NBER WORKING PAPER SERIES

USING AGGREGATED RELATIONAL DATA TO FEASIBLY IDENTIFY NETWORK STRUCTURE WITHOUT NETWORK DATA

Emily Breza
Arun G. Chandrasekhar
Tyler H. McCormick
Mengjie Pan

Working Paper 23491
http://www.nber.org/papers/w23491

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2017

We thank Liran Einav, Paul Goldsmith-Pinkham, Abhijit Banerjee, Esther Duflo, Ben Golub, Rema Hanna, Matthew Jackson, Michael Kremer, Rachael Meager, Betsy Ogburn, Elie Tamer, Tian Zheng and participants at various seminars/conferences who provided helpful comments. We also thank Shobha Dundi, Devika Lakhote, Ambika Sharma, Sneha Stephen, Tithee Mukhopadhyay, and Gowri Nagraj for outstanding research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data

Emily Breza, Arun G. Chandrasekhar, Tyler H. McCormick, and Mengjie Pan

NBER Working Paper No. 23491
June 2017
JEL No. C83,D85,L14

ABSTRACT

Social network data is often prohibitively expensive to collect, limiting empirical network research. Typical economic network mapping requires (1) enumerating a census, (2) eliciting the names of all network links for each individual, (3) matching the list of social connections to the census, and (4) repeating (1)-(3) across many networks. In settings requiring field surveys, steps (2)-(3) can be very expensive. In other network populations such as financial intermediaries or high-risk groups, proprietary data and privacy concerns may render (2)-(3) impossible. Both restrict the accessibility of high-quality networks research to investigators with considerable resources.

We propose an inexpensive and feasible strategy for network elicitation using Aggregated Relational Data (ARD) – responses to questions of the form "How many of your social connections have trait *k*?" Our method uses ARD to recover the parameters of a general network formation model, which in turn, permits the estimation of any arbitrary node- or graph-level statistic. The method works well in simulations and in matching a range of network characteristics in real-world graphs from 75 Indian villages. Moreover, we replicate the results of two field experiments that involved collecting network data. We show that the researchers would have drawn similar conclusions using ARD alone. Finally, using calculations from J-PAL fieldwork, we show that in rural India, for example, ARD surveys are 80% cheaper than full network surveys.

Emily Breza
Harvard University
Littauer Center, M28
1805 Cambridge Street
Cambridge, MA 02138
and NBER
ebreza@fas.harvard.edu

Tyler H. McCormick
Department of Statistics
Department of Sociology
University of Washington
Box 354322
Seattle, WA 98195-4322
tylermc@u.washington.edu

Arun G. Chandrasekhar
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
and NBER
arungc@stanford.edu

Mengjie Pan
Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195-4322
mpan1@uw.edu

1. Introduction

There has been a groundswell of empirical research on social and economic networks.[1] Nonetheless, a major barrier to entry into this space is access to network data, which is often extremely costly to collect. A typical network elicitation exercise requires, (1) enumerating every member of the network in a census, (2) asking each subject to name those individuals with whom they have a relationship and in what capacity, and (3) matching each individual's list of social connections back to the census. Practically speaking, this can be difficult and expensive. In the village network context, the matching step (3) can be particularly painstaking. Some teams offer full census lists to respondents and ask about links to every single fellow network member. Others collect numerous identifiers per individual and match responses to the census list, a process that often requires back-checks and multiple visits. In other contexts, such as measuring networks of financial intermediaries or high-risk populations, proprietary data and privacy concerns may render steps (2) and (3) impossible. Moreover, this process needs to be repeated across many networks to conduct convincing inference. With the help of J-PAL South Asia, we estimate that conducting full network surveys in 120 villages in India would cost approximately $190,000 and take over eight months. These high costs place significant limitations on conducting high-quality work in this space and discourage research, especially by those without access to considerable resources.

The contribution of this paper is to present a technique that makes network research scalable and accessible on a budget. We propose that researchers collect aggregated relational data (ARD). ARD are responses to questions of the form

> *"Think of all of the households in your village with whom you «INSERT ACTIVITY». How many of these have trait k?"*

ARD is considerably cheaper to obtain than full or even partial-network data. We show, using J-PAL South Asia cost estimates, that collecting ARD leads to a 70-80% cost reduction.[2]

Our proposed method is extremely intuitive and comes down to the following three simple observations. First, ARD is considerably cheaper and easier to collect than network data. Second, ARD provide the researcher with enough information to identify parameters of a oft-used and standard network formation model in the statistics literature (see e.g. Hoff et al. (2002)). The argument builds on prior work by McCormick and Zheng (2015), which shows how the network formation model is related to a likelihood that depends only on ARD. We

[1] See, e.g., Karlan, Mobius, Rosenblat, and Szeidl (2009); Centola (2010); Tontarawongsa, Mahajan, and Tarozzi (2011); Ligon and Schechter (2012); Cai, deJanvry, and Sadoulet (2013); Carrell, Sacerdote, and West (2013); Beaman, BenYishay, Magruder, and Mobarak (2016); Blumenstock, Eagle, and Fafchamps (2016); Alatas, Banerjee, Chandrasekhar, Hanna, and Olken (2016). Also see Chuang and Schechter (2015); Aral (2016); Boucher and Fortin (2016); Breza (2016) for overviews of empirical work using network data.

[2] While we present empirical evidence from village and neighborhood networks in India, the method can also be extended to other settings. See Section 7 for a discussion of applications to firm and banking networks.

describe this below. Third, this parametric model of network formation is sufficiently rich to capture a number of features of real-world social and economic network structure, as we demonstrate through a myriad of simulations and empirical exercises.

We examine the performance of our method for estimating functions of the graph in several ways. First, we show that the method works well under correct specification. Using a battery of simulations we show that we are able to guess what the underlying network structure looks like from the ARD, even as we vary the sparsity/density of the network, the size of the network, and the sampling share to reasonable degree.

Of course, real-world network data need not have been generated by the data generating process of our network-formation model. So we next consider an example where we have complete network data in nearly 16,500 households across 75 villages in Karnataka, India (Banerjee, Chandrasekhar, Duflo, and Jackson, 2016c). We show that had we collected ARD in these villages, even on a sample of 30%, we would have been able to estimate reasonably-well a variety of features of the network that economists care about.

We then provide two examples of recent research where either full or partial network data had been collected. Breza and Chandrasekhar (2016) study how the observation of one's savings behavior by more central individuals in the network leads to greater savings in order to maintain a reputation for being responsible. We show with constructed ARD, we can replicate the findings of this paper. Banerjee, Breza, Duflo, and Kinnan (2016a) study how the exposure to microcredit erodes social capital by reducing support, having collected partial network data. Having collected surveyed ARD in this sample, we show we can replicate the findings. Further, the ARD enables conclusions about how the microcredit exposure affected the overall slum informal financial network structure. These examples show the effectiveness of our approach across different contexts and how ARD would have helped in policy-relevant empirical work. Researchers could have reached their conclusions without having collected full network data, which also means that the financial barrier to entry for such research would be considerably lower, thereby democratizing in part this research frontier.

We present a sample budget for survey data collection of full network data in 120 villages. Collecting ARD reduces the costs by approximately 70-80%, depending on the sampling rate using sample budgets by J-PAL South Asia. While direct measurements of the network are always preferable to any estimation protocol, our calculations demonstrate that our proposed method can substantially expand the scope for and access to empirical networks research.

Overview of method. For the bulk of the paper, we consider settings where we have ARD for a randomly-selected subset of nodes in the network and a basic vector of covariates for the full set of nodes. ARD counts the number of links an agent has to members of different subgroups in the population. The core insight of our approach is that by combining ARD with a network formation model, we can derive the posterior distribution for the graph. To

do this, we assume a network formation model, which we refer to as the latent distance model, where the probability of a connection depends on individual heterogeneity and the positions of nodes in a latent social space (Hoff et al., 2002). The distance between nodes in the space is a pair-specific latent variable that is inversely related to the probability of a tie: nodes that are closer together in the latent space are more likely to form ties. The propensity to form ties across pairs is assumed conditionally independent given the latent variables. ARD gives us information on where different subgroups lie relative to one another in this latent space. That is, ARD allows us to triangulate the relative locations of nodes. In prior work, McCormick and Zheng (2015) show how to relate the network formation model to a likelihood that depends only on ARD. We extend that result and show how we can recover the parameters of the network formation model. In our case, this consists of both individual-level effects for every node in the sample as well as the location of all nodes in the latent-space. Using a Bayesian framework for inference, we show that the choice of prior distribution has minimal impact on our ability to accurately recover moments for a variety of network configurations. We note that, equipped with estimates of the degree distribution as well as the latent space locations in the ARD sample, we can use the demographic covariates for the entire sample to estimate the degree, fixed-effects, and latent locations for the entire population. We can then draw from the posterior distribution over graphs given the ARD response vector and compute network statistics based on these posterior samples.

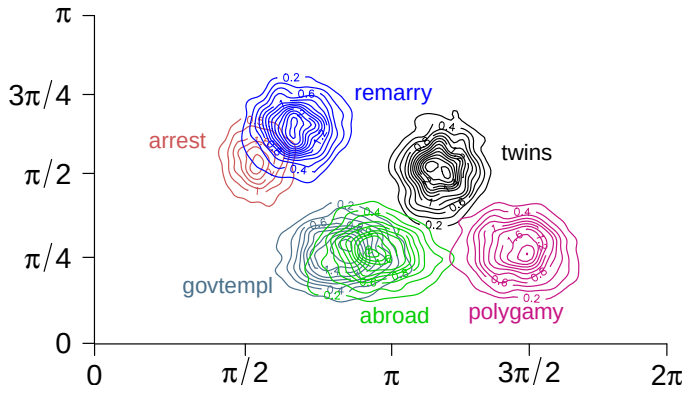


FIGURE 1. Plot of the posterior densities for six ARD characteristic groups from Hyderabad. The latent surface, a sphere, is represented by a cylindrical projection, with the vertical and horizontal axes representing latitude and longitude. Positions of the groups indicate similarity in the networks of respondents that report connections with the group. Concentration of the posterior density represents heterogeneity in the number known by respondents.

Figure 1 provides a simple illustration from one neighborhood in Hyderabad, India, where we collected ARD. The figure plots the positions on the latent surface, here a sphere, of six characteristic groups: households with histories of arrests, remarriages, members working abroad (likely in the Middle East), government employees, and twins. Several patterns

emerge in this example. First, people tend to have joint knowledge of households with arrests and remarriages, consistent with both characteristics carrying negative social stigma. Second, the arrested population is tightly correlated in space in comparison to other groups, indicating more extreme heterogeneity in the number of arrested individuals respondents know. Third, people who know individuals with government employment also often know people who have household members abroad, again consistent with the local context where both government jobs and foreign migration require connections and lead to higher incomes.

The attractive features of our approach are not without cost. A chief limitation of our approach is that it is parametric, relying on essentially guessing the network structure through the pseudo-true parameters of the latent distance formation model estimated from ARD. Thus, it can do no better than the best latent distance model can at capturing the likely distribution that generated the network. The model cannot, for example, represent clustering in a way that violates the triangle inequality. To see this, consider a two-dimensional Euclidean space with four groups that have equal probability of cross group interaction. If this is a feature of the data generating process, we will not capture it well. We discuss our limitations at the end of Section 2.

Relation to the literature. Our work contributes to and builds on several literatures. First, there is a nascent literature that seeks to apply the lessons from the economics of networks without having access to network data (e.g., Beaman et al. (2016), Banerjee et al. (2016c), and Chassang et al. (2017)). These methods are limited because they only speak to identifying central individuals or focus on proxies. Prior work shows that proxies such as geography or ethnic divisions do not capture the network well and augmenting sampled network data, which works, can still be expensive (Chandrasekhar and Lewis, 2014). Our approach does not restrict the researcher to inferences about one specific aspect of the data, instead providing a blueprint to recover a distribution over the entire graph at minimal cost.

Second, our work builds on a sizable literature on ARD, but expands both the context and inferential quantities of interest. In contrast to our work, most previous work on ARD focused on estimating the size of "hard-to-reach" populations (see e.g. Killworth et al. (1998) or Bernard et al. (2010)). These groups consist of individuals who are outside the sampling frame of most surveys. Rather than needing to reach these individuals directly, using ARD allows researchers to study individuals through their interactions with others who are captured by more traditional sampling strategies. Bernard et al. (2010) use ARD to estimate the number of individuals impacted by an earthquake whereas Kadushin et al. (2006) use ARD to estimate the number of individuals using heroine.[3]

[3]Perhaps the most common use of ARD is to estimate the number of individuals who are considered high risk for HIV/AIDS (e.g., Maghsoudi et al. (2014), Guo et al. (2013), Ezoe et al. (2012), Salganik et al. (2011)).

The primary tool for estimating population size with ARD is the Network Scale-up Method (N-Sum) and variations thereof. The general logic goes as follows. Say the goal is to estimate the number of injection drug users in the population. If a respondent reports knowing two injection drug users out of one-hundred total contacts, then approximately two percent of the respondent's network consists of individuals who are injection drug users. If the respondent's network is characteristic, then in a population of 300,000,000 individuals, this would mean there are about 6,000,000 injection drug users. Recent work has paid attention to estimating other features of the network[4], but the majority of work on ARD still focuses on estimating population sizes. Since we do not focus on populations that are hard-to-reach, we can ask directly about whether a respondent is a member of a group and, therefore, estimate population sizes directly. This distinction is essential for "scaling" a respondent's degree. If the size of each ARD group and the total population are known, we can use the N-Sum logic to estimate individuals' degrees.

The closest related work from the ARD literature is McCormick and Zheng (2015). Our work uses the same network formation model as McCormick and Zheng (2015) and builds on derivations that are the key contribution of McCormick and Zheng (2015). Specifically, McCormick and Zheng (2015) show that, for a specific formation model, it is possible to arrive at a likelihood that is informed by information in ARD. That is, McCormick and Zheng (2015) interpret and do inference on a likelihood for ARD. While we also have this likelihood, in our work it is merely an intermediate step. In our paper, we perform inferences about the parameters of the formation model itself. By explicitly making the link to the formation model, we can generate graphs and compute both graph and individual level statistics.

Third, our latent surface model[5] is closely related to the β-model (Holland and Leinhardt, 1981; Hunter, 2004; Park and Newman, 2004; Blitzstein and Diaconis, 2011) and the properties examined in Chatterjee et al. (2010) as well as Graham (2014). Every node has a fixed-effect. Links form conditionally independently given the fixed effects of the nodes involved, modulated by a function of distance between the nodes in a latent space. Relative to the Graham (2014) and Chatterjee et al. (2010) models, our model places nodes in a latent space (as in Hoff et al. (2002)), which we are trying to estimate, whereas the former only allows for observable covariates, and the latter has none. Further, whereas the previous approaches consider an asymptotic frame based on a growing graph, we consider an explicitly sampling-based framework. We empirically compare our proposed model to the beta model in Appendix C.

[4]Zheng et al. (2006), for example, estimate heterogeneity in the propensity to know members of certain groups, or overdispersion.

[5]In the context where the goal is inference about a regression coefficient that varies based on network connections, Auerbach (2016) presents a more general framework that links network formation to a function of distance between unobservable social characteristics that drive formation.

Organization. Section 2 presents the framework, model, and estimation algorithm. In Section 3 we present simulation exercises to demonstrate how our method varies with the nature of the underlying network. Section 4 shows how our method works when we apply it to 75 village where we have complete network data. In Section 5, we apply our results to two empirical examples. Section 6 demonstrates the 70-80% cost-savings based on J-PAL fieldwork that could be had using ARD instead of full networks surveys. Section 7 concludes.

2. Model and Estimation

In this section, we present a model to generate graphs using ARD. The general idea is as follows. We first propose a model for network formation based on latent variables. Then, we explicitly link the formation model to ARD, allowing us to use ARD to estimate the parameters of the formation model. Given these parameters, we can generate graphs that have features similar to those from which the respondents were sampled.

2.1. **Setup.** We begin by describing the underlying graph and the ARD. Let $g = (V, E)$ be an undirected, unweighted graph with vertex set V and edge set E, with $|V| = n$ nodes. We let $g_{ij} = \mathbf{1}\{ij \in E\}$. We also assume that researchers have a vector of demographic characteristics, X_i for every $i \in V$.

Finally, we assume that the researcher has an ARD sample of $m \leq n$ nodes which are selected uniformly at random. These could be the whole sample, with $m = n$, or a smaller share, and will depend on the context. It is useful to define V_{ard} to be the ARD sample set and $V_{non} = V \setminus V_{ard}$.

Formally, an ARD response is a count y_{ik} to a question "How many households with trait k do you know?" which we can write as

$$y_{ik} = \sum_{j \in G_k} g_{ij}$$

where $G_k \subset V$ is the set of nodes with trait k. That is, y_{ik} is a count of the number of households in group k that person i knows. Note that throughout we assume that we observe y_{ik} and, in some cases, additional information about the group of people with trait k (e.g., the number of households with this trait in the population), but we do not observe any links in the network.

It is easy to see how this could be applied to firm or banking network data. In the firm case, g is the directed, weighted supply-chain network, which is of course not observed by the researcher. G_k would be set of firms in sector k and g_{ij} would be volume of transaction between firms i and j. Here $y_{ik}^{out} = \sum_{j \in G_k} g_{ij}$ and $y_{ik}^{in} = \sum_{j \in G_k} g_{ji}$ are the total volume of directed transactions (inputs/outputs) between firm i and firms in sector k. For the remainder of the paper, we proceed with the example of a social network survey, however.

2.2. **Latent surface model.** The setup and model we use is from McCormick and Zheng (2015), motivated by, among others, Hoff et al. (2002). We model the underlying network as

$$\mathrm{P}(g_{ij} = 1|\nu_i, \nu_j, \zeta, z_i, z_j) \propto \exp(\nu_i + \nu_j + \zeta z_i' z_j), \tag{2.1}$$

where ν_i are person-specific random effects that capture heterogeneity in linking propensity. The set V of nodes occupy positions on the surface of a latent geometry. As in previous latent geometry models in the statistics and machine learning literatures, the distance between nodes on the latent surface is inversely proportional to their propensity for interaction, parsimoniously encoding homophily. Using a distance measure preserves the triangle inequality, thereby representing transitivity. That is, if the position of node i is close to that of node j and node j is close to node k, then the triangle inequality limits the distance between i and k. As we show below, equipped with the latent space terms, the model has features akin to random geometric graphs where clusters of nodes that are nearby are more likely to link, capturing realistic clustering patterns. For further discussion of the properties of this class of model see Hoff (2008). In our case, we use latent space positions on the surface of $p+1$ dimensional hypersphere, $\mathcal{Z} = \mathcal{S}^{p+1}$. As described below, the hypersphere has both conceptual and computational advantages when working with ARD. Finally, $\zeta > 0$ modulates the intensity of the latent component.

We use a Bayesian framework and, therefore, complete the model by specifying priors on the model components. We begin with the latent space. As in McCormick and Zheng (2015), we model priors for latent positions on $\mathcal{S}^{p+1}$ as

$$z_i|\upsilon_z, \eta_z \sim \mathcal{M}(\upsilon_z, 0) \text{ and}$$
$$z_{j\in G_k}|\upsilon_k, \eta_k \sim \mathcal{M}(\upsilon_k, \eta_k)$$

where $\mathcal{M}$ denotes the von Mises-Fisher distribution across $\mathcal{S}^{p+1}$. Here υ_k denotes the location on the sphere and η_k is the intensity: $\eta = 0$ means that the location is uniform at random, which makes sense since the ARD respondents are assumed to be drawn uniformly at random. The $z_{j\in G_k}$ terms describe the latent positions of individuals who have a particular trait k. For these groups, we estimate the center and spread of the distribution. The positions of these groups then triangulate the positions of individuals who have ARD. For individuals in the population without ARD data, we assign their positions based on the positions of individuals with ARD that have similar covariates.

Equipped with this, McCormick and Zheng (2015) show that the expected ARD response by i for category k can be expressed as

$$\lambda_{ik} = \mathrm{E}[y_{ik}] = d_i b_k \left(\frac{C_{p+1}(\zeta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1} \sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k \cos(\theta_{(z_i,\upsilon_k)})}} \right),$$

where d_i is the respondent degree and b_k is the share of ties made with members of group k, $C_{p+1}(\cdot)$ is the normalizing constant of the von Mises-Fisher distribution (which is a ratio depending on modified Bessel functions that is easy to compute with standard statistical software), $\theta_{(z_i,\upsilon_i)}$ is the angle between the two vectors (McCormick and Zheng, 2015). The expected number of nodes of type k known by i is roughly its expected degree scaled by the population share of the group, adjusted by a factor that captures the relative proximity of the node to the type in question in latent-space.

A key assumption in our formation model is that the propensities for individuals to form ties are conditionally independent given the latent variables. The likelihood for the formation model, conditional on the latent variables, is a Bernoulli trial for each pair. ARD, then, is the sum of (conditionally) independent Bernoulli trials, which we can approximate with a Poisson distribution. This allows us to compute the distribution of the ARD response, which will be distributed Poisson,

$$y_{ik} | d_i, b_k, \zeta, \eta_k, \theta_{(z_i,\upsilon_k)} \sim \text{Poisson}\left(\lambda_{ik}\right).$$

Though the likelihood above relies only on ARD, it does not uniquely identify the formation model since λ_{ik} estimates on the degree, d_i, rather than the individual heterogeneity parameter g_i. We can compute the expected degree as in (McCormick and Zheng, 2015),

$$d_i = n \exp(\nu_i) \mathrm{E}[\exp(\nu_j)] \left(\frac{C_{p+1}(0)}{C_{p+1}(\zeta)} \right). \tag{2.2}$$

The virtue here is that this allows us to estimate ν_i for $i \in V_{ard}$. The logic is similar to that in Chatterjee et al. (2010) or Graham (2014): in a model like the β-model, having a vector of degrees essentially provides the researcher with enough information to recover the vector of fixed-effects. If we take the above expression for each individual, then we have a system of n equations with $n+1$ unknown terms (n ν_i terms and one $\mathrm{E}[\exp(\nu_j)]$). Assuming that $\mathrm{E}[\exp(\nu_j)]$ is well-approximated by the average of the $\exp(\nu_i)$'s, we have a system with n equations and n unknowns and can, therefore recover individual ν_i terms using degree and the latent scaling term, ζ.

To complete the model, we need priors for the remaining parameters. We propose Gamma priors for ζ and η_k with conjugate priors on the hyperparameters. Then if $\boldsymbol{\theta}$ is the shorthand

for all parameters, the posterior is

$$\boldsymbol{\theta}|y_{ik} \propto \prod_{k=1}^{K}\prod_{i=1}^{n}\exp(-\lambda_{ik})\lambda_{ik}^{y_{ik}}\prod_{i=1}^{n}\text{Normal}(\log(d_i)|\mu_d,\sigma_d^2)$$
$$\times\prod_{k=1}^{K}\text{Normal}(\log(b_k)|\mu_b,\sigma_b^2)\prod_{k=1}^{K}\text{Normal}(\log(\eta_k)|\mu_{\eta_k},\sigma_{\eta_k}^2)\text{Gamma}(\zeta|\gamma_\zeta,\psi_\zeta).$$

Given the data, we can compute posteriors over degrees of nodes, their unobserved heterogeneity, population shares of categories, intensity of the latent space component in the network formation model, relative locations of categories on the sphere, and how intensely they are concentrated at these locations. So with any draw of $(z_1,...,z_n)'$, $(\nu_1,...,\nu_n)'$, and η, we can generate a graph from the distribution in (2.1).

2.3. **Intuition for Identification.** Before explaining how we go from the ARD sample to the full sample, we explain the intuition for identification of the parameters in the model. McCormick and Zheng (2015) presents a more extensive and formal discussion of identification for the latent space, as well as recommendations for the number of populations to fix based on the dimension of the hypersphere. Here we provide a simple intuition for the reader.

Figure 2 shows how the location υ_k and the concentration η_k for category k is intuitively identified assuming the latent geometry is a plane. Holding the location of three nodes fixed (here Tyler, Emily and Mengjie), and holding fixed their degree, the relative locations of categories (here Red, Green, and Blue) can be identified by placing their centers and controlling the concentration to match the Poisson rates observed in the ARD. Similarly the figure shows how the $\text{E}[d_{Tyler}]$ can be identified holding fixed the location and concentration of the various categories, since this affects $\lambda_{Tyler,k}$. Because the likelihood only depends on the latent space through the distances between individuals and groups, we fix the location of the center a small number of groups to address the invariance to distance-preserving rotations.

2.4. **From ARD sample to Non-ARD sample.** Thus far we only have posteriors for our ARD sample V_{ard}. We now turn to predicting ν_i and z_i for $j \in V_{non}$.

We could, of course, use a variety of machine learning tools to predict the latent variables for the non-ARD individuals. We use k-nearest neighbors to draw this distribution. Given demographic covariates X_i for all $i \in V$, we define a distance between two nodes in the feature space $d(X_i, X_j)$ for $i, j \in V$. For each $j \in V_{non}$, we pick $i' \in V_{ard}$ such that $d(X_{i'}, X_j)$ is among the k smallest distances. We then take a weighted average of $\nu_{i'}$ and $z_{i'}$ with weights

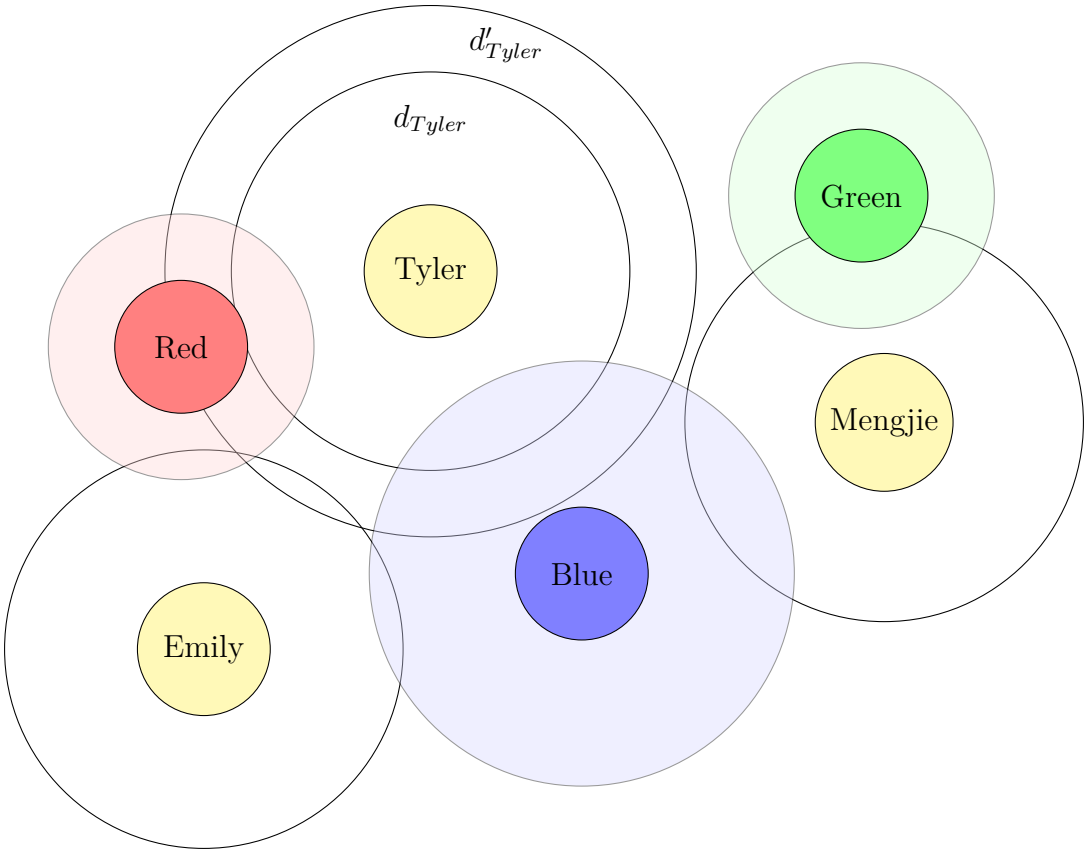


FIGURE 2. Identification of υ_k and η_k for $k \in \{\text{Red}, \text{Blue}, \text{Green}\}$ holding fixed locations and degrees of nodes in the ARD sample. Identification of $\text{E}[d_i]$ holding fixed locations and concentration parameters.

inversely proportional to $d(X_{i'}, X_j)$, to estimate ν_j and z_j, respectively. We normalize z_j such that $|z_j| = 1$ to map it to the surface of the sphere.

Taken together, we have described a framework that a researcher can use with only ARD data and demographic covariates to take a sample of draws from a network formation latent surface model.

2.5. **Drawing a graph.** We now describe the algorithm used to generate a distribution of graphs $\{g_s\}_{s=1}^{S}$. The algorithm for drawing graphs requires specifying the dimension of the latent hypersphere. Throughout the paper we follow McCormick and Zheng (2015) and use $p = 2$, for a three-dimensional hypersphere. This choice also facilitates visualising latent structure.

ALGORITHM **1** (Drawing Graphs)**.**

Input: $y_{ik} \ \forall i \in V_{ard}$, $X_i \ \forall i \in V$.

Assume ARD groups, $k = 1, ..., K$, *such that* $K \geq p$. *We propose fitting the model as follows (noting that steps 1 & 2 follow from McCormick and Zheng (2015)):*

(1) *For a subset of the ARD groups,* $k^{(s)} = 1, ..., K^{(s)}$, *fix* $\boldsymbol{\upsilon}_k^{(s)}$. *At each step we use these fixed positions in a Procrustes transformation (see Hoff et al. (2002)) to rotate the latent space back to a common orientation.*

(2) *Repeat to convergence for* $t = 1, ..., T$
   (a) *For each* i, *update* z_i *using a random walk Metropolis step with proposal* $z_i^* \sim \mathcal{M}(z_i^{(t-1)}, \textit{jumping scale})$. *Use the algorithm proposed by Wood (1994) to simulate proposals.*
   (b) *Update* $\boldsymbol{\upsilon}_k$ *using a conditionally conjugate Gibbs step (Mardia and El-Atoum, 1976; Guttorp and Lockhart, 1988; Hornik and Grün, 2013).*
   (c) *Update* d_i *with a Metropolis step with* $\log(d_i^*) \sim N(\log(d_i)^{(t-1)}, (\textit{jumping distribution scale}))$.
   (d) *Update* β *with a Metropolis step with* $\log(\beta^*) \sim N(\log(\beta)^{(t-1)}, (\textit{jumping distribution scale}))$.
   (e) *Update* η_k *with a Metropolis step with* $\eta_k^* \sim N(\eta_k^{(t-1)}, (\textit{jumping distribution scale}))$.
   (f) *Update* ζ *with a Metropolis step with* $\zeta^* \sim N(\zeta^{(t-1)}, (\textit{jumping distribution scale}))$.
   (g) *Update* $\mu_\beta \sim N(\hat{\mu}_\beta, \sigma_\beta^2)$ *where* $\hat{\mu}_\beta = \sum_{k=1}^K \beta_k / K$.
   (h) *Update* $\sigma_\beta^2 \sim \textit{Inv-}\chi^2(K-1, \hat{\sigma_\beta^2})$ *where* $\hat{\sigma_\beta^2} = \frac{1}{K-1} \sum_{k=1}^K (\beta_k - \mu_\beta)^2$.
   (i) *Update* $\mu_d \sim N(\hat{\mu}_d, \sigma_d^2)$ *where* $\hat{\mu}_d = \sum_{i=1}^n d_i / n$.
   (j) *Update* $\sigma_d^2 \sim \textit{Inv-}\chi^2(n-1, \hat{\sigma_d^2})$ *where* $\hat{\sigma_d^2} = \frac{1}{n-1} \sum_{i=1}^n (d_i - \mu_d)^2$.

(3) *Repeat for* $s \in \{T/2 + 1, ..., T\}$
   (a) *Calculate* $\nu_i^t \ \forall i \in V_{ard}$ *such that* ν_i^t *satisfies* $(d_i)^t = \exp(\nu_i^t) \sum_i \exp(\nu_i^t) \left(\frac{C_{p+1}(0)}{C_{p+1}(\zeta)}\right)$.
   (b) *Use method described in Section 2.4 to estimate* ν_j^t *and* $z_j^t \ \forall j \in V_{non}$.
   (c) *Sample graph* g_t *using the the procedure described below.*

Output: $\{g_s\}_{s=1}^S$

To generate graphs, recall that the formation model has $\mathrm{P}(g_{ij} = 1 | \nu_i, \nu_j, \zeta, z_i, z_j) \propto \exp(\nu_i + \nu_j + \zeta z_i' z_j)$. We estimate ζ and z_i, z_j using the likelihood derived in McCormick and Zheng (2015). The expression (2.2) relates degree to the unobserved gregariousness parameters, ν_i. If we approximate $\mathrm{E}[\exp(\nu_j)]$ as the average of the ν_i's, then we can view (2.2) as a system with n equations and n unknowns and obtain estimates for ν_i for each respondent.

We then normalize the $\exp(\nu_i + \nu_j + \zeta z_i^{t\,\prime} z_j^t)$ terms to produce probabilities. Define

$$\mathrm{P}(g_{ij} = 1 | z_i, z_j, \nu_i, \nu_j) = \frac{\exp(\nu_i + \nu_j + \zeta z_i' z_j) \sum_i d_i}{\sum_{i,j} \exp(\nu_i + \nu_j + \zeta z_i' z_j)}.$$

Normalizing in this way ensures $\sum_i d_i \triangleq \sum_i \sum_j \mathrm{P}(g_{ij} = 1|z_i, z_j, \nu_i, \nu_j)$. Since the formation model assumes that the propensities to form a ties between pairs are conditionally independent given the latent variables, we can now generate graphs by taking draws from a Bernoulli distribution for each pair with probability defined by $\mathrm{P}(g_{ij} = 1|z_i, z_j, \nu_i, \nu_j)$.

2.6. **Sensitivity to choice of prior distributions.** A natural question in any Bayesian analysis is how the modelers' choices about prior distributions impact posterior inferences. In our context, the priors are influential in two settings. First, as explained above, we put priors directly on the parameters of the ARD likelihood. The ARD likelihood parameters then, in turn, determine the parameters for the network formation model. To evaluate the influence of the prior distributions on our ability to estimate the parameters of the ARD likelihood (and therefore formation model), we conduct a series of experiments presented in Appendix D. For the scalar and vector parameters (e.g., the individual degree, d_i) we examine the posterior distribution after varying the spread and center of the distribution of the prior. For the latent space locations, recall that we fix some population centers for identification. To ensure that our results are not sensitive to these choices, we perform experiments where we randomly choose both which ARD population centers we fix and where these groups are positioned on the sphere's surface.

A second consideration in exploring our prior choices is the way that priors on the ARD likelihood parameters imply (via the formation model) priors on our network moments of interest. That is, we do not explicitly put a prior on centrality. The prior on centrality (and the other network moments) is, however, implied by the prior distribution placed on the parameters in the ARD likelihood. Appendix D presents a second set of results that show how the priors used for our model relate to the network moments of interest. We begin by simulating networks using the procedure above without any observed data. That is, we generate a series of networks entirely from the specified prior distributions. This series of networks demonstrates the wide range of possible networks that are supported by our formation model and the priors we specified. For context, we also plot the distribution of network moments from our estimated posterior distribution and from the observed data in Section 4.1.

2.7. **Discussion.** We have provided a simple algorithm to go from ARD questions to draws from the posterior distribution of the graph that would have given rise to ARD answers by respondents with characteristics similar to those we observed in the data. The model leverages a latent surface model similar to Hoff et al. (2002), used in McCormick and Zheng (2015), which is intimately related to the β-model studied in Chatterjee and Diaconis (2011) and Graham (2014). One issue that has arisen from both the Bayesian and frequentest perspectives is the notion of density in the limit, or the rate at which the number of edges

grows compared to the number of nodes. The Bayesian paradigm uses the Aldous-Hoover Theorem (Hoover, 1979; Aldous, 1981) for node-exchangeable graphs to justify representing dependence in the network through latent variables. This exchangeability assumption implies that a graph can be sparse if and only if it is empty (Lovász and Szegedy, 2006; Diaconis and Janson, 2007; Orbanz and Roy, 2015; Crane and Dempsey, 2015). From a frequentist perspective, Chatterjee and Diaconis (2011) show that the individual fixed effects (corresponding to, for example, gregariousness) can only be consistently estimated when the network sequence is dense.

In contrast to this previous work, however, we assume that our sample of egos arises from a population with fixed n. That is, in our paradigm there is a network of finite size, n, and we observe a small m number of actors. We see the reliance on this assumption in, for example, our expression relating degree to the individual heterogeneity parameters, ν_i. Put a different way, there is no asymptotic sequence of networks. The number of edges in a graph still impacts estimation, however. Even when the number of nodes is large, we do not expect d_i to uniformly converge to $\mathrm{E}[d_i]$ if the graph is not dense. This additional variability propagates through the model and inflates the posteriors of ν_i. These may be quite poor in practice, though it is difficult to derive the finite sample distribution. Nonetheless, what this suggests is that in cases where the network is too sparse, the ARD approach may be uninformative, and the researcher will see this plainly. This is the case for two reasons. First, by definition, anyone in the ARD sample will know fewer alters with trait k since the network has fewer links on average. Second, there will be too much variation in our location estimates and degree estimates, which then will also affect our node heterogeneity estimates. This means that when the researcher faces rather diffuse posteriors, the network may be too sparse to convey much information. We explore these issues in simulations below.

Another natural question to ask is what features does this approach capture well and what features of the network does it not capture well? Intuitively the answer to this lies in a composition of two parts: how well could the latent surface model we use do to capture real world network features and how much information about the latent surface model can we glean from ARD? We explore this below in our simulation and empirical results. Our model, of course, carries with it some of the limitations of random geometric graph sorts of models: conditional on locations on the surface, it is unlikely for very distant nodes to ever link, making so-called "short-cuts" rather rare events. Further, clustering in the network (e.g. homophily based on a given characteristic) is accomplished through the positions of particular individuals in the latent space. If there is a clear cleavage in the network (and the ARD questions asked on the survey also make it possible to detect this), then our model will generate graphs that faithfully reflect this distinction. If, however, there is a weak preference for connection within rather than between groups, this will be more difficult to detect.

3. Simulations

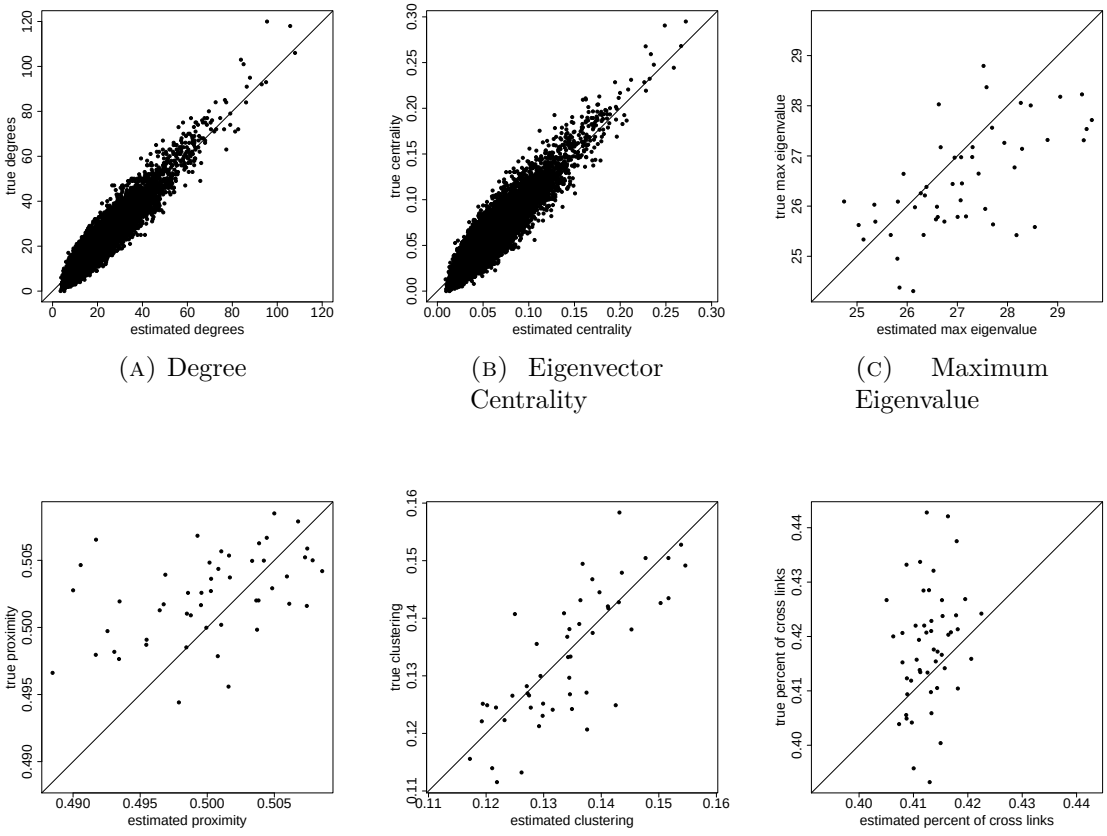
In this section we conduct simulation exercises to explore the efficacy of our procedure. The goal is several-fold. First, we want to show that the approach works in a parametric setting without any mis-specification. Second, we want to explore how the effectiveness of the approach varies with the degree of the sparsity of the network since empirical research has shown that real-world networks can be sparse. Third, we explore how well the method does in what we call a rural environment (a smaller graph of 200-500 nodes) and an urban setting (thousands of nodes). In particular, we are interested in how well the procedure works as we vary the share of the sample for which we have ARD.

To get the results in this section, we simulate ARD on a graph generated from the network formation model. We manipulate the parameter values in the formation model to produce graphs with varying levels of sparsity and sampling. As we show below, the performance of our method reflects the ability of the formation model with the parameters we specify to produce a reasonable facsimile to actual data. To better understand the performance of our method for graphs we observe in a practice, we repeat these simulations with data from Karnataka, India in the next section.

3.1. **Simulation Model.** The simulation procedure is as follows:

(1) We randomly generate n locations on $\mathcal{S}^{p+1}$ uniformly at random to get $(z_i)_{i=1}^n$.
(2) We randomly generate ν_i i.i.d. from a Normal distribution with parameters μ, σ^2.
(3) We generate a graph
$$\mathrm{P}(g_{ij} = 1 | z_i, z_j, \nu_i, \nu_j)$$
(4) We then pick K features which we maintain to be binary.
   (a) Features are located with centers distributed uniformly at random over $\mathcal{S}^{p+1}$ at sites υ_k.
   (b) Each feature k is distributed with concentration parameter η_k.
   (c) A given site i at location z_i receives feature k i.i.d. with probability $\mathrm{P}(i \in G_k) = \mathbf{1}\{u_{ik} < f(z_i|\upsilon_k, \eta_k)\}$ where u_{ik} is a uniform random variable and $f(z_i|\upsilon_k, \eta_K)$ is the von Mises-Fisher density value at location z_i.
(5) Constructed ARD responses are built using features of one's neighbors and the underlying graph.

Unless otherwise stated, we set $n = 250$, $\zeta = 0.3$, $\mu = -1.27$, $\sigma = 0.5$, and $K = 12$, which are chosen to generate graphs that resemble our empirical network data in terms of average degree 20, clustering 0.13, proximity (defined as the harmonic mean of path lengths) 0.50, average path length 2.15, and the maximal eigenvalue 26.51 of the network.



(A) Degree

(B) Eigenvector Centrality

(C) Maximum Eigenvalue

(D) Proximity

(E) Network level Clustering

(F) Eigenvector Cut

FIGURE 3. Node level and network level measures estimation for 50 simulations at core simulation set-up. These plots show scatterplots of estimated measure on the x-axis and true measure on the y-axis. There is a strong correlation between estimated statistic and statistic obtained from the true underlying graph, with the exception of eigenvector cut. The weak correlation in eigenvector cut comes from the fact that we sample individuals' and ARD subgroups' latent positions uniformly, as there is no strong separation of two groups in the true simulated graph.

3.2. **Core simulation results.** Figure 3 presents the results of our procedure using synthetic ARD data from graphs generated at these parameters. We see that the procedure works well. Throughout the paper, we look at the degree, eigenvector centrality, and clustering at the node level, as well as the maximal eigenvalue, average path length, clustering,

and eigenvector cut at the graph-level.[6] The figure shows a strong correlation between the true value in the simulation and that predicted by the ARD sample, except that the correlation is weak in eigenvector cut. The eigenvector cut takes a narrow range of values in the underlying graph, however, because we simulate both the locations of both individuals and groups uniformly across the surface of the sphere. That is, there is no cut structure in the underlying formation model.

3.3. **Varying sparsity.**

3.3.1. *Varying* $\mathrm{E}[\nu_i]$. We next hold all the parameters fixed, including $\zeta = 0.3$ at its original value, but now vary the distribution of the node effects. In particular we change the mean of the effect μ, with $\mu \in \{-1.96, -1.62, -1.27, -0.92, -0.58\}$. This varies the expected degree from 5 to 80, holding fixed $n = 250$.

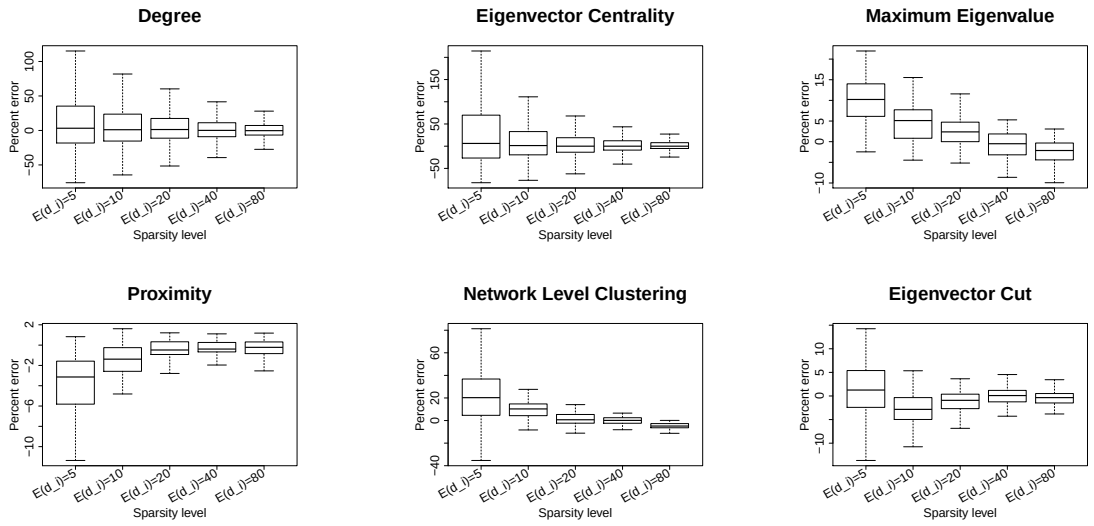


FIGURE 4. Node level and network level measures estimation for 50 simulations at each sparsity level. The plots show boxplots of percentage errors for estimated statistic, with outliers not shown on the graph. For node level measures, the bias is near zero at all sparsity levels, and variance decrease with increasing sparsity. For network level measures, the bias is overall small. Even for network level clustering estimation at the most sparse level, the middle 50% has less than 40 percent error.

We define the percentage error as the difference between the estimated and true measure divided by the true measure. At each sparsity level, we pool simulations and make plots of mean $\pm$ standard deviation of percent error. Figure 4 shows how well our algorithm estimates

[6]The eigenvector cut metric is defined by the eigenvector with the second smallest eigenvalue of the laplacian matrix. Using the median of the eigenvector to partition the graph gives us two balanced groups of equal size. We plot the fraction of links that cross group boundaries.

these measures at varying sparsity levels. As the graph becomes less sparse, we have smaller bias and variation in the estimation of degree and centrality. For maximum eigenvalue, proximity, and clustering, the bias in estimation has a monotone pattern. For proximity and clustering, we have less variation as the graph becomes less sparse. For eigenvector cut, the bias is very small at all sparsity levels and the variation decreases as the graph becomes less sparse.

3.3.2. *Sparse with thick tails.* Our next exercise is to approximate networks that exhibit heavy tails. That is, the network may mostly be sparse but some nodes may have extremely high degree. To operationalize this, we hold all the parameters fixed as before, but now draw ν_i from a Normal distribution with $\mu = -0.92, \sigma = 0.3$ with probability λ and from a Normal distribution with $\mu = -1.96, \sigma = 0.3$ with probability $1 - \lambda$. The high centrality nodes have, on average, expected degrees of 40, while the rest have, on average, expected degrees of 5. We pick $\lambda = 0.1$ so the average number of high centrality nodes is 25, but the actual number may vary in each simulation. The goal of this exercise is to study whether we can pick out which members of the network have high eigenvector centrality, which is important in a diffusion process for instance, even though the graph is extremely sparse.

		Estimated top decile		
		Yes	No	
True top decile	Yes	18.16	6.84	25
	No	6.84	218.16	225
		25	225	250

TABLE 1. Confusion matrix of top decile eigenvector centrality estimation

Table 1 shows the confusion matrix for the top decile eigenvector centrality estimation average over 50 simulations. With a 73% true positive rate and a 27% false positive rate, we successfully recover the majority of high centrality nodes. We note that the actual number of high centrality nodes varies in each simulation, which results in some noise in our estimation.

3.4. **Varying network size and sampling share.**

Next we study what happens as we move from what we call a rural environment to an urban environment. That is, what happens as the number of nodes in the population gets larger, and when we have to reduce the ARD sampling share. In particular we vary $n \in \{250, 500, 1000\}$. We also vary the share in the ARD sample, $\psi \in \{0.2, 0.5, 1\}$. When $\psi < 1$, we sample demographic features X for all nodes with $X_{i1} \sim N(\nu_i, \sigma)$. We construct X_{i2} such that X_{i2} is in one of eight categories depending on the sign of each coordinate of z_i.

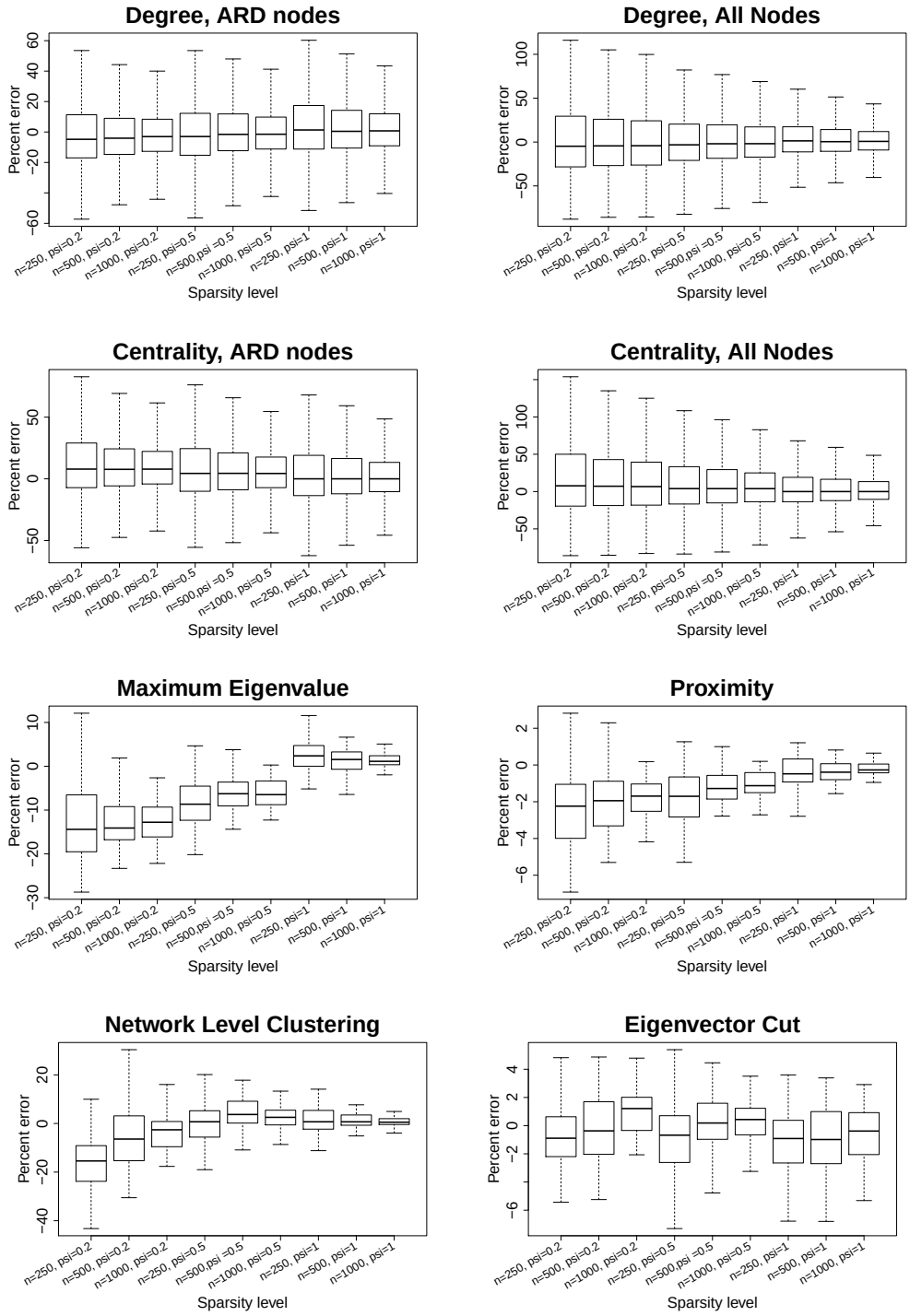


FIGURE 5. Node level and network level measures estimation for 50 simulations at each combination of $\psi \in \{0.2, 0.5, 1\}$ and $n \in \{250, 500, 1000\}$. The plots show boxplots of percentage errors for estimated statistic, with outliers not shown on the graph. The typical bias for node level statistic estimation is near zero at all levels of ψ and n, and variance decreases as we increase ψ and n. Our estimation of network level statistics improve with increasing ψ and n, with the exception of eigenvector cut. The estimated percentage of cross links has low bias and variance at all levels of ψ and n.

Figure 5 presents estimation results when we vary n and ψ. When ψ is fixed, in general we have less bias and variation as we increase n. When n is fixed, performances of degree and centrality estimation on ARD nodes are similar at various ψ. As we expect, increasing the share of ARD nodes increases the precision of node level measures estimation for all nodes.

We underestimate maximum eigenvalue when we do not have 100% ARD sampling, and we overestimate maximum eigenvalue when we do have a 100% ARD sample. We underestimate average path length at all n and ψ; the bias in estimation decreases as we increase n and ψ. Our estimation of network level clustering is within 20% of true value most of the time, and our estimation of the percentage of cross links using eigenvector cut is mostly within 5% of the true values.

4. Simulations with Real-World Networks

The goal of this section is to take the technique to the field and see how well, in a real, empirically-relevant context, we might have done using ARD in place of full network data. After all, our ARD technique can only do as well as the latent surface model does at capturing network structure.

Section 5.2 will present an exercise where we actually use surveyed ARD from the field in an empirical application, the expansion of microfinance to various slums in Hyderabad, India, to illustrate the sorts of conclusions researchers would have reached had they collected ARD from scratch.

4.1. **Setting and Data.**

We aim to show the potential for ARD to be used in place of detailed social network maps. To do this, we begin with the rich network data collected by Banerjee et al. (2016c). This consists of network data from 89% of 16,476 households across 75 villages in Karnataka, India. Thus, in the undirected, unweighted graph, we have information about 98% of all potential links. The survey asks about 12 types of interactions: (1) whose house the respondent visits; (2) who visits the respondent's house; (3) kin in the village; (4) non-relatives with whom the respondent socializes; (5) who provides help on medical decisions; (6) from whom the respondent borrows money; (7) to whom the respondent lends money; (8) from whom the respondent borrows material goods such as kerosene or rice; (9) to whom the respondent lends such material goods; (10) from whom the respondent receives advice before an important decision; (11) to whom the respondent gives advice; and (12) with whom the respondent goes to temple, mosque or church. We use a graph which is undirected and unweighted, taking a link as the union over all the above dimensions. The ratio of average degree over network size ranges from 0.04 to 0.21, with a median of 0.08. The sparsity level is the same as our core simulation, where ratio of expected degree over network size is 20/250=0.08.

We asked 12 additional questions in a follow-up survey 12 months later to a random sample of approximately 30% of households, covering traits such as owning a tractor, having met with an accident, illness incidents, birth of twins, educational attainment and family composition. We use 8 of these 12 traits as the basis for the ARD analysis. The other four questions are deleted because they are rare or non-informative of sampled households' positions in the network.

Our first goal is a proof of concept for the use of ARD and the latent distance model to generate a posterior distribution for each graph. To do this, we construct ARD responses for the 30% sample: what would be the aggregate counts these respondents would have given us had we asked them ARD questions? It also allows us to abstract from errors in knowledge or in recall by survey respondents.[7]

For what follows, the 29% of the households with supplemental surveys form our ARD sample, while the remaining 71% of households are non-ARD nodes. Because we construct ARD responses for households who answer supplemental surveys in each village, the actual percentage of households with constructed ARD responses varies by village. One village only has a 6.7% sampling rate and therefore gets dropped, increasing the sampling rate across all villages used to 30%. Recall that we observe a set of demographic covariates collected in the census of Banerjee et al. (2016c) for all nodes and we can use these covariates to predict ν_i and z_i for nodes not in the ARD sample.

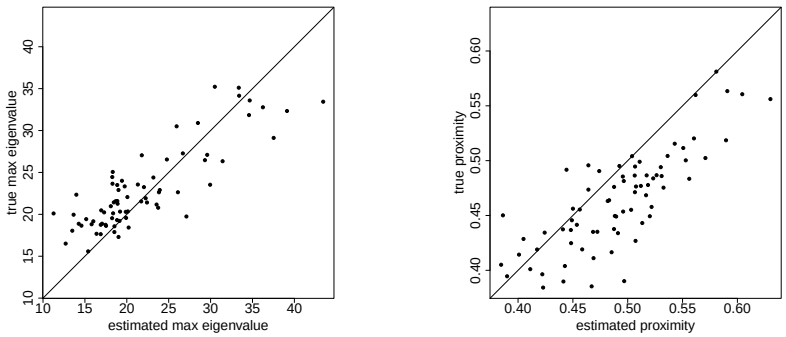
4.2. **Network Level Results.** We begin by looking at the same network-level statistics that we have focused on throughout the paper: $\lambda_1(g)$, social proximity, clustering, and eigenvector cut.

Figure 6 plots the result. In particular, each panel plots the posterior mean for the network statistic in question against the true value in the data, for each of the 75 villages. We see, rather remarkably, that these global network features are rather well-captured by the ARD procedure. The procedure is weakest for clustering but note that though there is clearly a bias, it is small and out-performs many off-the-shelf models of network formation (Chandrasekhar and Jackson, 2016).

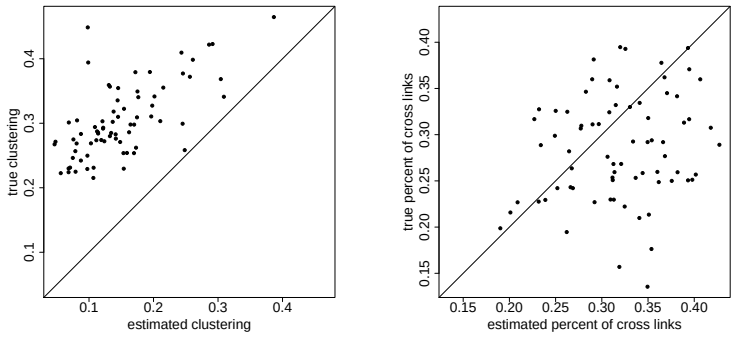
4.3. **Node Level Results.** Next we turn to node-level results. We again focus on degree, eigenvector centrality, and clustering.

Figure 7 presents the results for the ARD sample and Figure 8 presents the results for the entire sample. We see from Figure 7 that the estimated degree, eigenvector centrality, and clustering coefficient are strongly correlated with the true values in the data (Panels A, B,

[7]For example, we know the tractor ownership of each individual in the 30% sample. We can then construct the number of links of each ARD respondent to others in the ARD sample who have a tractor. This gives us the ARD responses for the induced subgraph. To estimate the number of links to tractor-owning households in the full graph, we can simply scale by the sampling rate.



(A) $\lambda_1(g)$ (B) Social Proximity



(C) Clustering (D) Eigenvector Cut

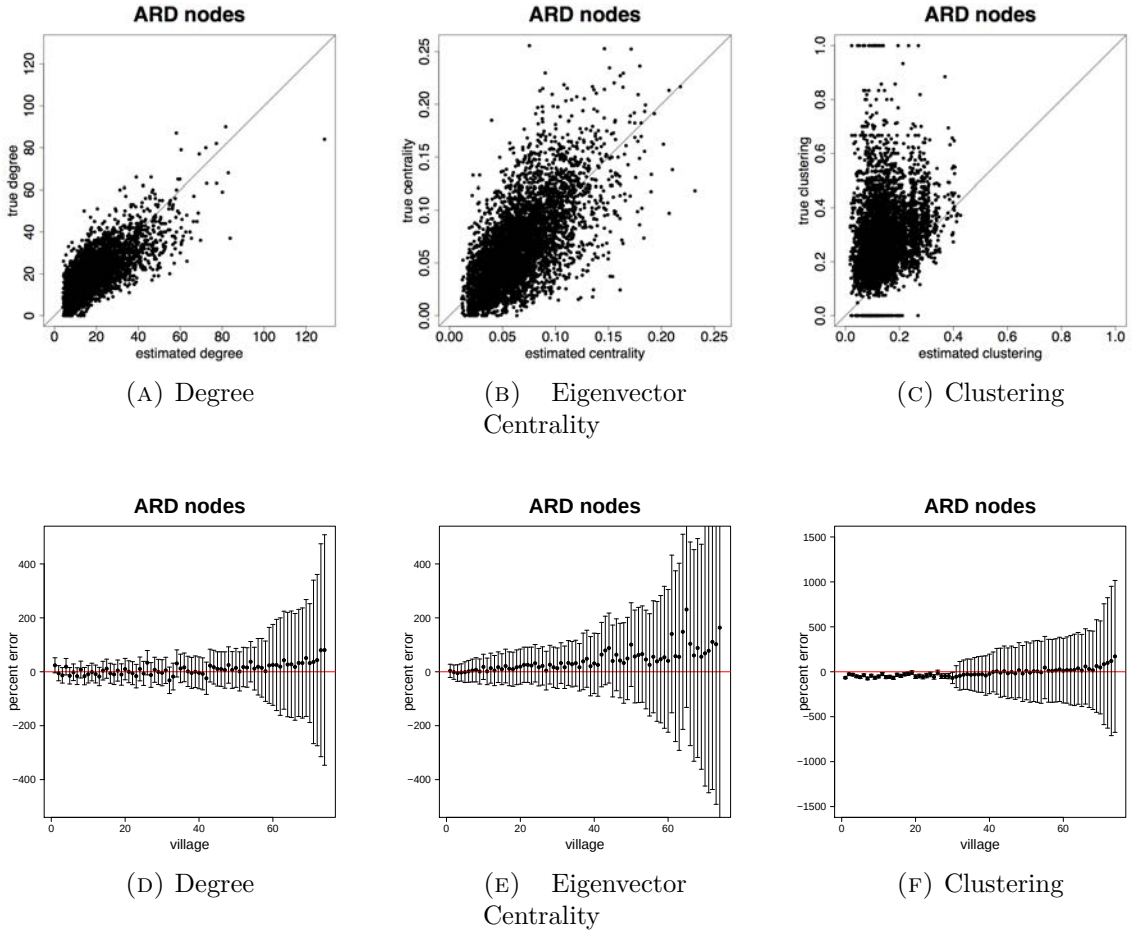
FIGURE 6. Network level measures estimation for households in villages in Karnataka. These plots show scatterplots across all villages with the estimated network level measure on the x-axis and the measure from the true underlying graph on the y-axis. There is correlation between the estimated statistic and the true statistic, even though there is some bias for clustering.

C). Furthermore, in Panels D, E, and F we plot the percent error averaged over all nodes in the sample by village, plotted by village ordered by standard deviation of percent errors.

Table 2 presents a confusion matrix to look at the probability that a node picked by a researcher using ARD is in the top decile of the centrality distribution, which is a 47% true positive rate. For comparison, this is a comparable rate to that in Banerjee et al. (2016c) using the "gossip survey" technique to elicit nominations from the village as to who is central if the nominee is also a social or political leader in the village.

Figure 8 repeats the above results for the entire sample. The results are largely similar to the ARD sample alone, though clearly there is more noise, as expected, when including the non-ARD sample.

Table 3 presents the confusion matrix for the entire sample, with a 29% true positive rate. We have a 16% true positive rate even when we pick top decile centrality nodes from



(A) Degree (B) Eigenvector Centrality (C) Clustering

(D) Degree (E) Eigenvector Centrality (F) Clustering

FIGURE 7. Node level measures estimation for households in villages in Karnataka. These plots show results using only nodes with ARD. Plots in the top row show scatterplots across all villages with the estimated node level measure on the x-axis and the measure from the true underlying graph on the y-axis. The bottom row shows mean +/− standard deviation of percent errors of the estimated node level measure across all villages. We see that overall there is strong correlation between the statistic on the underlying graph and the one estimated with ARD, with the exception of clustering. With clustering as a measure of triadic closure and the specified form of our generative model, it is not surprising that node level clustering estimation is a little weak.

non-ARD sample. For context, this is about as high as a non-nominated leader (Banerjee et al., 2016c), whom a microfinance institution might specifically pick to diffuse information widely.

4.4. **Discussion.** Taken together, our results suggest that ARD with the latent distance model and the procedure proposed here is a useful tool because the researcher will have

		Estimated top decile		
		Yes	No	
True top decile	Yes	234	271	505
	No	271	4012	4283
		505	4283	4788

TABLE 2. Confusion matrix of top decile eigenvector centrality estimation for ARD nodes

		Estimated top decile		
		Yes	No	
True top decile	Yes	470	1167	1637
	No	1167	13262	14429
		1637	14429	16066

TABLE 3. Confusion matrix of top decile eigenvector centrality estimation for all nodes

reasonable estimates of a number of network features. As is unsurprising for a model of the form specified here, it is a little bit weak when it comes to clustering.

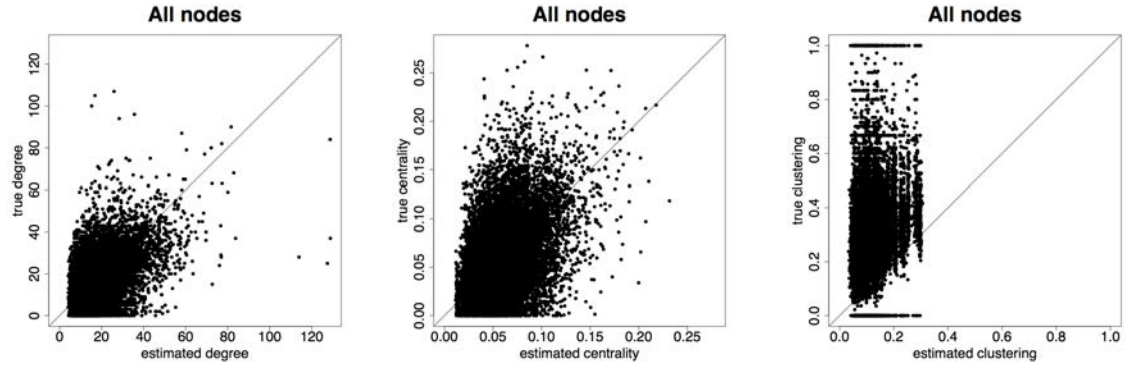
5. Empirical Applications

We now present two empirical applications that use ARD techniques. They build upon prior work by the authors, in part. The goal is to illustrate here that a researcher could have done this sort of economic analysis using ARD only, equipped with our method.

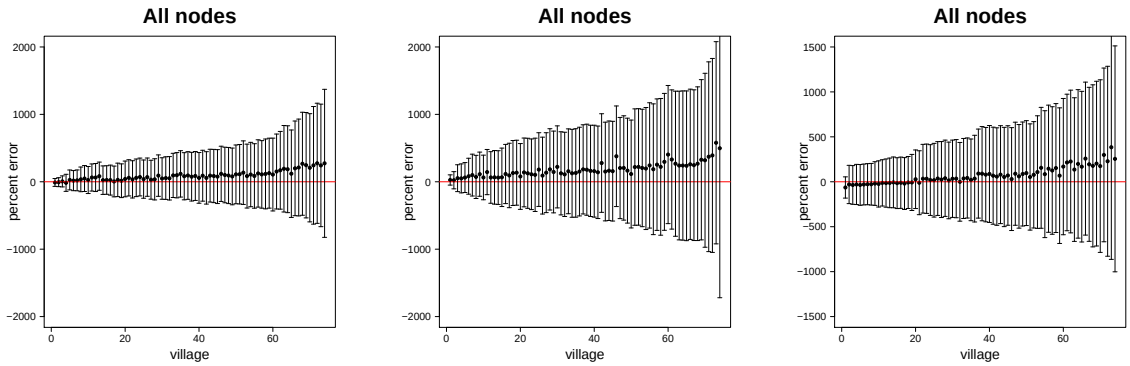
The first example looks at what would have happened if the researchers had obtained ARD for an experiment on savings and reputation. The second example actually looks at a setting where survey ARD was collected.

5.1. **Encouraging savings behavior in rural Karnataka.**

Our first application builds on Breza and Chandrasekhar (2016). The authors study social reputation through the lens of savings. In a field experiment, savers set 6-month targets for themselves. They do so knowing they may be assigned a "monitor," a villager who will be notified biweekly about their progress. Progressing towards a self-set target exhibits more responsibility, providing an avenue for the saver to build reputation with the monitor and others in the community. In 30 villages, monitors are randomly assigned to a subset of savers. This generates variation in the position of the monitor in the network. Because the monitor is free to talk to others, information about the saver's progress and reputation may spread.



(A) Degree (B) Eigenvector Centrality (C) Clustering



(D) Degree (E) Eigenvector Centrality (F) Clustering

FIGURE 8. Node level measures estimation for households in villages in Karnataka. These plots show results using all nodes. Plots in the top row show scatterplots across all villages with the estimated node level measure on the x-axis and the measure from the true underlying graph on the y-axis. The bottom row shows mean +/− standard deviation of percent errors of the estimated node level measure across all villages. We see that overall there is weak correlation between the statistic on the underlying graph and the one estimated with ARD. The weak correlation for non-ARD nodes comes from the noisy mapping from demographic covariates to ν and z_i.

A signaling model on a network guides the analysis: if the saver is more central, information can spread more widely, and if the saver is more proximate to the monitor, information likely spreads to those with whom the saver is more likely to interact in the future.

Breza and Chandrasekhar (2016) have near-full network data in 60 villages (from the Banerjee et al. (2016c) sample) and find that randomly-selected monitors increase household savings across all accounts by 35%. Breza and Chandrasekhar (2016) present a simple signaling model to investigate which randomly assigned monitors would be more effective at generating savings by the saver out of reputational concern. For saver i and monitor j, the

model show that the network matters for signaling through the quantity[8]

$$q_{ij} = \frac{1}{n}\text{Monitor Centrality} \times \text{Saver Centrality} + n \cdot \text{Proximity of Saver-Monitor}.$$

Consistent with the model, a one-standard deviation increase in q_{ij} leads to an additional 29.6% increase in total savings. Additionally, 15 months after the end of our savings period, they show that reputational information spread: randomly selected individuals surveyed about savers in the study were more likely to have updated correctly about a saver's responsibility when the saver was randomly assigned a more central monitor. Breza and Chandrasekhar (2016) also show that the savings increase persisted, and in the intervening 15 months, monitored savers were better able to cope with shocks. To benchmark the results, in 30 other villages savers chose their monitors. Monitored savers save similar amounts and non-monitored savers increase their savings relative to their random-assignment village counterparts.

Here we demonstrate how the core conclusions of this study could have been measured using only ARD data. So, if one took the perspective of a policymaker and used ARD data to guess which monitors would make for the best matches to potential savers, the procedure would have worked well.

TABLE 4. Log total savings across all household accounts regressed on monitor centrality

	(1) Log Total Ending Savings	(2) Log Total Ending Savings
Signaling value of monitor with full network data (q_{ij}), Standardized	0.259	
	(0.0937)	
Predicted signaling value of monitor with ARD (q_{ij}), Standardized		0.179
		(0.0826)
Observations	422	422
R-squared	0.148	0.138
Number of villages	30	30

Notes: Standard errors clustered at village level in parentheses.

Table 4 presents regressions of the log of total household savings across all household accounts against the model-based measure of how much signaling value the monitor provides the saver, q_{ij}. In the experiment we showed that a one standard deviation increase in q_{ij} due to random assignment of the monitor led to a 29.6% increase in total household savings

[8]Formally Breza and Chandrasekhar (2016) shows

$$q_{ij} = \frac{1}{n}\sum_k p_{jk}\sum_k p_{ik} + n \cdot \text{cov}(p_{\cdot i}.p_{\cdot j})$$

Here $p_{ij} \propto \left[\sum_{t=1}^{T}(\theta g)^t\right]$ is the probability that a unit of information that begins with i is sent to j, where transmission across each link happens with probability θ. Banerjee et al. (2016c) shows that for sufficiently high T, $\sum_k p_{jk}$ converges to the eigenvector centrality of j. Breza and Chandrasekhar (2016) shows that in equilibrium, only when q_{ij} is sufficiently high does the saver actually save.

(column 1). In column 2 we show that even if we did not have the network data, if we had ARD alone for a 30% sample, we would have had a very similar conclusion, inferring that a one standard deviation increase in predicted q_{ij} corresponds to a 19.6% increase in total household savings across all accounts. Said differently, we could have used ARD questions to easily pick good monitor-saver pairs.

5.2. **Impact of microfinance in Hyderabad.**

The goal of our final example is to demonstrate to the reader a context in which we collected and use only ARD survey questions in our analysis. We first demonstrate that the researcher could have obtained the same conclusions using the ARD instead of the network data that was collected in this study. But because the network data was incomplete (specifically the authors only measured degree – the number of links but not the identities – and support – how many links had a friend in common), the researchers could not ask how their intervention impacted the network more generally. Using ARD techniques, we show what conclusions the researchers could have learned about how the network was affected by the intervention only using the ARD survey data and estimates from the surveys of each neighborhood's average degree.

This example concerns the introduction of microfinance in Hyderabad, India. A recent literature has examined the effects that introducing microfinance to previously unbanked communities can have ambiguous and heterogenous effects on the underlying social and economic networks that facilitate informal risk-sharing. On the one hand, as in Feigenberg et al. (2013), links may be built between microfinance members and there may be an increased incentive to build links to relend (Kinnan and Townsend, 2012). On the other hand, the fact that individuals who have now become banked have less of a need to rely on informal insurance may nudge them to break links with others, and this can have local or even general equilibrium effects on the network, which can reduce density and increase paths among all nodes (Banerjee et al., 2016b).

In Banerjee et al. (2015), the authors study a randomized controlled trial where microfinance was introduced randomly to 52 out of 104 neighborhoods in Hyderabad. Banerjee, Breza, Duflo, and Kinnan (2016a) look at longer run outcomes 6 years after the intervention. This example is useful for two reasons. First, it is an urban setting where the researchers have no hope of obtaining full network data. Second, it shows how we may measure the effect of economic interventions on social network structure, as predicted by theory, despite not having network data.

In the original paper, Banerjee et al. (2016a) measure each node's degree and support, defined as the fraction of links between the respondent and a connection such that there exists a third person who is linked to both nodes in the pair. They find that both degree

and support decrease with the treatment. Note that they did not get any subgraph data since the links were not matched to a household listing: degree and support can be thought of as just two numbers.

Banerjee et al. (2016a) also collected ARD data, which we use here. In particular, a sample of approximately 55 nodes in every neighborhood was surveyed and demographic covariates as well as ARD were collected for this entire sample. As before, we fit a network formation model using the ARD data and this sample of nodes.[9] A complete list of ARD questions used in this survey is in Appendix B.

We explore whether the introduction of microfinance affected the structure of the social network by regressing

$$y_v(g) = \alpha + \beta \text{Treatment}_v + \epsilon_v$$

where v indexes neighborhood and Treatment_v is a dummy for treatment neighborhoods. Our outcome variable $y_v(g)$ of interest is the rate of support.

Theory is silent on whether density should increase or reduce, whether triadic closure (clustering or support) should increase or reduce, which can depend on a number of things: for instance, whether relending or autarky forces affect the incentives to maintain risk-sharing links (Jackson et al., 2012).

TABLE 5. Network statistics regressed on treatment

	(1) Percent Supported (Data)	(2) Percent Supported (Estimate)	(3) Graph-level Proximity (Estimate)
Treatment Neighborhood	-0.0655	-0.0575	-0.0504
	(0.0164)	(0.0250)	(0.0296)
Observations	3,514	3,598	62

Notes: Standard errors clustered at village level in parentheses. Sample includes neighborhoods with estimated sampling rate $\geq 20\%$. For large number of excluded low sampling rate neighborhoods, the population count is top-coded at 500 households. For these very large neighborhoods, we calculate the sampling rate using a population of 500.

Table 5 reports the regression results. Column 1 replicates the findings from Banerjee et al. (2016a) that past exposure decreased support. Column 2 presents the same regression, but using estimated support. The estimates of the treatment effects are quite similar, but the level of support is somewhat underestimated. We view this exercise as a "validation" of the ARD-based model. Given that the estimated treatment effect looks quite similar using the different support measures, in Column 3, we present the results of a graph-level regression, using proximity (the average inverse path length in the network) as the outcome variable. Note that it was not possible for the authors to collect such a statistic using their surveys.

[9]In this application we use the survey responses for degree and input each graph's estimated average degree directly into the model.

We also find that estimated proximity decreases, meaning that the decline in links due to microfinance exposure lead to larger average distances between households in the community. This exercise demonstrates how our method may be useful to researchers seeking to study the evolution of networks, without requiring full network data.

6. Cost Savings Using ARD

We have demonstrated that our approach for estimating network statistics has the potential to serve as a replacement for the collection of full network data. Namely, we show above that we can replicate the findings of Breza and Chandrasekhar (2016) and Banerjee et al. (2016a) with our ARD-based estimates alone. While it is always preferable to collect the underlying graph data, one important benefit from ARD is that it is substantially easier and cheaper to collect.

Table 6 presents a comparison of the costs associated with a full network survey with those of an ARD exercise for a target sample of 120 villages. Panel A summarizes the major differences in the budget assumptions between the two methods. We assume that a census is conducted in both methodologies, though household members need only be enumerated in the full network surveys. We also assume that the full network data is collected from 100% of households, while the ARD protocol samples from 30% of households. Importantly, the ARD method does not require the time consuming matching of a household's reported links with the enumerated census. Given these assumptions, Panel B of Table 6 shows that ARD is substantially cheaper, costing approximately 80% less than the full network surveys.

In Figure 9, we show that these dramatic cost reductions are not only a bi-product of the 30% sampling rate assumption. Even with 100% sampling, ARD surveys are still over 70% cheaper than the full network alternative. This sample budget highlights that using ARD estimates could indeed expand the feasibility of empirical network research.

It should go without saying that should a researcher be able to afford it, full network data is the gold-standard, and even partial network data could help being used in conjunction with ARD. The findings of this paper suggests that the Hoff (2008) model is good enough at capturing relevant features of the network. Therefore, while the network formation model can be estimated using ARD, certainly having more information about a subgraph will aid the researcher in both estimating the network formation model and integrating over the missing data in order to recover features of interest to the researcher as argued in Chandrasekhar and Lewis (2014).

TABLE 6. Cost Comparison: Full Network vs. ARD Surveys

PANEL A: ASSUMPTIONS	Traditional Network Survey	ARD Survey
Project Duration (months)	8.2	3.2
Number of Villages	120	120
Census Sampling Rate	100%	100%
Fully Enumerated Census	Yes	No
Network / ARD Survey Sampling Rate	100%	30%

Panel B: COSTS		Traditional Network Survey		ARD Survey	
	Description	Total Cost($)	Per village cost($)	Total Cost($)	Per village cost($)
Variable	Census	29,904	249	12,816	107
	Networks Survey	84,954	708	4,486	37
	Data Entry and Matching	14,284	119	-	0
	Tablet Rentals	8,584	72	1,026	9
Fixed	Project Staff Salaries	20,185	168	7,959	66
	Travel	1,617	13	638	5
	J-PAL Training/Staff Meetings	1,916	16	1,886	16
	Office Expenses	3,047	25	1,201	10
OH	J-PAL IFMR OH (15%)	24,674	206	4,502	38
	Total Cost	**189,164**	**1,576**	**34,512**	**288**

Notes: This cost comparison was prepared by J-PAL South Asia, the organization that implemented the network surveys for Banerjee et al. (2013) in Karnataka, India.

7. Conclusion

We have shown that by adding a very simple set of questions to standard survey instruments, researchers and policymakers can retrieve powerful information about the underlying social network structure. This information is easy to obtain in standard instruments and therefore can be employed in a cost-effective way.

We suggest a simple blueprint for researchers and policymakers in the field to obtain network data. If possible, researchers should add five to ten ARD questions to the census, as a standard demographic variable that would be recorded just like geographic data. If not, then researchers should at least ask ARD questions for a sample of respondents. We discuss how one might collect ARD data for use in our model in Appendix A.

There are several avenues for future research. The first would involve optimizing and standardizing ARD question design. What sorts of ARD questions should be asked? What would provide the most information to make better inferences about network structure? This has been in part the subject of work by, for example, Feehan et al. (2016) in the sociology and epidemiology literatures. Another avenue for future work builds upon the recent interest in

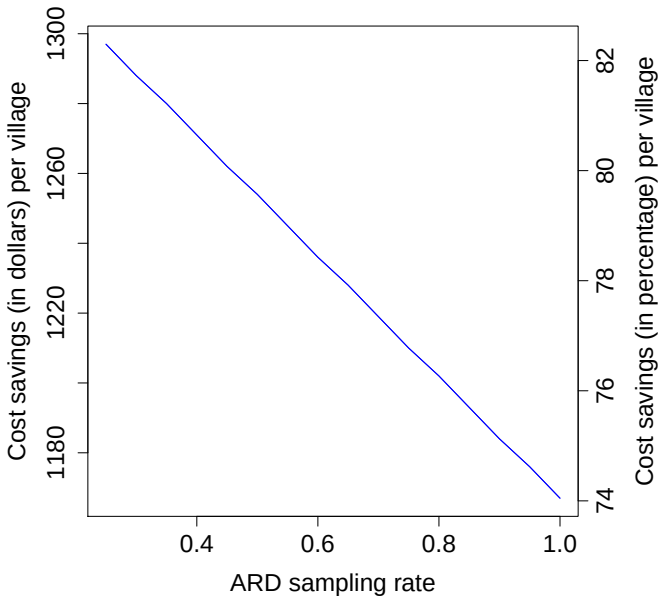


FIGURE 9. Cost Savings of ARD vs. Full Network Surveys by ARD Sampling Rate

trying to control for unobservables that both drive network structure and outcome variables of interest, the ARD approach might allow us to identify and control for latent variables.

A final avenue for future research involves looking beyond the survey network setting. Predominantly, the literature on ARD has been focused on surveyed social networks. However, we note here that our entire framework readily extends to any network context where the researchers naturally have aggregated data about links between nodes and categories of other nodes. To see this, consider the two most common economic network applications outside of social networks: inter-sectoral linkages (Acemoglu et al., 2012; Barrot and Sauvagnat, 2016; Carvalho et al., 2016) and banking (Acemoglu et al., 2015; Elliott et al., 2014).

Let us consider the simple example of a dataset where the researcher has a sample of firms and input-output data. So the researcher sees a collection of firms and then transactions the firm has with other (sub-)sectors. One can reinterpret this as simply "How many links does the firm have to firms with trait k?" where many links will now just be a weighted (by, for example the volume of trade) conditional degree instead of a conditional degree and trait k is just (sub-)sector k. This is just ARD for a weighted and directed graph.[10]

What this immediately implies is that questions of interest such as whether firm-level shocks propagate or get absorbed in their production networks (e.g., Barrot and Sauvagnat

[10]The model presented above in the paper is for cases when the underlying network is unweighted (binary) and undirected. The formation model we use is unnormalized, however, making the extension to the weighted case straightforward. We could extend the method to address directed graphs by introducing an asymmetric distance measure as suggested in, for example, Hoff et al. (2002).

(2016)) or whether if theory suggest that certain supply chains should be more robust than others to shocks, could be probed even with limited ARD data, using the techniques developed in this paper. There is nothing specific to survey network data in our statistical framework, rather it applies more broadly to any context where there are measurements of aggregate interactions between connected units.

Similarly, if we consider a dataset where the researcher sees aggregated data from bank loans, where the bilateral inter-bank loan is unavailable, but aggregated loans are (e.g., by type of bank), the methodology applies once again. Thus, our technique suggests an avenue for regulators and agencies, such as the Federal Reserve, to release anonymized data in aggregates that still allow researchers to get at important network economic questions.

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012): "The network origins of aggregate fluctuations," *Econometrica*, 80, 1977–2016. 7

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2015): "Systemic risk and stability in financial networks," *The American Economic Review*, 105, 564–608. 7

Alatas, V., A. Banerjee, A. G. Chandrasekhar, R. Hanna, and B. A. Olken (2016): "Network structure and the aggregation of information: Theory and evidence from Indonesia," *The American Economic Review*, 106, 1663–1704. 1

Aldous, D. J. (1981): "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, 11, 581–598. 2.7

Aral, S. (2016): "Networked Experiments," *Oxford Handbook on the Economics of Networks, Oxford: Oxford University Press*. 1

Auerbach, E. (2016): "Identification and estimation of models with wndogenous network formation," *Working Paper*. 5

Banerjee, A., E. Breza, E. Duflo, and C. Kinnan (2016a): "Do credit constraints limit entrepreneurship: Heterogeneity in the returns to microfinance," *Working Paper*. 1, 5.2, 5.2, 6, B

Banerjee, A., A. Chandrasekhar, E. Duflo, and M. Jackson (2013): "Diffusion of Microfinance," *Science*, 341, 1–7. ??

——— (2016b): "Changes in social network structure in response to exposure to formal credit markets," *Working Paper*. 5.2

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2016c): "Gossip: Identifying central individuals in a social network," . 1, 1, 4.1, 4.3, 4.3, 5.1, 8, A

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015): "The miracle of microfinance? Evidence from a randomized evaluation," *American Economic Journal: Applied Economics*, 7, 22–53. 5.2

Barrot, J.-N. and J. Sauvagnat (2016): "Input specificity and the propagation of idiosyncratic shocks in production networks," *The Quarterly Journal of Economics*, 1543–1592. 7

Beaman, L., A. BenYishay, J. Magruder, and A. M. Mobarak (2016): "Can network theory based targeting increase technology adoption?" *Working Paper*. 1, 1

Bernard, H. R., T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelniciuc, et al. (2010): "Counting hard-to-count populations: the network scale-up method for public health," *Sexually transmitted infections*, 86, ii11–ii15. 1

Blitzstein, J. and P. Diaconis (2011): "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet Mathematics*, 6, 489–522. 1

Blumenstock, J. E., N. Eagle, and M. Fafchamps (2016): "Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters," *Journal of Development Economics*, 120, 157–181. 1

Boucher, V. and B. Fortin (2016): "Some challenges in the empirics of the effects of networks," *Oxford Handbook on the Economics of Networks, Oxford: Oxford University Press*. 1

Breza, E. (2016): "Field experiments, social networks, and development," *The Oxford Handbook on the Economics of Networks, Oxford: Oxford University Press*. 1

Breza, E. and A. Chandrasekhar (2016): "Social Networks, Reputation and Commitment: Evidence from a Savings Monitors Experiment," *Working Paper*. 1, 5.1, 8, 6

Cai, J., A. deJanvry, and E. Sadoulet (2013): "Social networks and the decision to insure," *University of Michigan Working Paper*. 1

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013): "From natural variation to optimal policy? The importance of endogenous peer group formation," *Econometrica*, 81, 855–882. 1

Carvalho, V. M., M. Nirei, Y. U. Saito, and A. Tahbaz-Salehi (2016): "Supply chain disruptions: Evidence from the great east Japan earthquake," . 7

Centola, D. (2010): "The spread of behavior in an online social network experiment," *Science*, 329, 1194–1197. 1

Chandrasekhar, A. and R. Lewis (2014): "Econometrics of sampled networks," Stanford Working Paper. 1, 6

Chandrasekhar, A. G. and M. O. Jackson (2016): "A network formation model based on subgraphs," *Stanford University Working Paper*. 4.2

Chassang, S., P. Dupas, C. Reardon, and E. Snowberg (2017): "Selective trials for technology evaluation and adoption," *Working Paper*. 1

Chatterjee, S. and P. Diaconis (2011): "Estimating and Understanding Exponential Random Graph Models," *Arxiv preprint arXiv:1102.2650*. 2.7

Chatterjee, S., P. Diaconis, and A. Sly (2010): "Random graphs with a given degree sequence," *Arxiv preprint arXiv:1005.1136*. 1, 2.2

Chuang, Y. and L. Schechter (2015): "Social networks in developing countries," *Annual Review of Resource Economics*, 7, 451–472. 1

Crane, H. and W. Dempsey (2015): "A framework for statistical network modeling," *arXiv preprint arXiv:1509.08185*. 2.7

Diaconis, P. and S. Janson (2007): "Graph limits and exchangeable random graphs," *arXiv preprint arXiv:0712.2749*. 2.7

Elliott, M., B. Golub, and M. O. Jackson (2014): "Financial networks and contagion," *The American Economic Review*, 104, 3115–3153. 7

Ezoe, S., T. Morooka, T. Noda, M. L. Sabin, and S. Koike (2012): "Population size estimation of men who have sex with men through the network scale-up method in Japan," *PLOS ONE*, 7, 1–7. 3

Feehan, D. M., A. Umubyeyi, M. Mahy, W. Hladik, and M. J. Salganik (2016): "Quantity versus quality: A survey experiment to improve the Network Scale-up Method," *American Journal of Epidemiology*, 183, 747–757. 7

Feigenberg, B., E. Field, and R. Pande (2013): "The economic returns to social interaction: Experimental evidence from microfinance," *The Review of Economic Studies*. 5.2

Graham, B. S. (2014): "An econometric model of link formation with degree heterogeneity," *National Bureau of Economic Research Technical Report*. 1, 2.2, 2.7

Guo, W., S. Bao, W. Lin, G. Wu, W. Zhang, W. Hladik, A. Abdul-Quader, M. Bulterys, S. Fuller, and L. Wang (2013): "Estimating the size of HIV key affected populations in Chongqing, China, using the network scale-up method," *PLOS ONE*, 8, e71796. 3

Guttorp, P. and R. A. Lockhart (1988): "Finding the location of a signal: a Bayesian analysis," *Journal of the American Statistical Association*, 83, 322–330. 2b

Hoff, P. (2008): "Modeling homophily and stochastic equivalence in symmetric relational data," in *Advances in Neural Information Processing Systems*, 657–664. 2.2, 6

Hoff, P., A. Raftery, and M. Handcock (2002): "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97:460, 1090–1098. 1, 1, 1, 2.2, 1, 2.7, 10

Holland, P. W. and S. Leinhardt (1981): "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical association*, 76, 33–50. 1

Hoover, D. N. (1979): "Relations on probability spaces and arrays of random variables," *Preprint, Institute for Advanced Study, Princeton, NJ*. 2.7

Hornik, K. and B. Grün (2013): "On conjugate families and Jeffreys priors for von Mises-Fisher distributions," *Journal of statistical planning and inference*, 143, 992–999. 2b

Hunter, D. R. (2004): "MM algorithms for generalized Bradley-Terry models," *Annals of Statistics*, 384–406. 1

Jackson, M. O., T. R. Rodriguez-Barraquer, and X. Tan (2012): "Social capital and social quilts: Network patterns of favor exchange," *American Economic Review*, 102, 1857–1897. 5.2

Kadushin, C., P. D. Killworth, H. R. Bernard, and A. A. Beveridge (2006): "Scale-up methods as applied to estimates of heroin use," *Journal of Drug Issues*, 36, 417–440. 1

Karlan, D., M. Mobius, T. Rosenblat, and A. Szeidl (2009): "Trust and Social Collateral," *The Quarterly Journal of Economics*, 24, 1307–1361. 1

Killworth, P. D., C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen (1998): "Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach," *Evaluation review*, 22, 289–308. 1

Kinnan, C. and R. Townsend (2012): "Kinship and financial networks, formal financial access, and risk reduction," *The American Economic Review*, 102, 289–293. 5.2

Ligon, E. and L. Schechter (2012): "Motives for sharing in social networks," *Journal of Development Economics*, 99, 13–26. 1

Lovász, L. and B. Szegedy (2006): "Limits of dense graph sequences," *Journal of Combinatorial Theory, Series B*, 96, 933–957. 2.7

Maghsoudi, A., M. R. Baneshi, M. Neydavoodi, and A. Haghdoost (2014): "Network scale-up correction factors for population size estimation of people who inject drugs and female sex workers in Iran," *PLOS ONE*, 9, e110917. 3

Mardia, K. V. and S. A. M. El-Atoum (1976): "Bayesian inference for the von Mises-Fisher distribution," *Biometrika*, 63, 203–206. 2b

McCormick, T. H., M. J. Salganik, and T. Zheng (2010): "How many people do you know?: Efficiently estimating personal network size," *Journal of the American Statistical Association*, 105, 59–70. 3a, C

McCormick, T. H. and T. Zheng (2015): "Latent surface models for networks using Aggregated Relational Data," *Journal of the American Statistical Association*, 110, 1684–1695. 1, 1, 1, 2.2, 2.2, 2.3, 2.5, 1, 2.5, 2.7

Orbanz, P. and D. M. Roy (2015): "Bayesian models of graphs, arrays and other exchangeable random structures," *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, 37, 437–461. 2.7

Park, J. and M. E. Newman (2004): "Statistical mechanics of networks," *Physical Review E*, 70, 066117. 1

Salganik, M. J., D. Fazito, N. Bertoni, A. H. Abdo, M. B. Mello, and F. I. Bastos (2011): "Assessing network scale-up estimates for groups most at risk of HIV/AIDS: evidence from a multiple-method study of heavy drug users in Curitiba, Brazil," *American Journal of Epidemiology*, 174, 1190–1196. 3

Tontarawongsa, C., A. Mahajan, and A. Tarozzi (2011): "(Limited) Diffusion of Health-protecting Behaviors: Evidence from Non-beneficiaries of a Public Health Program in Orissa (India)," *Working Paper*. 1

Wood, A. T. A. (1994): "Simulation of the von mises fisher distribution," *Communications in statistics-simulation and computation*, 23, 157–64. 2a

Zheng, T., M. J. Salganik, and A. Gelman (2006): "How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks," *Journal of the American Statistical Association*, 101, 409–423. 4

Appendix A. Implementation Blueprint

The goal of this section is to provide a researcher or policymaker with a practical blueprint for collecting the data to use in our latent distance model. We propose this method in situations when the researchers want to estimate social network characteristics but when full social network maps are either infeasible or prohibitively expensive to collect.

To help organize thoughts, we discuss two separate cases that we call "census feasible" and "census infeasible." The census feasible case covers a situation such as a rural village, where typically there is enumeration done and basic demographics are taken at the census level. The census infeasible case covers a situation such as an urban slum where it is infeasible to expect the researcher to collect any census information on the entire population of interest.

Finally, irrespective of whichever case the researcher or policymaker is in, we encourage them to collect "gossip questions" as in Banerjee, Chandrasekhar, Duflo, and Jackson (2016c).

A.1. **Census Feasible.** In this subsection we assume that the researcher has access to a census of the population and has a possibly small vector of attributes for every unit (e.g., household or individual) in the population.

(1) Calculate what share of households will be assessed ARD.
   - This is simply a budgetary computation.

(2) Decide on which traits will enter the ARD questionnaire.
   - The traits should be from information the researcher has at the census level for the entire population.
   - The traits should satisfy the core assumptions of the model: that in a latent space sense they are located predominantly in one region (the distribution of individuals' latent positions is single-peaked). Thus, having twins may not be a very appropriate question because it is spread widely through the population whereas being of a certain sub-caste may be an appropriate question. Traits that do not have individuals in common are also helpful.
   - The traits should likely be observable by others (because eliciting the information in a survey relies on the observations of the respondent) and should not be subject to much measurement error (respondents should not know so many people with the trait that it is difficult for them to recall everyone, for example).
   - The list should not be very long, both to avoid survey fatigue and keep costs low.[11]

(3) Eliciting the ARD.

[11]However, recall that the method requires fixing the positions of three groups on the surface. Therefore, the number should be larger than five.

(a) Ask the subject to reflect on their friends (or links in whatever manner the researcher is trying to collect data in).
    - This can be recorded by the enumerators. The number of links gives the degree for each ARD node.
    - If the number of links is expected to be too large for respondents to reliably count, use a N-Sum like method (see e.g. McCormick et al. (2010)).
    - If cost permits, the names can also be matched to the census later. But if this is not cheap in terms of time or matching costs, it can be omitted.

(b) For every trait in the ARD list, ask the subject to count within their list of friends (links) how many have each trait.
    - Again, if cost permits, recording the exact identities would of course be helpful. But it is not necessary.

The end result should equip the researcher with

- ARD responses for ψ share of the population.
- The respondent's degree for ψ share of the population (if the friend list was recorded).
- Population shares by trait (computable in census data).
- Predictors for every household in the non-ARD sample.

If the researcher also had the resources or time to collect network data itself, then the researcher may have a star subgraph for the population of ψ nodes.

A.2. **Census Infeasible.** In this subsection we assume that the researcher does not have access to a census of the population and has a vector of attributes for every unit (e.g., household or individual) in the population. Intuitively, the core difference between this context and the prior context is that the researcher does not have the population share by type from the census itself.

(1) Calculate what share of households will be assessed ARD.
    - This is simply a budgetary computation.

(2) Decide on what traits will enter the ARD questionnaire.
    - The researcher must ask every node in the sample whether they have this trait as well, in order to compute an estimate of the population share.
    - The traits should satisfy the core assumptions of the model: that in a latent space sense they are located predominantly in one region (the distribution is single-peaked). Thus, having twins may not be a very appropriate question whereas being of a certain sub-caste may be an appropriate question.

- The traits should likely be observable by others (since the way one is eliciting the information relies on observations of others) and should not be subject to much measurement error.

(3) Population estimates
- These will be derived by the fact that the researcher is visiting a random sample of households and asking them each the trait questions. This population is the same as the ARD sample itself.

(4) Eliciting the ARD.
(a) Ask the subject to reflect on their friends (or links in whatever manner the researcher is trying to collect data in).
- This can be recorded by the enumerators. The number of links gives the degree for each ARD node.
- Now for every trait in the ARD list, ask the subject to count among that list of friends how many of them have each trait.

The end result should equip the researcher with

- ARD responses for ψ share of the population.
- The respondent's degree for ψ share of the population (if the friend list was recorded).
- Population share estimates by trait (computable in census data).

Appendix B. ARD Questions

This section presents the ARD questions used in Banerjee, Breza, Duflo, and Kinnan (2016a) that we use in Section 5.2.

How many other households do you know in your neighborhood ...

(1) where a woman has ever given birth to twins?
(2) where there is a permanent government employee?
(3) where there are 5 or more children?
(4) where any child has studied past 10th standard?
(5) where any adult has had typhoid, malaria, or cholera in the past six months?
(6) where any adult has been arrested by the police?
(7) where at least one woman has had a second marriage?
(8) where at least one man currently has more than one wife?

Appendix C. Comparing Latent Model to a Beta Model

We compare our model to the beta model to illustrate how adding latent positions to our model fitting procedure affects the precision of our estimation.

To fit a beta model, we first run (McCormick et al., 2010) to get a posterior distribution of estimated degrees for ARD nodes. Then taking $\zeta = 0$ in Equation (2.2), we get a posterior distribution of ν_i. As with the latent case, we generate graph using $\mathrm{P}(g_{ij} = 1|\nu_i, \nu_j) \propto \exp(\nu_i)\exp(\nu_j)$ and average measures over simulated graphs.

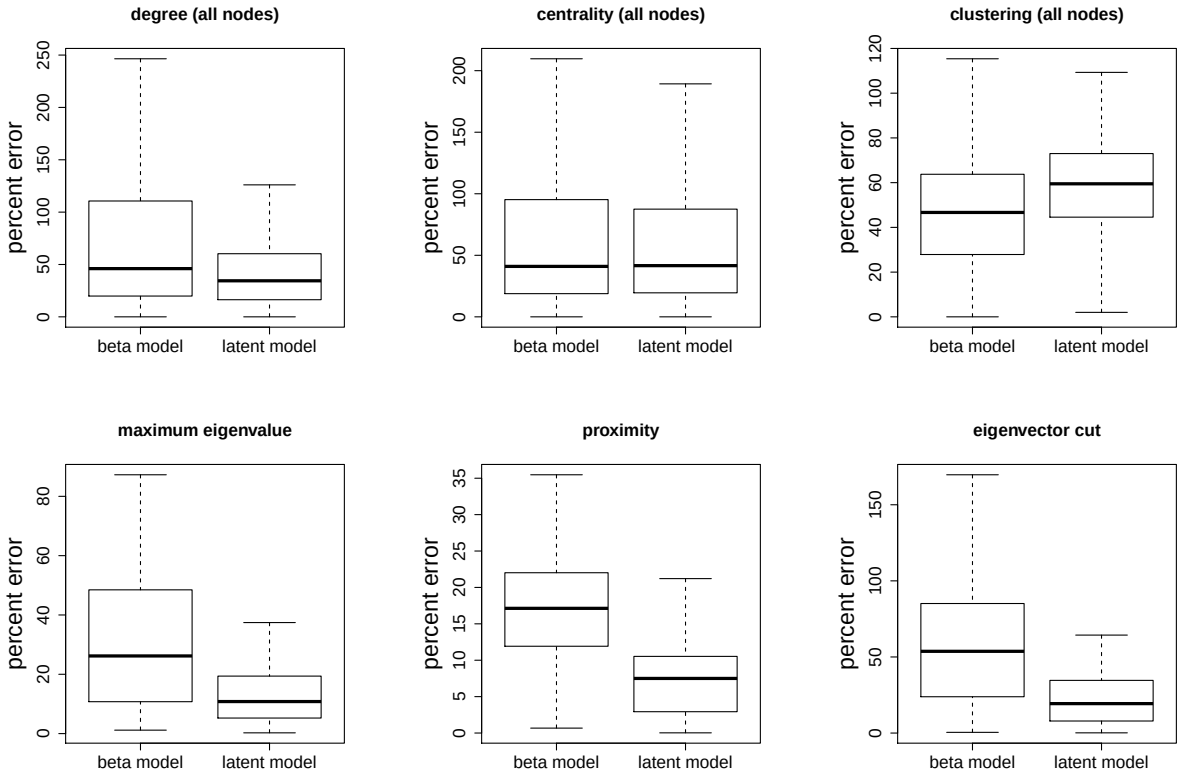


Figure C.1. Comparison of using beta model and latent model to estimate node level measures for all nodes and network level measures. These plots show boxplot of absolute percentage error for each statistic. Latent model outperforms beta model on all network level measures, and has similar performances on node level measures.

We compare beta model and latent model on degree, centrality, and clustering estimation on all nodes, as well as maximum eigenvalue, proximity, and eigenvector cut. Because the absolute percentage errors are very right skewed, we present boxplots that show the distribution for each measure (Figure C.1), with outliers omitted from the plot. The beta

model performs slightly better in estimating clustering, but performs worse in degree, proximity, maximum eigenvalue, and eigenvector cut. The mean absolute percentage error with eigenvector cut using latent model is approximately two thirds of the one using beta model. This illustrates one advantage of using a latent surface model. The propensity of forming an edge not only depends on the popularity of two nodes, but also on their distance on the latent surface. So the simulated graphs resemble the true graph's partitioning better than the simulated graphs from a beta model.

Appendix D. Prior Experiments

We show how the choice of priors and fixed subpopulations affect our results. The priors we use in Section 4 are: uniform hyperpriors for μ_d, σ_d^2, Gamma(0.5,0.5) for ζ, and Gamma(5,0.1) for η_k, and this is what "base model" in Figures D.1-D.5 refers to. We have experimented with the following alternate priors: $\mu_d \sim \mathcal{N}(0,5)$, $\mu_d \sim \mathcal{N}(2,5)$, and $\mu_d \sim \mathcal{N}(4,5)$; σ_d^2 follow inverse-chi-squared distribution with parameters (1,0.5) and (1,3); $\zeta \sim$ Gamma(2,0.5) and Uniform(0.001,10); $\eta_k \sim$ Gamma(10,0.1) and Uniform(0.1,150).

We perform two types of sensitivity analyses. First, we show that the quality of our estimates is consistent across a wide set of choices of prior values. Second, we directly examine the influence of the prior by comparing three sets of densities: the density in the observed Karnataka data, the posterior density, and the density from the prior. Additionally we consider two different ways of fixing positions of a subset of subpopulations on the latent space. In Section 4 we fix subpopulations based on their caste information and the fact that people in the same caste are more likely to know each other. Here we experiment with choosing randomly positions and which subpopulations to fix ("mukfixRandom" in Figure D.5), as well as intentionally fixing subpopulations very close to each other ("mukfixClose" in Figure D.5).

Similar to Figure C.1, Figures D.1-D.5 show the distribution of absolute percentage errors for each measure with outliers omitted. We see from these figures that changing priors and fixed subpopulations have no impact on the performances of our proposed method, although the prior on ζ has slight impact on the estimation of maximum eigenvalue.

Moving now to our second set of sensitivity analyses, Figure D.6 shows density plots for five different network features. In each of the plots we see three histograms. The green histograms represent the density of the network feature that arises from the prior distribution choices we use in Section 4. These densities arise from generating networks from the prior distributions. That is, they describe the types of networks our formation model would produce in the absence of data. As a contrast, we plot the densities from the (estimated) posterior, which includes information from both the prior and from ARD constructed using the Karnataka data. For comparison, we also included the observed density from the Karnataka data, or the "true" density.

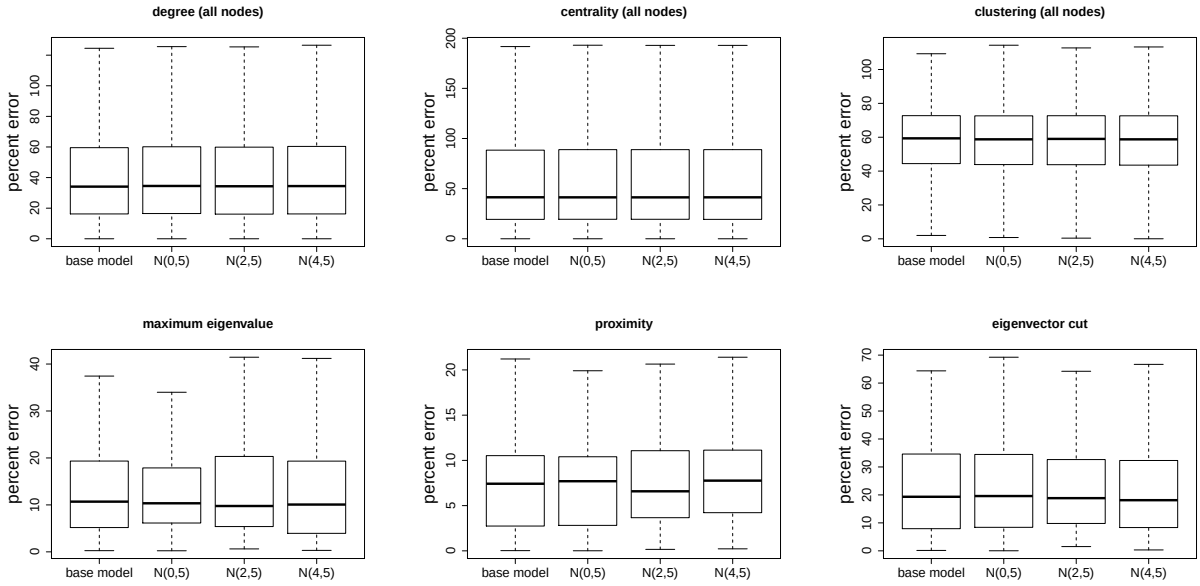


FIGURE D.1. Comparison of using uniform, $\mathcal{N}(0,5), \mathcal{N}(2,5), \mathcal{N}(4,5)$ priors for hyperparameter μ_d to estimate node level measures for all nodes and network level measures. These plots show boxplot of absolute percentage error for each statistic. Prior of μ_d do not have an impact on the results.

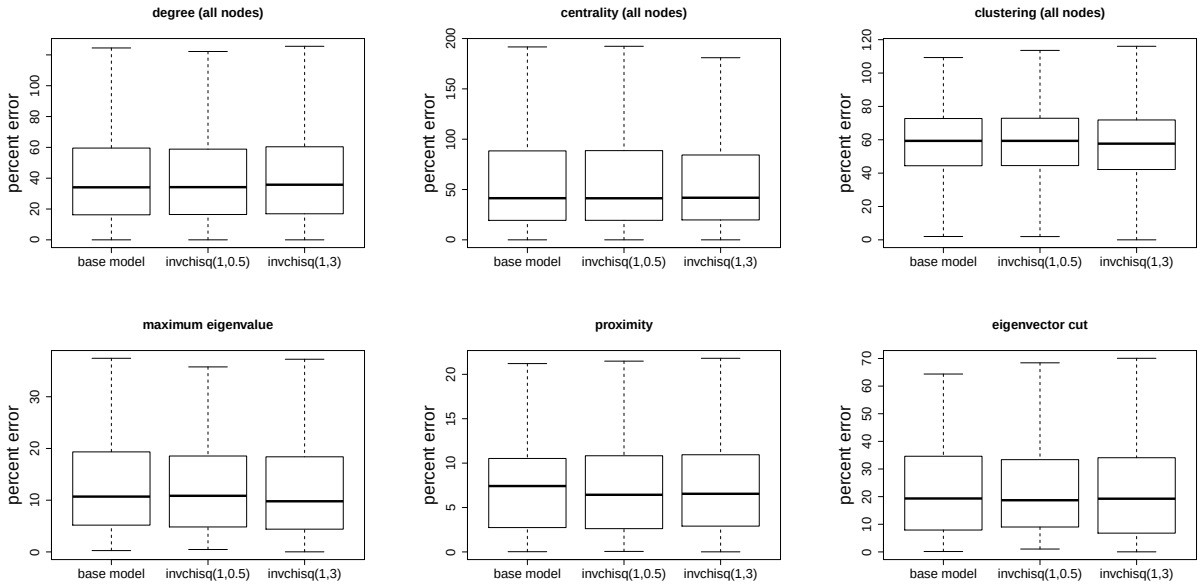


FIGURE D.2. Comparison of using uniform, inverse-chi-squared distribution with parameters (1,0.5) and (1,3) priors for hyperparameter σ_d^2 to estimate node level measures for all nodes and network level measures. These plots show boxplot of absolute percentage error for each statistic. Prior of σ_d^2 do not have an impact on the results.

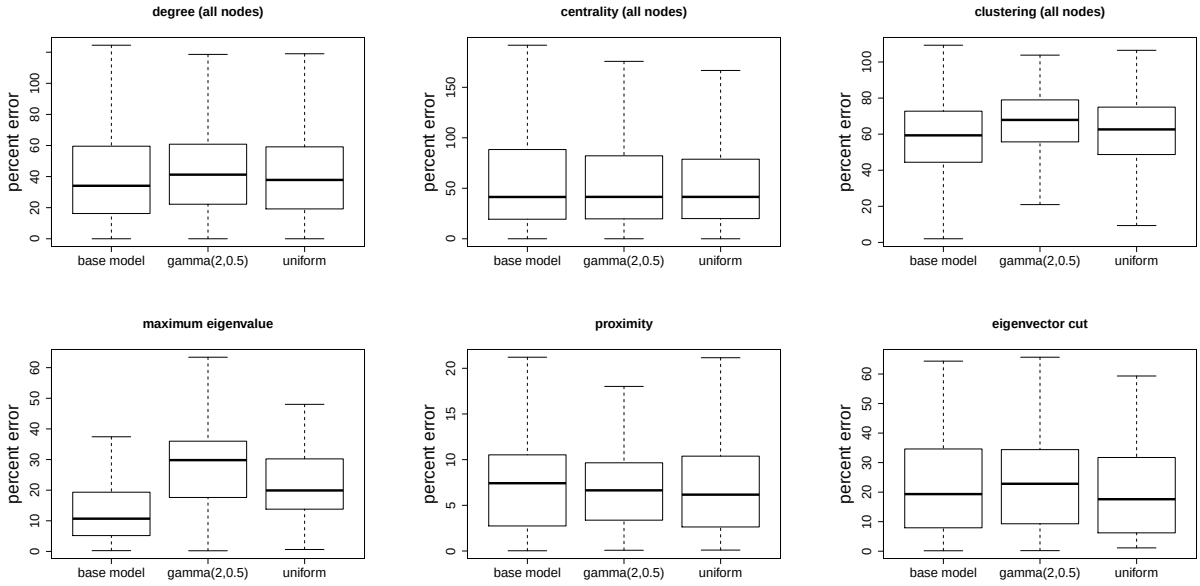


FIGURE D.3. Comparison of using Gamma(0.5,0.5), Gamma(2,0.5) and Uniform(0.001,10) priors for ζ to estimate node level measures for all nodes and network level measures. These plots show boxplot of absolute percentage error for each statistic. Prior of ζ impacts maximum eigenvalue and clustering slightly.

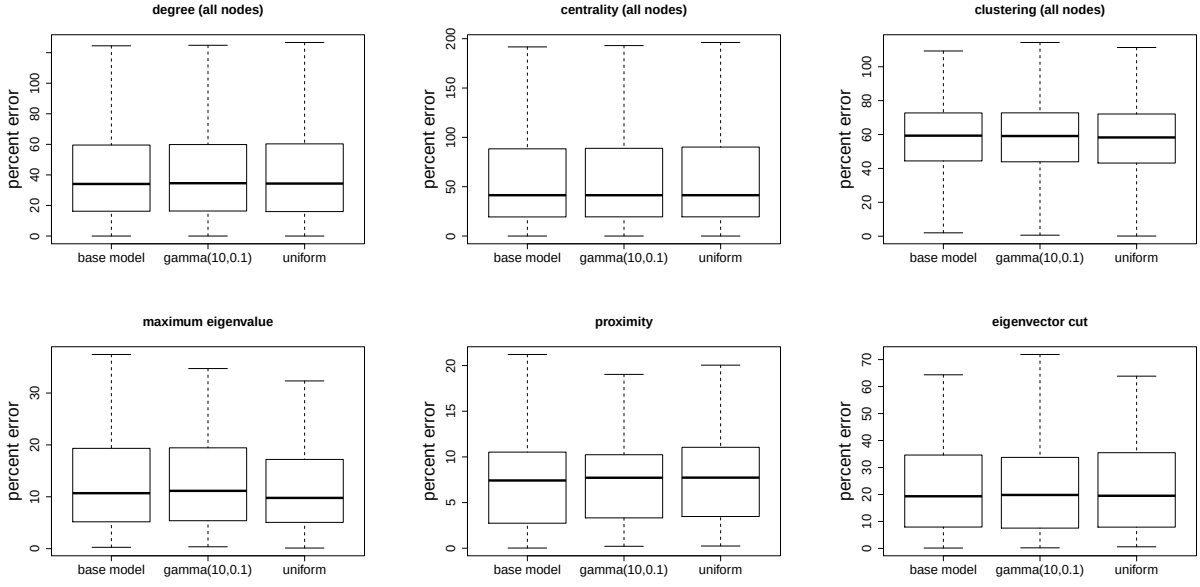


FIGURE D.4. Comparison of using Gamma(5,0.1), Gamma(10,0.1) and Uniform(0.1,150) priors for η_k to estimate node level measures for all nodes and network level measures. These plots show boxplot of absolute percentage error for each statistic. Prior of η_k do not have an impact on the results.

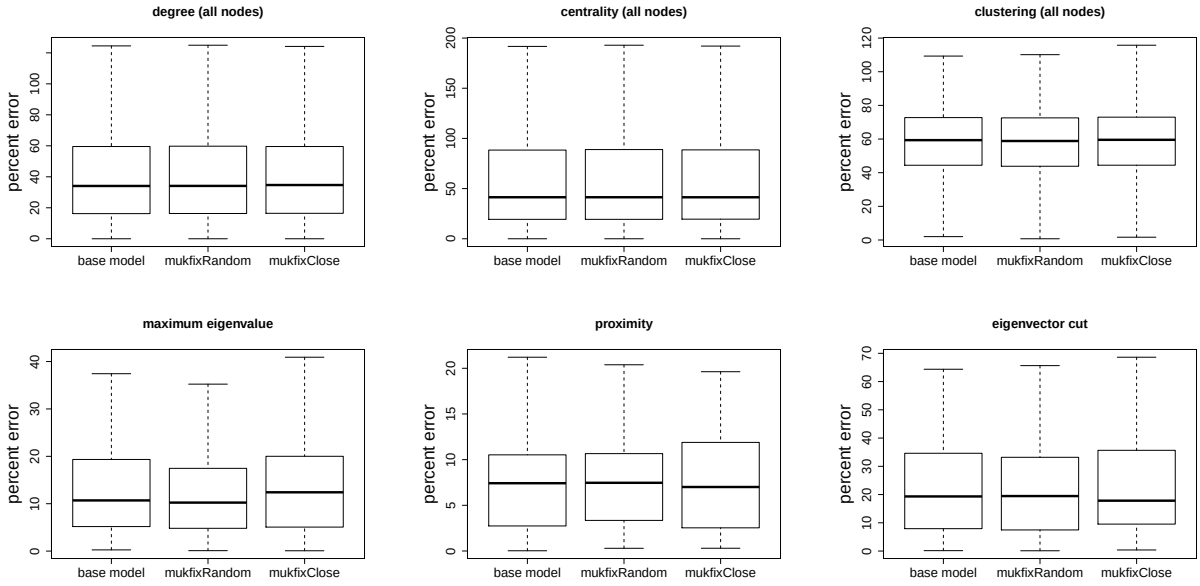


FIGURE D.5. Comparison of results from models fixing subpopulations based on caste information, fixing subpopulations randomly, and intentionally fixing subpopulations close. These plots show boxplot of absolute percentage error for node level measures for all nodes and network level measures. These three ways of fixing a subset of subpopulations do not have an impact on the results.

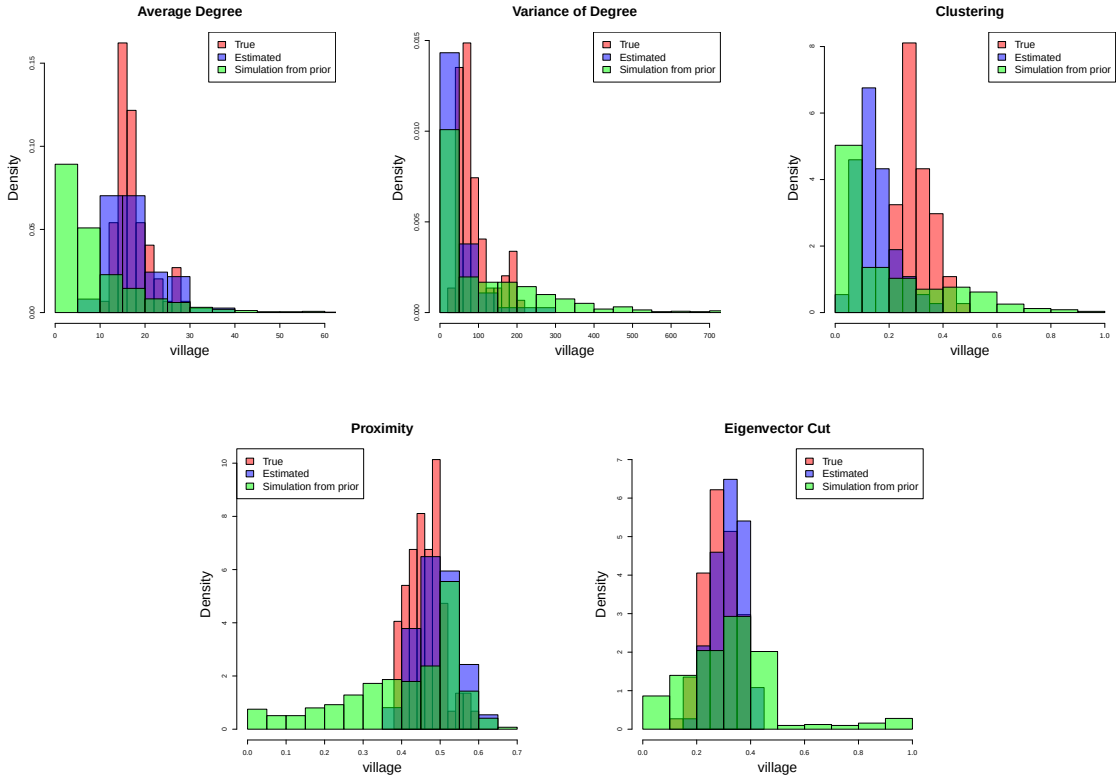


FIGURE D.6. Density of average degree, variance of degree, network-level clustering, proximity, and eigenvector cut. The histograms labeled "True" show the density observed in the Karnataka networks. The "Estimated" histograms are the density estimated from fitting our model to this data and the "Prior" histograms are the density from networks simulated using our chosen prior distributions. Overall, the "Prior" histograms have higher variance and are, in many cases, centered in different places than estimated (posterior) densities, indicating that information in the ARD data are driving estimation, rather than the prior.