NBER WORKING PAPER SERIES

NETWORK STRUCTURE AND THE AGGREGATION OF INFORMATION: THEORY AND EVIDENCE FROM INDONESIA

Vivi Alatas
Abhijit Banerjee
Arun G. Chandrasekhar
Rema Hanna
Benjamin A. Olken

Working Paper 18351
http://www.nber.org/papers/w18351

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2012

We thank Ritwik Sarkar, Prani Sastiono, Ririn Purnamasari, Hendratno Tuhiman, Matthew Wai-Poi, and Chaeruddin Kodir for outstanding research assistance and thank Mitra Samya, SurveyMeter the Indonesian Central Bureau of Statistics for their cooperation implementing the project. Most of all we thank Lina Marliani for her exceptional work leading the field implementation teams. Funding for this project came from a World Bank – Royal Netherlands Embassy trust fund. Chandrasekhar is grateful for support from the National Science Foundation GRFP. We thank Matthew Elliott, Ben Golub, Matthew O. Jackson, Chris Udry, participants at the Calvo-Armengol Workshop and NEUDC 2010 for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the World Bank, the Royal Netherlands Embassy, the National Bureau of Economic Research, or any other organization.

Network Structure and the Aggregation of Information: Theory and Evidence from Indonesia
Vivi Alatas, Abhijit Banerjee, Arun G. Chandrasekhar, Rema Hanna, and Benjamin A. Olken
NBER Working Paper No. 18351
August 2012
JEL No. D83,D85,H23,O12

ABSTRACT

We use a unique data-set from Indonesia on what individuals know about the income distribution in their village to test theories such as Jackson and Rogers (2007) that link information aggregation in networks to the structure of the network. The observed patterns are consistent with a basic diffusion model: more central individuals are better informed and individuals are able to better evaluate the poverty status of those to whom they are more socially proximate. To understand what the theory predicts for cross-village patterns, we estimate a simple diffusion model using within-village variation, simulate network-level diffusion under this model for the over 600 different networks in our data, and use this simulated data to gauge what the simple diffusion model predicts for the cross-village relationship between information diffusion and network characteristics (e.g. clustering, density). The coefficients in these simulated regressions are generally consistent with relationships suggested in previous theoretical work, even though in our setting formal analytical predictions have not been derived. We then show that the qualitative predictions from the simulated model largely match the actual data in the sense that we obtain similar results both when the dependent variable is an empirical measure of the accuracy of a village's aggregate information and when it is the simulation outcome. Finally, we consider a real-world application to community based targeting, where villagers chose which households should receive an anti-poverty program, and show that networks with better diffusive properties (as predicted by our model) differentially benefit from community based targeting policies.

Vivi Alatas
World Bank
Jakarta Stock Exchange Building
Tower 2, 12th & 13th Floor
Jakarta, Indonesia
valatas@worldbank.org

Abhijit Banerjee
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
banerjee@mit.edu

Arun G. Chandrasekhar
Microsoft Research New England
agc2104@gmail.com

Rema Hanna
Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
and NBER
Rema_Hanna@hks.harvard.edu

Benjamin A. Olken
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
bolken@mit.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w18351

1. Introduction

Economists are increasingly conscious of the influence that our neighbors and friends exert on our choices. In particular, there is a growing interest in how information is aggregated within the community. Many individuals may have information that is useful to others in their community, but does this information get transmitted to those who need it, either through direct communication between parties or through the observation of their choices? And how does the answer to this question vary with the nature of the social network within the community? Being able to answer these types of questions is important for policy design: for example, recent evidence suggests that the speed with which new agricultural technologies are adopted depends on who talks to whom about what (e.g., Munshi (2004), Bandiera and Rasul (2006), Duflo et al. (2004), and Conley and Udry (2010)). Likewise, social connections have been shown to be important in spreading information about jobs, microfinance, and public health (e.g., Munshi (2003), Bandiera et al. (2009), Banerjee et al. (2012), Kremer and Miguel (2007)).

The increasing trend in developing countries towards the decentralization of policy to the local level – e.g. community monitoring of teachers and health professionals or decentralized budgeting of local public goods – is predicated, in part, on the idea that communities have more information and can more effectively aggregate that information than central governments. For example, decentralization has become increasingly popular for *targeting* the poor for government assistance programs.[1] The idea is that it is costly for the central government to identify the poorest people within a village, whereas the community may have a good sense of who they are, simply by virtue of living next to them. In designing these types of community-based targeting systems, it is crucial to understand how information about poverty flows within villages and how it is aggregated through intra-village processes.

However, despite a number of important and insightful theoretical contributions to this question – many of which are discussed below – laying out a general relationship between a network's characteristics and the extent of information sharing remains challenging due to the mathematical complexity of networks: they can differ along many dimensions, and how each individual network characteristic relates to the degree of information aggregation within the network can depend both on the network structure and the underlying model of social learning.[2] To illustrate this point,

[1] The Bangladesh Food-For-Education (Galasso and Ravallion, 2005), Albanian Economic Support safety net (Alderman and Haque, 2006), and BRAC Ultra-Poor program (Bandeira et al., 2012) are examples of community targeted programs.

[2] Due to the difficulty of describing transitional learning dynamics, much of the social learning literature has focused on asymptotic learning. The early literature on observational learning, where agents observe others' actions and attempt to learn the state of the world through these observations, showed how even Bayesian agents may inefficiently herd and ignore their own information (Banerjee, 1992; Bikhchandani et al., 1992). More recently, Acemoglu et al. (2011) show that under sequential observational learning in stochastic networks, provided that agents have expanding observations, asymptotic learning occurs. Gale and Kariv (2003) look at a special case in which a finite set of individuals in a network each simultaneously take an action in every period having observed their neighbors' actions in previous periods, which Mueller-Frank (2011) extends considerably. Under myopic Bayesian behavior, they provide conditions under which a consensus emerges, making use of the martingale convergence theorem. Mossel et al. (2011) show in a world with binary uncertainty that with probability tending to one a sequence of growing networks which lead to consensus have consensus on the right state of the world. That is, not only is there agreement, but that individuals agree while learning the truth. Meanwhile, another strand of literature studies various rule of thumb social learning processes. For instance Golub and Jackson (2010) and Golub and Jackson (forthcoming) look at DeGroot learning

consider the fact that while more connections typically facilitate better communication, having a higher average number of connections, i.e. higher average degree, is not enough to guarantee better information aggregation. This is made clear by Jackson and Rogers (2007), who require first order stochastic dominance of the degree distribution (which is much stronger than higher average degree) to ensure greater diffusion of information in a meeting model where nodes meet other nodes with probability proportional to their degree. To see why, consider the possibility that there could be a group of people in the community who are all connected with each other (leading to a high average degree), but are entirely disconnected from the rest of the network, making information aggregation very inefficient relative to a network where average degree is lower but there is little clustering in any one part of the network.[3]

In this simple example, the networks differ on both degree and clustering patterns. This suggests that if we want, for example, a general prediction on the effect of degree, we might want to only compare networks that have both similar clustering patterns and other network features. However, there is no one measure of clustering that summarizes all the relevant information, just as no one measure of degree is sufficient (i.e. the variance of degree matters, as do higher moments). In particular, controlling for the average amount of clustering in the network is not sufficient (see, for instance, Jackson (2010), Watts and Strogatz (1998), among others). In the example above, one can even imagine cases where the average clustering in the two networks is the same because everyone outside the one densely connected component in the first network is not connected at all. More generally, real networks differ on so many dimensions that the theoretical results may fail to provide clear predictions as to which networks will experience better information diffusion except in special cases.

In this paper, we take a more rough and ready approach to the problem. As we said above, we are interested in whether it is possible to predict the degree of information aggregation based on network characteristics. We start from an unusual data set which is in many ways ideal for this purpose. We have network data from 631 villages in Indonesia that we collected as a part of a study on the effectiveness of different targeting methodologies. It is very rare to have network data for so many separate networks and without that data it is hard to do a credible cross-village comparison of the kind we attempt here. We also have a natural measure of information aggregation: for a sample of villagers in each village we know how they rank a set of others in their village in terms of relative incomes (i.e. which of the two households is richer). Finally we have a good measure of the actual incomes of those households, from which we generate the “correct” ranking of these households. We use the accuracy of households in ranking others in their village in terms of income as our measure of information diffusion, and ask how this relates to various network characteristics

Our empirical analysis starts with some reduced form evidence that network position is correlated with what people know. This is similar to the results in the literature (see for example, Munshi

on networks and speed of convergence in a model in which, in every period, individuals average the beliefs of their network neighbors and communicate their updated beliefs to their neighbors in the following period. In contrast, Jackson and Rogers (2007), in their seminal paper on this subject, study information transmission as a percolation or contagion process in a model in which the network directs the probability of individuals meeting others; this setup is closest to describing our environment and thus a key focus of our paper.

[3] See also the echo-chamber effect discussed in Golub and Jackson (forthcoming) describing how information aggregation may be slower in more segregated networks.

(2004), Bandiera and Rasul (2006), Kremer and Miguel (2007), Duflo et al. (2004), and Conley and Udry (2010)). Note, however, we do not claim to make any special progress on the important and difficult identification issues (Manski, 1993). Nonetheless, the patterns are strikingly clear and strong. We show that better connected households are better at ranking other households, especially if we measure being better connected by average degree. Similarly households that are socially closer (in terms of path length) to their ranker are more likely to be more accurately ranked. Therefore, there is at least prima facie evidence for the importance of network channels for information transmission.

The main focus of the empirical analysis is, however, on the cross-village comparisons. Ideally, we would like clear theoretical predictions from a diffusion model about what network characteristics should matter. Unfortunately, there is no analytical theory rich enough for our setting. Therefore, rather than getting the predictions of what network characteristics matter (and how) from theorems, we get them by using what we call numerical theorizing.

Specifically, we take the following approach: we use the *within village v*ariation in our data to estimate parameters of a model of diffusion and use that model to predict the extent of information diffusion in every village. We then run regressions to estimate the *cross-village* correlations between network characteristics and the extent of information diffusion. We then use these predicted correlations as our benchmark for what we find when we do the empirical cross-village comparisons. This is what we mean by numerical theorizing: by comparing the reduced form regression estimates to the counterparts generated by simulated data, we can see whether the patterns we pick up in the data are qualitatively similar to those predicted by standard models from network theory.[4]

The model we estimate is inspired by the existing literature that tries to relate information transmission to network characteristics. The closest connections are to Jackson and Yariv (2007); Jackson and Rogers (2007); Galeotti and Vega-Redondo (2011); López-Pintado (2008) – which, in turn, are motivated by Pastor-Satorras and Vespignani (2001), among others. The core idea of these models is that information transmission is like an infection – individuals transmit information with some probability to those they are connected to but also forget information with some probability.[5] We estimate a two parameter information diffusion model based on this assumption; the two parameters are the probability of information transmission from a household that has the information to its neighbor on the network in any period, and the probability that a household that has the information will forget it in any given period. Then we simulate the model for each

[4] Simulations have also been used to study other network phenomena that are too complicated to solve analytically. See, for example, Golub and Jackson (Forthcoming), who use simulations to complement an analytic study of a homophily-based link formation model.

[5] Our modeling strategy therefore respects the local structure of the network: people only get information from people close to them in the network. This in contrast with Jackson and Rogers (2007) who study an approximation of this sort of process on a network. Specifically, they consider a dynamic model in which individuals meet other individuals with probability proportional to their popularity (degree). To gain analytic tractability, they make a mean-field approximation, which means that the local information rate in an individual's neighborhood matches the global average. This approximation implies that the fraction of people that each individual meets in a given period that are informed is equal to the population average of that quantity. As noted by the authors, this means that there is no sense in which individuals that are socially closer to one another are more likely to be informed about each other's information than individuals who are socially farther away. In our set-up, however, this property is clearly violated: as we discuss below in Section 3, our reduced form evidence suggests that local connections matter greatly. For example, households do indeed have more information about those whom they are socially close to.

of the 631 individual networks in our data to generate a predicted level of information aggregation for each network. We then regress the predicted information aggregation on a number of commonly used network statistics (size, average degree, average clustering, first eigenvalue of adjacency matrix, link density and fraction of nodes in giant component) separately as well as jointly, to generate theoretically motivated predictions for the relationship between those networks statistics and the extent of information aggregation.[6] We can then test these theoretical predictions by asking whether the empirical cross-network correlation between the observed degree of information aggregation and network characteristics that we observe in actual data ("the reduced form results") are qualitatively similar to the predictions from the simulated predictions.

The empirical results suggest that the observed patterns match up reasonably well with what our theory predicts for the estimated parameter values. In particular, we show that the Jackson and Rogers (2007) result, on stochastic dominance of the degree distribution described earlier, holds up both in our model and in the data. To the best of our knowledge this is the first "test" of that theory. Moreover we find that for the most part whenever either the predicted (simulated) or the actual (empirical) correlations are significantly different from zero, they have the same sign and this sign matches what we would have expected based on existing theoretical research.[7] For example, networks with larger first eigenvalues exhibit lower error rates, both in the predicted and actual data.

However, we also see interesting divergences from what we might have intuitively expected: For example, the effect of higher average degree on information aggregation, *controlling for other network characteristics,* is negative both in our "numerical theoretical predictions," as well as in the reduced form empirical results. Though there is a standard intuition that more connections are better, this is not true as a conditional correlation.

To make sure that our results are not driven by the specific parameter values we estimate in the diffusion model (especially since the bounds on estimates are not very tight), we redo the cross village simulation and regression exercise for a wide interval of parameter values more or less centered around the estimated values. While the parameter values matter – some correlations change when we approach the boundaries of the interval – the basic predictions turn out to be remarkably robust. This is reassuring in the sense that it suggests our conclusions may be portable. However, the exercise also sheds light on what set of parameter values would overturn the usual intuitions – the qualitative results of the model are largely simliar except for the case when the diffusion process has a very low transmission rate but a higher rate of forgetting.

Finally, we look at how actual policy decisions match up with the predictions from the theory. This data-set comes from an experiment in which villages were randomly assigned to determine eligibility for an anti-poverty program using either community-based targeting, in which a village

[6] The choice of these network characteristics is inspired by important analytical results in the literature on the determinants of information aggregation in networks (even though they cannot be directly applied to our context). For example, Jackson and Rogers (2007), as mentioned above, focus on the effects of first order shifts in the degree distribution, and Bollobás et al. (2010) focus on the role of the first eigenvalue of the adjacency matrix. More generally, though, there are an enormous number of ways of summarizing the properties of the adjacency matrix, so it was impossible to consider all such permutations.

[7] In some instances, the theoretical claims that we have in mind are based on intuitive discussions rather than formal proofs.

meeting ranked households from poorest to richest and assigned benefits to the poorest, or using proxy-means tests (PMT), which assign benefits based on a deterministic function of a household's assets. If we expect that information is efficiently aggregated in the village, we would expect that better connected networks would be relatively better at community-based targeting. This is indeed the case: we show that villages that our network model predicts should have better information passing properties do better in the sense that community targeting better reflects people's self-assessment of their poverty.

Our overall findings are useful for at least two reasons. First, they suggest that the standard intuitions about what the key differences between networks are may not be so far from the truth, despite the absence of general analytical results behind them, at least if the way we model transmission is broadly correct. For example, networks that have higher first eigenvalues of their adjacency matrices do seem to aggregate information better, and probably for reasons that we understand from previous theoretical work (since our network model is closely related to the tractable simplified models used in the literature).[8] Second, the findings highlight the role of social networks in actual community decision making, thus offering insights into policy design problems where governments aim to seek out and harness aggregate local information (e.g. to whom to provide a loan, where local infrastructure should be built) or those that rely on understanding the ways that information spreads within a network (e.g. public health campaigns, agricultural extension programs). They suggest the possibility of using standard network statistics to predict whether in a particular context we would expect effective information aggregation, or conversely, whether some outside intervention will be needed to supplement information flows through the network.

The paper is organized as follows. Section 2 describes the data. Section 3 presents reduced form evidence at the individual level and Section 4 establishes the framework and describes the predictions of the numerical model. Section 5 describes our main empirical results. Section 6 makes the connection with targeting. Section 7 concludes.

2. Context and Data

2.1. **Context.** This study stems from a broader data collection effort that was designed to study the efficacy of different targeting methodologies in Indonesia. Between November 2008 and March 2009, we conducted a randomized evaluation to compare the accuracy of three key common methods to identify beneficiaries for targeted social programs: proxy-means testing (PMT), wherein one collects asset and demographic information on everybody in the census and uses the data to predict consumption; a community targeting approach, wherein decisions on beneficiaries are made in a communal meeting; and a methodology that combined both community and PMT methods (Hybrid). A detailed description and findings from this study are described in Alatas et al. (2012).

In this paper, we utilize the detailed data that we collected on social networks in that study, as well as data on individuals' reports about the relative incomes of other villagers. Below, we first discuss the sample construction. In Section 2.3, we then provide a detailed description of the

[8] Bollobás et al. (2010) show that the appearance of a giant component of a percolation process (wherein most nodes become informed) on a sequence of dense graphs emerges only if the transmission probability is at least as large as the inverse of the first eigenvalue.

survey data collected, describe the construction of the network, and then briefly discuss the design of the targeting experiment. Finally, we report key sample statistics in Section 2.4.

2.2. **Sample Description.** The initial sample consists of 640 hamlets spread across three Indonesian provinces: North Sumatra, South Sulawesi, and Central Java. The provinces were chosen to be broadly representative of Indonesia's diverse geography and ethnic makeup, with one province located on each of the three most populous islands (Sumatra, Sulawesi, and Java). Within these three provinces, we randomly selected a total of 640 villages, stratifying the sample to consist of approximately 30 percent urban and 70 percent rural locations. For each village, we obtained a list of the smallest administrative unit within it (a *dusun* in North Sumatra and a *Rukun Tetangga* (RT) in South Sulawesi and Central Java), and randomly selected one of these units (henceforth "hamlets") for the experiment. The hamlets are best thought of as neighborhoods. Each hamlet has an elected or appointed administrative head, whom we refer to as the hamlet head, and contains an average of 54 households. We make use of 631 hamlets that have network data available.

2.3. **Data.**

2.3.1. *Data Collection.* We primarily use data that was collected as part of the baseline survey for the experiment. SurveyMeter, an independent survey organization, administered the baseline survey in the field in November to December 2008, before any mention of the experiment or the social program were made to villages. For each randomly selected hamlet in the village, we constructed a census of households and then randomly selected eight households to be surveyed. In addition, we always surveyed the hamlet head to obtain the "leadership" perspective. From this survey, we used information on social networks and on both the perceived and actual income distribution within the village.

To construct the social networks (discussed in Section 2.3.2), we used two forms of social connections data. First, we used a series of data on familial relationships within each hamlet. Specifically, we asked each of the surveyed households to name all other households in the hamlet to whom they were related (either through blood or marriage).[9] We then asked the respondent to name the formal and informal leaders, the five poorest households in the hamlet, and five richest households in the hamlet, along with all of the relatives of each person named. Second, we asked each respondent to name the social groups that each household member participated in within the hamlet, and prompted them with various types of groups to ensure a complete list. The social groups included, but were not limited to, neighborhood associations, religious groups, school groups, ROSCAs, farmers' associations, etc.

In this study, we are concerned with how accurately information about the income distribution diffuses within a hamlet. Thus, we needed to construct a measure of each household's beliefs about the income distribution, and needed to compare it to a measure of the "true" income distribution within the hamlet. To collect data on the perceived distribution of incomes, as part of the survey we conducted a poverty ranking exercise where we asked each household to rank the other eight

[9] On the forms, there was room to list up to 10 households in the village. If households were related to more than 10 households in the hamlet, the enumerator could add additional related households to the survey. On average, households reported that they were related to about 3.1 households in the hamlet.

households that were interviewed from their hamlet from the "most well-off" (*paling mampu*) and to the "poorest" (*paling miskin*). Note that this was done before any of the targeting treatments were implemented or even discussed in the village, so individual responses should not be affected by the subsequent targeting experiment.

We then collected two measures of the "true" income distribution of households. First, we collected a measure of actual per capita expenditures levels at the time of the baseline survey, using the standard 28-question Indonesian SUSENAS expenditure module. Second, we asked households to self-assess their own poverty status. Specifically, each household was asked "Please imagine a six-step ladder where on the bottom (the first step) stand the poorest people and on the highest step (the sixth step) stand the richest people. On which step are you today?" Each respondent responded with a number from 1 to 6. We can then construct an error rate for each household's knowledge of the income distribution. Specifically, we compute this measure as the fraction of times that the surveyed household makes an error in the (8 choose 2) comparisons that it makes during the poverty ranking exercise, where the right answer is either per capita consumption or the household self-assessment.[10] Note that we construct a village level error rate analogously.

2.3.2. *Network Data.* The networks utilized in this paper are undirected, unweighted graphs that are constructed from the familial and social group data in a way we now describe. Specifically, we first construct edges between the households that we sample and those that they identify as their family members. Second, we consider each household that was named as one of the poorest or richest, or as a leader by any household we surveyed, and then construct an edge between the named household and all of their named relatives. Moreover, we construct an edge between each pair of these relatives (i.e. if household i is named as being in the same extended family as household j, and household j is separately named (potentially by another respondent) as being in the same extended family as household k, we construct edge (i, k) in addition to (i, j) and (j, k). Third, we construct an edge between any two households who are registered as part of the same social group. Finally, we take the union of these graphs.

Two specifics are worth mentioning. First, the data consists of a set of subgraphs of the target graphs that we are ultimately interested in. As noted in Chandrasekhar and Lewis (2012), regression analysis on partial samples of network data can show biases due to non-classical measurement error.[11] However, on average, we have complete family data on 65 percent of households in each hamlet. In addition, for a number of key quantities and specifications, for instance the first order stochastic dominance of a village's degree distribution against another's, our results are conservative

[10] Note that if a respondent was unable to rank a household during the poverty ranking exercise (i.e. since he or she did not know members from the household or anything about their income level), we assigned this as an "error," i.e., they were unable to correctly rank the households. An alternative would have been to assume the household could have guessed, and gotten it right with a 50 percent probability; the main results look similar if we model error in this manner (see Appendix E), but this introduces more noise into the model.

[11] Most of the bias correction solutions discussed in Chandrasekhar and Lewis (2012) are not applicable as they rely on missing-at-random data. In addition, the estimates in our structural model described in Section 4 are generated by fitting a diffusion process taking place on sub-graphs of the true underlying network which then, in turn, are likely to affect the relationship between the network regressors and the simulated outcomes. We discuss in footnote 18 how this affects our qualitative predictions.

as the bias will generate attenuated coefficients.[12] Second, our data is unique in terms of the sheer number of networks we have at our disposal. Typical papers have very few graphs in their sample (closer to 5 than 50). Having a sample with over 600 networks puts us in a unique position to shed light on questions about how cross-network variation in social structure affects the outcome of a diffusion process.

2.3.3. *Aggregation of Data in Community Based Targeting.* Whether to decentralize "targeting" – the selection of beneficiaries to social programs aimed towards the poor – to local communities has become a key policy question in recent years as household income is challenging and costly to measure. The data used in the paper was collected prior to an experiment in which we compared community targeting with nationally-imposed, data driven approaches. Specifically, in each hamlet, the Central Statistics Bureau (BPS) and Mitra Samya, an Indonesian NGO, implemented an unconditional cash transfer program, where a fixed number of households would receive a one-time, Rp. 30,000 (about $3) cash transfer. The amount of the transfer is equal to about 10 percent of the median beneficiary's monthly per-capita consumption, or a little more than one day's wage for an average laborer. Each hamlet was randomly allocated to one of three main targeting treatments: PMT, Community or Hybrid. In the PMT treatment, program beneficiaries were determined through a regression-based formula that mapped easily observable household characteristics into a single index. In the community treatment, the hamlet residents determine the list of beneficiaries through a poverty-ranking exercise at a public meeting. In the hybrid treatment, the community ranking procedure was done first, followed by a subsequent PMT verification. Additional details of these three procedures can be found in Appendix C and in Alatas et al. (2012).

Using intuitions from network theory on information aggregation, we can test whether the network characteristics that are typically associated with a better informed population also predict where community-based targeting does better, i.e. where the community will do better at ranking people when collectively entrusted to do so. Following Alatas et al. (2012), we create two metrics to assess the degree to which these methods correctly assign benefits to poor households. First, we compute the rank correlation between the results of the targeting experiment (the "targeting rank list") and per capita consumption. Second, we compute the rank correlation of the targeting experiment with respondents' self-assessment of poverty, as reported in the baseline survey. To assess the degree to which different network structures affect the targeting outcomes, we can examine whether the difference in these rank correlations between community / hybrid treatments (which use community information) and the PMT treatment (which does not) is greater in villages with network structures that should lead to better information transmission.

2.4. **Sample Statistics.** Table 1 reports descriptive statistics for the primary network and outcome variables used in the study (Appendix A provides definitions of each network variable). Panel A provides the statistics for the hamlet level variables, while Panel B provides corresponding household level statistics. We report variable means in in Column 1 and standard deviations in Column 2.

[12] Note that, conditional on sign-consistency, any *standardized* effect has to decrease even with non-classical measurement error. Following Cauchy-Schwarz it is easy to show that $\beta_0 \cdot \sigma_x > \text{plim}\,\widehat{\beta} \cdot \sigma_{\bar{x}}$ as $\sigma_x \sigma_{\bar{x}} > \text{cov}\,(x_i, \bar{x}_i)$ where $\widehat{\beta}$ is the estimated regression coefficient, β_0 is the true value, x is the true regressor, and $\bar{x}$ is the mismeasured regressor.

The sampled hamlets tend to be small (Panel A). The average network consists of about 53 households. The number of connections per household, called a household's *degree*, averages 8.25. Villages exhibit significant *clustering,* with a mean of 0.41; this means that about 41 percent of an individual's contacts are also linked themselves. The average *path length* is about 2, which suggests that two randomly chosen households will be separated by one household in between, conditional on being in the same component. The networks have an average *fraction of nodes in the giant component* of only 0.50, which means that about half of the households are interconnected to each other through some chain of connections.[13]

Households struggle with making wealth-based comparisons. The mean average error rate at the village level based on consumption is 0.502, while the mean error rate based on the self-assessment is about 0.463. However, there is heterogeneity in the error rate across villages – the standard deviation for both variables is about 0.2, which means that in the very best villages the error rate is at little as 0.1.[14] Panel B provides corresponding sample statistics at the household level. Most notable is the fact that the average clustering coefficient is 0.64. This differs from the aggregated data in Panel A because we have more information about sampled individuals than we have about the rest, which is natural because everything we know about non-sampled individuals comes from reports from the sampled group.

3. Reduced Form Analysis at the Household Level

3.1. **Household Level.**

In this section, we provide prima facie evidence of information diffusion through the network. To begin, we explore how a household's place in the network is correlated with their ability to rank others within the hamlet (section 3.1.1). We then explore whether households are better at ranking those who are more connected to them (section 3.1.2).

3.1.1. *Network Position of those Ranking Others.* We begin by asking whether individuals that are more central within the network have a lower error rate in ranking other households in the hamlet based on their well-being. Specifically, we estimate:

$$Error_{ir} = \beta_0 + \beta_1' W_{ir} + X_{ir}'\delta + \epsilon_{ir} \tag{3.1}$$

where i is the household doing the ranking, r is a hamlet, $Error_{ir}$ is household i's error rate in ranking, W_{ir} are i's network characteristics, ϵ_{ir} is the error term, and X_{ir} are covariates for household i (log consumption, years of education of the respondent, and dummy variables that indicate whether the household is a leader within the village, whether the household is from an ethnic minority, whether the household is from a religious minority, and whether the respondent is female). Table 2A reports the results with no covariates (i.e. constraining δ to be zero) and Table 2B reports

[13] It is likely that the true underlying network is in fact fully connected, and the fact that this number differs greatly from 1 comes from the sampling of the graph. Note that more dense graphs will exhibit a higher fraction of nodes in the giant component under sampling. There is considerable variation in the fraction of households in the giant component, with a standard deviation of 0.244, which implies that there is significant heterogeneity in the sparsity of the underlying true graphs. As discussed in Footnote 18, despite the sampling problem the correlations of the data are still in line with those predicted from the model.

[14] The 5th percentile for these variables are 0.254 and 0.138, respectively.

them when we include a full set of covariates (X_{ir}).[15] The considered network characteristics are degree (Column 1), which is the number of links to other households; the clustering coefficient (Column 2), which is the fraction of a household's neighbors that are themselves neighbors; and the eigenvector centrality (Column 3), where eigenvector centrality is a measure of the node's importance defined, recursively, to be proportional to the sum of her neighbors' importances. Formal definitions are included in Appendix A. In Column 4, we estimate the effect of each of these three network characteristics, conditional on one another. In Columns 5 - 8, we replicate the analysis in Columns 1 - 4, but additionally include hamlet fixed effects. This allows us to estimate the effect of the household's characteristics within the network conditional on others within the network. In Panel A of each table, the error rate is based on per-capita consumption, while it is based on the self-assessment in Panel B. Panel C of each table shows simulations from the model, which will be discussed in Section 4.3 below. All equations are estimated using OLS, with standard errors clustered at the hamlet level.

Overall, households that are more connected within the network have an easier time ranking other households. Using consumption as the measure of the truth (Panel A of Table 2A), the univariate regressions (Columns 1-3) show that households that have higher number of links with other households in the network (degree), that have more interwoven social neighborhoods (clustering), and that households that are a more important node in the network (eigenvector centrality) are less likely to make errors in ranking others. Conditional on each other, we find that a one standard deviation increase in average degree is associated with a 5pp drop in the error rate of a household and similarly a one standard deviation increase in the clustering coefficient is associated with a 1.2pp drop in the error rate (Column 4). Holding constant the fixed effect of the hamlet, degree (Column 5) and eigenvector centrality (Column 7) continue to predict a household's error rate (both at the 1 percent level), but clustering is no longer significant. When all three measures are included in Column 8, we find that a one standard deviation increase in degree corresponds to roughly a 1.03pp decrease in the error rate (significant at the 5 percent level). However, the clustering and eigenvalue centrality are no longer significant, though the magnitude of clustering remains similar and the magnitude of eigenvector centrality drops.

Similarly, as Panel B illustrates, households that are more connected also have an easier time ranking other households as compared against their self-assessment. In fact, the coefficient estimates of all models in Panel B are very similar to those in Panel A, both in terms of sign and magnitude. In Column 8, we find that a one standard deviation increase in degree corresponds to roughly a 1.4 pp decrease in the error rate (significant at the 5 percent level).

The results in Table 2B, which include a large number of additional characteristics of the household that is doing the ranking, are generally quite similar to the results in Table 2A. This suggests that the results are not driven by observable household characteristics. For example, the coefficient estimates in Column 4, Panel A, imply that one standard deviation increases in degree and clustering are associated with 4pp and 1.2pp declines, respectively, in the consumption error rate; the analogous impacts from Table 2A (with no covariates X_{ir}) were 5pp and 1.2pp. Again, the

[15] The remainder of the tables in the paper present results conditional on covariates, unless otherwise noted, though we include appendix versions without covariates.

patterns in the data look similar using self-assessment (Panel B) as the measure of the truth rather than consumption (Panel A).

In sum, the evidence thus far suggests that a household's position within the network is predictive of its ability to accurately rank the income distribution within the hamlet.

3.1.2. *Connections Between Ranker and Rankee.* The preceding analysis explored how one's place in the network affected the accuracy of the ranking. We now test whether the ranker is more accurate when he or she is more connected to the the households that he or she is ranking. Specifically, in Table 3, we address whether a household i does a better job of ranking nodes j versus k if the pair is closer to i. To measure distance on the network, we use the shortest path length. However, because of sampling, many nodes cannot be connected by any path and therefore have infinite distance between them. To address this, we include a term for the average reachability between (i, j) and (i, k) as well as the average of the distances between (i, j) and (i, k).[16] Specifically, we estimate:

$$Error_{ijkr} = \beta_0 + \beta_1' W_{ijkr} + X_{ijkr}' \delta + \epsilon_{ijkr} \tag{3.2}$$

where $Error_{ijkr} = \mathbf{1}\{i \text{ ranks } j \text{ versus } k \text{ incorrectly}\}$ (which is done for all $j < k$, $j \neq i$, $k \neq i$), W_{ijkr} is the average network characteristics of the households that are being ranked (j and k), and X_{ijkr} are physical covariates. In Column 1, we show the basic correlations between the error rate, average distance from i to j and k, and average reachability conditional on demographic controls. In Column 2, we introduce additional network characteristics (average degree, average clustering coefficient and average eigenvector centrality, where once again, the average is across the two people being ranked). In Column 3 and 4, we include hamlet fixed effects and ranker fixed effects, respectively. Note that all standard errors are clustered at the village level.

Average reachability and distance tend to be highly predictive of the accuracy in the ranking. Using consumption as the measure of truth (Panel A), if both are on the same connected component as i as compared to neither being on the same component, then household i is 6 to 13 percentage points less likely to rank them incorrectly, and if the average distance of the ranked pair increases by one standard deviation, then there is a resulting increase of 1 to 1.5 percentage points in the probability that household i ranks them incorrectly. These results are generally robust to using physical covariates (Columns 2-4), hamlet fixed effects (Columns 3-4), and ranker fixed effects (Column 4). Using self-assessment as the truth (Panel B), the average reachability and distance predict the error of the ranked pairs with physical controls and hamlet fixed effects (Column 3). However, when controlling for ranker fixed effects (Column 4), it is no longer significant at conventional levels, although the sign and magnitudes of the coefficients are generally similar to Column 3. Panel C of each table once again shows simulations from the model, which will be discussed in Section 4.3

[16] If there is no path from node i to j, the distance is by convention infinite. In regression, then, we use instead of distance two terms: whether i is reachable from j and a second term which is an indicator function of reachability times the distance between i and j (where infinite paths are replaced by any arbitrarily high finite number). Note that distance is not interpretable without reachability and therefore we always include them both in the regression analysis.

4. Framework

The results thus far suggest that a network-based model may plausibly describe how information is spread, since a household's characteristics within the network predict how much it know about others. In this section, we carry out what we have previously described as numerical theorizing. We begin by describing a simple model of information transmission on a network that captures the basic features of our environment. We derive the expressions for the village level error rates and the cross-village rankings of village level error rates (our main outcomes of interest) as a function of the parameters of that simple model. In the next sub-section we turn towards structurally estimating the parameters of the model using within-village variation. Subsection 4.3 then confirms that the model does generate the cross-individual patterns that we found in the reduced form analysis in Section 3(the effect of distance, etc.). The last sub-section reports on our cross-village simulations which give us our numerical propositions: We estimate the effects of various village characteristics on village level error rates that we generate by simulating our model.

4.1. **Model.** We consider a simple variation on the standard Susceptible-Infected-Susceptible (SIS) model. While this model originates in the epidemiological literature (see e.g., Pastor-Satorras and Vespignani (2001)), it has also been extended to study the diffusion of information through the network (e.g. Jackson and Yariv (2007); Jackson and Rogers (2007); Galeotti and Vega-Redondo (2011); López-Pintado (2008)). As it is not easy to analytically analyze, the literature typically models the network diffusion process by an approximation wherein nodes independently meet other nodes with probability proportional to their degree. The authors use a mean-field approximation to compute the steady state information rate in the network. In a mean-field approximation, essentially the heterogeneity in local information in the neighborhoods of households is assumed away and replaced with steady-state mean values.[17] This makes the problem analytically tractable. As the authors of the literature note, the price paid for gaining analytical tractability is assuming away much of the rich local structure – a phenomenon that we believe is particularly important in our setting.

We depart from this literature in two main ways. First, our problem is fundamentally multi-dimensional. We are interested in whether or not an individual node has two distinct pieces of information. They need to know the income status of two different households in order to rank them. Second, since our goal is not to recover an analytic approximation to the steady-state distribution, we take the literal environment wherein nodes pass information to their network neighbors and do not make any mean-field approximation. The mean-field approximation would force us, by assuming the average neighborhood information rate is also the one that any node faces locally in her neighborhood, to ignore some of the richness encoded in the process that we are trying to study. We describe the model below and, in greater detail, in Appendix B.

Let $G = (V, E)$ be a graph, which consists of a set of vertices, V, and a set of (undirected, unweighted) edges, E. The graph can be described by its adjacency matrix $A := A(G)$ where $A_{ij} = \mathbf{1}\{ij \in E\}$ and $A_{ii} = 0$.

[17] When a node meets a collection of other nodes in a given period, the share of its partners that are informed is the same as the population average share of neighbors that are informed. This homogeneity makes the problem tractable.

We use I to denote the set of nine chosen households in the hamlet. The wealth of these households in I will be the information being transmitted through the graph via a simple diffusion process. We assume that a household i can correctly rank the wealth of j versus k if and only if i knows both j and k's wealths.

Let $S_t^{(j)}$ be an n-vector indicating whether individuals at time t know j's wealth. That is, $S_{ti}^{(j)} = 1$ indicates that i knows j's wealth at time t while $S_{ti}^{(j)} = 0$, says that i does not know j's wealth at time t. We are interested in the evolution of this information on our networks. In particular, we are interested in an $|I|$-dimensional transmission process where each household's wealth information is transmitted through the graph.

To have a steady state where some, but not all nodes, are fully informed, we follow the literature and model the interplay of two forces. First, a household that has information about some individual's wealth will transmit information to a neighbor with some probability in any given period. Second, a household that knows some information may forget it in a given period. We establish three simple rules for the process:

(1) If it is informed, household $i \in V$ transmits the wealth of j to neighbors with probability p, independently of each other.
(2) Household $i \in I$ never forgets its *own* wealth.
(3) Household $i \in V \setminus I$ forgets the wealth of j with probability δ, independently of each other.

It helps to define Δ to be a random matrix with entries Δ_{ij} which are independent Bernoulli random variables taking on 1 with probability δ and impose the restriction $\Delta_{ii} = 0$. Also, let $X_{ti}^{(j)}$ be distributed as Bernoulli with probability $p_{ti}^{(j)} = 1 - (1-p)^{A_{i \cdot} S_{t-1}^{(j)}}$. Note that $(1-p)^{A_{i \cdot} S_{t-1}^{(j)}}$ is the probability that that none of i's neighbors who at time $t-1$ are informed about j actually inform i about j's wealth. Therefore $p_{ti}^{(j)}$ is the complementary probability.

The behavior of the diffusion processes is then given by the following system of stochastic evolution equations:

$$S_{t,i}^{(j)} = X_{t,i}^{(j)}(1 - S_{t-1,i}^{(j)}) + S_{t-1,i}^{(j)}(1 - \Delta_{ij}) \ \forall j \in I, \ \forall i \in V.$$

In vector form: for every $j \in I$,

$$S_t^{(j)} = \mathrm{Diag}\left(X_t^{(j)}\right)\left(\iota_n - S_{t-1}^{(j)}\right) + S_{t-1}^{(j)}{}'\left(I_n - \mathrm{Diag}\left(\Delta^{(j)}\right)\right),$$

where $\iota_n = (1, ..., 1)'$. This generates a well-defined Markov process, albeit one that is difficult to characterize analytically.

Since we are interested in whether i can rank the wealth of j versus k, we define $\mathrm{D}_{ti}^{(j,k)}$ as:

$$\mathrm{D}_{ti}^{(j,k)} := S_{ti}^{(j)} S_{ti}^{(k)}.$$

Therefore, $\mathrm{D}_{ti}^{(j,k)}$ is a random variable which describes whether at period t, i knows whether j or k is wealthier – note that as assumed above $\mathrm{D}_{ti}^{(j,k)}$ is only equal to 1 when both $S_{ti}^{(j)}$and $S_{ti}^{(k)}$are equal to one.

In what follows, we use the empirical analogues of $\mathrm{D}_{ti}^{(j,k)}$ and functions of $\mathrm{D}_{ti}^{(j,k)}$ to construct the outcome measures:

(1) Hamlet level error rate for hamlet r:

$$Error_r := \frac{1}{|I|\binom{|I|-1}{2}} \sum_i \sum_{j<k:\ j\neq i, k\neq i} \left(1 - D_{ti}^{(j,k)}\right).$$

(2) Whether the error rate of hamlet I exceeds that of J:

$$Error_{I>J} := \mathbf{1}\{Error_I > Error_J\}.$$

(3) Household level error rate for household i in hamlet r:

$$Error_{ir} := \frac{1}{\binom{|I|-1}{2}} \sum_{j<k:\ j\neq i, k\neq i} \left(1 - D_{ti}^{(j,k)}\right).$$

(4) Whether household i ranked j versus k correctly (all in hamlet r):

$$Error_{ijkr} := 1 - \mathrm{D}_{ti}^{(j,k)}.$$

4.2. **Structural Estimation and Numerical Propositions.** In this section, we estimate the diffusion model that we detailed in the previous section. The parameter estimates of $\gamma = (p, \delta)$ are interesting in their own right as they represent the underlying transmission and forgetting probabilities. More importantly, having a structural estimate of the model also enables us to simulate out information transmission and then study the behavior of the reduced form regressions under these simulations. This is useful because the model itself is analytically intractable, i.e. it does not allow for clear predictions as to what regression coefficients ought to theoretically look like if we conduct regressions of information outcomes on network statistics for data generated by this model. The exercise conducted here provides a method of numerical theorizing: by comparing the reduced form regression estimates to the counterparts generated by simulated data, we can see whether the patterns we pick up in the data are qualitatively similar to those predicted by standard models from network theory.[18]

To estimate the model, we use simulated method of moments (SMM). We use two moments, so the model is just-identified. The first moment is the error rate for the graph. The second moment is a weighted version of an error rate, where when considering how i ranks j versus k we weight by how well connected i is to each of j and k. The key difference between the two moments is this weighting.

Specifically, define $m_1(Z_r)$ as the empirical error rate for graph r, $Error_r$ as defined in Section 4.1, and set:

$$\psi_1(Z_{rs};\gamma) := m_1(Z_{rs};\gamma) - m_1(Z_r)$$

where $m_1(Z_{rs};\gamma)$ is the error rate for graph r under simulation s. Let $m_2(Z_r)$ be the error rate among the i ranking j versus k, weighted by the number of paths between i and j as well as i and

[18] We are focusing on the qualitative as opposed to the quantitative predictions from the model. Given that for much of the parameter space we retain the same predictions, we argue that our findings should typically be conservative. Note that for our simulation exercise to be misleading, the following must be true. It must be the case that the relationship between the simulated outcomes based on a diffusion process on the induced subgraph and the sampled network statistics has to have a different sign than the relationship between the true outcomes (generated by a diffusion process on the entire graph) and the sampled network statistics.

k, and put

$$\psi_2(Z_{rs};\gamma) := m_2(Z_{rs};\gamma) - m_2(Z_r),$$

where $m_2(Z_{rs};\gamma)$ is the analogue from simulation s. Intuitively, the differential variation in ψ_2 versus ψ_1 identifies δ under our model, while ψ_1 identifies p given δ. (In practice, of course, they are jointly estimated.) We estimate:

$$\widehat{\gamma} = \underset{\gamma\in[0,1]^2}{\operatorname{argmin}} \left\| \frac{1}{R}\sum_{r=1}^{R}\frac{1}{S}\sum_{s=1}^{S}\psi(Z_{rs};\gamma)\right\|^2$$

where $\psi = (\psi_1, \psi_2)'$.

For some intuition on identification, consider the following example where i is ranking j versus k as well as j' versus k'. Assume $d(i,j') = d(i,k') = 2$ and there are many such path between i and each of j' and k'. Meanwhile assume $d(i,j) = d(i,k) = 2$ but there is only one such path between i and each of j and k. Then, ceteris paribus, i is more likely to hear about the wealths of j' and k' as compared to the wealths of j and k. However, notice that if i has $\mathrm{D}_{ti}^{(j,k)} = 1$ and $\mathrm{D}_{ti}^{(j',k')} = 1$, the probability i forgets either the wealth of j' or k' is the same as the probability that i forgets either the wealth of j or k: each occurs with probability $1-(1-\delta)^2$.

Using this approach, we find that $\widehat{\gamma} = (0.4, 0.35)$, with standard errors of $(0.21, 0.21)$. While the standard errors are quite large, in Section 5.3 we show robustness of our approach to a variety of other parameter values.

4.3. **Simulation Results at the Individual Level.** We begin by exploring the predictive capabilities of these parameter estimates at the individual level.

Given the parameter $\widehat{\gamma}$, we simulate out a diffusion process in the following manner: Every individual in I, the set of randomly chosen households, is thought to know their own wealth. We take 100 draws from the invariant distribution of the diffusion process described above by running the transmission process out 100 times after a burn-in phase of 50 rounds. For every draw, $s = 1, ..., 100$, we compute an error dummy, $Error^{s}_{ijkr}$ indicating whether i ranked j versus k wrong in hamlet r. We then compute the expected error rate across the 100 simulations, $\overline{Error}_{ijkr} := \frac{1}{S}\sum_s Error^{s}_{ijkr}$. To generate predictions corresponding to Table 2, we use as an outcome variable $Error^{SIM}_{ir} := \binom{|I|-1}{2}^{-1}\sum_{j<k:\ j\neq i,k\neq i}\overline{Error}_{ijkr}$ and to generate predictions corresponding to Table 3, we use $\overline{Error}_{ijkr}$ directly. Note that this is described in more detail in Appendix B.

We then regress the simulated outcomes on the various network characteristics of interest to observe what the qualitative relationship between the network characteristic and the error rate should have been under the null of our model. Specifically, we rerun the same regressions as in Tables 2A, 2B, and 3 using the simulated data from the model. The results using the simulated data are shown in Panel C of each table.[19]

[19] We note that it may be the case that projecting a complex diffusion process into a specific linear regression specification may itself generate unintuitive coefficient estimates. However, as our method compares the signs of those generated by simulations from the model and the real data, if the model is a good description of the information transmission process, the unintuitive projections should be similar across both the simulations and the real data. That is, even still, it is the case that comparing two regressions – one with a simulated outcome variable and another with an empirical one – turns out to be a reasonable test of whether the real-life process is like the model process.

By and large, Panel C of Tables 2A and 2B confirms our intuitions. Households that have a higher degree are associated with lower error rates, households that have higher clustering are associated with lower error rates, and households that are more eigenvector central are associated with lower error rates. Similarly, the distance and reachability results also conform to our intuitions. In Panel C of Table 3, we find that being in the same component as those who an individual is ranking reduces the error rate while being several steps farther away increases the error rate. Qualitatively, the patterns all match the actual empirical results shown in Panels A and B of both tables, though the simulated magnitudes for clustering and eigenvector centrality are larger in the simulated data than in the actual data in Tables 2A and B and the simulated magnitudes for distance and reachability are larger in the simulated data than in the actual data in Table 3.[20]

4.4. **Simulation Results at the Village Level: Numerical Propositions.** A key question is how network-level characteristics affect information diffusion across the network. We start from the important analytical result in Jackson and Rogers (2007) showing that if network I's degree distribution and neighbor degree distribution first-order stochastic dominates network J's degree distribution and neighbor degree distribution, respectively, then in steady state of a mean-field approximation to the matching process described above, network I should have a higher equilibrium information rate than network J.[21]

As noted above, this result unfortunately cannot be directly applied to our context for at least two reasons. First, as discussed in Section 4.1, the model uses a mean-field approximation to a matching process, which itself tries to approximate the contagion process described above, to gain analytic tractability. However, we are precisely interested in the cases where the mean-field approximation may not be apt, i.e. where we do not believe that everyone's local neighborhood essentially contains the same average information as the global average. The approximation does not work well when, for instance, each node does not have a proportion of neighbors who are infected equal to the average neighbor infection rate. We would imagine this not to be true, for instance, when a household does not forget its own wealth. Second, to rank households, each node needs to have two pieces of information, whereas there is only thing to learn in Jackson and Rogers (2007). While one can readily extend their model and use a two mean-field approximations while tracking two independent diffusion processes, again the aforementioned local patterns will be lost.

We therefore use the numerical simulations of our model to test whether we should expect the equivalent result to hold in our context. The simulations are described in detail in Appendix B. As discussed above, we generate $\overline{Error}_{ijkr}$ via the aforementioned simulation process and, in this case, we construct hamlet level error rates by averaging over the individual level error rates $Error_{ir}^{SIM}$. Then, we compute the share of times $Error_{I}^{SIM} > Error_{J}^{SIM}$ for hamlets I and J. We regress this variable on whether I stochastically dominates J or vice versa. The results, which are reported in

[20] One reason for this is that the simulations assume that our sampled network is the true network, whereas in fact it is a subset of the true network. There is therefore more measurement error in the true network measures (in Panels A and B) than in the simulated network measures (Panel C). As discussed by Chandrasekhar and Lewis (2012), this problem is likely to be least severe for degree and most severe for eigenvector centrality, since centrality is the most global network feature and thus most sensitive to measuring the entire network. This is consistent with the empirical results.

[21] The neighbor degree distribution is the empirical cdf of the number of links a neighbor has, taken over all neighbors as we count over all nodes.

Panel C of Table 4, suggest that the Jackson and Rogers (2007) pattern holds in our context and a complete discussion of the table is provided in Section 5.1.

We can also apply the same methodology to test other claims about the role of other fundamental network characteristics. We choose six standard measures used in various related, but otherwise different, models – network size, average degree, average clustering, first eigenvalue of adjacency matrix, link density and fraction of nodes in giant component – and simulate how they affect diffusion within our estimated model, described in detail in Appendix B. As discussed above, we generate $\overline{Error}_{ijkr}$ via the aforementioned simulation process and we then construct hamlet level error rates by averaging over the individual level error rates $Error_{ir}^{SIM}$.

Given these simulation-based hamlet level error rates, we estimate:

$$Error_r^{SIM} = \beta_0 + W_r'\beta_1 + X_r'\delta + \epsilon_r \tag{4.1}$$

where $Error_r^{SIM}$ is the average error rate in hamlet r from the simulations and W_r is a vector of graph level statistics including average degree, average clustering, the number of households in the hamlet, first eigenvalue, link density, and fraction of nodes in giant component. Together with the set of hamlet-level covariates X_r, we include many potentially correlated variables in the specification of the regression model. It is not ex ante obvious that the conditional correlations of network features with the outcome variables will behave the same as the unconditional correlations, and this is what we look at here.

The results are reported in Panel C of Table 5. When the network characteristics W are included one by one, most of the network statistics of interest have significant effects on the error rate and they all go in the "intuitive" direction: there are lower error rates in villages where the average degree is higher, clustering is higher, the first eigenvalue of adjacency matrix is bigger, the link density is higher and there are more households that are in the giant component.[22] The inclusion of hamlet level covariates make no difference (see Appendix F, Table F.5). When we jointly estimate the relationship of all of these network variables with the error rate, we observe some counterintuitive patterns (Column 7, Panel C). In particular, while most of the effects remain significant, average degree and average clustering now have the "wrong" sign, suggesting that even with more than 600 hamlets, we may not have enough independent variation to properly estimate these effects jointly.[23] This suggests that it is hard to separately identify the effects of these very interconnected variables or that or that some of our intuition is off when we are considering the variables conditional on others.

5. Cross hamlet comparisons

5.1. **Stochastic Dominance Results.**

We are interested in testing the Jackson and Rogers (2007) predictions about stochastic dominance in the degree distribution. In addition to being interesting

[22] We do not interpret the effect of hamlet size because it is difficult to do so in our framework.

[23] A natural worry is that average degree, number of households, and link density (which amounts to average degree over number of households) may be generating too much collinearity. However, conditional on the other covariates in column 7, omitting link density makes no difference to the "wrong" sign that degree takes on in the the regression. It appears, instead, that conditioning on the first eigenvalue and clustering leaves average degree to not matter in an obvious way. A table documenting this is available upon request.

in its own right, focusing on stochastic dominance has a major advantage in our context. Working with a sampled graph, rather than the full network, may result in bias that could lead to sign reversals in the estimates. An advantage of working with FOSD is that while we would expect attenuation bias in our estimates, we would not expect a sign reversals (Chandrasekhar and Lewis, 2012).[24] As such, our results would provide a lower bound of the predictive capabilities of the network.

To our knowledge, the claim about stochastic dominance has not been empirically tested before due to data limitations. In order to do so, a large sample of both locality networks and information diffusion is necessary. The data collected for this study provides a plausibly large enough sample (631 hamlets) to do so. We begin by estimating a regression of whether the error rate of the hamlet I exceeds the error rate of hamlet J ($Error_{I>J}$) on dummy variables that indicate whether hamlet I stochastically dominates hamlet J ($\mathbf{1}\{I \succ J\}$) and vice versa ($\mathbf{1}\{J \succ I\}$):[25]

$$Error_{I>J} = \beta_0 + \beta_1 \cdot \mathbf{1}\{I \succ J\} + \beta_2 \cdot \mathbf{1}\{J \succ I\} + X'_{IJ}\delta + \epsilon_{IJ}. \tag{5.1}$$

The omitted category is when hamlet I's and hamlet J's degree distribution are not comparable.[26] We can also estimate regressions where we drop hamlets that are not comparable:

$$Error_{I>J} = \beta_0 + \beta_1 \cdot \mathbf{1}\{I \succ J\} + X'_{IJ}\delta + \epsilon_{IJ}. \tag{5.2}$$

Table 4 presents the results of these regressions. Column 1 presents the results from estimating equation (5.1), while Column 2 presents the results from estimating equation (5.2). For both models, we include stratification group fixed effects, estimate with OLS, and specify two-way clustered standard errors, for hamlet I and hamlet J. Once again, we compute error rates with consumption as measure of truth (Panel A) and with self-assessment as measure of truth (Panel B). The results validate the implications of the model: if a hamlet's degree distribution first-order stochastic dominates another hamlet's distribution, it will have lower error rates in ranking the income distribution of the hamlet (for both measures of truth). Specifically, as Panel B, Column 2 shows, if hamlet I dominates J, then I has on average a 17pp lower error rate than J (significant at the 1 percent level). Columns 3 and 4 repeat the exercises of Columns 1 and 2, respectively, adding in physical controls to which the results are robust and the coefficients remain stable.

5.2. **General Cross-Hamlet Results.** We now present the general hamlet level regression. Our theoretical benchmarks are the numerical simulations described above and presented in Panel C of Table 5. We thus present equivalent reduced form analysis in Panel A and B of Table 5. In Columns 1 to 6, we present the univariate regressions while in Column 7 we present the multivariate regression. Once again, we consider error rates based on consumption (Panel A) and the self-assessment metric (Panel B).

[24] Sign-switching would be possible only when over half of the categorizations of I dominating J become flipped due to sampling, which probabilistically does not happen.

[25] Stochastic dominance was determined at the decile level. If the distribution function for the degree of hamlet I was weakly lower than J at all deciles (and was strict for at least one), then we say that I dominates J.

[26] Stochastic dominance establishes a partial ordering only.

The results look very similar regardless of using consumption or the self-assessment metric. The univariate regressions tend to have the expected sign. For instance, an increase in the average degree of the hamlet is associated with a lower error rate (column 1), an increase in the average clustering coefficient is associated with a lower error rate (column 2), and an increase in the number of households is not associated with the error rate (column 3, Panel A) or associated with a higher error rate (column 3, Panel B). In addition, as seen in Column 4, a higher first eigenvalue of the adjacency matrix is associated with a considerable reduction of the error rate (a one standard deviation increase is associated with a 6pp drop in error rate). Column 5 shows that a higher fraction of nodes being in the giant component is associated with an extremely lowered error rate. As expected, Column 6 of Panel B shows that a higher density of links corresponds to a lower error rate, though we do not find this in Panel A. Column 7 yields the expected results except for the case of average degree (for both outcome variables). However, this is likely to be the case because the first eigenvalue is highly correlated (0.88) with average degree and that error rates are really a function of the entire distribution of links. As discussed in footnote 23, it appears that conditional on the first eigenvalue (a global property that captures a key aspect of the diffusion process), average degree need not enter in the obvious way.

When we run the regressions one by one, the reduced form results match up quite closely with our numerical predictions. Whether it is Panel A or Panel B that we compare with Panel C and irrespective of whether the covariates get included (see Appendix F, Table F.5 for comparison without covariates), whenever both coefficients are significant (which is most of the time), they always have the same sign. Including all network variables in the regression model (Column 7), we once again find a good match between the *actual* and *simulated* results. Strikingly, higher average degree appears to be a positive and significant predictor of error rate (higher degree means more errors) across both our reduced form and simulated results in this column. The first eigenvalue of the adjacency matrix, the link density and the fraction of nodes in the giant component all come negative (and mostly significant) in Panels A and B, exactly as our simulations would have had us expect and confirming intuitions that come from Bollobás et al. (2010), among others. The one exception is clustering which comes in with the "right" sign in the data, but was positive in the simulations.

5.3. **Robustness.** To study the robustness of our numerical propositions, especially since the bounds on estimates are not very tight, we redo the cross village simulation and regression exercise for a wide interval of parameter values more or less centered around the estimated values. While the parameter values matter – some correlations change when we approach the boundaries of the interval – the basic predictions turn out to be remarkably robust.

For ease of presentation, we report the results on four chosen parameter values: $(p, \delta) \in \{0.1, 0.8\}^2$. We present these because it allows us to document four conceptually distinct cases: high p/δ ratio wherein the effective rate of transmission is high, low p/δ ratio wherein the effective rate of transmission is low, and p/δ ratio of 1 with high transmission levels versus low transmission levels. These last two cases are ones in which the *level* of transmission (0.1 versus 0.8) and the network architecture may generate different patterns despite having the same effective transmission rate. We present our results in a collection of tables in Appendix (D).

While in principle we only needed the numerical propositions to be confirmed by the data local to the parameter value estimated under the model, there are several reasons why we like this exercise. First, if we find that the qualitative predictions are robust to (p, δ), then the conclusions of our specific exercise are more likely to be portable to outside contexts. Second, we are interested in the qualitative predictions of the structural model for several reasons: we are working with sampled network data, we do not add disturbance terms into the model and we refrain from adding high dimensional complexities such as allowing for heterogeneous transmission probabilities (e.g., $p(X; \theta_p)$ and $\delta(X; \theta_\delta)$ as functions). As such, it would be more reassuring if we found that the patterns held for a larger region of the parameter space. Third, such a robustness exercise will provide us with insights as to when and why our standard intuitions are likely to fail.

Ultimately, as seen in Appendix (D), we find that the aforementioned results are essentially robust; the qualitative predictions are mostly invariant to the parameter values (p, δ). While some of the magnitudes change (higher δ tends to correspond to lower correlations between network characteristics and outcomes), the basic qualitative predictions remain the same. In addition, there is power against unreasonable regions of the parameter space. For instance, we often get qualitatively different predictions in the low transmission, high forgetting regime $(0.1, 0.8)$. This is a case in which information essentially travels very short distance as the probability of forgetting is very high relative to the transmission rate and therefore diffusion is greatly handicapped. At this boundary, we find, for example, that the *conditional effects* of eigenvector centrality of the ranker (Table D.2), eigenvector centrality of the rankees (Table D.3), average degree of a village, the first eigenvalue of the adjacency matrix and the fraction of nodes in the giant component (Table D.5) all switch signs relative to both the effects we find in the data and also the other parameter values.[27]

6. Application: Targeting

In this section, we test whether the network characteristics affect the quality of real-world decisions that rely on communal information. To do so, we examine the targeting experiment discussed in Section 2.3.3 above. In particular, we test whether community-based targeting is relatively better than proxy-means testing at identifying the poor in networks that we expect to be better at diffusing information about poverty. If communities efficiently aggregate information, we would expect that this would be the case, since community-based targeting utilizes local information and the findings thus far have shown that better networked communities hold more accurate information. However, if there are distortions in the processes through which information is aggregated, this may not necessarily be the case.

We estimate regressions of the form:

$$y_r = \alpha + \beta_C \mathbf{1}\{r \in C\} \cdot \rho_r + \beta_H \mathbf{1}\{r \in H\} \cdot \rho_r + \tau_c \mathbf{1}\{r \in C\} + \tau_h \mathbf{1}\{r \in H\} + \rho_r + \epsilon_r, \tag{6.1}$$

where y_r is the rank correlation between the targeting program's assessment of poverty and the benchmark of true poverty (either based on per-capita consumption or based on the self-assessment), $\mathbf{1}\{r \in C\}$ and $\mathbf{1}\{r \in H\}$ are dummies for experimental assignment of hamlet r to either community

[27] The stochastic dominance results, however, remain completely robust across all presented parameter values (Table D.4).

or a hybrid treatment (the omitted category is PMT), and ρ_r is a measure (discussed below) of how diffusive a network is. We are mostly interested in β_C, and to a lesser extent, β_H. Given that higher ρ_r indicates that a network is better at spreading information, then we would expect that $\beta_C > 0$, or in other words, we expect community-based targeting to perform better relative to a proxy-means test when ρ_r is higher.

We take two approaches to computing ρ_r, the diffusiveness of a hamlet network. In Table 6, we first compute ρ_r by using principal-components to aggregate the six measures of network diffusiveness from Table 5: average degree, clustering, first eigenvalue, number of households, link density and fraction of nodes in the giant component. We then take the first principal component vector corresponding to the covariance matrix between these six network attributes and define $\rho_r = \sum_{k=1}^{6} v_k W_{k,r}$ where v_k are the entries of the principal component vector and $\{W_{k,r}\}$ are the six network features for hamlet r. Panel A of Table 6 shows that β_C and β_H are not distinguishable from zero when we take y_r to be the rank correlation using consumption data, i.e. we do not observe that community targeting is more accurate in more diffusive communities relative to the PMT (Columns 2-5 of Panel A of Table 6). However, when we take y_r to be the rank correlation using self-assessment data, we find positive and significant estimates of β_C and β_H (Columns 2-5 of Panel B of Table 6). Conditional on community targeting, a one standard deviation increase in diffusiveness corresponds to a 0.06 increase in the rank correlation of the targeting outcome with the self-assessment benchmark (which has a mean of 0.4) relative to the PMT. Not surprisingly, when we pool the treatments, the relationship persists.The fact that ρ_r only matters for the effectiveness of community targeting when assessed using self-assessment is consistent with the experimental findings in Alatas et al. (2012), which also showed that in general, community meetings increased the rank correlation with self-assessment, but not with per-capita consumption.

A second approach is to use the model from Section 4 and simulations as discussed in Appendix B to compute ρ_r. Specifically, we use the simulated error rate for a hamlet, $Error_r^{SIM}$, as an inverse measure of its diffusiveness since, by definition, networks that are better at spreading information should exhibit lower error rates. Table 7 replicates the exercises in Table 6, but now uses the simulated error rate as ρ_r. Since higher $Error_r^{SIM}$ means less diffusiveness, i.e. lower, ρ_r, we instead hypothesize that $\beta_C < 0$. As expected, we find that community targeting differentially works better when a hamlet has lower error rates when measured using self-assessment. A one standard deviation increase in the simulated error rate, conditional on community targeting, a 0.075 decrease in the rank correlation of the targeting outcome with the self-assessment benchmark, relative to the PMT.[28]

Taken together, the findings show that the network structure and diffusion model not only accurately predict how information spreads, but are also useful in understanding how real decisions are made using that information.

[28] We note that in Panel B of both Tables 6 and 7, a more diffusive network is correlated with worse targeting under PMT when measured by the correlation with self-assessment. In fact, we can show that the covariance between consumption based wealth ranking and self-assessment based wealth ranking decreases as we look at more diffusive hamlets. Therefore, it seems that high ρ_r hamlets (for the case of the principal component) makes the self-assessment based notions of poverty harder to detect by conventional means. However, it seems that the community does know more about who is poor by this criterion; as the community also puts weight on this criterion, the community pulls the outcome closer to the self-assessment metric.

7. Conclusion

In this paper, we estimate a simple model of how information about poverty status is transmitted within the network, and then use the estimated model to predict the relationship between the network characteristics and how information on poverty status is aggregated within the network. We then compare our predictions with empirical evidence from a unique dataset of 631 villages, where we have both detailed social network data and measures of how accurately households can describe the poverty status of other households. The empirical results match up nicely with the model predictions: the characteristics that predict better information aggregation in the model also do so in the data. In particular, we provide evidence supporting the Jackson and Rogers (2007) claim that if a network's degree distribution first-order stochastic dominates another's distribution, it will have an overall lower error rates in ranking the income distribution of the hamlet. We then show that the network can be utilized for real-world policy decision making: communities that are better networked are more likely to accurately choose the beneficiaries in community-based targeting relative to more traditional data-driven approaches.

The results are encouraging because they suggest that possibility of using standard network statistics to predict whether in a particular context we would expect effective information aggregation, or conversely, whether some outside intervention will be needed to supplement information flows through the network. Moreover the results give us some confidence that we are not very far off in using very simple infection style models to study communications in networks.

One limitation of what we do here comes from the data: while we had access to over 600 networks, a relatively small fraction of each network was sampled and as a result, our measures of the network characteristics are likely to be biased. This no doubt limits the precision of our predictions in the cross-village data – here we only ask whether the correlations ar e qualitatively similar to those observed in the data. One important next step would be to find/collect a dataset that has much more detailed network data and see how much better we do in quantitatively predicting network outcomes.

References

Acemoglu, D., M. Dahleh, I. Lobel, and A. Ozdaglar (2011): "Bayesian learning in social networks," *The Review of Economic Studies*, 78, 1201–1236.

Alatas, V., A. Banerjee, R. Hanna, B. Olken, and J. Tobias (2012): "Targeting the Poor: Evidence from a Field Experiment in Indon," *American Economic Review*, 102.

Alderman, H. and T. Haque (2006): "Countercyclical safety nets for the poor and vulnerable," *Food Policy*, 31, 372–383.

Bandeira, O., R. Burgess, S. Gulesci, I. Rasul, and M. Sulaiman (2012): "Can Entrepreneurship Programs Transform the Lives of the Poor?"," Tech. rep., LSE.

Bandiera, O., I. Barankay, and I. Rasul (2009): "Social connections and incentives in the workplace: Evidence from personnel data," *Econometrica*, 77, 1047–1094.

Bandiera, O. and I. Rasul (2006): "Social networks and technology adoption in northern mozambique*," *The Economic Journal*, 116, 869–902.

Banerjee, A. (1992): "A simple model of herd behavior," *The Quarterly Journal of Economics*, 797–817.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. Jackson (2012): "The Diffusion of Microfinance," MIT working paper.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992): "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, 100.

Bollobás, B., C. Borgs, J. Chayes, and O. Riordan (2010): "Percolation on dense graph sequences," *The Annals of Probability*, 38, 150–183.

Chandrasekhar, A. and R. Lewis (2012): "Econometrics of sampled networks," MIT working paper.

Conley, T. and C. Udry (2010): "Learning about a new technology: Pineapple in Ghana," *The American Economic Review*, 100, 35–69.

Duflo, E., M. Kremer, and J. Robinson (2004): "Understanding technology adoption: Fertilizer in Western Kenya, preliminary results from field experiments," *Unpublished manuscript, Massachusetts Institute of Technology.*

Galasso, E. and M. Ravallion (2005): "Decentralized targeting of an antipoverty program," *Journal of Public Economics*, 89, 705–727.

Gale, D. and S. Kariv (2003): "Bayesian learning in social networks," *Games and Economic Behavior*, 45, 329–346.

Galeotti, A. and F. Vega-Redondo (2011): "Complex networks and local externalities: A strategic approach," *International Journal of Economic Theory*, 7, 77–92.

Golub, B. and M. Jackson (2010): "Naive Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, 2, 112–149.

——— (Forthcoming): "Does Homophily Predict Consensus Times? Testing a Model of Network Structure via a Dynamic Process," *Review of Network Economics.*

——— (forthcoming): "How homophily affects learning and diffusion in networks," *Quarterly Journal of Economics.*

Jackson, M. and B. Rogers (2007): "Relating network structure to diffusion properties through stochastic dominance," *The BE Journal of Theoretical Economics*, 7, 1–13.

Jackson, M. and L. Yariv (2007): "Diffusion of behavior and equilibrium properties in network games," *The American economic review*, 97, 92–98.

Jackson, M. O. (2010): *Social and Economic Network*, Princeton.

Kremer, M. and E. Miguel (2007): "The Illusion of Sustainability*," *The Quarterly Journal of Economics*, 122, 1007–1065.

López-Pintado, D. (2008): "Diffusion in complex social networks," *Games and Economic Behavior*, 62, 573–590.

Manski, C. (1993): "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 60, 531–542.

Mossel, E., A. Sly, and O. Tamuz (2011): "From agreement to asymptotic learning," *Arxiv preprint arXiv:1105.4765*.

Mueller-Frank, M. (2011): "A general framework for rational learning in social networks," .

Munshi, K. (2003): "Networks in the Modern Economy: Mexican Migrants in the US Labor Market*," *Quarterly Journal of Economics*, 118, 549–599.

——— (2004): "Social learning in a heterogeneous population: technology diffusion in the Indian Green Revolution," *Journal of Development Economics*, 73, 185–213.

Pastor-Satorras, R. and A. Vespignani (2001): "Epidemic spreading in scale-free networks," *Physical review letters*, 86, 3200–3203.

Watts, D. and S. Strogatz (1998): "Collective dynamics of small-world networks," *Nature*, 393, 440–442.

Tables

Table 1: Descriptive Statistics

	Mean (1)	Standard Deviation (2)
Panel A: Village level		
Number of households	53.08	27.32
Average degree	8.25	2.65
Variance of degree distribution	16.35	13.63
Average clustering coefficient	0.41	0.18
Fraction of nodes in giant component	0.50	0.24
Average path length	2.02	0.50
First eigenvalue	8.57	3.13
Inequality	1.02	0.39
Error rate (consumption)	0.50	0.17
Error rate (self-assessment)	0.46	0.22
Panel B: Household level		
Degree	8.35	4.91
Clustering coefficient	0.64	0.30
Eigenvector centrality	0.23	0.14
Error rate (consumption)	0.50	0.20
Error rate (self-assessment)	0.45	0.26

Notes: Panel A provides sample statistics on the network characteristics of the 631 villages in the sample. It also provides information on the average level of competency in the village in assessing the poverty level of other members of the village. Panel B provides equivalent sample statistics for the 5,633 households in the sample.

Table 2A: The Correlation between Household Network Characteristics and the Error Rate in Ranking Income Status of Households

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: Error Rate Based on Consumption				
Degree	-0.0098***			-0.0103***	-0.0029***			-0.0021**
	(0.0010)			(0.0013)	(0.0006)			(0.0010)
Clustering		-0.0525***		-0.0415***		-0.0087		-0.0062
		(0.0128)		(0.0119)		(0.0079)		(0.0087)
Eigenvector Centrality			-0.152***	0.0582			-0.0833***	-0.0354
			(0.0327)	(0.0427)			(0.0209)	(0.0339)
R-squared	0.055	0.006	0.010	0.058	0.662	0.660	0.662	0.662
				Panel B: Error Rate Based on Self-Assessment				
Degree	-0.0130***			-0.0141***	-0.0039***			-0.0028**
	(0.0012)			(0.0016)	(0.0007)			(0.0012)
Clustering		-0.0568***		-0.0459***		-0.0028		0.0002
		(0.0158)		(0.0145)		(0.0010)		(0.0108)
Eigenvector Centrality			-0.174***	0.107**			-0.103***	-0.0439
			(0.0415)	(0.0545)			(0.0247)	(0.0408)
R-squared	0.061	0.004	0.008	0.065	0.674	0.672	0.673	0.674
				Panel C: Error Rate Based on Simulation				
Degree	-0.0280***			-0.0251***	-0.0127***			-0.0081***
	(0.0011)			(0.0011)	(0.0008)			(0.0012)
Clustering		-0.227***		-0.193***		-0.0856***		-0.0830***
		(0.0145)		(0.0113)		(0.0106)		(0.0095)
Eigenvector Centrality			-0.746***	-0.219***			-0.445***	-0.223***
			(0.0346)	(0.0376)			(0.0259)	(0.0381)
R-squared	0.314	0.068	0.163	0.387	0.676	0.655	0.683	0.690
Village Fixed Effect	No	No	No	No	Yes	Yes	Yes	Yes

Notes: This table provides estimates of the correlation between a household’s network characteristics and its ability to accurately rank the poverty status of other members of the village. The sample comprises 5,633 households. The mean of the dependent variable in Panel A (a household’s error rate in ranking others in the village based on consumption) is 0.50, while the mean of the dependent variable in Panel B (a household's error rate in ranking others in the village based on a household's own self-assessment of poverty status) is 0.46. Details of the simulation procedure for Panel C are contained in Appendix B. Standard errors are clustered by village and are listed in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 2B: The Correlation between Household Network Characteristics and the Error Rate in Ranking Income Status of Households, Controlling for Household Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: Error Rate Based on Consumption				
Degree	-0.0080***			-0.0082***	-0.0023***			-0.0016
	(0.0009)			(0.0012)	(0.0006)			(0.0009)
Clustering		-0.0489***		-0.0389***		-0.0087		-0.0063
		(0.0120)		(0.0116)		(0.0078)		(0.0086)
Eigenvector Centrality			-0.135***	0.0344			-0.0683***	-0.0318
			(0.0310)	(0.0411)			(0.0206)	(0.0331)
R-squared	0.081	0.052	0.055	0.083	0.666	0.665	0.666	0.666
				Panel B: Error Rate Based on Self-Assessment				
Degree	-0.0102***			-0.0109***	-0.0030***			-0.0021*
	(0.0012)			(0.0015)	(0.0007)			(0.0012)
Clustering		-0.0517***		-0.0423***		-0.0028		0.0001
		(0.0146)		(0.0141)		(0.0097)		(0.0106)
Eigenvector Centrality			-0.148***	0.0709			-0.0819***	-0.0376
			(0.0394)	(0.0529)			(0.0243)	(0.0398)
R-squared	0.098	0.066	0.069	0.100	0.679	0.677	0.678	0.679
				Panel C: Error Rate Based on Simulation				
Degree	-0.0277***			-0.0245***	-0.0127***			-0.0081***
	(0.0012)			(0.0011)	(0.0008)			(0.0012)
Clustering		-0.228***		-0.192***		-0.0866***		-0.0833***
		(0.0143)		(0.0114)		(0.0106)		(0.0095)
Eigenvector Centrality			-0.746***	-0.233***			-0.444***	-0.224***
			(0.0348)	(0.0379)			(0.0258)	(0.0379)
R-squared	0.316	0.096	0.188	0.390	0.676	0.656	0.683	0.691
Village FE	No	No	No	No	Yes	Yes	Yes	Yes

Notes: This table provides estimates of the correlation between a household's network characteristics and its ability to accurately rank the poverty status of other members of the village, controlling for the household's characteristics. The sample comprises 5,630 households. The mean of the dependent variable in Panel A (a household's error rate in ranking others in the village based on consumption) is 0.50, while the mean of the dependent variable in Panel B (a household's error rate in ranking others in the village based on a household's own self-assessment of poverty status) is 0.46. Details of the simulation procedure for Panel C are contained in Appendix B. Standard errors are clustered by village and are listed in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: The Correlation Between Inaccuracy in Ranking a Pair of Households in a Village and the Average Distance to Rankees

	(1)	(2)	(3)	(4)
Panel A: Consumption Metric				
Average Reachability	-0.128***	-0.0858***	-0.0657***	-0.0577**
	(0.0181)	(0.0180)	(0.0127)	(0.0278)
Average Distance	0.0183*	0.00845	0.0156***	0.0184*
	(0.0099)	(0.0094)	(0.0057)	(0.0098)
Average Degree		-0.0056***	0.0010	0.0007
		(0.0015)	(0.0026)	(0.0026)
Average Clustering		-0.0102	0.0162	0.0140
Coefficient		(0.0211)	(0.0219)	(0.0229)
Average Eigenvector		0.0485	-0.0512	-0.0590
Centrality		(0.0538)	(0.0726)	(0.0757)
R-squared	0.015	0.016	0.112	0.168
Panel B: Self-Assessment Metric				
Average Reachability	-0.152***	-0.102***	-0.0722***	-0.0397
	(0.0239)	(0.0230)	(0.0149)	(0.0324)
Average Distance	0.0209*	0.0113	0.0163**	0.0152
	(0.0126)	(0.0120)	(0.0070)	(0.0124)
Average Degree		-0.0071***	-0.0003	-0.0010
		(0.0020)	(0.0032)	(0.0033)
Average Clustering		-0.0346	-0.0098	-0.0127
Coefficient		(0.0262)	(0.0271)	(0.0273)
Average Eigenvector		0.115	0.0381	0.0098
Centrality		(0.0719)	(0.0911)	(0.0952)
R-squared	0.035	0.038	0.187	0.269
Panel C: Simulation				
Average Reachability	-0.758***	-0.624***	-0.578***	-0.726***
	(0.0080)	(0.0103)	(0.0142)	(0.0290)
Average Distance	0.0934***	0.0588***	0.0592***	0.0419***
	(0.0063)	(0.0059)	(0.0069)	(0.0094)
Average Degree		-0.0098***	-0.0079***	-0.0073***
		(0.0006)	(0.0016)	(0.0016)
Average Clustering		-0.0982***	-0.112***	-0.0577***
Coefficient		(0.0103)	(0.0143)	(0.0146)
Average Eigenvector		-0.130***	-0.199***	-0.0844
Centrality		(0.0214)	(0.0509)	(0.0516)
R-squared	0.599	0.619	0.650	0.868
Physical Controls	No	Yes	Yes	Yes
Village FE	No	No	Yes	Yes
Ranker FE	No	No	No	Yes

Notes: This table provides an estimate of the correlation between the accuracy in ranking a pair of households in a village and the characteristics of the households that are being ranked. In Panel A, the dependent variable is a dummy variable for whether person *i* ranks person *j* versus person *k* incorrectly based on using consumption as the metric of truth (the sample mean is 0.497). In Panel B, the self-assessment variable is the metric of truth (the sample mean is 0.464). In Panel C, the outcome variables are whether *i* ranks *j* versus *k* incorrectly in the simulated data, described in greater detail in Appendix B. The sample is comprised of 155,751 ranked pairs in Panel A, 117,157 in Panel B, and 139,420 in Panel C. Standard errors are clustered by village and are listed in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Stochastic Dominance

	(1)	(2)	(3)	(4)
	Panel A: Consumption Metric			
I fosd J	-0.0853***	-0.123***	-0.0902***	-0.125***
	(0.0196)	(0.0306)	(0.0195)	(0.0291)
J fosd I	0.0394**		0.0499***	
	(0.0185)		(0.0182)	
Observations	192,510	141,755	192,510	141,755
	Panel B: Self-Assessment Metric			
I fosd J	-0.102***	-0.174***	-0.0795***	-0.129***
	(0.0180)	(0.0268)	(0.0182)	(0.0264)
J fosd I	0.0748***		0.0621***	
	(0.0170)		(0.0169)	
Observations	192,510	141,755	192,510	141,755
	Panel C: Simulation			
I fosd J	-0.160***	-0.348***	-0.175***	-0.356***
	(0.0163)	(0.0224)	(0.0165)	(0.0223)
J fosd I	0.190***		0.204***	
	(0.0178)		(0.0166)	
Observations	179,101	130,739	179,101	130,739
Non-Comparable	Yes	No	Yes	No
Physical Controls	No	No	Yes	Yes
Stratification Group FE	Yes	Yes	Yes	Yes

Notes: In these regressions, the outcome variable is a dummy for whether the error rate of village *I* exceeds the error rate of village *J* . When included, physical controls are differences between the standard controls for villages *I* and *J* . Panel A presents results for error rates using the consumption metric. Panel B presents results for error rates using the self-assessment metric. Panel C presents results for error rates using simulated data, as described in Appendix B. Standard errors in parentheses, two-way clustered at I and J. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Correlation between Village Network Characteristics and Village-Level Error Rate

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Panel A: Consumption Metric			
Average Degree	-0.0073**						0.0206*
	(0.0035)						(0.0110)
Average Clustering		-0.211***					-0.265**
		(0.0632)					(0.106)
Number of Households			0.0006				0.0004
			(0.0004)				(0.0004)
First Eigenvalue of Adjacency Matrix				-0.0053*			-0.0103*
				(0.0028)			(0.0061)
Fraction of Nodes in Giant Component					-0.156***		-0.139**
					(0.0447)		(0.0599)
Link Density						-0.0808	0.183
						(0.0852)	(0.133)
R-squared	0.261	0.280	0.260	0.260	0.277	0.255	0.295
				Panel B: Self-Assessment Metric			
Average Degree	-0.0127***						0.0117
	(0.0033)						(0.0125)
Average Clustering		-0.311***					-0.321***
		(0.0640)					(0.113)
Number of Households			0.0012***				0.0007
			(0.0004)				(0.0005)
First Eigenvalue of Adjacency Matrix				-0.0063**			-0.0053
				(0.0023)			(0.0065)
Fraction of Nodes in Giant Component					-0.223***		-0.106
					(0.0461)		(0.0831)
Link Density						-0.233***	0.204
						(0.0751)	(0.135)
R-squared	0.316	0.337	0.319	0.308	0.332	0.311	0.340
				Panel C: Simulation			
Average Degree	-0.0274***						0.0424***
	(0.0037)						(0.0100)
Average Clustering		-0.303***					0.295**
		(0.0570)					(0.113)
Number of Households			0.0003				0.0001
			(0.0003)				(0.0004)
First Eigenvalue of Adjacency Matrix				-0.0243***			-0.0376***
				(0.0032)			(0.0051)
Fraction of Nodes in Giant Component					-0.379***		-0.578***
					(0.0478)		(0.0820)
Link Density						-0.381***	-0.261**
						(0.0768)	(0.113)
R-squared	0.583	0.544	0.506	0.603	0.606	0.534	0.670

Notes: This table provides village network characteristics and the error rate in ranking others in the village. Columns 1-6 show the univariate regressions, while column 7 provides the multvariate regressions. Physical covariates include consumption, education, PMT score, agricultural share, education of household head and RT head, urban dummy, stratification group FE, and inequality. The sample comprises 631 villages. Panel A presents results for error rates using the consumption metric. Panel B presents results for error rates using the self-assessment metric. Panel C presents results for error rates using simulated data, as described in Appendix B. Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Rank Correlation on Targeting Type Interacted with Diffusiveness

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Rank Correlation (Consumption)						
Community x Diffusiveness		-0.0101	-0.0102	-0.0109	-0.0125	
		(0.0164)	(0.0165)	(0.0172)	(0.0176)	
Hybrid x Diffusiveness		-0.0104	-0.0110	-0.0161		
		(0.0152)	(0.0153)	(0.0169)		
Community	-0.0587*	-0.0622*	-0.0588*	-0.0617*	-0.0594*	
	(0.0319)	(0.0323)	(0.0323)	(0.0336)	(0.0340)	
Hybrid	-0.0592*	-0.0608*	-0.0579*	-0.0515		
	(0.0327)	(0.0329)	(0.0331)	(0.0348)		
Diffusiveness		-0.0065	-0.003	0.0010	0.0057	0.0008
		(0.0109)	(0.0114)	(0.0140)	(0.0162)	(0.0140)
Community or Hybrid						-0.0571*
						(0.0294)
Community or Hybrid x Diffusiveness						-0.0135
						(0.0146)
R-squared	0.014	0.015	0.019	0.096	0.150	0.096
Panel B: Rank Correlation (Self-Assessment)						
Community x Diffusiveness		0.0323**	0.0321**	0.0278*	0.0266	
		(0.0154)	(0.0155)	(0.0164)	(0.0167)	
Hybrid x Diffusiveness		0.0368**	0.0357**	0.0353**		
		(0.0150)	(0.0151)	(0.0160)		
Community	0.108***	0.108***	0.114***	0.114***	0.112***	
	(0.0321)	(0.0323)	(0.0321)	(0.0338)	(0.0342)	
Hybrid	0.0842**	0.0793**	0.0852**	0.0846**		
	(0.0330)	(0.0334)	(0.0332)	(0.0346)		
Diffusiveness		-0.0311***	-0.0239**	-0.0272*	-0.0290*	-0.0267*
		(0.0111)	(0.0116)	(0.0143)	(0.0157)	(0.0142)
Community or Hybrid						0.100***
						(0.0299)
Community or Hybrid x Diffusiveness						0.0311**
						(0.0142)
R-squared	0.033	0.029	0.043	0.127	0.163	0.126
Keca FE	No	No	No	Yes	Yes	Yes
Demographic Covariates	No	No	No	Yes	Yes	Yes

Notes: The outcome variable is the rank correlation. Panel A presents rank correlation with community using the consumption metric. Panel B presents rank correlation with community using the self-assessment metric. Diffusiveness is the predicted value based on the first principal component vector of the covariance matrix of the network characteristics described in Table 4. Physical covariates include consumption, education, PMT score, agricultural share, education of household head and RT head, urban dummy, stratification group FE, and inequality. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7: Rank Correlation on Targeting Type Interacted with Simulated Error Rate

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Rank Correlation (Consumption)						
Community x Simulated Error Rate		-0.181	-0.166	-0.101	-0.113	
		(0.166)	(0.166)	(0.180)	(0.183)	
Hybrid x Simulated Error Rate		-0.195	-0.199	-0.126		
		(0.161)	(0.161)	(0.170)		
Community	-0.0587*	0.0484	0.0444	0.00288	0.0155	
	(0.0319)	(0.100)	(0.100)	(0.107)	(0.108)	
Hybrid	-0.0592*	0.0507	0.0583	0.0207		
	(0.0327)	(0.104)	(0.104)	(0.109)		
Simulated Error Rate		0.155	0.142	0.0896	0.117	0.0901
		(0.0993)	(0.100)	(0.129)	(0.137)	(0.128)
Community or Hybrid						0.0117
						(0.0915)
Community or Hybrid						-0.113
x Simulated Error Rate						(0.145)
R-squared	0.014	0.010	0.016	0.094	0.150	0.094
Panel B: Rank Correlation (Self-Assessment)						
Community x Simulated Error Rate		-0.378**	-0.355**	-0.364**	-0.398**	
		(0.169)	(0.167)	(0.176)	(0.176)	
Hybrid x Simulated Error Rate		-0.247	-0.257	-0.238		
		(0.180)	(0.177)	(0.182)		
Community	0.108***	0.330***	0.324***	0.328***	0.345***	
	(0.0321)	(0.102)	(0.102)	(0.107)	(0.108)	
Hybrid	0.0842**	0.235**	0.248**	0.236**		
	(0.0330)	(0.111)	(0.110)	(0.112)		
Simulated Error Rate		0.394***	0.374***	0.324**	0.330**	0.321**
		(0.117)	(0.117)	(0.150)	(0.165)	(0.149)
Community or Hybrid						0.283***
						(0.0950)
Community or Hybrid						-0.302**
x Simulated Error Rate						(0.152)
R-squared	0.033	0.040	0.055	0.138	0.178	0.137
Keca FE	No	No	No	Yes	Yes	Yes
Demographic Covariates	No	No	No	Yes	Yes	Yes

Notes: The outcome variable is the rank correlation. Panel A presents rank correlation with community using theconsumption metric. Panel B presents rank correlation with community using the self-assessment metric. The simulated error rate is described inAppendix B. It is the expected predicted value of the error rate in a hamlet under the estimated parameters of the diffusion model. Physical covariates include consumption, education, PMT score, agricultural share, education of household head and RT head, urban dummy, stratification group FE, and inequality. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix A. Network definitions

In this section, we provide basic definitions and interpretations for the different network characteristics that we consider. At the household level, we study:

- Degree: the number of links that a household has. This is a measure of how well connected a node is in the graph.
- Clustering coefficient: the fraction of a household's neighbors that are themselves neighbors. This is a measure of how interwoven a household's neighborhood is.
- Eigenvector centrality: recursively defined notion of importance. A household's importance is defined to be proportional to the sum of its neighbors' importances. It corresponds to the i^{th} entry of the eigenvector corresponding to the maximal eigenvalue of the adjacency matrix. This is a measure of how important a node is, in the sense of information flow. We take the eigenvector normalized with $\|\cdot\|_2 = 1$.
- Reachability and distance: we say two households i and j are reachable if there exists a path through the network which connects them. The distance is the length of the shortest such path.

At the hamlet level, we consider:

- Average degree: the mean number of links that a household has in the hamlet. A network with higher average degree has more edges on which to transmit information.
- Average clustering: the mean clustering coefficient of households in the hamlet.This measures how interwoven the network is.
- Average path length: the mean length of the shortest path between any two households in the hamlet.Shorter average path length means information has to travel less (on average) to get from household i to household j.
- First eigenvalue: the maximum eigenvalue of the adjacency matrix.This is a measure of how diffusive the network is. A higher first eigenvalue tends to mean that information is generally more transmitted.
- Fraction of nodes in the giant component: the share of nodes in the graph that are in the largest connected component. Typically, realistic graphs have a giant component with almost all nodes in it. Thus, the measure should be approaching one. For a network that is sampled, this number can be significantly lower. In particular, networks which were tenuously or sparsely connected, to begin with, may "shatter" under sampling and therefore the giant component may no longer be giant after sampling. In turn, it becomes a useful measure of how interwoven the underlying network is.
- Link density: the average share of connections (out of potential connections) that a household has. This measure looks at the rate of edge formation in a graph.

Appendix B. Model and Structural estimation

In this section we formally describe the model and the estimation procedure.

Model Algorithm.

(1) $t = 0$:

Every $i \in I$ is informed about i's own wealth (and only i's own wealth) and every $j \in V \backslash I$ is completely uninformed. Therefore, $S_{0i}^{(i)} = 1$ for $i \in I$ and $S_{0i}^{(j)} = 0$ for every $j \in V$ with $j \neq i$.

(2) At $t = 1$:

Every $i \in I$ informs each neighbor $j \in N_i$ about its wealth with probability p.

(3) At each $t \geq 2$:

- For every $i \in V \backslash \{j\}$ with $S_{t-1,i}^{(j)} = 1$, i informs every $k \in N_i$ about j's wealth with iid probability p.
  - If k becomes informed, then we have $S_{t,k}^{(j)} = 1$.
- For every $i \in V \backslash \{j\}$ with $S_{t-1,i}^{(j)} = 1$, $S_{t,i}^{(j)} = 0$ with probability δ.

Estimation Algorithm. Let $\Gamma = [0,1]^2$ be the parameter space and Ξ a grid on Γ. As described in Section, we use moment functions $\psi(Z;\gamma) = (\psi_1(Z;\gamma), \psi_2(Z;\gamma))'$. Also, put B as the number of bootstraps and S as the number of draws from the invariant distribution of the diffusion process used to construct the simulated moment. This nests the case with $B = 1$ when we just find the minimizer of the objective function

(1) Pick lattice $\Xi \subset \Gamma$.

(2) For $\xi \in \Xi$ on the grid:

(a) For each hamlet $r \in [R]$, compute

$$d(r,\xi) := \frac{1}{S} \sum_{s \in [S]} \psi(Z_{rs};\xi).$$

(b) For each $b \in [B]$, compute

$$D(b) := \frac{1}{R} \sum_{r \in [R]} \omega_r^b \cdot d(r,\xi)$$

where $\omega_r^b := e_{br}/\bar{e}_r$ with e_{br} iid exp(1) random variables and $\bar{e}_r := \frac{1}{R}\sum e_{br}$ if we are conducting bootstrap and $\omega_r^b = 1$ if we are just finding the minimizer.

(c) Find $\xi^{\star b} = \operatorname{argmin} Q^{\star b}(\xi)$, with

$$Q^{\star b}(\xi) := D(b)'D(b).$$

(3) Obtain $\{\xi^{\star b}\}_{b \in B}$.

(4) For conservative inference on $\widehat{\gamma}_j$, the j^{th} component, consider the $1-\alpha/2$ and $\alpha/2$ quantiles of the $\xi_j^{\star b}$ marginal empirical distribution.

We use the grid $\Xi = [0.1,\ 0.2,\ 0.3,\ 0.35,\ 0.4,\ 0.45,\ 0.5,\ 0.55,\ 0.6,\ 0.65,\ 0.7,\ 0.8,\ 0.9]^2$ and set $B = 1000$ and $S = 100$.

Numerical Propositions. We now describe the procedure by which we generated the Panel C's in Tables 2-5.

(1) Run the model described above at $\widehat{\gamma}$ for $t_0 = 50$ burn in periods.
(2) Take $S = 100$ draws from the distribution of $\mathrm{D}_{ti}^{(j,k)}$ from the diffusion process.
(3) Compute $Error_{ijkr}^{s}$ for each $i, j, k \in I$ for $s = 1, ..., S$.
(4) Compute $\overline{Error}_{ijkr} := \frac{1}{S} \sum_s Error_{ijkr}^{s}$.
(5) Compute outcome variables used in equations (3.1), (3.2), (5.1), (5.2) and the cross-hamlet regressions from $\overline{Error}_{ijkr}$ for $i, j, k \in I$ and $r \in [R]$.

Appendix C. Details on Poverty Targeting Procedures

This appendix briefly describes the poverty targeting procedures used to allocate the transfer program to households. Additional details can be found in Alatas et al. (2012).

- PMT Treatment: the government created formulas that mapped 49 easily observable household characteristics into a single index using regression techniques.[29] Government enumerators collected these indicators from all households in the PMT hamlets by conducting a door-to-door survey. These data were then used to calculate a computer-generated predicted consumption score for each household using a district-specific PMT formula. A list of beneficiaries was generated by selecting the pre-determined number of households with the lowest scores in each hamlet, based on quotes determined by a geographic targeting procedure.
- Community Treatment: To start, a local facilitator visited each hamlet to publicize the program and invite individuals to a community meeting.[30] At the meeting, the facilitator first explained the program. Next, he or she displayed the list of all households in the hamlet (which came from the baseline survey). The facilitator then spent about 15 minutes having the community brainstorm a list of characteristics that differentiate the poor from the wealthy households in their community. The facilitator then proceeded with the ranking exercise using a set of randomly-ordered index cards that displayed the names of each household in the neighborhood. He or she hung a string from wall to wall, with one end labeled as "most well-off" (paling mampu) and the other side labeled as "poorest" (paling miskin). Then, he or she started by holding up the first two name cards from the randomly-ordered stack and asking the community, "Which of these two households is better off?" Based on the community's response, he or she attached the cards along the string, with the poorer household placed closer to the "poorest" end. Next, the facilitator displayed the third card and asked how this household ranked relative to the first two households. The activity

[29] The chosen indicators encompassed the household's home attributes (wall type, roof type, etc), assets (TV, motorbike, etc), household composition, and household head's education and occupation. The formulas were derived using pre-existing survey data: specifically, the government estimated the relationship between the variables of interest and household per-capita consumption. While the same indicators were considered across regions, the government estimated district-specific formulas due to the perceived high variance in the best predictors of poverty across regions

[30] On average, 45 percent of households attended the meeting. Note, however, that we only invited the full community in half of the community treatment hamlets. In the other half (randomly selected), only local elites were invited, so that we can test whether elites are more likely to capture the community process when they have control over the process.

continued with each card being positioned relative to the already-ranked households one-by-one until complete. Before the final ranking was recorded, the facilitator read the ranking aloud so adjustments could be made if necessary. After all meetings were complete, the facilitators were provided with "beneficiary quotas" for each hamlet based on the geographic targeting procedure. Households ranked below the quota were deemed eligible.

- Hybrid Treatment: This method combines the community ranking procedure with a subsequent PMT verification. The ranking exercise, described above, was implemented first. However, there was one key difference: at the start of these meetings, the facilitator announced that the lowest-ranked households would be independently checked by the government enumerators before the beneficiary list was finalized. After the community meetings were complete, the government enumerators indeed visited the lowest-ranked households to collect the data needed to calculate their PMT score. The number of households to be visited was computed by multiplying the "beneficiary quotas" by 150 percent. Households were ranked by their PMT score, and those below the village quota became beneficiaries of the program. Thus, it was possible that some households could become beneficiaries even if they were ranked as slightly wealthier than the beneficiary quota cutoff line on the community list. Conversely, some relatively poor-ranked households on the community list might become ineligible.