summaries such as the diameter, average path length, average degree and average clustering.

Let w_i(G) denote node-level network statistics for node i and network G, such as local degree, clustering, betweenness centrality and path length. It is useful to define these terms, which are common in the network literature but jargon elsewhere. The degree of node i, d_i(g), is the number of edges that involve i or in brief the number of connections mentioned by i. A directed graph has two measures of degree: outdegree, which is the number of links mentioned by i and indegree, the number of links who mention i. Undirected graphs would thus have the same out and in degree. The degree distribution, P(d), of a network is a description of relative frequencies of nodes that have different degrees d. Let l (i, j) denote the length of the shortest path between node i and j (or the distance

[1] http://www.ifpri.org/dataset/ethiopian-rural-household-surveys-erhs-1989-2004

[2] The edge weight g_{ij}> 0 can also take on non-binary values, representing the intensity of the interaction, where (N, g) is a weighted graph.

between i and j). The diameter of a network is the largest distance between any two nodes in the network, while the average path length is the average distance between any two nodes in the network and is bounded above by the diameter. Clustering measures the extent to which my friends are friends with each other too and captured by the overall clustering coefficient Cl(g), which measures the fraction of triples that have their third edge filled in to complete the triangle. It is defined as Cl(g)= $\frac{3(\text{number of triangles in G})}{\text{number of connected triples of nodes}}$ where a "connected triple" refers to a node with edges to an unordered pair of nodes. The equivalent individual clustering for a node i is $Cl_i(g) = \frac{\text{number of triangles connected to vertex i}}{\text{number of triples centered at i}}$, with the average defined over all nodes n. Another useful summary is centrality, which captures the importance of a node's position in the network. Thus degree centrality is given by $d_i(g)/n - 1$, while closeness centrality captures how close a given node is to any other node: so for node i , one such measure is $\frac{n-1}{\sum_{j \neq i} I(i,j)}$, and betweenness centrality captures how well situated a node is in terms of the paths that it lies on between nodes.

It is clear from the definitions above that the subgraph might easily fail to describe the full graph. All these measures require information on both degree and the paths connecting nodes. The first is often underestimated, particularly in directed graphs, while the sparseness of paths means that measures such as clustering and betweeness measured on a subgraph might bear little relation to the complete graph. In what follows, we begin with a description of these measures in both illustrations of the networks we examine, followed by a discussion of the implications for network formation and economic outcomes. We do so, based on two kinds of data. In the first, we use a sample of a set of m nodes, which represent the households from the survey sample where each sampled node was asked about their connections, which were then matched where possible with the other m -1 nodes in that data set. This is the subgraph which restricts the network among those who are sampled. This (sub)network is contrasted with the full set of M nodes of all resident households from the village (network), which we call the census: each node was asked to name the entire set of social connections to anyone in the entire village (network). The latter may be regarded as a sample where network links reach outside the village but the number of such links are few relative to the difference between the subgraph and the (almost) complete graph and will be neglected below. In this respect, it might be useful to think of the village as akin to Crusoe's island.

The data derive from a household survey in Ethiopia in 2004 (Ethiopian Rural Household Survey, ERHS), part of a panel survey of households begun in 1994 and repeated every 5 years, with the last round in 2009. We use the cross section in 2004, because the census of networks was also conducted in 2004, shortly after the end of the household survey. Apart from the data on multiple informal networks for sampled households, data on network connections was also obtained for all households in a single village in the sample, Sirbana Godeti. In brief, we have complete data on network connections and basic household endowments for all informal networks for the 250 households in this village, as well as more detailed data on households including consumption and incomes for the sample of 76 households in the village. In what follows we study three issues: first, we examine the role of measurement error in links; second, we examine the difficulties of ascribing unilateral and bilateral links and finally, we examine the relationship between networks and economic outcomes, contrasting the use of limited sample information with data from the censuswe study three issues: first, we examine the role of measurement error in links; second, we examine the difficulties of ascribing unilateral and bilateral links and finally, the relationship between insurance networks and economic outcomes.. In each case, we examine the role of mis-measurement in links and/or limited information.

It should be noted that the census data on networks were collected roughly 4 months after the end of the household survey. The same team that interviewed the households also took on the interviews of all the households and their network connections. In each of these cases, there is little evidence to indicate that enumerators were more or less careful: in fact, enumerators knew many of the households well and were themselves puzzled by the many discrepancies in data that they collected. It is to these discrepancies that we now turn.

2 Summary measures of networks

We begin with a brief description of the two main types of networks we focus on here. The list of relatives in the village in the census is obtained as answers to the question," "List the households in the village that you consider relatives (zemed). Specify if they are also relatives by blood or marriage". In this context, the term for relatives, *zemed*, also includes godparents and god children In the household survey, for each of the informal networks, respondents were asked if the person named as a link was a relative and the list of relatives was backed out of these answers. The second network is usually treated as a measure of access to informal insurance and is made up of those links whom the node could rely on. The precise question posed in the census is " List the households whom you can rely on in case of need. If they are not from the village, specify whether they are related to you by blood or marriage", while that in the survey was posed as "We want to know about the FIVE most important people you can you rely on in time of need for support, both within the village or elsewhere". Respondents were also asked to say what they knew about the person, including information about demographics, land, oxen owned, whether they were neighbours or relatives and whether they had any other links and exchanges with them. In brief, even if the individual links were not part of the survey, there was information on their key endowments. This is similar to data collected in other surveys such as that by Udry and others in Ghana[3] but goes further since we also know the endowments

[3]The data on network links provides us with a random sample.... The data are 'egocentric' in the parlance of network theory: we know (in principle) about the links from our sample individuals to other people. If one of these links happens to be to another individual in our sample then we know a great deal about both nodes of that link, and about further connections along the network. "However, if the individual at the other end of the link is not a member of our sample, then our information is quite limited. In particular, we know nothing about further links along the network." ([18] pages 5-6)

of those links not actually part of the household survey.

Tables 1 and 2 offer a simple summary of the discrepancies between the survey and the census. The comparison of survey and census responses is revelatory. There are two important sources of mis-measurement in these data. The first, as shown in Table 1, is simply the underreporting of links in the survey as opposed to the census and this is to be expected and has been commented on, particularly in the sociological literature on networks (see [2] [?]). The second is the fact that there is a substantial difference between directed and undirected links as in Table 2,and again this is similar to other recent studies ([17], [6] [7]), but less studied. However, even more striking here is that even within the networks of relatives, we have a clear distinction between directed and undirected links which might be rather odd since marriage, blood and godparents are bilateral relationships. The measurement or censoring error in listing of links is compounded by the difficulty that it cannot be assumed that even obviously bilateral links such as relatives are often directed[4]. The data suggest that even links with relatives, require some investment to be maintained as an acknowledged link. Finally there is the usual difficulty of those claiming no links at all which in turn might well be spurious. To the extent that both the census and the survey support each other on these lack of links, there is little to be done. This is the classification error writ large.

Tables 1a and 1b thus present the simple evidence of the two main difficulties in the data. Table 1a suggests that the classification error seems to be asymmetric and while this asymmetry is driven by the censoring bias explicit in the survey question, it is not the case that all errors are in this direction. There are also claims of links in the survey that have not been repeated in the larger census and they are a quarter of the discrepancies observed. Table 1b shows that only 20% of relatives claim reciprocal links and an even smaller fraction of about 11% of insurance links are reciprocated. Tables 2a and 2b examine the degree distribution if we make the assumption that links of relatives or those who are in informal insurance arrangements can be assumed to be undirected. A useful comparison here might be between the undirected distribution in the sample survey and the census distributions, whether directed or not. One could argue that the censoring or recall bias can be adequately dealt with by assuming that the aggregated links across nodes, assuming reciprocity can capture the censoring bias or go some way in accounting for it. However, a comparison of the distributions suggests this is unlikely and a simple test of independence between the distributions cannot be rejected at between 5 and 10% significance, depending on the test statistic used[5].

Figure 1a offers a view of the clustered nature of insurance networks and contrasts this with the rather sparse picture in Figure 1b obtained by answers to the question in the household survey that censored responses to utmost 5

[4] There is a small literature on the econometrics of misclassification ([4]). We examine the potential for remedying misclassification in an extension of this paper.

[5] One could use various tests for ordinal data such as Kendall's tau-b. The test results range from Pearson chi2(64) (Pr >p)= 0.03 to Goodman's gamma with an ASE = 0.124 or Kendall's tau-b = with ASE = 0.102.

households. It is useful to examine this picture together with Figure 2 which describes the degree distribution in the census. The sharp fall in the frequency of degrees past 5 suggests that the censored distribution should not have resulted in quite as sparse a picture of links as that seen in Figure 1b. and is perhaps due to the sampling frame based on households.

Table 2 offers a different perspective on the same issue. We examine whether basic features of the network such as degree, clustering, betweenness and number of nodes can be captured by random sampling of network nodes and whether snowball sampling does any better. The snowball sample consists of 2 waves. In the first wave 25% of village households are sampled and in the second, links of first-stage households are sampled. The evidence is suggestive :simple random sampling of nodes fares rather poorly relative to the snowball particularly on capturing degree and the number of nodes. In the next section we investigate whether the pattern of discrepancies can be written off as random noise or whether there is a systematic pattern in the omission of links.

3 Are the differences in measured links random?

We begin by examining the relationship between a declared link in the survey and the link declared in the census. These are both directed links and we reserve discussion of whether they are reciprocated or not for later in this section. This is a dyadic regression, where we use all potential pairs of links using households in the sample survey of the village, which gives us approximately 78X78 potential links, a fraction of which are mentioned as an existing link in the survey. This vector of pairs makes up the dependent variable[6]. The question is whether, for this set, the bulk of explanation is taken up by the links actually mentioned in the census of all households, where there was no censoring of the number of links that could be mentioned by households. If the reason for censoring was simply recall bias, one would not expect a consistent pattern in the remainder of the variation that could be described by the endowments of the households or those of their potential partners. Note that we have full information on basic endowments such as demographics, land, livestock and whether they are also relatives for all potential partners from the data from the census. The demographic variables include age, sex, whether the household head was born in the village (which captures their migration status and hence their within-village connectivity), and whether the potential links are relatives (as acknowledged by the node or ego i). We also use the sum and differences of their endowments of land, the number of adult males (which captures labour supply), and livestock[7].

[6] The actual number of observations is lower becuse of missing information on some endowments.

[7] Livestock is both a productive asset and a measure of liquid wealth in this context. A more direct measure of productive assets is the number of oxen and in what follows we use one of thes measures. Ethiopian households only have user rights to land and thus while land does affect incomes, it is not a direct measure of wealth. It is also often allocated on the basis of household size though certainly by 2004 the reallocation of land on this basis had been suspended.

We also have information on the health of each partner, proxied by whether they can hoe a field for a day[8].

Table 4 thus presents the results of a logistic regression, where we include both total endowments of all partners and their differences in endowments (i-j) in the tradition of directed dyadic regressions[10] . The encouraging news here is that the survey is more likely to pick up links when they are also reported in the census and more interesting, when they are supported by a relative recognised by the node as well as when there are differences in age. There are strong positive level effects from all the endowments as well as sharing similarities such as both heads of households being non-migrants to the village, which also raise the likelihood of returning a link in the survey, But the striking result here is that differences in livestock value and land reduce the likelihood of reporting a link in the survey. Differences in wealth are clearly vital and where ego i is wealthier than his partner j, he is more likely to censor him from his list. Note that we do not investigate this mis-measurement again with the data on relatives since that information was not asked directly in the survey but was backed out from the relationship mentioned when questioning households about their informal networks. This is in contrast to the census where each household was asked to list their entire set of relatives in the village.

A similar exercise is to examine the differences in degree as reported in the survey and the census. Clearly, the censoring matters but it should also be noted that most respondents censored their list well below the maximum of 5. In Table 5, we present the results of a right censored Poisson regression, examining the relationship between the out-degree measured in the survey and that in the census, again as a function of own endowments and the average endowments of partners. Note that this is not a dyadic regression but is estimated at the household level, where we now have the smaller sample size of 68. We also lose some variables that are perfectly collinear due in part to the small sample and because we are aggregating across all endowments of stated links and using the average.

The main result is that those with more land and livestock are also likely to underreport the links as are those who are more educated. This pattern is echoed in relation to the partner's endowments and so reported degree is falling in the average land and livestock of the partners. However, being a relative strongly increases the reported degree. This is entirely consistent with the pattern found earlier in the kinds of links that are likely to be omitted and suggests that measured degree is not going to suffer from non-random measurement error. This is a serious issue since it is also the key variable usually used to capture network characteristics and hence a potential source of bias [3]. We now turn to the question of whether links are reciprocal and if not, why they might not be so.

[8] This was the final and most testing of five questions aimed at assessing fitness and included questions such as whether the person could stand up unaided after sitting down.

4 What affects reciprocity of insurance links and reciprocity for relatives?

We now turn to the second type of dissonance in the data, namely whether links in both types of networks are reciprocated or not. The reason for thinking of this as dissonant is that in theory, an insurance link - and even more sensibly a relative, ought to be reciprocal, .but as in Table 2, the majority of relationships are directed. There are now three papers that explore these difficulties ([6], [17] [7]) and we proceed in a similar fashion. Schecter and Yukusavage[17] examine credit relationships, which in principle might be either reciprocated or not, while Comola and Fafchamps[6] [7] examine rely-on or insurance networks of the kind discussed here. In addition, as mentioned before, we also examine networks of relatives which ought to be reciprocal but are clearly not so, with a margin that rules out the possibility of recall or other error. In fact, in both networks, as in Schecter and Yukusavage's credit network, the differences are systematic and offer room for explanation.

We examine the correlates of reciprocal or undirected links versus directed links for both insurance and links of relatives in Tables 6 and 7[9]. Here, we again examine dyadic links, this time dividing them into three groups, where the base group is not having a link relative to either a directed or undirected link. Both tables report the result of a multinomial logit regression with the main covariates being total endowments including the total age, education, land and livestock as well as the sex of the partners (almost always the heads of the households) and the strength and fitness of the partners. We use the variable whether both spouses were born in the village to capture attachment to the village and also define a variable to capture similar status in terms of occupying a position of influence, whether formal or informal. Differences in endowments enter in absolute terms in determining reciprocal links but are (signed) differences in determining directed links [17]. We also report the relevant marginal effects.

The pattern in Table 6 is similar to the finding for credit networks mentioned earlier and might also be in accord with Comola and Fafchamps' explanation of the 'desire to link' driving directed links. First, directed links are more likely if the ego (node i) has less land, livestock, male adults and is less powerful. The directed link is also likely to be a directed relative link. Furthermore, the effects of total land suggest that this is not wealth calling to wealth and is asymmetric in wealth. For undirected or reciprocated links, the strongest predictor is a reciprocal relative link and the likelihood of such a link is also increasing in total livestock, relative to the difference in livestock. The impact of the relative link can be thought of as support for the insurance link as predicted in theoretical models of network formation[10]. The effect of power is weakly asymmetric but might be affected by the small number of equally powerful and suggests simply

[9] We also did the same exercise for the sample and the results are similar to the ecnsus so have not been reported,

[10] Jackson et al. [11] show that networks with two-way flows should exhibit high levels of support. Bala and Goyal [1] and Galeotti[13] show that networks with one-way flows may exhibit a star-like structure.

that at least one of the partners is someone important in the village but probably not very different in other endowments. In brief, asymmetry in links is driven by asymmetry in wealth and status.

Table 7 provides a similar set of covariates to explain asymmetry in links to relatives. Again, directed links are less likely to be claimed if node i is wealthier than his partner. For reciprocated links, large differences make a link less likely, but symmetry in other endowments or high total endowments encourage reciprocity.

To summarise, asymmetry is driven largely by asymmetry in wealth and other endowments. As in Schecter and Yukusavage[17], we believe that the theory underlying directed link formation offers support for the patterns we see here.

5 Does mis-measurement matter for outcomes?

We now turn to examining the effect of mis-measured degree on outcomes. In what follows, we examine the effect of undirected degree on outcomes, to capture the total number of potential links whether relied upon or relying on. The idea here is that these outcomes are general and do not depend obviously on the direction of links. Table 8 offers two different kinds of outcomes. The first measures subjective well-being and is the answer to "“Suppose we say that the top of a ladder represents the best possible life for you and the bottom represents the worst possible life for you. Where on the 9-step ladder do you feel you personally stand at the present time?” (Circle the selected number)."[11]". The second is the answer to whether they would be able to raise 100 birr if it was necessary. Note that 100 birr represents twice the per capita poverty line or twice the ability to command a minimal basket of goods. The first regression is estimated using an ordered logit while the second is an unordered logit regression. The estimation is at the household level, using the households survey, giving us 65 observations. The main covariates, apart from degree are household endowments that might be thought to affect subjective well-being or the ability to raise a significant sum in a hurry. We include demographic variables such as sex, household size, and age of the head of household. Also included are measures of wealth such as land and livestock and whether a powerful member of the community to capture status. We also include the number of relatives underlying the 'rely on@ network which is perhaps less exogenous than one might wish for but is used to capture the degree of support or social collateral[15].

The main comparison here is between the coefficients on measured degree. The census measure of degree suggests that the higher the number of total connections, the higher is subjective well-being while that on the survey measure is insignificant and near zero in size. Both more land and more adult males in

[11] We find similar results for an alternative measure "Please imagine again a nine-step ladder, where on the bottom, the first step, are those who are totally unable to change their lives, while on step 9, the highest step, stand those who have full control over their own life. On which step are you?".

the family raise well-being and are similar in both the sample based regression and that of the census. Relatives supporting the informal insurance network also raise well-being. The examination of whether a person can raise 100 birr in a hurry suggests that the higher the degree, the more likely the person can do so while the survey measure is insignificant again. The other coefficients are largely sensible and again, do not differ between the two sets of estimates. They suggest that more adult men in the family are also helpful in raising such a sum. These effects are largely suggestive since they are based on a small sample but the aim was merely to ask whether simple summary measures of networks such as degree might lead as astray in a sample and the answer appears to be a definite yes. In each of these examples, the network connections would simply have appeared not to affect outcomes at all.

6 Conclusion

We have presented a very brief summary of the difficulties bound up in empirical work on networks. The key point here is that measurement error is far from random and is related systematically to the incentives in forming and maintaining links. There are clearly large pitfalls in trying to extrapolate both network features and behaviour from survey data. The key features of networks including the role of reciprocity and network characteristics that might carry over from survey evidence require careful investigation.

References

[1] Bala, V. & Goyal, S. (2000), ‘A noncooperative model of network formation’, *Econometrica* 68(5), 1181–1229.

[2] Borgatti, S. P., Carley, K. M. & Krackhardt, D. (2006), ‘On the robustness of centrality measures under conditions of imperfect data’, *Social Networks* 28, 124–136

[3] Chandrasekhar, A. & Lewis, R. (2011), Econometrics of sampled networks. Unpublished Manuscript

[4] Chen, X, H. Hong, and D. Nekipelov. 2011. "Nonlinear Models of Measurement Errors." *Journal of Economic Literature,* 49(4): 901-37.

[5] Chuang, Y and Schecter, L., 2014, "Social Networks in Developing Countries", Forthcoming

[6] Comola, M. and M. Fafchamps (2014), ‘Testing Unilateral and Bilateral Link Formation’, *The Economic Journal,* 124: 954-976.

[7] Comola, M. and M. Fafchamps (2014), Estimating Mis-reporting in Dyadic Data: Are Transfers Mutually Beneficial?, Manuscript

[8] De Weerdt, J. & Dercon, S. (2006), 'Risk-sharing networks and insurance against illness',*Journal of Development Economics* 81(2), 337–356.

[9] Durlauf, S. N. and M. Fafchamps (2004). Social capital. *Handbook of economic growth*. S.Durlauf and P. Aghion.

[10] Fafchamps, M. & Gubert, F. (2007), 'The formation of risk sharing networks', *Journal of Development Economics* 83(2), 326–350.

[11] Jackson, M. O., Rodriguez-Barraquer, T. & Tan, X. (2012), 'Social capital and social quilts: Network patterns of favor exchange', *American Economic Review* 102(5), 1857–1897.

[12] Fafchamps, M. & Lund, S. (2003), 'Risk-sharing networks in rural Philippines', *Journal of Development Economics* 71(2), 261–287.

[13] Galeotti, A. (2006), 'One-way flow networks: The role of heterogeneity', *Economic Theory* 29, 163–179.

[14] Hogset, H. & Barrett, C. B. (2010), 'Social learning, social influence, and projection bias: A caution on inferences based on proxy reporting of peer behavior', *Economic Development and Cultural Change* 58(3), 563–589.

[15] Karlan, D., Mobius, M., Rosenblat, T. & Szeidl, A. (2009), 'Trust and social collateral',*Quarterly Journal of Economics* 124(3), 1307–1361

[16] Santos, P. & Barrett, C. B. (2008), What do we learn about social networks when we only sample individuals? Not much. Unpublished Manuscript.

[17] Schechter, L. & Yuskavage, A. (2012), 'Inequality, reciprocity, and credit in social networks',American Journal of Agricultural Economics 94(2), 402–410

[18] 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

Table 1: Mis-measured links

Number of links mentioned in census but not in sample survey	197
Number of links mentioned in sample survey but not in census	64

Table 2: Distribution of links in census data: Directed versus Undirected

	Relatives	Insurance
Unilateral (directed)	974	597
Bilateral (undirected)	298	66
Total	1570	729

Figure 1a: Census of insurance links[1]

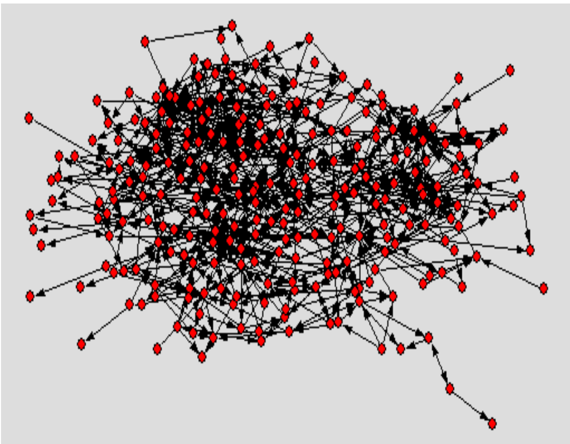
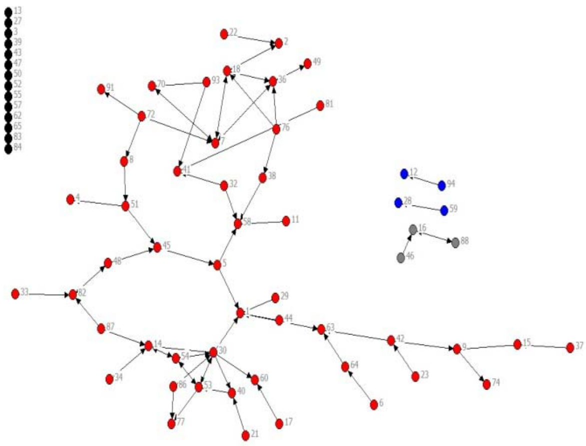


Figure 1b: Sample of insurance links[2]



[1] This is based on the answers to the following question in the census of networks: "**List the households whom you can rely on in case of need. If they are not from the village, specify whether they are related to you by blood or marriage**" In addition, for each name mentioned, respondents were asked about the main endowments of the person (age, education, number of male adults, number of oxen, land, whether a neighbour or relative, whether linked for other reasons).

[2] This is based on responses from Part IV, section 3: Networks, Round 6(2004) of the ERHS, to the question, "**We would like to know about the FIVE most important people you can you rely on in time of need for support, both within the village or elsewhere."**

Figure 2a: (Census) Degree distribution of relatives[3]

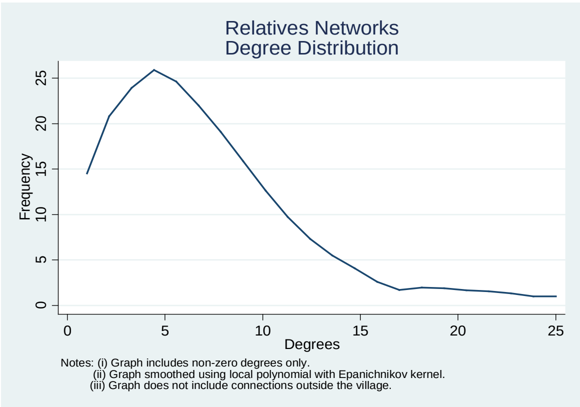


Figure 2b: (Census) Degree distribution of insurance links

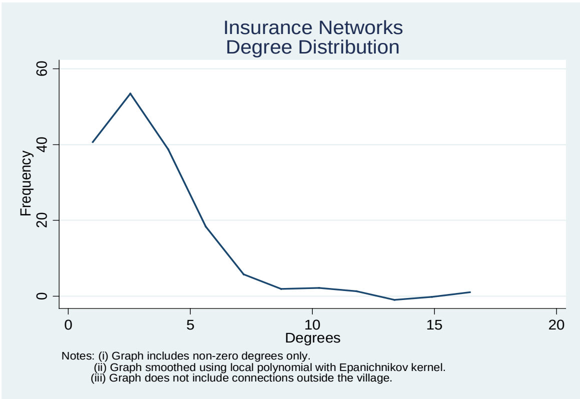
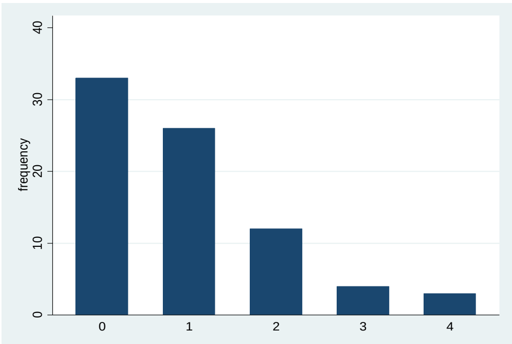
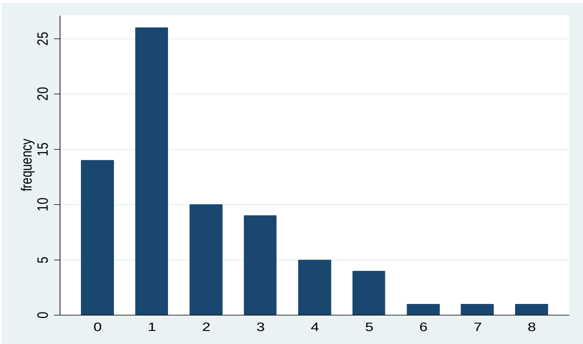


Figure 3: Degree distribution of directed and undirected insurance links: Household Survey sample versus Census

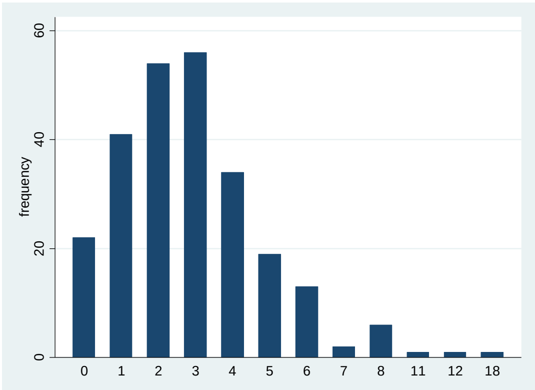
Household Survey: Directed links



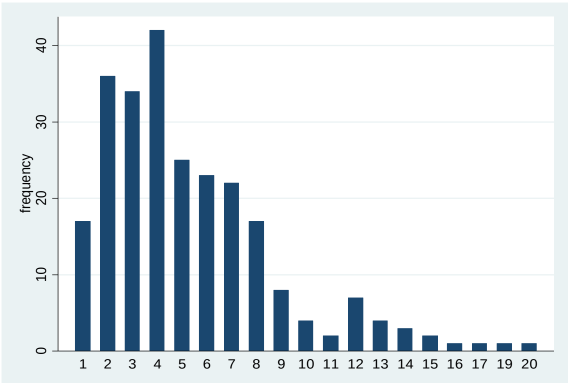
Household survey: Undirected links



Census: Directed links



Census: Undirected links



[3] This is based on the answers to the following question in the census of networks: "**List the households in the village that you consider relatives (zemed). Specify if they are also relatives by blood or marriage**".

Table 3: Robustness of network characteristics to sampling

	Relatives			Insurance		
	Census	Sample (from census)	Snowball	Census	Sample (from census)	Snowball
Average degree	6.65	4.69	6.26	3.53	4.69	3.00
Density (loops)	0.014	0.14	0.014	0.01	0.01	0.01
Diameter	10	12	10	12	12	13
Clustering	0.0	0.0	0.0	0.0	0.0	0.0
Centrality (betweenness)	0.05	0.09	0.07	0.10	0.09	0.13
Number of Links	1,615	788	1,340	772	380	490
Number of nodes	486	336	428	438	277	326

Notes: (i) Note that the network does not include connections outside the village

(ii) Sample consists of 50% randomly drawn households in the village. Their out-of-sample connections are included

(iii)The snowball sample consists of 2 waves. In the first wave 25% of village households are sampled and in the second, links of first-stage households are sampled. (Out-of-sample connections of the second stage households are included.)

Table 4: Are differences in being linked in census and sample random? No

Logistic regression: Whether linked in survey sample (directed link)				
	Coefficient	Std. Error	z	Pr>z
Whether linked in census (directed link)	1.78***	0.51	3.48	0.00
Difference in age	0.02**	0.01	1.71	0.08
Difference in education	0.02	0.05	0.38	0.70
Difference in land	-0.09**	0.05	-1.73	0.08
Difference in number of adult men	0.07	0.10	0.69	0.49
Difference in livestock value	-0.53***	0.19	-2.78	0.01
Directed relative link	0.96***	0.46	2.11	0.04
Maximum (spatial) distance between links	0.07	0.33	0.21	0.83
Total age	0.00	0.01	0.48	0.63
Total education	0.09**	0.05	1.83	0.07
Total land	0.11***	0.04	2.64	0.01
Total livestock value	0.41***	0.17	2.47	0.01
Both heads born in village	0.96***	0.37	2.57	0.01
Same sex of head of household	0.26	0.30	0.85	0.39
Both heads able to hoe field for a day	0.62***	0.28	2.19	0.03
Same status (power)	0.07	0.28	0.25	0.80
Constant	-15.42	3.74	-4.13	0.00
Number of observations: 4818	**Wald Chi2(16): 108.8**			
	Pseudo R^2: 0.17			

Note: *** denotes significance at 5% or better, **at 10% and * at 15%.

Table 5: Are differences in measured degree in survey and census random ? No.

Right-censored Poisson regression				
Number of directed links (survey)	Coefficient	Std. Err.	t	P>t
Own endowments				
Number of directed links (census)	0.08	0.05	1.43	0.15
Sex of household head	0.83**	0.45	1.85	0.06
Education in years	-0.29***	0.13	-2.16	0.03
Land	-0.23*	0.15	-1.59	0.11
Ability to hoe field for a day	-0.38	0.36	-1.03	0.30
Number of adult males	-0.24	0.24	-1.01	0.31
Value of livestock	-0.00*	0.00	-1.59	0.11
Partners' average endowments				
Education in years	-40.72***	15.06	-2.7	0.01
Land	-17.23	17.02	-1.01	0.31
Number of adult males	-65.95**	38.01	-1.73	0.08
Oxen	-99.32***	37.39	-2.66	0.01
Relative	13.30***	6.09	2.19	0.03
Number of observations: 68	**LR chi2(14)=30.8, Pr(>chi)=0.1 Pseudo R^2=0.17**			

Note: *** denotes significance at 5% or better, **at 10% and * at 15%.

Table 6: Why are insurance links not reciprocal?

Type of link	Coefficient	Std. Error	z	P>z	Marginal effects	
Directed link: Δ denotes actual differences					effect	p-value
Total age	0.004	0.00	1.33	0.18		
Total education	-0.003	0.01	-0.23	0.82		
Total land	-0.017	0.01	-1.63	0.10	-0.0001	0.09
Total livestock	0.020	0.01	1.83	0.07	0.0002	0.10
Spouses born in village	-0.027	0.10	-0.26	0.79	-0.0003	0.76
Same sex of head	-0.002	0.10	-0.02	0.98	-0.0001	0.90
Same strength	-0.023	0.11	-0.2	0.84		
Same power	-0.647	0.11	-5.69	0.00	-0.0053	0.00
Δage	-0.005	0.00	-2.2	0.03	0.000	0.45
Δeducation	-0.051	0.01	-3.83	0.00	-0.0004	0.00
Δland	-0.026	0.01	-2	0.05	-0.0002	0.05
Δlivestock	-0.020	0.01	-1.79	0.07	-0.0002	0.07
Δmale adults	-0.116	0.04	-3.01	0.00	-0.0010	0.00
Δfemale adults	-0.064	0.05	-1.26	0.21		
Distance	0.087	0.06	1.37	0.17	0.0007	0.20
Relative	3.895	0.10	40.24	0.00	0.0315	0.00
cons	-5.387	0.36	-14.9	0.00		
Undirected link: Δ denotes absolute differences						
Total age	0.004	0.01	0.8	0.43		
Total education	0.013	0.03	0.4	0.69		
Total land	0.023	0.02	1	0.32	0.0001	0.27
Total livestock	0.047	0.04	1.15	0.25	0.0001	0.28
Same sex	0.130	0.23	0.56	0.58	0.0006	0.21
Similar strength	0.326	0.26	1.27	0.20		
Spouses born in village	0.159	0.25	0.63	0.53	0.0003	0.56
Same power status	-0.682	0.23	-2.98	0.00	-0.0012	0.01
Δage	0.000	0.01	0.03	0.98	0.0000	0.98
Δeducation	-0.010	0.04	-0.23	0.82	0.0000	0.82
Δland	-0.009	0.04	-0.25	0.81	0.0000	0.81
Δlivestock	-0.036	0.05	-0.69	0.49	-0.0001	0.49
Δmale adults	-0.179	0.11	-1.64	0.10		
Δfemale adults	-0.001	0.12	-0.01	1.00		
Distance	0.139	0.12	1.19	0.24	0.0002	0.28
Relative (reciprocal)	4.594	0.22	21.22	0.00	0.0079	0.00
cons	-8.280	1.00	-8.24	0.00		
Observations: 56693		Wald chi^2(32)=2600.8, Pr(>chi)=0				

Note: *** denotes significance at 5% or better, **at 10% and * at 15%.

Table 7: Why are relatives' links not reciprocal?

Type of link	Coefficient	Std. Error	z	P>z	Marginal effects	
Directed link: Δ denotes actual differences					effect	p-value
Total age	0.003	0.00	1.47	0.14	0.000	0.156
Total education	0.023	0.01	2.32	0.02	0.000	0.03
Total land	0.005	0.01	0.55	0.58	0.000	0.60
Total livestock	0.022	0.01	2.74	0.01	0.000	0.01
Same sex of head	0.270	0.08	3.41	0.00	0.004	0.001
Same strength	0.168	0.08	2.00	0.05	0.003	0.05
Spouses born in village	0.486	0.08	6.35	0.00	0.007	0
Same power status	-0.468	0.09	-5.36	0.00	-0.007	0
Δage	-0.003	0.00	-1.74	0.08	-0.000	0.083
Δland	-0.013	0.01	-1.30	0.19	-0.000	0.194
Δeducation	-0.019	0.01	-1.99	0.05	-0.000	0.047
Δlivestock	-0.020	0.01	-2.40	0.02	-0.000	0.017
Δmale adults	-0.139	0.03	-5.35	0.00	-0.002	0.001
Δfemale adults	-0.106	0.03	-3.21	0.00		
Distance	0.124	0.05	2.59	0.01	0.002	0.01
Constant	-5.218	0.28	-18.76	0.00		
Undirected link: Δ denotes absolute differences						
Total age	0.009	0.00	3.06	0.00	0.000	0.00
Total education	0.079	0.02	4.30	0.00	0.001	0.00
Total land	0.022	0.01	1.98	0.05	0.000	0.05
Total livestock	0.038	0.02	2.45	0.01	0.000	0.02
Same sex	0.545	0.12	4.67	0.00	0.005	0.00
Similar strength	0.304	0.13	2.34	0.02	0.003	0.02
Spouses born in village	0.805	0.11	7.16	0.00	0.007	0.00
Same power status	-0.487	0.12	-4.18	0.00	-0.004	0.00
Δage	0.004	0.00	1.01	0.31	0.000	0.31
Δeducation	-0.019	0.02	-0.79	0.43	-0.000	0.43
Δland	-0.038	0.02	-2.12	0.03	-0.010	0.03
Δ livestock	-0.040	0.02	-1.96	0.05	-0.000	0.05
Δmale adults	-0.06	0.05	-1.25	0.21		
Δfemale adults	0.04	0.06	0.60	0.55		
Distance	0.004	0.06	0.06	0.95		
Constant	-7.078	0.42	-16.69	0.00		
Observations:56693	Wald $chi^2(28)$=548.13, Pr(>chi)=0					

Note: *** denotes significance at 5% or better, **at 10% and * at 15%.

Table 8: Impact of measured degree on outcomes: Census versus sample

Ladder of best possible life (ologit)	Sample			Census		
	Coefficient	Std. Error	Pr(>z)	Coefficient	Std. Error	Pr(>z)
Degree	0.09	0.13	0.51	0.16**	0.10	0.09
Number of adult males	0.58***	0.19	0.00	0.59***	0.19	0.00
Household size	-0.16	0.14	0.24	-0.16	0.14	0.25
Spouse born in village	-0.29	0.38	0.44	-0.35	0.37	0.34
Status	-1.19	0.85	0.16	-1.63***	0.79	0.04
Able to hoe for a day	-1.21*	0.79	0.13	-1.17*	0.75	0.12
Education	0.12	0.14	0.39	0.09	0.13	0.50
Land	0.57***	0.19	0.00	0.49***	0.19	0.01
Age	-0.01	0.02	0.69	0.00	0.03	0.94
Livestock value	0.52	0.38	0.17	0.49	0.34	0.16
Relatives in rely on network	0.15	0.05	0.00	0.14	0.05	0.00
Observations	65			65		
Wald Chisq(11), Pr (>chi)=0	52			52		
Pseudo R^2	0.16			0.18		

Can raise 100 birr	Sample			Census		
Logistic regression	Coefficient	Std. Error	Pr(>z)	Coefficient	Std. Error	Pr(>z)
Degree	-0.10	0.27	0.70	0.33**	0.17	0.06
Adult males	1.15***	0.53	0.03	1.34***	0.51	0.01
Household size	-0.15	0.19	0.43	-0.16	0.21	0.44
Born in village	-0.20	0.73	0.79	0.22	0.71	0.75
Spouse born in village	-0.63	0.46	0.17	-0.58	0.46	0.21
Land	0.29	0.28	0.29	0.10	0.32	0.75
Status (power)	-1.07	1.28	0.40	-2.25	1.59	0.17
Education	-0.09	0.21	0.67	-0.31	0.25	0.21
Age	0.00	0.03	0.93	0.01	0.04	0.83
Livestock value	0.41	0.38	0.28	0.44	0.42	0.30
Relatives in rely on network	-0.05	0.09	0.59	-0.06	0.09	0.49
Constant	-2.25	3.61	0.53	-5.09	4.15	0.22
Observations	64			64		
Wald chi2(10) (Pr>chi)	10.48 (0.39)			18.07 (0.05)		
Pseudo R^2	0.18			0.27		

Note: *** denotes significance at 5% or better, **at 10% and * at 15%.