American Economic Review 2020, 110(8): 2454–2484
https://doi.org/10.1257/aer.20170861

Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data[†]

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

Social network data are often prohibitively expensive to collect, limiting empirical network research. 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 links have trait k?" Our method uses ARD to recover parameters of a network formation model, which permits sampling from a distribution over node- or graph-level statistics. We replicate the results of two field experiments that used network data and draw similar conclusions with ARD alone. (*JEL* C81, C93, D85, Z13)

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 are often extremely costly to collect. A typical network elicitation exercise requires (i) enumerating every member of the network in a census, (ii) asking each subject to name those individuals with whom they have a relationship and in what capacity, and (iii) matching each individual's list of social connections back to the census. In field work, this can be difficult and expensive. Further, in other contexts, such as measuring networks of financial intermediaries or high-risk populations, proprietary data and privacy concerns may render steps (ii) and (iii) impossible. Moreover, this process needs to be repeated across many networks to conduct convincing inference. These barriers place significant

* Breza: Department of Economics, Harvard University (email: ebreza@fas.harvard.edu); Chandrasekhar: Department of Economics, Stanford University (email: agc@stanford.edu); McCormick: Departments of Statistics and Sociology, University of Washington (email: tylermc@uw.edu); Pan: Department of Statistics, University of Washington (email: mpan1@uw.edu). Pinelopi Goldberg was the coeditor for this article. We thank Paul Goldsmith-Pinkham, Abhijit Banerjee, Axel Gandy, Ben Golub, Rema Hanna, Guy Harling, Rachel Heath, Jeff Eaton, 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, Varun Kapoor, Devika Lakhote, Ambika Sharma, Sneha Stephen, Tithee Mukhopadhyay, and Gowri Nagraj for outstanding research assistance. This work is partially supported by grants SES-1326661 and SES-1559778 from the National Science Foundation and grant number K01 HD078452 from the National Institute of Child Health and Human Development (NICHD). Breza and Chandrasekhar are supported by the Sloan Foundation. This material is based upon work supported by, or in part by, the US Army Research Laboratory and the US Army Research Office under grant number W911NF-12-1-0379.

[†] Go to https://doi.org/10.1257/aer.20170861 to visit the article page for additional materials and author disclosure statements.

[1] See, e.g., Karlan et al. (2009); Centola (2010); Tontarawongsa, Mahajan, and Tarozzi (2011); Ligon and Schechter (2012); Cai, de Janvry, and Sadoulet (2013); Carrell, Sacerdote, and West (2013); Beaman et al. (2016); Blumenstock, Eagle, and Fafchamps (2016); Alatas et al. (2016). Also see Chuang and Schechter (2015); Aral (2016); Boucher and Fortin (2016); and Breza (2016) for overviews of empirical work using network data.

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 percent cost reduction.[2]

Our proposed method is intuitive and comes down to the following three simple observations. First, ARD is considerably cheaper and easier to collect than network data. Second, ARD provides the researcher with enough information to identify parameters of an oft-used and standard network formation model in the statistics literature (see, e.g., Hoff, Raftery, and Handcock 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 describe this and present an identification argument. We prove consistency of the estimation of the model parameters in Breza et al. (2020a).

Third, this parametric model of network formation is sufficiently rich to capture a number of features of real-world network structures. We provide two examples of recent research where either full or partial network data had been collected. Breza and Chandrasekhar (2019a) studies 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 paper's findings. Banerjee et al. (2019c) used network data to study how exposure to microcredit erodes social capital by reducing support. The authors in part, in their Hyderabad sample, collected survey ARD and we show we can replicate the findings. Further, the ARD enables conclusions about how microcredit exposure affected the neighborhood-level 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 collecting 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 percent, depending on the sampling rate, using budgets prepared 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.

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

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, Raftery, and Handcock 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) shows 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 posterior distributions of the degrees, fixed effects, and latent locations for the entire population. We can then generate graphs from the posterior distribution over formation model parameters given the ARD response vector and compute network statistics for each generated graph.

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), polygamy, 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 costs. Our approach is parametric, relying on guessing the network structure through the pseudo-true parameters of the latent distance formation model estimated from ARD. It can do no better than the best latent distance model at capturing the likely distribution that generated the network. It cannot, for example, represent clustering in a way that

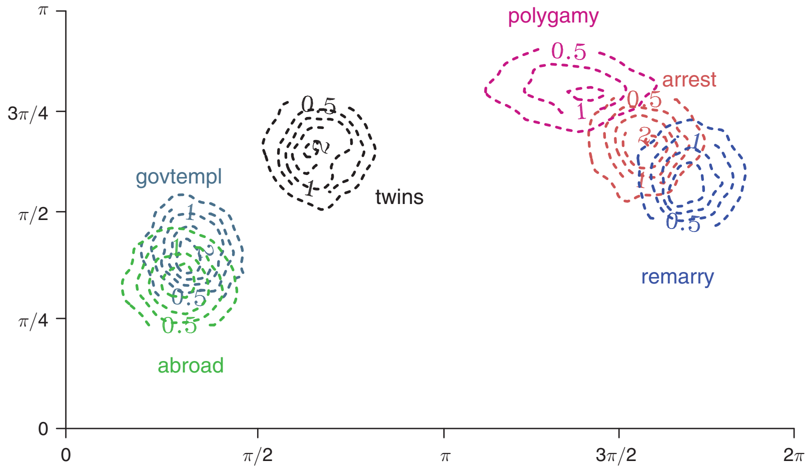

Figure 1. Plot of the Posterior Densities for Six ARD Characteristic Groups from Hyderabad

Notes: 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.

violates the triangle inequality.[3] To see this, consider a two-dimensional Euclidean space with four groups that have equal probability of cross-group interaction. If the data-generating process has this feature, we will not capture it well. Importantly, the approach can generate clustering patterns among nodes in close proximity in the latent space so whether this is sufficient to mimic real-world data is an empirical question. Further, the parameters of the formation model give a distribution over possible graphs that are consistent with the observed ARD. This is, of course, a distinct exercise from recovering the single, realized graph of connections between individuals.

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. 2019a, 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 2016). 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

[3] For an example of a network formation model which can do this, see Chandrasekhar and Jackson (2016).

(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) uses ARD to estimate the number of individuals impacted by an earthquake whereas Kadushin et al. (2006) uses ARD to estimate the number of individuals using heroin.[4]

The primary tool for estimating population size with ARD is the Network Scale-up Method (N-Sum) and variations thereof. Say the goal is to estimate the number of injection drug users in the population. If a respondent reports knowing 2 injection drug users out of 100 total contacts, then approximately 2 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,[5] but the majority of work on ARD still focuses on estimating population sizes. As 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 to estimate population sizes. 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): here, we use the same network formation model and build on derivations that are the key contribution of that work. Specifically, McCormick and Zheng (2015) shows that, for a specific formation model, it is possible to arrive at a likelihood that is informed by information in ARD. That is, they 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[6] 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, Diaconis, and Sly (2010) and Graham (2017). 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 (2017) and Chatterjee, Diaconis, and Sly (2010) models, our model places nodes in a latent space (as in Hoff, Raftery, and Handcock 2002), which we are trying to estimate, whereas the former only allows for observable covariates, and the latter has none. Whereas previous approaches consider an asymptotic frame based on a growing graph, we consider an explicitly sampling-based framework.

[4] 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).

[5] Zheng, Salganik, and Gelman (2006) estimates heterogeneity in the propensity to know members of groups, or overdispersion.

[6] 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.—We begin with an overview of our method for an applied researcher in Section I. Section II presents the full framework, model, and estimation algorithm. In Section III, we apply our results to two empirical examples. Section IV demonstrates the 70–80 percent cost savings of ARD versus full network elicitation. Section V provides a discussion of how an applied researcher could navigate the model's limitations. Section VI concludes.

I. Overview of Method

We begin with a simple overview of the proposed method. Suppose that a researcher is interested in studying networks in a set of rural villages. A village network with n households is given by $\mathbf{g}$, which is a collection of links ij where $g_{ij} = 1$ if and only if households i and j are linked and $g_{ij} = 0$ otherwise. To fix ideas, suppose that the researcher wants to learn how some outcome variable W is related to a network statistic (or a vector of statistics) of interest $S(\mathbf{g})$. Or, perhaps the researcher is interested in how a treatment (such as exposure to microcredit) affects features of network structure, $S(\mathbf{g})$.

Our procedure takes five steps.

(i) **Conduct ARD Survey:** Sample a share ψ (e.g., 30 percent) of households. Have each enumerate a list of their network links.[7] Ask 5–8 ARD questions, such as

> *How many households among your network list do you know where any adult has had typhoid, malaria, or cholera in the past six months?*

The ARD response for a household i is

$$y_{ik} = \sum_j g_{ij} \cdot \mathbf{1}\{j \text{ has had one of those diseases in past 6 mo.}\}$$

where trait k denotes the disease question. This just adds up all friends who have had the diseases over the last six months. We include a sample ARD questionnaire in online Appendix Section A.

(ii) **Conduct Census Exercise:** Obtain basic information about the full set of households in the village in a very rapid survey (denoted X_i for all $i = 1, \ldots, n$).

- Minimal demographics: e.g., GPS coordinates, caste/subcaste.
- ARD traits: e.g., whether the household has had typhoid, malaria, or cholera in the past six months.

A sample census questionnaire is in online Appendix Section A.

[7] Note that this gives a direct estimate of the respondent's degree. The method laid out in Section II does not require this and can also produce estimates for expected degree based on the ARD responses alone.

(iii) **Estimate Network Formation Model with ARD:** Use the information from the ARD survey and the population counts from the census to estimate the parameters of a network formation model. In this model, the probability that two households i and j are linked depends on household fixed effects (ν_i), and distance in some latent space (latent locations z_i) with

$$\Pr\big(g_{ij} = 1 | \nu_i, \nu_j, \zeta, z_i, z_j\big) \propto \exp\big(\nu_i + \nu_j + \zeta \cdot \text{distance}\big(z_i, z_j\big)\big).$$

- Fit a model to predict ν_i, z_i using X_i in the ARD sample.
- Predict ν_i, z_i using X_i for all households in the census but not in the ARD sample.

Equipped with estimated fixed effects and latent locations for all n households in the network, the probability of any network $\mathbf{g}$ being drawn is fully computed. The code is freely available and discussed in online Appendix Section B.

(iv) **Compute Network Statistics of Interest:** Use the estimated probability model (using ζ, fixed effects ν_i and latent locations z_i) to compute $E[S(\mathbf{g})|\mathbf{Y}]$. The code is freely available and discussed in online Appendix Section B.[8]

(v) **Estimate Economic Parameter of Interest:** e.g., run regressions such as

$$W_v = \alpha + \beta' E\big[S(\mathbf{g}_v)|\mathbf{Y}_v\big] + \epsilon_v \quad \text{or} \quad E\big[S(\mathbf{g}_v)|\mathbf{Y}_v\big] = \alpha + \beta \mathit{Treatment}_v + \epsilon_v,$$

though clearly one can do more complex exercises once one has estimated the network formation model above.

II. Model and Estimation

In this section, we present formally the procedure outlined above. This includes defining ARD, introducing the network formation model, linking explicitly the formation model to the ARD, and finally, outlining how to generate graphs from that network formation model. The result is a distribution over graphs (and therefore graph statistics) based on the observed ARD.

A. *Setup*

We begin by describing the underlying graph and the ARD. Let $\mathbf{g} = (V, E)$ 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 (where we define $\psi = m/n$). These could be the

[8] Note that here, the method produces estimates of the latent locations of each node, which may themselves be useful for some research questions.

whole sample, with $\psi = 1$, or a smaller share, and will depend on the context. It is useful to define V_{ard} to be the ARD sample set and $V_{non} = V \backslash V_{ard}$.

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. Further, G_k would be set of firms in sector k and g_{ij} would be the volume of transactions 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.

B. *Latent Surface Model*

The setup and model we use is from McCormick and Zheng (2015), motivated by, among others, Hoff, Raftery, and Handcock (2002). We model the underlying network as

$$(1) \qquad \Pr\left(g_{ij} = 1 | \nu_i, \nu_j, \zeta, z_i, z_j\right) \propto \exp\left(\nu_i + \nu_j + \zeta z_i' z_j\right),$$

where ν_i are person-specific random effects that capture heterogeneity in linking propensity.[9] 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 generating likely triadic closure. 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 (Penrose 2003). 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, $\mathcal{Z} = \mathcal{S}^{p+1}$, centered at the origin. As described below, the hypersphere has both conceptual and

[9] While we develop our methodology for this specific network formation model, we should note that it is likely possible to use ARD and other components of our method alongside a range of other formation models. While generalizing the method is outside the scope of this paper, we do view it as an avenue for future work, especially in real-world settings where researchers have a strong preference for alternative models.

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

$$z_i | v_z, \eta_z = 0 \sim \mathcal{M}(v_z, 0) \quad \text{and} \quad z_j | j \in G_k, v_k, \eta_k \sim \mathcal{M}(v_k, \eta_k),$$

where $\mathcal{M}$ denotes the von Mises-Fisher distribution across $\mathcal{S}^{p+1}$.[10] Here v_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 $z_j | j \in G_k$ 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) shows that the expected ARD response by i for category k can be expressed as

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

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), $\theta_{(z_i, v_i)}$ 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. Note that, in the expression above, both the distance between an individual and the center of the latent trait distributions as well as the concentration of the latent trait distribution influence the (expected) number of individuals know. Recall that our formation model only relies on the distance between individuals in the latent space. The positions of individuals, however, are estimated using the likelihood above, meaning that both the position and concentration are relevant for our formation model.

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

[10] Informally, the von Mises-Fisher distribution can be thought of as follows. If the concentration parameter is large, it is similar to a normal distribution on the sphere in that it is unimodal and symmetrically dissipating in distance from the center (though it should not be confused with the wrapped normal distribution or other projection of the normal to a sphere). If the concentration parameter is small, it is essentially uniform over the sphere's surface. Formally, the probability distribution function is given by $\mathcal{M}(z; v, \eta) = C_{p+1}(\eta)\exp(\eta v' z)$, where the normalizing constant depends on the modified Bessel function and simplifies to $\eta/2\pi(e^{\eta} - e^{-\eta})$ when $p = 2$. Here v is the mean direction, which we call center for simplicity, and η is the concentration parameter. See also Fisher, Lewis, and Embleton (1993) or Mardia and Jupp (2009) for a formal definition and properties.

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,

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

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 ν_i. We can compute the expected degree as in McCormick and Zheng (2015),

$$d_i = n \exp(\nu_i) E\left[\exp(\nu_j)\right] \left(\frac{C_{p+1}(0)}{C_{p+1}(\zeta)} \right). \tag{3}$$

The virtue here is that this allows us to estimate ν_i for $i \in V_{ard}$.[11] The logic is similar to that in Chatterjee, Diaconis, and Sly (2010) or Graham (2017): 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 expression above for each individual, then we have a system of n equations with $n + 1$ unknown terms ($n\nu_i$ terms and one $E[\exp(\nu_j)]$). Assuming that $E[\exp(\nu_j)]$ is well approximated by the average of the $\exp(\nu_i)$ terms, 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

$$\boldsymbol{\theta} | y_{ik} \propto \prod_{k=1}^{K} \prod_{i=1}^{n} \exp(-\lambda_{ik}) \lambda_{ik}^{y_{ik}} \prod_{i=1}^{n} \text{Normal}\left(\log(d_i) \,|\, \mu_d, \sigma_d^2\right)$$

$$\times \prod_{k=1}^{K} \text{Normal}\left(\log(b_k) \,|\, \mu_b, \sigma_b^2\right) \prod_{k=1}^{K} \text{Normal}\left(\log(\eta_k) \,|\, \mu_{\eta_k}, \sigma_{\eta_k}^2\right) \text{Gamma}\left(\zeta | \gamma_\zeta, \psi_\zeta\right).$$

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 $(z_1, \ldots, z_n)'$, $(\nu_1, \ldots, \nu_n)'$, and η, we can generate a graph from the distribution in (1).

[11] Note that if in our ARD elicitation, we also collect information on each node's degree, which we recommend, then we can use that information here, without needing to first estimate d_i above.

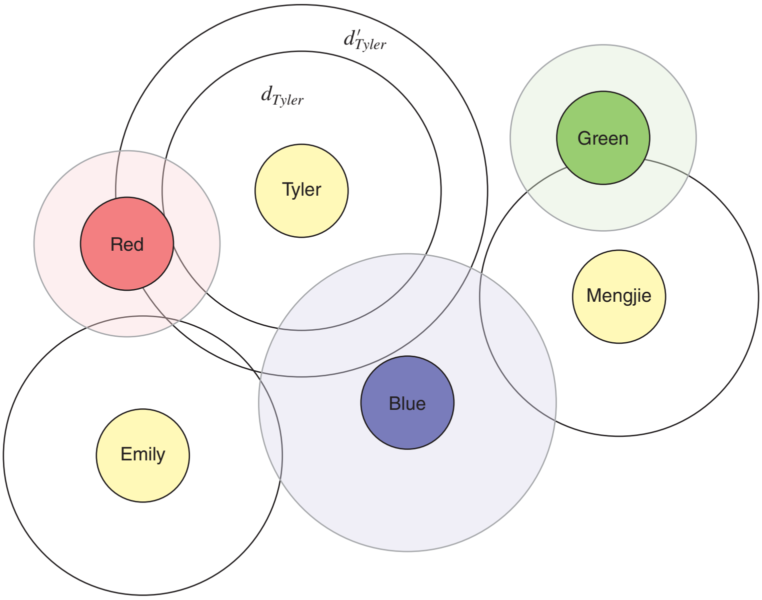

Figure 2

Notes: 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 $E[d_i]$ holding fixed locations and concentration parameters.

C. *Identification*

Before explaining how we go from the ARD sample to the full sample, we explain identification of the parameters in the model.[12] Here we provide a simple intuition, followed by a formal statement with proof in the Appendix.

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. To see that the concentrations of the Red, Green, and Blue trait groups are identified, consider what would happen if we changed the concentration of one of the groups. If we increased the concentration of the Blue group (i.e., decreased the variance), then we would need to move Mengjie (and Tyler and Emily) closer to the Blue group to preserve the overlap between Emily's disc and the Blue group. Moving Emily closer to the Blue group, though, necessitates moving her away from the Red group, reducing her overlap with the Red group. We could try to compensate by decreasing the concentration (increasing the variance) of the Red group. We can't do this, though, because doing so would change the overlap between Tyler's disc and the Red group. Similarly the figure shows how the $E[d_{Tyler}]$ can be identified holding fixed the location and concentration of the

[12] Also see McCormick and Zheng (2015) for a discussion of identification as well as recommendations for the number of populations to fix based on the dimension of the hypersphere.

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 of a small number of groups to address the invariance to distance-preserving rotations.

To see the formal statement, it is useful to recall that we say two points on a sphere are *antipodal* if there are indefinitely many great circles passing through them.

ASSUMPTION 1: $K > 3$ *and the centers of the von Mises-Fisher distributions representing three of the alter groups are fixed.*

ASSUMPTION 2: *The fixed centers are not all on the same great circle.*

ASSUMPTION 3: *For some* k, k', $\eta_k \neq \eta_{k'}$.

ASSUMPTION 4: $\zeta > 0$.

THEOREM 1: *Under Assumptions 1–4, for any n by K matrix of ARD responses* $\mathbf{Y}$, *we have that* $\mathcal{L}(d_i, b_k, \zeta, \eta_k, \theta_{(z_i, v_k)}; \mathbf{Y}) = \mathcal{L}(d_i, b_k, \zeta', \eta_k', \theta'_{(z_i, v_k)}; \mathbf{Y})$ *only if* $\eta_k = \eta_k'$, $\theta_{(z_i, v_k)} = \theta'_{(z_i, v_k)}$, $\zeta = \zeta'$, $\nu_i = \nu_i'$, *and* $z_i = z_i'$.

We provide a formal proof of the theorem in the Appendix.

Assumption 1, that $K > 3$ and are fixed, is innocuous. The content of Assumption 2 is as follows. Let the traits be "red," "blue," and "green." If you know the likelihood of say a "red" and a "blue" type linking on average (i.e., distance between the centers) and you know the likelihood of a "red" and a "green" type linking on average, it does not entirely determine the likelihood of a "blue" and "green" linking. Practically this means that essentially knowing two features (someone having a migrant, someone having a tenth standard pass family member) does not determine the third (on average). Assumption 3 requires that at least one trait has a different concentration parameter. In some sense both Assumptions 2 and 3 can be interpreted as ruling out "measure zero" events if one thinks of trait centers and concentration parameters themselves being generated according to any smooth distribution on a sphere. Assumption 4 means that the latent space has content for the model (by assumption $\zeta \neq 0$): distance in the space indeed reduces the odds of being linked. Put another way, it means that there is network structure not explained by the individual effects.[13]

D. *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 use k-nearest neighbors to draw this distribution.

[13] We could check this assumption using the residuals obtained by fitting a model with no latent space term (i.e., where the expected number known in trait k is $d_i b_k$). In a world where $\zeta = 0$, the residual number known in the "green" group shouldn't depend on whether person i is in the "red" or "blue" group. If, however, there is an increased linking probability between the "green" and "red" groups then the same residual exercise should reveal, overall, larger residuals for "red" group individuals asked about the "green" group. We could formalize this intuition by running regressions where the outcome is the aforementioned residual and the regressors are traits.

Given demographic covariates X_i for all $i \in V$, we define a distance between 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 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. Thus, 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.

E. *Drawing a Graph*

We now describe the algorithm used to generate a distribution of graphs $\{\mathbf{g}_s\}_{s=1}^S$. 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.[14] This choice also facilitates visualizing latent structure. The posterior distribution is not available in closed form. We therefore use a Metropolis-within-Gibbs algorithm to obtain samples from the posterior. In the description below, the jumping scale[15] is tuned adaptively throughout the course of sampling. Specifically, every 50 draws we look at the acceptance rate of these draws and then adjust the scale of the jumping distribution. We follow the guidelines given in Gelman et al. (2013) and perform checks to ensure that our sampler has converged.

ALGORITHM 1 (Drawing Graphs): *Input:* $y_{ik}\ \forall i \in V_{ard}$, $X_i\ \forall i \in V$. *Assume ARD groups,* $k = 1, \ldots, K$, *such that* $K \geq p$. *We propose fitting the model as follows (noting that steps (i) and (ii) follow from McCormick and Zheng 2015):*

(*i*) *For a subset of the ARD groups,* $k^{(s)} = 1, \ldots, K^{(s)}$, *fix* $\boldsymbol{\upsilon}_k^{(s)}$.

(*ii*) *Repeat to convergence for* $t = 1, \ldots, T$.

(*a*) *For each* i, *update* z_i *using a random walk Metropolis step with proposal* $z_i^* \sim \mathcal{M}(z_i^{(t-1)}$, *jumping distribution scale). Use the algorithm proposed by Wood (1994) to simulate proposals implemented in the R package Rfast (Papadakis et al. 2017).*

(*b*) *Update* $\boldsymbol{\upsilon}_k$ *using a conditionally conjugate Gibbs step* $\boldsymbol{\upsilon}_k \sim \mathcal{M}(\mathbf{m}_k/||\mathbf{m}_k||_2, ||\mathbf{m}_k||_2)$, *where* $\mathbf{m}_k = \eta_k \sum_{j \in k} z_j$. *(See, e.g., Mardia and El-Atoum 1976, Guttorp and Lockhart 1988, Hornik and Grün 2013, Straub et al. 2015).*

[14] We also investigated the performance of the method in real-world networks for $p = 3$ and $p = 4$ and found similar performance. We thus use $p = 2$ to facilitate visualization.

[15] For Metropolis steps in the Markov chain Monte Carlo (MCMC), the procedure for updating parameters involves proposing a new potential value based on the current parameter value and then probabilistically deciding whether to accept or reject the new candidate value. The jumping scale refers to the variance of the distribution we used to propose a new parameter. Intuitively, selecting a larger jumping scale will produce candidates that are less similar to the current value (and possibly larger differences between the current and subsequent parameter values).

(*c*) *Update* d_i *with a Metropolis step with* $\log(d_i^*) \sim N(\log(d_i)^{(t-1)}$, *jumping distribution scale*).

(*d*) *Update* b_k *with a Metropolis step with* $\log(b_k^*) \sim N(\log(b_k)^{(t-1)}$, *jumping distribution scale*).

(*e*) *Update* η_k *with a Metropolis step with* $\eta_k^* \sim N(\eta_k^{(t-1)}$, *jumping distribution scale*).

(*f*) *Update* ζ *with a Metropolis step with* $\zeta^* \sim N(\zeta^{(t-1)}$, *jumping distribution scale*).

(*g*) *Update* $\mu_b \sim N(\hat{\mu}_b, \sigma_b^2)$ *where* $\hat{\mu}_b = \sum_{k=1}^K \log(b_k)/K$.

(*h*) *Update* $\sigma_b^2 \sim$ *Inv*-$\chi^2(K-1, \hat{\sigma}_b^2)$ *where* $\hat{\sigma}_b^2 = \frac{1}{K-1}\sum_{k=1}^K (\log(b_k) - \mu_b)^2$.

(*i*) *Update* $\mu_d \sim N(\hat{\mu}_d, \sigma_d^2)$ *where* $\hat{\mu}_d = \sum_{i=1}^n \log(d_i)/n$.

(*j*) *Update* $\sigma_d^2 \sim$ *Inv*-$\chi^2(n-1, \hat{\sigma}_d^2)$ *where* $\hat{\sigma}_d^2 = \frac{1}{n-1}\sum_{i=1}^n (\log(d_i) - \mu_d)^2$.

(*iii*) *Repeat for* $t \in \{T/2+1, \ldots, T\}$.

(*a*) *Calculate* $\nu_i^t\ \forall i \in V_{ard}$ *such that* ν_i^t *satisfies* $(d_i)^t = \exp(\nu_i^t)\sum_i \exp(\nu_i^t) \times \left(C_{p+1}(0)/C_{p+1}(\zeta)\right)$.

(*b*) *Use method described in Section IID to estimate* ν_j^t *and* $z_j^t\ \forall j \in V_{non}$.

(*c*) *Sample graph* $\mathbf{g}_t$ *using the the procedure described below*.

Output: $\{\mathbf{g}_s\}_{s=1}^S$.

To generate graphs, recall that the formation model has $\Pr(g_{ij} = 1 | \nu_i, \nu_j, \zeta, z_i, z_j) \propto \exp(\nu_i + \nu_j + \zeta z_i' z_j)$. We estimate ζ and z_i, z_j using the likelihood derived in McCormick and Zheng (2015). The expression (3) relates degree to the unobserved gregariousness parameters, ν_i. If we approximate $E[\exp(\nu_j)]$ as the average of the ν_i terms, then we can view (3) as a system with n equations and n unknowns and obtain estimates for ν_i for each respondent.

We then normalize the $\exp(\nu_i + \nu_j + \zeta z_i^{t'} z_j^t)$ terms to produce probabilities. Define

$$\Pr\left(g_{ij} = 1 | z_i, z_j, \nu_i, \nu_j\right) = \frac{\exp\left(\nu_i + \nu_j + \zeta z_i' z_j\right)\sum_i E[d_i]}{\sum_{i,j}\exp\left(\nu_i + \nu_j + \zeta z_i' z_j\right)}.$$

Normalizing in this way ensures $\sum_i E[d_i] \triangleq \sum_i \sum_j \Pr(g_{ij} = 1 | z_i, z_j, \nu_i, \nu_j)$. Since the formation model assumes that the propensities to form 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 $\Pr(g_{ij} = 1 | z_i, z_j, \nu_i, \nu_j)$.

F. *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, Raftery, and Handcock (2002), used in McCormick and Zheng (2015), which is intimately related to the β-model studied in Chatterjee and Diaconis (2011) and Graham (2017). 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, though this theorem only gives the existence of a latent variable representation and not the specific form we use. The 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) shows 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 $E[d_i]$ 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.

III. How Well Does the Procedure Perform with Real Data?

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.

A. *Encouraging Savings Behavior in Rural Karnataka*

Our first application builds on Breza and Chandrasekhar (2019a). The authors study social reputation through the lens of savings. In a field experiment, savers set six-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 toward 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 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. For saver *i* and monitor *j*, the model shows that the network matters for signaling through the quantity[16]

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

Breza and Chandrasekhar (2019a) has near-full network data (from the Banerjee et al. 2019a sample), allowing them to calculate $q_{i,j}$. They find that randomly selected monitors increase household savings across all accounts by 35 percent. Consistent with the model, a one standard deviation increase in q_{ij} leads to an additional 29.6 percent 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. Moreover, the savings increase persisted, and in the intervening 15 months, monitored savers were better able to cope with shocks.

How would our conclusions have changed if Breza and Chandrasekhar (2019a) only had access to ARD and not the full network maps? Table 1 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}. We construct ARD estimates by taking samples from the posterior distribution and then using the average estimated q_{ij} across those posterior draws. In the experiment we showed that a 1 standard deviation increase in q_{ij} due to random assignment of the monitor led to a 24.8 percent increase in total household savings (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 percent sample, we would have had a very similar conclusion, inferring that a one standard deviation increase in predicted q_{ij} corresponds

[16] Formally, Breza and Chandrasekhar (2019a) shows

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

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. (2019a) shows that for sufficiently high T, $\sum_k p_{jk}$ converges to the eigenvector centrality of j. Breza and Chandrasekhar (2019a) shows that in equilibrium, only when q_{ij} is sufficiently high does the saver actually save.

Table 1—log Total Savings across All Household Accounts Regressed on Monitor Signaling Value

	log total ending savings	
	(1)	(2)
Signaling value of monitor with full network data (q_{ij}), standardized	0.254 (0.0869)	
Predicted signaling value of monitor with ARD (q_{ij}), standardized		0.185 (0.0925)
Observations	422	422
Number of villages	30	30

Note: Standard deviation of village-level block bootstrap in parentheses.

to a 18.1 percent increase in total household savings across all accounts. Said differently, we could have used ARD questions to easily pick good monitor-saver pairs.

As a further examination of our approach, we repeat the same exercise using another specification from Breza and Chandrasekhar (2019a). Table 2 shows the results of a regression where the outcome is the respondent's belief about the saver's responsibility and the regressor is the monitor's centrality. Observing the complete network, a unit increase in the monitor's centrality corresponds to about a 5 percent increase respondent's belief about saver responsibility. Using ARD, we would estimate an increase of about 3.4 percent, leading (as in the previous example) to the same substantive conclusions.

This application also gives us an opportunity to visualize how network characteristics map to the latent space representation. In Figure 3, we plot the locations and concentrations of the ARD traits for four sample villages that were part of the Breza and Chandrasekhar (2019a) savings study. We then overlay the positions in the latent space of the individuals participating in the experiment as monitors, depicted as rings. The size of the ring depicts the monitor's eigenvector centrality. Finally, we color the monitor rings to indicate the savings performance of the saver to whom each monitor was randomly allocated: darker shades depict higher levels of savings.

As Breza and Chandrasekhar (2019a) finds, there appears to be a relationship between monitor centrality (here denoted by larger rings) and the saver's performance (here given by darker colors). This is consistent with the theory that more central monitors under the signaling model generate larger incentives for the saver to save. Furthermore, the visualization demonstrates that the larger rings tend to be located closer to the centers of traits or between centers of traits. That is, they are closer to the center of masses of clusters of types of individuals. This makes sense as this means that the latent location of a central monitor will tend to be closer to many more other individuals, ceteris paribus.

B. *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

Table 2—Beliefs about Savers and Monitor Centrality

	Belief about saver's responsibility	
	(1)	(2)
Monitor centrality with full network data, standardized	0.0500 (0.0146)	
Predicted monitor centrality with ARD, standardized		0.0340 (0.0160)
Observations	4,743	4,743
Number of villages	30	30

Notes: Standard deviation of village-level block bootstrap in parentheses. *Responsibility* is constructed as $\mathbf{1}$(saver reached goal) × $\mathbf{1}$(respondent indicates saver is good or very good at meeting goals) + (1 − $\mathbf{1}$(saver reached goal)) × $\mathbf{1}$(respondent indicates saver is mediocre, bad or very bad at meeting goals). See Breza and Chandrasekhar (2019a) for further details.

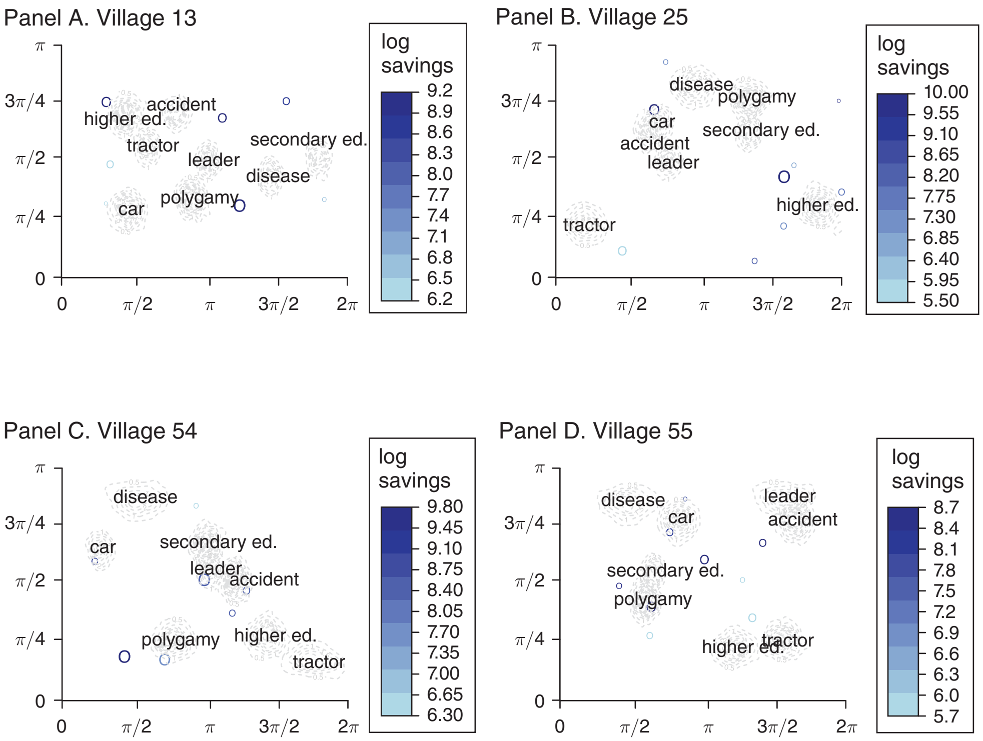

Figure 3. Sample Latent Locations of Randomly Assigned Monitors by Centrality and the Savings of Their Respective Savers

Notes: Monitors with higher eigenvector centrality have larger rings. The color of the ring indicates the savings performance of the saver to whom each monitor was randomly assigned, with darker colors indicating higher savings levels. This illustrates the pattern that more central monitors corresponded to higher levels of savings.

of the network data that were collected in this study. But because the network data were 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 heterogeneous effects on the underlying social and economic networks that facilitate informal risk-sharing. On the one hand, as in Feigenberg, Field, and Pande (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. 2019c).

In Banerjee et al. (2015, 2019d), the authors study a randomized controlled trial where microfinance was introduced randomly to 52 out of 104 neighborhoods in Hyderabad, India. Banerjee et al. (2019c) looks at long-run outcomes on network structure 6 years after the intervention as one of two empirical exercises. This example is useful for two reasons. First, it is an urban setting where the researchers have no hope of obtaining full network data.[17] 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.

Banerjee et al. (2019c) measures each node's within-neighborhood 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.

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 (Banerjee et al. 2019e). As before, we fit a network formation model using the ARD data and this sample of nodes.[18] A complete list of ARD questions used in this survey is in online Appendix Section C.

We explore whether microfinance affects network structure by regressing

$$y_v(g) = \alpha + \beta Treatment_v + \epsilon_v ,$$

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.

[17] We thank an anonymous referee for noting that we could also tweak our surveys in urban settings to measure ARD responses separately within the respondent's own neighborhood and also across neighborhoods. While mapping an entire urban space likely requires an infeasible number of surveys, putting some structure on relationships within and across neighborhoods might allow for better urban network maps. We leave such an application to future work.

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

Table 3—Network Statistics Regressed on Treatment

	Percent supported (data) (1)	Percent supported (estimate) (2)	Graph-level proximity (estimate) (3)
Treatment neighborhood	−0.0655	−0.0901	−0.0500
	(0.0319)	(0.0541)	(0.0164)
Constant	0.4427	0.4364	0.3642
	(0.0628)	(0.093)	(0.0108)
Mean of the response variable	0.3880	0.3125	0.3375
Observations	3,514	3,598	62

Notes: Standard deviation of village-level block bootstrap in parentheses. Sample includes neighborhoods with estimated sampling rate $\geq$ 20 percent. 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. The outcome variable of columns 1 and 2 is the share of links that are supported and in column 3 it is the average proximity in the graph.

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, Rodriguez-Barraquer, and Tan 2012).

Table 3 reports the regression results. Column 1 replicates the specification from Banerjee et al. (2019c) that past exposure decreased support. Column 2 presents the same regression, but using estimated support. The estimates of the treatment effects along with the levels of support (the regression constant) are quite similar. We view this exercise as a "validation" of the ARD-based model. The fact that estimated support matches measured support quite well is especially reassuring given that triadic closure is exactly the type of network statistic that the Hoff model may have a hard time replicating.[19]

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. We 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.

IV. 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

[19] Recall that our latent space model can accommodate clustering for groups of individuals located in close proximity to one another. One interpretation of this result is that for the Hyderabad setting, the form clustering we observe is well captured by the model.

Table 4—Cost Comparison: Full Network versus ARD Surveys

	Traditional network survey	ARD survey
Panel A. Assumptions		
Project duration (months)	8.2	3.2
Number of villages	120	120
Census sampling rate (percent)	100	100
Fully enumerated census	Yes	No
Network/ARD survey sampling rate (percent)	100	30

	Traditional network survey		ARD survey	
	Total cost($)	Per village cost($)	Total cost($)	Per village cost($)
Panel B. Costs				
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 percent)	24,674	206	4,502	38
Total cost	189,164	1,576	34,512	288

Note: 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.

(2019a) and Banerjee et al. (2019c) 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 4 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 are collected from 100 percent of households, while the ARD protocol samples from 30 percent 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 4 shows that ARD is substantially cheaper, costing approximately 80 percent less than the full network surveys.

In Figure 4, we show that these dramatic cost reductions are not only a byproduct of the 30 percent sampling rate assumption. Even with 100 percent sampling, ARD surveys are still over 70 percent 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 are the gold standard, and even partial network data could help being used

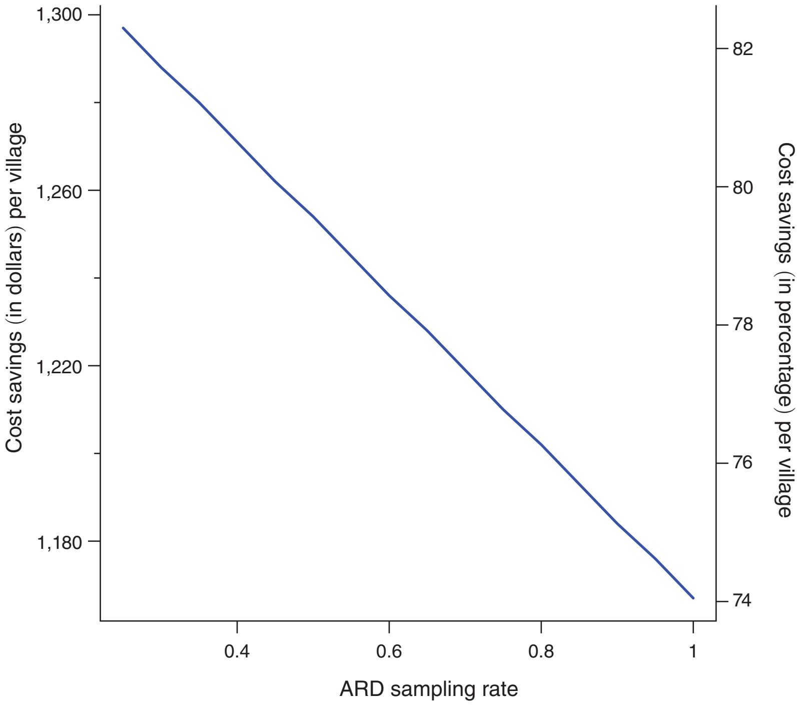

Figure 4. Cost Savings of ARD versus Full Network Surveys by ARD Sampling Rate

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 (2016).

V. Discussion and Limitations

Our method is not without limitations, and we have highlighted two issues that should be considered when using our method for applied research. While a detailed theoretical study is beyond the scope of the paper, we discuss briefly how applied researchers might navigate these limitations.

First, the method produces a distribution of networks that are consistent with the estimated network formation model: we do not learn about the specific realization that generated the observed graph. For example, of course we can never say whether a given link exists. This means that network features that rely on the existence of a specific links will not be captured well. If the research question requires knowledge of specific links, then the researchers should ask about these relationships directly when possible.[20] This intuition also suggests that features such as betweenness

[20] The method can easily be adjusted to incorporate information about any known links.

centrality, which rely on specific paths, may be hard to capture with the method.[21] If the researcher has a contextual reason to know that a specific node may be of outsize importance based on its observable characteristics (e.g., Amazon or Walmart in a retail supply-chain network), the researchers can incorporate that information into the research design, either using those characteristics as ARD traits or by collecting information about the realized links to that specific node. Finally, appealing to the result of Chandrasekhar and Lewis (2016), if inference is being conducted across many independent networks, then these issues are of much less concern.[22]

Second, the method relies upon a parametric network formation model. If that model is not a good representation of the network of interest, then the resulting ARD estimates may be biased. As mentioned above, one might be particularly concerned about the ability of the Hoff model to capture the level of clustering. However, as we show in our Hyderabad microfinance example, the model actually does well in practice at predicting the change in the level of link support, a related notion. We recommend that applied researchers follow this empirical example and also elicit network support directly from survey respondents. The researcher can then "validate" the ARD method for the specific applied context by estimating support using ARD and comparing the estimates to the true values.[23]

Other examples of network structures that our parametric model may be ill suited to capture include hierarchical trees as well as bipartite networks. Hierarchical trees are naturally embedded in latent spaces with negative curvature, hyperbolic surfaces, rather than the sphere with positive curvature. Further, bipartite networks place complicated restrictions on the latent space as well, not naturally representable on the sphere.

Finally, in our companion paper (Breza et al. 2020a), we study how the quality of the estimation of network features varies by statistic. We prove consistency of the maximum likelihood estimates of all parameters of the latent space model from ARD. Further, we investigate how the estimates depend on the density of the network formation distribution, the sampling rates used by the researcher, and so on. To summarize these results briefly, we find that the method works quite well for many empirically relevant network features both at the node and network level. At the network level it performs well when we look at degree, path length, maximal eigenvalue of the adjacency matrix, graph-level clustering, whereas it does poorer when estimating the number of components in the network. At the node level, degree, eigenvector centrality, among other features perform well; path lengths tend to be underestimated whereas node-level clustering performs the worst and are systematically estimated to be near the expectation irrespective of node. When varying the tail-thickness of the degree distribution of the sample, we are still able to systematically pick out the most central nodes. Our sampling simulations show that the

[21] Our Savings Monitors example shows that the model can do well at capturing more recursive notions of centrality such as diffusion and eigenvector centrality. Thus, the method should still do well in cases where betweenness centrality is highly correlated with these other measures.

[22] Chandrasekhar and Lewis (2016) shows that by the law of iterated expectations, when conducting inference across many graphs, there is no difference between working with the realized graph- or node-level object or the expectation of that object.

[23] Again, the survey measures of support are not used directly in the ARD estimation procedure.

estimates are of high quality so long as we have at least a 20 percent sample when looking at networks of size 200–1,000.

VI. 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.

There is a prior literature as to whether a researcher could simply ask individuals from the network. For instance, Banerjee et al. (2019a) shows that simply asking "gossip" questions can be used to identify eigenvector central individuals. However, there are no results for other features such as such as clustering, path length, cut in the network, and so on.[24] Further, we have reason to believe this sort of procedure likely would not work for other network features. For instance Friedkin (1983), Krackhardt (1987, 2014), among others in sociology, and also our own work in Breza, Chandrasekhar, and Tahbaz-Salehi (2017), all document such biases. They show that network knowledge decays in distance, that degrees are systematically misestimated, and that individuals are more likely to think their friends are friends, among other things.

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 online Appendix Section 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 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. Yet another direction would provide guidelines for picking the dimension of the latent space, or the latent geometry in general. In particular, we could use fraction of overlap between traits to restrict the set of feasible latent dimensions.[25]

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

[24] Note that part of the insight in Banerjee et al. (2019a) was to realize that eigenvector seems complicated but if you know who you hear gossip about frequently, this mechanically corresponds to central individuals. This is a unique trait for centrality, not all statistics.

[25] To see the intuition for this, consider the case where there are three groups A, B, and C. Each of these groups would need to be placed on a sphere in such a way as to reflect the overlaps between individuals in one or more of the groups (a person who is a member of A and B should go in the disc of both groups, for example). The configuration implied by these overlaps may not be possible in all dimensions. Fosdick et al. (2019) points out a similar restriction arising because of the triangle inequality for latent spaces on the plane.

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: intersectoral linkages (Acemoglu et al. 2012; Barrot and Sauvagnat 2016; Carvalho et al. 2016) and banking (Acemoglu, Ozdaglar, and Tahbaz-Salehi 2015; Elliott, Golub, and Jackson 2014; Gandy and Veraart 2016, 2019; Upper and Worms 2004).

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.[26]

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 2016) or whether if theory suggests 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.

Appendix. Proofs

A. *Identification*

In this section, we formally discuss identification. Essentially, we need three latent group centers to be fixed and to have distinct positions on the hypersphere. We also need to know the trait status of at least some individuals and for there to be at least some individuals with more than one trait. This is sufficient to identify the parameters governing the locations of each of the types and the concentration parameters. If we assume that trait status is unrelated to gregariousness (which is necessary for the derivation of the likelihood anyway) then we can identify the coefficient zeta. Based on zeta and degree (which is identified as described in McCormick and Zheng 2015 using the latent trait group sizes) we can identify the individual gregariousness

[26] The model presented above 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. One could extend the method to address directed graphs by introducing an asymmetric distance measure as suggested in, for example, Hoff, Raftery, and Handcock (2002).

parameters. All that is left are the individual-level latent positions, which we show can be identified based on the previously described parameters.

We begin by defining terms necessary to describe the spherical geometry and then provide the necessary conditions. Throughout the proofs here we will assume a latent sphere centered at the origin.

PROOF OF THEOREM 1:

Under Assumptions 1–4, this is a direct corollary to Propositions 1, 2, and 3. ∎

PROPOSITION 1: *Considering Assumptions 1–4, trait centers υ_k for $k = 4, \ldots, K$, concentration parameters η_k for $k = 1, \ldots, K$, and ζ are identified.*

PROOF:

The von Mises-Fisher distribution is a symmetric unimodal distribution with probability mass declining in distance from the center, υ, tuned by concentration parameter η. For each individual we know their latent trait group(s). This is a fundamental distinction between our setting and that of McCormick and Zheng (2015), who typically do not assume this information is known. We can think of the positions of each individual as draws from one or more of the von Mises-Fisher distributions on the sphere. An individual who belongs to two trait groups has to be at the intersection of the densities of the two trait groups. Knowing the fraction of individuals who have both traits, therefore, intuitively tells us something about the overlap between the densities of the two trait groups. Throughout this proof keep in mind that we are not using the specific locations of individuals (which we only show is identified in a subsequent proposition), but rather the density defined by the overlap between trait groups.

More formally, define the lens, $\ell(A,B)$, as the expected share of individuals drawn from this distribution who have traits A and B. Equivalently, we can think of this as the volume of the overlap between the densities of the two distributions for all individuals up to a prespecified, but arbitrary,[27] cumulative probability. In general let $\ell(A_1, \ldots, A_k)$ denote the expected share of individuals drawn who have all traits. We can treat all lenses as observed in the data because for a large m, we know the traits that every node has.

For notational convenience and without loss of generality, we will assume that the fixed group centers correspond to the first three latent trait groups, $\upsilon_1, \upsilon_2, \upsilon_3$. Observe that this immediately implies all three η_k for $k = 1, \ldots, 3$ are identified. For the sake of argument assume that η_1 is known. Then from $\ell(1,2)$ we have that η_2 is identified. Given η_2, from $\ell(2,3)$, we have η_3 identified. But we can of course identify η_1 similarly from η_3. This logic applies because we can map the overlapping section, $\ell(1,2)$, into specific values of the cumulative distribution function of the von Mises-Fisher distributions. If we change η_2, then the location of individuals' latent positions that are draws from this distribution must also change. Changing these locations changes the boundary of $\ell(1,2)$. Similarly, changing the boundary

[27] We could define the lens for example as the area of the overlap in bands that represent that ninety-fifth percentile of the distribution. We need to specify a cutoff because the densities are continuous across the surface. The choice is arbitrary so long as the discs are sufficiently wide to include the overlap between densities.

of $\ell(1,2)$ implies a change in the densities of the von Mises-Fisher distributions for the first and second traits. Since the centers of these distributions are fixed, any change in the distribution must come through the concentration parameter.

Further, this solution is unique. To see this, assume that we are at some unique solution η_1, η_2, η_3. Consider an alternative value of any combination of concentration parameters. Clearly all concentration parameters cannot increase because then the lenses would not match the true lenses. Consider then the case where at least one η_k declines. In this case, if $\eta_{k'}$ were not to increase, then $\ell(k,k')$ would not match the expectation observed in the data. Consequently, $\eta_{k'}$ must increase. In this case, should $\eta_{k'}$ increase, then $\eta_{k''}$ must decline to preserve $\ell(k',k'')$. But in this case, the lens $\ell(k,k'')$ must increase as both concentration parameters have declined. Therefore the solution is unique.

To see why ζ is identified, consider any two k,k' with $\eta_k \neq \eta_{k'}$. Because we know the respective von Mises-Fisher distributions for each trait, we can compute the ratios of the expectations of (2) conditional on each type k and k', plugging in for d_i from (3). Because the individual effects are drawn independently of trait by assumption, all terms that depend on ν_i drop since the distribution of ν_i is independent of trait type, so they have the same expectations irrespective of k or k'. As such,

$$\frac{E_i[\lambda_{ik} | i \in G_k]}{E_j[\lambda_{jk} | j \in G_{k'}]} = f(b_k, b_{k'}, \eta_k, \eta_{k'}, \zeta),$$

where the right-hand side is a known function that comes from taking these ratios. The only unknown is ζ. There is a unique solution to the equation, we leave the algebra to the reader, but can be seen from the fact that the link probability is monotonically declining in ζ and faster for lower η_k, holding all else fixed, so the ratio term also is monotone in ζ. ∎

PROPOSITION 2: *Considering the conditions above*, ν_i *for* $i = 1,\ldots,m$, *individual gregariousness effects for the entire ARD sample*, *are identified.*

PROOF:

By Proposition 1, the υ_k and η_k and ζ are identified. By (2), d_i can be obtained and by (3) we have for every $i = 1,\ldots,m$ in the ARD sample an equation relating the fixed effect ν_i to the degree. We have m equations and m unknowns.

To see why the solution is unique consider fixing for the moment some ν_1 without loss of generality. In this case, we can write $\nu_i = h_i \nu_1$ for every i, where h_i is the ratio of the degrees between person i and person 1. Then we can write

$$\exp(\nu_1)\left(\frac{1}{n}\sum_i \exp(h_i \nu_1)\right) = \frac{d_1}{m \cdot \frac{C_{p+1}(0)}{C_{p+1}(\zeta)}}.$$

This is a monotone function in ν_1 and has a unique solution, which then identifies the remainder of the ν_i as well scaling by h_i. ∎

PROPOSITION 3: *Considering the conditions above*, *the latent locations* z_i *for* $i = 1,\ldots,m$ *for the entire ARD sample*, *are identified.*

PROOF:

From Propositions 1 and 2, we have identified all parameters except for z_i. To show this result, we first state two results from spherical geometry. The proofs of these results are available in standard texts (e.g., Biringer 2015).

Result: The great circle between two points is unique unless the points are antipodal.

Result: There are exactly three isomorphisms for spherical geometry.

The first result defines a unique distance from each respondent latent position and at least two of the three latent group means. A respondent position can be antipodal with one of the three fixed groups, but then cannot be with the two others because the three groups are not on the same great circle.

The second result limits the number of possible operations that threaten identifiability. Recall that, if an operation changes the latent distance between a point and the center of a group, then the operation will also change the likelihood. Thus, if we show that we cannot perform any of the three possible distance-preserving transformations on the sphere after fixing group centers, then we have also completed the proof.

We consider two cases: the first takes an arbitrary point that is not antipodal to any of the latent centers, whereas the second case considers any point that is antipodal with one latent center.

Case 1: Since we fix three centers which are not on a great circle, we cannot do any reflections of points without changing the distance to one of the centers. For rotations, consider centers v_1 and v_2, and a point z_i. Since v_1 and v_2 are not antipodes, if we rotate z_i around center v_1 and keep $d(z_i, v_1)$ the same, it is possible that $d(z_i, v_2)$ changes. The points z_i, z_i' such that $d(z_i, v_1) = d(z_i', v_1)$ and $d(z_i, v_2) = d(z_i', v_2)$ are reflections over the plane that intersects v_1 and v_2 in a great circle. Thus, z_i and z_i' have equal distance to any point on this great circle, and unequal distance to any point not on this great circle. Since the third center v_3 is not on this the great circle that intersects v_1 and v_2, $d(z_i, v_3) \neq d(z_i', v_3)$.

Case 2: When we change the point's position, then the distance between that point and the antipodal latent center decreases. ■

REFERENCES

Acemoglu, Daron, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2012. "The Network Origins of Aggregate Fluctuations." *Econometrica* 80 (5): 1977–2016.

Acemoglu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2015. "Systemic Risk and Stability in Financial Networks." *American Economic Review* 105 (2): 564–608.

Alatas, Vivi, Abhijit Banerjee, Arun G. Chandrasekhar, Rema Hanna, and Benjamin A. Olken. 2016. "Network Structure and the Aggregation of Information: Theory and Evidence from Indonesia." *American Economic Review* 106 (7): 1663–1704.

Aldous, David J. 1981. "Representations for Partially Exchangeable Arrays of Random Variables." *Journal of Multivariate Analysis* 11 (4): 581–98.

Aral, Sinan. 2016. "Networked Experiments." In *The Oxford Handbook of the Economics of Networks*, edited by Yann Bramoullé, Andrea Galeotti, and Brian Rogers, 376–411. Oxford: Oxford University Press.

Auerbach, Eric. 2016. "Identification and Estimation of Models with Endogenous Network Formation." Unpublished.

Banerjee, Abhijit, Emily Breza, Arun Chandrasekhar, Esther Duflo, Matthew O. Jackson, and Cynthia Kinnan. 2019c. "Changes in Social Network Structure in Response to Exposure to Formal Credit Markets." Unpublished.

Banerjee, Abhijit, Emily Breza, Esther Duflo, and Cynthia Kinnan. 2019d. "Can Microfinance Unlock a Poverty Trap for Some Entrepreneurs?" Unpublished.

Banerjee, Abhijit, Emily Breza, Esther Duflo, and Cynthia Kinnan. 2019e. "Can Microfinance Unlock a Poverty Trap for Some Entrepreneurs?: Dataset." Unpublished.

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2013. "The Diffusion of Microfinance." *Science* 341 (6144).

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2019a. "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials." *Review of Economic Studies* 86 (6): 2453–90.

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2019b. "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials: Dataset." *Review of Economic Studies*.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 7 (1): 22–53.

Barrot, Jean-Nöel, and Julien Sauvagnat. 2016. "Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks." *Quarterly Journal of Economics* 131 (3): 1543–92.

Beaman, Lori, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak. 2016. "Can Network Theory-Based Targeting Increase Technology Adoption?" Unpublished.

Bernard, H. Russell, Tim Hallett, Alexandrina Iovita, Eugene C. Johnsen, Rob Lyerla, Christopher McCarty, Mary Mahy, et al. 2010. "Counting Hard-to-Count Populations: The Network Scale-Up Method for Public Health." *Sexually Transmitted Infections* 86 (S2): ii11–ii15.

Biringer, Ian. 2015. "Geometry in Two Dimensions." Unpublished.

Blitzstein, Joseph, and Persi Diaconis. 2011. "A Sequential Importance Sampling Algorithm for Generating Random Graphs with Prescribed Degrees." *Internet Mathematics* 6 (4): 489–522.

Blumenstock, Joshua E., Nathan Eagle, and Marcel Fafchamps. 2016. "Airtime Transfers and Mobile Communications: Evidence in the Aftermath of Natural Disasters." *Journal of Development Economics* 120: 157–81.

Boucher, Vincent, and Bernard Fortin. 2016. "Some Challenges in the Empirics of the Effects of Networks." In *The Oxford Handbook of the Economics of Networks*, edited by Yann Bramoullé, Andrea Galeotti, and Brian W. Rogers, 277–302. Oxford: Oxford University Press.

Breza, Emily. 2016. "Field Experiments, Social Networks, and Development." In *The Oxford Handbook of the Economics of Networks*, edited by Yann Bramoullé, Andrea Galeotti, and Brian Rogers, 412–39. Oxford: Oxford University Press.

Breza, Emily, and Arun G. Chandrasekhar. 2019a. "Social Networks, Reputation and Commitment: Evidence from a Savings Monitors Experiment." *Econometrica* 87 (1): 175–216.

Breza, Emily, and Arun G. Chandrasekhar. 2019b. "Social Networks, Reputation and Commitment: Evidence from a Savings Monitors Experiment: Dataset." *Econometrica.*

Breza, Emily, Arun G. Chandrasekhar, Tyler H. McCormick, and Mengjie Pan. 2020a. "Consistently Estimating Graph Statistics Using Aggregated Relational Data." Unpublished.

Breza, Emily, Arun G. Chandrasekhar, Tyler H. McCormick, and Mengjie Pan. 2020b. "Replication Data for: Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E110841V1.

Breza, Emily, Arun G. Chandrasekhar, and Alireza Tahbaz-Salehi. 2017. "Seeing the Forest for the Trees? An Investigation of Network Knowledge." Unpublished.

Cai, Jing, Alain de Janvry, and Elisabeth Sadoulet. 2013. "Social Networks and the Decision to Insure." Unpublished.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81 (3): 855–82.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Alireza Tahbaz-Salehi. 2016. "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake." Unpublished.

Centola, Damon. 2010. “The Spread of Behavior in an Online Social Network Experiment.” *Science* 329 (5996): 1194–97.

Chandrasekhar, Arun G., and Matthew O. Jackson. 2016. “A Network Formation Model Based on Subgraphs.” Unpublished.

Chandrasekhar, Arun G., and Randall Lewis. 2016. “Econometrics of Sampled Networks.” Unpublished.

Chassang, Sylvain, Pascaline Dupas, Catlan Reardon, and Erik Snowberg. 2017. “Selective Trials for Technology Evaluation and Adoption.” Unpublished.

Chatterjee, Sourav, and Persi Diaconis. 2011. “Estimating and Understanding Exponential Random Graph Models.” arXiv:1102.2650.

Chatterjee, Sourav, Persi Diaconis, and Allan Sly. 2010. “Random Graphs with a Given Degree Sequence.” arXiv:1005.1136.

Chuang, Yating, and Laura Schechter. 2015. “Social Networks in Developing Countries.” *Annual Review of Resource Economics* 7: 451–72.

Crane, Harry, and Walter Dempsey. 2015. “A Framework for Statistical Network Modeling.” arXiv:1509.08185.

Diaconis, Persi, and Svante Janson. 2007. “Graph Limits and Exchangeable Random Graphs.” arXiv:0712.2749.

Elliott, Matthew, Benjamin Golub, and Matthew O. Jackson. 2014. “Financial Networks and Contagion.” *American Economic Review* 104 (10): 3115–53.

Ezoe, Satoshi, Takeo Morooka, Tatsuya Noda, Miriam Lewis Sabin, and Soichi Koike. 2012. “Population Size Estimation of Men Who Have Sex with Men through the Network Scale-Up Method in Japan.” *PLOS ONE* 7 (1): e31184.

Feehan, Dennis M., Aline Umubyeyi, Mary Mahy, Wolfgang Hladik, and Matthew J. Salganik. 2016. “Quantity versus Quality: A Survey Experiment to Improve the Network Scale-Up Method.” *American Journal of Epidemiology* 183 (8): 747–57.

Feigenberg, Benjamin, Erica Field, and Rohini Pande. 2013. “The Economic Returns to Social Interaction: Experimental Evidence from Microfinance.” *Review of Economic Studies* 80 (4): 1459–830.

Fisher, N. I., T. Lewis, and B. J. J. Embleton. 1993. *Statistical Analysis of Spherical Data.* Cambridge: Cambridge University Press.

Fosdick, Bailey K., Tyler H. McCormick, Thomas Brendan Murphy, Tin Lok James Ng, and Ted Westling. 2019. “Multiresolution Network Models.” *Journal of Computational and Graphical Statistics* 28 (1): 185–96.

Friedkin, Noah E. 1983. “Horizons of Observability and Limits of Informal Control in Organizations.” *Social Forces* 61 (1): 54–77.

Gandy, Axel, and Luitgard A. M. Veraart. 2016. “A Bayesian Methodology for Systemic Risk Assessment in Financial Networks.” *Management Science* 63 (12): 4428–46.

Gandy, Axel, and Luitgard A. M. Veraart. 2019. “Adjustable Network Reconstruction with Applications to CDS Exposures.” *Journal of Multivariate Analysis* 172: 193–209.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis.* Boca Raton, FL: CRC Press.

Graham, Bryan S. 2017. “An Econometric Model of Network Formation with Degree Heterogeneity.” *Econometrica* 85 (4): 1033–63.

Guo, Wei, Shuilian Bao, Wen Lin, Guohui Wu, Wei Zhang, Wolfgang Hladik, Abu Abdul-Quader, Marc Bulterys, Serena Fuller, and Lu Wang. 2013. “Estimating the Size of HIV Key Affected Populations in Chongqing, China, Using the Network Scale-Up Method.” *PLOS ONE* 8 (8): e71796.

Guttorp, Peter, and Richard A. Lockhart. 1988. “Finding the Location of a Signal: A Bayesian Analysis.” *Journal of the American Statistical Association* 83 (402): 322–30.

Hoff, Peter. 2008. “Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data.” In *Advances in Neural Information Processing Systems 20*, edited by J. C. Platt et al., 657–64. Vancouver: NIPS.

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. 2002. “Latent Space Approaches to Social Network Analysis.” *Journal of the American Statistical Association* 97 (460): 1090–98.

Holland, Paul W., and Samuel Leinhardt. 1981. “An Exponential Family of Probability Distributions for Directed Graphs.” *Journal of the American Statistical Association* 76 (373): 33–50.

Hoover, Douglas N. 1979. “Relations on Probability Spaces and Arrays of Random Variables.” Unpublished.

Hornik, Kurt, and Bettina Grün. 2013. “On Conjugate Families and Jeffreys Priors for von Mises–Fisher Distributions.” *Journal of Statistical Planning and Inference* 143 (5): 992–99.

Hunter, David R. 2004. “MM Algorithms for Generalized Bradley–Terry Models.” *Annals of Statistics* 32 (1): 384–406.

Jackson, Matthew O., Tomas R. Rodriguez-Barraquer, and Xu Tan. 2012. "Social Capital and Social Quilts: Network Patterns of Favor Exchange." *American Economic Review* 102 (5): 1857–97.

Kadushin, Charles, Peter D. Killworth, H. Russell Bernard, and Andrew A. Beveridge. 2006. "Scale-Up Methods as Applied to Estimates of Heroin Use." *Journal of Drug Issues* 36 (2): 417–40.

Karlan, Dean, Markus Mobius, Tanya Rosenblat, and Adam Szeidl. 2009. "Trust and Social Collateral." *Quarterly Journal of Economics* 124 (3): 1307–61.

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 (2): 289–308.

Kinnan, Cynthia, and Robert Townsend. 2012. "Kinship and Financial Networks, Formal Financial Access, and Risk Reduction." *American Economic Review* 102 (3): 289–93.

Krackhardt, David. 1987. "Cognitive Social Structures." *Social Networks* 9 (2): 109–34.

Krackhardt, David. 2014. "A Preliminary Look at Accuracy in Egonets." In *Contemporary Perspectives on Organizational Social Networks,* Vol. 40, edited by Daniel J. Brass et al., 277–93. Bingley, UK: Emerald Group Publishing.

Ligon, Ethan, and Laura Schechter. 2012. "Motives for Sharing in Social Networks." *Journal of Development Economics* 99 (1): 13–26.

Lovász, László, and Balázs Szegedy. 2006. "Limits of Dense Graph Sequences." *Journal of Combinatorial Theory,* Series B 96 (6): 933–57.

Maghsoudi, Ahmad, Mohammad Reza Baneshi, Mojtaba Neydavoodi, and AliAkbar 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 (11): e110917.

Mardia, Kanti V., and S. A. M. El-Atoum. 1976. "Bayesian Inference for the von Mises–Fisher Distribution." *Biometrika* 63 (1): 203–206.

Mardia, Kanti V., and Peter E. Jupp. 2009. *Directional Statistics.* New York: John Wiley & Sons.

McCormick, Tyler H., and Tian Zheng. 2015. "Latent Surface Models for Networks Using Aggregated Relational Data." *Journal of the American Statistical Association* 110 (512): 1684–95.

Orbanz, Peter, and Daniel M. Roy. 2015. "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2): 437–61.

Papadakis, Manos, Michail Tsagris, Marios Dimitriadis, Ioannis Tsamardinos, Matteo Fasiolo, Giorgos Borboudakis, and John Burkardt. 2017. "Rfast: Fast R Functions." R package version, 1 (5).

Park, Juyong, and M. E. J. Newman. 2004. "Statistical Mechanics of Networks." *Physical Review E* 70 (6): 066117.

Penrose, Mathew. 2003. *Random Geometric Graphs.* Oxford: Oxford University Press.

Salganik, Matthew J., Dimitri Fazito, Neilane Bertoni, Alexandre H. Abdo, Maeve B. Mello, and Francisco 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 (10): 1190–96.

Straub, J., T. Campbell, J. P. How, and J. W. Fisher. 2015. "Small-Variance Nonparametric Clustering on the Hypersphere." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 334–42.

Tontarawongsa, Chutima, Aprajit Mahajan, and Alessandro Tarozzi. 2011. "(Limited) Diffusion of Health-Protecting Behaviors: Evidence from Nonbeneficiaries of a Public Health Program in Orissa (India)." Unpublished.

Upper, Christian, and Andreas Worms. 2004. "Estimating Bilateral Exposures in the German Interbank Market: Is There a Danger of Contagion?" *European Economic Review* 48 (4): 827–49.

Wood, Andrew T. A. 1994. "Simulation of the von Mises Fisher Distribution." *Communications in Statistics—Simulation and Computation* 23 (1): 157–64.

Zheng, Tian, Matthew J. Salganik, and Andrew 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 (474): 409–23.