PNAS RESEARCH ARTICLE | SOCIAL SCIENCES STATISTICS

Consistently estimating network statistics using aggregated relational data

Emily Breza[a], Arun G. Chandrasekhar[b], Shane Lubold[c], Tyler H. McCormick[c,d,1], and Mengjie Pan[c]

Edited by H. Russell Bernard, Arizona State University, Gainesville, FL; received April 25, 2022; accepted March 21, 2023

Collecting complete network data is expensive, time-consuming, and often infeasible. Aggregated Relational Data (ARD),which ask respondents questions of the form "How many people with trait X do you know?" provide a low-cost option when collecting complete network data is not possible. Rather than asking about connections between each pair of individuals directly, ARD collect the number of contacts the respondent knows with a given trait. Despite widespread use and a growing literature on ARD methodology, there is still no systematic understanding of when and why ARD should accurately recover features of the unobserved network. This paper provides such a characterization by deriving conditions under which statistics about the unobserved network (or functions of these statistics like regression coefficients) can be consistently estimated using ARD. We first provide consistent estimates of network model parameters for three commonly used probabilistic models: the beta-model with node-specific unobserved effects, the stochastic block model with unobserved community structure, and latent geometric space models with unobserved latent locations. A key observation is that cross-group link probabilities for a collection of (possibly unobserved) groups identify the model parameters, meaning ARD are sufficient for parameter estimation. With these estimated parameters, it is possible to simulate graphs from the fitted distribution and analyze the distribution of network statistics. We can then characterize conditions under which the simulated networks based on ARD will allow for consistent estimation of the unobserved network statistics, such as eigenvector centrality, or response functions by or of the unobserved network, such as regression coefficients.

social networks | aggregated relational data | consistency | survey methods

Significance

Collecting full network data is often infeasible, costly, or limited by privacy concerns. Aggregated Relational Data (ARD), where researchers collect the number of connections to different groups, can sometimes save over 80complete network data. Little is known about when ARD can be reliably used. We show that ARD contain sufficient information to estimate parameters of common generative network models. We characterize conditions for network statistics—such as the degree of a node or its eigenvector centrality—or responses in the network to an intervention or vice versa, to be accurately estimated based solely on ARD. Our analysis shows that inexpensive data can be used to reliably conduct estimation in a host of socioeconomic settings.

The empirical study of social networks has grown rapidly across a variety of disciplines, including but not limited to economics, psychology, public health, sociology, and statistics. The aim ranges from researchers trying to understand features of the network structure across populations, to parameters in models of network formation, to how network features affect socioeconomic behavior, to how interventions can affect the structure of the social network. Studying network structure and its relationship to other phenomena can be demanding particularly in contexts where survey-based research methods are used: Obtaining high-quality network data from large populations can be expensive and often infeasible for cost, privacy, or logistical reasons. The challenges associated with collecting complete network data mean that researchers must choose to either i) reuse one of a handful of existing full graph datasets, likely not designed with their particular research goals in mind or ii) postpone their research agenda while raising sufficient capital.

One recent approach to address these issues is known as Aggregated Relational Data (ARD), which solicit summaries of respondents' connections by asking for the number of people a respondent knows with a given trait. ARD questions take the form "How many people with trait k are you linked to?" and can be integrated into standard probability-based survey sampling schemes because they do not directly solicit any connections in the graph (1). One major advantage of collecting ARD over more traditional network surveys is the reduced cost. In the context of one large-scale randomized controlled trial studying the relationship between network structure and household finance in 60 villages, ref. 1 showed that ARD implementation is shorter (3 vs. 8 mo) and cheaper ($34,000 vs. $189,000) compared to full network enumeration and yet delivered the same economic conclusions that would have been obtained using the full network data. Because it is cheaper to collect, ARD also enable practitioners to collect panel data across multiple networks.

ARD were originally proposed to estimate the size of hard-to-reach populations, such as the number of HIV-positive men in the United States (2–4). Since then,

Author affiliations: [a]Department of Economics, Harvard University, Cambridge, MA 02138; [b]Department of Economics, Stanford University, Stanford, CA 94305; [c]Department of Statistics, University of Washington, Seattle, WA 98195; and [d]Department of Sociology, University of Washington, Seattle, WA 98195

Author contributions: E.B., A.G.C., and T.H.M. designed research; E.B., A.G.C., S.L., T.H.M., and M.P. performed research; E.B., A.G.C., S.L., T.H.M., and M.P. analyzed data; and E.B., A.G.C., S.L., T.H.M., and M.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: tylermc@u.washington.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2207185120/-/DCSupplemental.

Published May 16, 2023.

the use of ARD has expanded significantly, particularly in the social sciences (5–7) where ARD enable researchers to estimate core features of respondent networks (e.g., a respondent's centrality or the extent of clustering). In terms of methodology for analyzing ARD, ref. 8 connects a model for ARD responses to network models of the fully observed graph. Specifically, ref. 8 established a connection between ARD and the latent distance model, a common statistical approach for modeling fully observed network data. The key result is that ARD are sufficient to identify parameters in a generative model for graphs, allowing inference about the distribution of graphs that plausibly correspond to the ARD. Ref. 1 exploits this connection to generate a distribution over network statistics, such as the centrality of an individual or the average path length of the graph, and shows examples where using statistics generated from ARD gives similar results to using statistics from the completely observed graph. ARD have also been used to estimate common econometric models and outcomes, such as the linear-in-means model (9), choosing the optimal seeding for maximal information flow (10), and can be used to assess network model goodness-of-fit (11).

Despite its increasingly widespread use, there is still little understanding of when or why ARD contain sufficient information to estimate model parameters or estimate network properties of the unobserved network. We provide such a characterization in two steps. First, we show that we can consistently estimate the parameters of a rich class of generative network models using only ARD. This fact relies on a simple but powerful observation that if the cross-type link probabilities allow us to identify the model parameters, then ARD are sufficient for consistent estimation. Critically, this insight allows us to sidestep maximizing the complicated log-likelihood directly and instead solve a system of equations based on the cross-trait linking probabilities. We show that three common generative models fall into this class. The graphs we consider all have the property that the average degree of the graph grows as the graph size increases. Practically, this means that the results we present may not be suitable for extremely large, very sparse graphs (e.g., social media communications, massive cities, or lower scale peer-to-peer financial networks). We refer readers to ref. 12 for a discussion of density in empirically observed network data. We provide evidence in this paper using simulations on graphs with properties similar to the networks observed in ref. 13, which has an average degree of 17 and an average size of 213 nodes. We also conduct simulations to show that we are able to recover graph and individual-level statistics when the average degree is 9 and the density is .036. These simulation results are in *SI Appendix*, section S11. Extending this work to additional generative models that do not have this property could be a fruitful avenue for future work.

Next, we provide sufficient conditions to consistently estimate features of the underlying, unobserved network using ARD. The intuition is that, for sufficiently large graphs, some statistics of graphs converge to their expected value, where the expectation is taken over graphs from the same generative process. In such cases, ARD suffice to recover the value of the graph statistics, so long as the statistics are not too sensitive to error introduced by using estimates for the generative network models parameters. In such cases, the information in ARD is sufficient to consistently estimate generative model parameters as well as graph statistics of interest.

We investigate this both theoretically and empirically in two settings. The first is when researchers can consistently estimate features of the underlying, unobserved network structure itself. Examples include centrality or clustering measures for nodes. This analysis studies the case of a single large network. The second is when researchers can consistently estimate response functions of or by the network. That is, how do changes in network features correspond to changes in socioeconomic outcomes or how might an intervention affect the structure of the network. This analysis studies the case of many networks. Our evaluation includes experiments with previously analyzed data, with IRB approval from MIT IRB COUHES # 1010004040.

Aggregated Relational Data

We begin by defining ARD formally. Take an undirected, unweighted graph $g = (V, E)$ with vertex set V and edge set E. There are $n = |V|$ nodes, so we sometimes write g_n to emphasize the graph size, and $g_{ij} = \mathbf{1}\{ij \in E\}$ denotes that i and j are connected in the graph. Suppose each node has one of K traits, where K is fixed and $K > 3$. Let G_k denote the nodes with trait k, for $k = 1, \ldots, K$, where $n_k = |G_k|$ is the number of nodes with trait k. We write $t_i^\star = k$ to denote that node i has trait k. We suppose that the traits are binary and mutually exclusive, so every node has one of K traits. Researchers using existing data will need to check this assumption and may need to use a subset of ARD questions (e.g., first names) on existing data. We recommend researchers collecting new data construct mutually exclusive ARD questions. Imagine there are L characteristics (e.g., rural/urban or college educated/not), and for simplicity, assume that these are binary. Then, it is clear that we can always construct K traits, mutually exclusive, with $K = 2^L$. The extension to multivalued characteristics is straightforward. Additionally, traits constructed through intersecting characteristics (e.g., men with a given occupation below a particular age) also reduce the size of the target population, which can limit recall bias (14).

To collect Aggregated Relational Data (ARD), the researcher asks m randomly chosen nodes "How many people with trait k are you linked to?" for each of these K traits. Linking is typically defined as knowing a potential connection (e.g., having interacted with the person in the past 2 y or recognizing the person if passing on the street). Ref. 15 provides an extensive discussion and experimental evidence regarding the definition of linking. To simplify exposition, we will set $m = n$, meaning we have ARD from all nodes. Our results also apply when $m << n$, as is common in practice. In such cases, we would either need to impute parameters for nodes without ARD (ref. 1, for example) or make an assumption about node equivalence. For example, under a stochastic block model, all nodes in a given community have the same linking behavior. So, by collecting ARD from at least one node in each community, we can then estimate the parameters of all nodes.

Let y_{ik} denote node i's response to this question about trait k, with $y_{ik} = \sum_{j \in G_k} g_{ij}$. Critically, when collecting ARD, the researcher does not observe any edges, just how many edges are present between a node i and people of a given trait. We use $\mathbf{y}$ to denote the $m \times K$ matrix of ARD responses. Since the K traits are mutually exclusive, ARD responses count distinct alters across each trait group. Otherwise, if trait group A and trait group B overlap, then a person in both groups would be counted twice, once in response to the ARD question about trait A and once about B.

To model the network, we consider a general graph model $\mathbb{P}(g_n|\theta^\star)$, where edges form independently in the network, conditional on the unknown parameter vector $\theta^\star$. We call such models conditional edge-independent graph models. The number of elements in the vector $\theta^\star$ can depend on the graph

size n (to accommodate node-level heterogeneity parameters, for example), but we omit this dependence to simplify the notation ($\theta^\star = \theta_n^\star$). In most settings, each component of the vector $\theta^\star$, which we denote by $\theta_i^\star$ for $i = 1, \ldots, n$, is independently and identically drawn from F. In other cases, sometimes, the distribution of $\theta_i^\star$ depends on the traits that node i possesses, which we write as $\theta_i^\star | t_i^\star = k \sim F_k$. This conditional independence representation relies on exchangeability among nodes and, thus, implies that the resulting asymptotic sequence of networks generated by these models are dense, meaning that the average degree for a given n is a constant times n (16).

We define $p_{ij} = p_{ij}(\theta)$ to be the probability that i and j connect, given the model parameters. Proving consistency of a maximum likelihood estimate of the model parameters ($\hat{\theta}_n := \arg\max_\theta \mathcal{L}_n(\mathbf{y} \mid \theta)$) is challenging due to the complex nature of the log-likelihood, since each parameter θ_i appears in n terms of the likelihood. We explore a different approach to estimate $\theta^\star$. Specifically, we work directly with the probability that node i connects to an arbitrary node with trait k, P_{ik}, which is

$$P_{ik} := \mathbb{P}(g_{ij} = 1 | \theta_i^\star, j \in G_k) = \int_{\Theta_k} \mathbb{P}(g_{ij} = 1 | \theta_i^\star, \theta_j) dF_k(\theta_j),$$

where again $\theta_j \sim F_{\theta,k}$ for nodes i with trait k, and Θ_k denotes the support of F_k. In the latent space model, Θ_k might be p-dimensional Euclidean, spherical, or hyperbolic space, and node locations are drawn according to a mixture model along the surface of the latent space (17–19). In the beta-model, Θ_k is a subset of the real line. For any node i,

$$\frac{y_{ik}}{n_k} = \frac{1}{n_k} \sum_{j \in G_k} g_{ij} \xrightarrow{p} P_{ik}, \qquad [1]$$

as $n_k \to \infty$, where P_{ik} is again the probability that node i connects with someone of trait k and n_k is the number of nodes with trait k. Here, we have assumed that the weak law of large numbers applies to the average y_{ik}/n_k, as is the case for conditionally edge-independent graphs. In the conclusion, we discuss extensions for settings where edges could be correlated or where edge probability scales with the graph size.

Supposing that (**1**) holds, we can then equate the vector of normalized ARD responses with their respective edge probabilities $P_{ik}(\theta^\star)$ and use an estimating equation approach to estimate the model parameters. Supposing that this system has a unique solution in θ (or unique up to an isometry, as in the latent space model), this general approach allows us to derive estimators of model parameters and prove uniform convergence of these estimators in a host of rich and frequently used network models. When this system does have a unique solution, we say informally that such a model "identifies" the model parameters.

By equating observed ARD responses and the probability of connection between a node and nodes in a given trait group, we can invert that equation to solve for the parameters $\theta_i^\star$. In the next three sections, we consider three common generative network models and derive consistent estimates of the parameters in each model using this intuition.

Beta-Model

We first consider the generalized beta-model (20, 21). The original version of this model states that an edge forms between nodes i and j with probability $\text{expit}(\nu_i^\star + \nu_j^\star)$ for some sequence of parameters $\nu_1^\star, \ldots, \nu_n^\star$ that encode the popularity of nodes. Here, $\text{expit}(x) = \exp(x)/(1 + \exp(x))$. The generalized beta-model includes a term that measures the effect of dyad-level covariates $X_{ij} \in \mathbb{R}^p$ on linking probability, so

$$\mathbb{P}(g_{ij} = 1 | \nu_i^\star, \nu_j^\star, \beta^\star) = \text{expit}(\nu_i^\star + \nu_j^\star + \beta^\star X_{ij}).$$

Refs. 20 and 21 propose estimates of the parameters using a fixed-point procedure using the full network data. This procedure only requires the degree of a node. Suppose that we observe ARD about a collection of traits that are mutually exclusive and exhaustive. Then, the degree of node i is $d_i = \sum_{k=1}^K y_{ik}$. Let $\hat{\nu}_i$ and $\hat{\beta}$ denote the fixed-point estimates of ν_i and β from refs. 20 and 21 computed using the ARD, which is by the preceding comments equivalent to the estimate computed from the full network data.

In the theorem below, we require that the support of the parameters in the beta-model be compact subsets of $\mathbb{R}$. This regularity condition was also imposed in ref. 21.

Theorem 1. *Suppose the support of each node effect $\nu_i^\star$ and of $\beta^\star$ are compact subsets of $\mathbb{R}$. Then, with probability $1 - O(1/n^2)$,*

$$\max_{1 \le i \le n} |\hat{\nu}_i - \nu_i^\star| \le C \sqrt{\frac{\log(n)}{n}},$$

for some constant C that does not depend on n. In addition, $\hat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$.

Here, we have not made any assumption about the relationship between traits and the distribution of the node parameters.

In cases where ARD are collected at the characteristic level and not at the trait level (which creates a mutually exclusive partition), or when the mutually exclusive traits do not exhaust the space, $\sum_{k=1}^K y_{ik}$ does not need to equal d_i, the degree of node i. In these cases, we can estimate the degree of a node via other methods. One such example is the network scale-up method (2, 22), which assumes that given a node's degree, ARD responses are modeled as $y_{ik} | d_i \sim \text{Binomial}(d_i, \frac{n_k}{n})$. This leads to the estimator $\hat{d}_i = n \sum_{k=1}^K y_{ik} / \sum_{k=1}^K n_k$, where n_k is the size of group k and n is the total size of the population (2, 14, 23). Typically, these ARD questions are based on characteristics with known group sizes, so that each n_k is known. We can then plug in $\hat{d}_i$ in place of d_i in the estimation procedures from refs. 20 and 21 to estimate the model parameters.

Stochastic Block Model

We consider a generalized version of the stochastic block model (SBM), in which observable traits are dependent on, but potentially distinct from, latent community structure. Edges are determined by latent community structure. This setting corresponds to a case where nodes belong to unobserved communities and a researcher observes traits that are (imperfectly) associated with community membership. We show that ARD allow us to use links to observable groups to infer latent community membership.

We begin by describing the model for edge formation, based on latent communities, and then relate this model to the observable traits. First, assign node communities $c_i^\star$ independently with probabilities $\pi_1, \ldots, \pi_C$. Conditioned on these parameters, edges are generated independently with probabilities

$$\mathbb{P}(g_{ij} = 1 | c_i^\star = c, c_j^\star = c') = P_{cc'},$$

where P is a $C \times C$ matrix of within- and cross-community edge probabilities. Here, we suppose that the graph is undirected, so

that P is assumed to be symmetric. The intuition for this model is that the probability that two nodes connect depends only on their latent group membership. Typically, the community structure is unknown a priori and unobservable. We show that it is possible to recover this community structure with observable traits, so long as people with similar traits play similar roles in the network. We let the $C \times K$ matrix Q encode the probability of having trait k, given that a node is in community c, so

$$\mathbb{P}(t_i^\star = k | c_i^\star = c) = Q_{ck}. \quad [2]$$

Since traits are mutually exclusive, each node is assigned exactly one of the K traits with probabilities in Eq. **2**. The intuition behind this model is that nodes with the same traits form edges in a similar way.

We suppose that the ARD we have access to are about these K traits and not about the unobserved community structure. To estimate the parameters in the SBM, we begin by making the following assumption, which allows us to consistently cluster the ARD to estimate community structure in the unobserved graph. Specifically, we assume that no two communities have the same linking pattern to all other traits, which is clearly required for identification.

Assumption 1. *The following condition holds:*

$$\min_{c,c'} ||Z_c - Z_{c'}|| > 0,$$

where $Z_c := (\tilde{P}_{c1}, \ldots, \tilde{P}_{cK})$ *and* $\tilde{P}_{ck} := \mathbb{P}(g_{ij} = 1 | c_i = c, t_j = k)$ *is the probability that a node in community c connects to a node with trait k:* $\tilde{P}_{ck} = \left(\sum_{\ell=1}^{C} Q_{\ell k}\pi_\ell\right)^{-1} \sum_{\ell'=1}^{C} P_{c\ell'} Q_{\ell' k}\pi_{\ell'}$.

To understand this assumption, let us consider a simple case when $C = K = 2$. Assumption 1 then requires that

$$\begin{pmatrix} \tilde{P}_{11} \\ \tilde{P}_{12} \end{pmatrix} \neq \begin{pmatrix} \tilde{P}_{21} \\ \tilde{P}_{22} \end{pmatrix}.$$

We now explore further when these equalities do not hold. If the probabilities of belonging to community 1 and 2 are equal ($\pi_1 = \pi_2$), the first inequality is then equivalent to requiring that

$$(P_{11} - P_{21})Q_{11} + (P_{12} - P_{22})Q_{21} \neq 0.$$

If $P_{11} = P_{12} = P_{22} = P_{22}$, which corresponds to no community structure in the model, then Assumption 1 is not satisfied for any Q matrix. If $Q_{12} = Q_{21}$, which means that there is no relationship between traits and community membership, then Assumption 1 is satisfied whenever $P_{11} - P_{21} \neq P_{22} - P_{21}$, which occurs in undirected networks whenever $P_{11} \neq P_{22}$. In this case, even if there is no relationship between traits and network structure, Assumption 1 is satisfied provided that communities behave differently in the network (i.e., there is meaningful community structure).

We now provide a classification algorithm to estimate the community membership of nodes. This procedure does not require us to know the number of communities. We initialize $W = V$, the set of nodes in the sample, so $|W| = n$. Let $\tilde{y}_i = (y_{i1}/n_1, \ldots, y_{iK}/n_K)$. While $W \neq \emptyset$, do the following, which we refer to as Algorithm 1:

1. Select a node i randomly from W. Set $W = W \setminus \{i\}$.
2. For any $j \in W$: If $||\tilde{y}_i - \tilde{y}_j||^2 \leq n^{-1}\log(n)$, assign node j to be in the same community as i, and second set $W = W \setminus \{j\}$.

This procedure returns a consistent estimate of the community membership and the number of communities. The distribution of ARD responses for people in a given community c collapses to a point mass as the sample size grows, and so, clustering in our problem is easier than clustering in general clustering problems, where the distribution of data does not need to change with the sample size. We therefore propose the algorithm above, over more standard clustering algorithms, because our clustering algorithm lends itself easily to concluding the uniform consistency in Theorem 2 that we need later in Theorems 4 and 5.

We prove in Theorem 2 that this classification algorithm returns consistent community labels. Given the community memberships $\hat{\mathbf{c}}$, let $\hat{C}_c$ denote the set of nodes in our sample that are estimated to be in community c, under the membership vector $\hat{\mathbf{c}}$, with $|\hat{C}_c| =: m_c(n)$. The term $\sum_{k=1}^{K} y_{ik} \times \mathbb{P}(c_j = c' \mid t_j = k)$ is the expected number of connections that node i has to a node in community c'. Since the communities are latent, we now relate $\mathbb{P}(c_j = c' \mid t_j = k)$ to terms that we are more familiar with:

$$\begin{aligned}\mathbb{P}(c_j = c' \mid t_j = k) &= \frac{\mathbb{P}(t_j = k, c_j = c')}{\mathbb{P}(t_j = k)} \\ &= \frac{\mathbb{P}(t_j = k \mid c_j = c')\mathbb{P}(c_j = c')}{\mathbb{P}(t_j = k)} \\ &= \frac{Q_{c'k} \times \pi_{c'}}{\mathbb{P}(t_j = k)}.\end{aligned}$$

Now, these terms we can estimate from data. Given an estimated community membership vector $\hat{c}$, we can estimate $\hat{\mathbb{P}}(c_j = c') = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{c}_i = c'\}$. We can estimate Q_{ck} with

$$\hat{Q}_{ck} = \frac{1}{m_c(n)} \sum_{i \in \hat{C}_c} \mathbf{1}\{t_i = k\},$$

where t_i is the observed trait of node i, and we can estimate π with entries $\hat{\pi}_c = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{c}_i = c\}$. Let $\hat{\mathbb{P}}(c_j = c' \mid t_j = k)$ denote the estimate of $\mathbb{P}(c_j = c' \mid t_j = k)$ based on the estimated probabilities, as described above. Then, we can estimate

$$\hat{P}_{cc'} = \begin{cases} \frac{1}{m_c(n)} \sum_{i \in \hat{C}_c} \sum_k \frac{y_{ik}\hat{\mathbb{P}}(c_j = c' | t_j = k)}{n_{c'}}, & c \neq c' \\ \frac{1}{m_c(n)} \sum_{i \in \hat{C}_c} \sum_k \frac{y_{ik}\hat{\mathbb{P}}(c_j = c' | t_j = k)}{n_{c'} - 1}, & c = c' \end{cases}.$$

Here, y_{ik} is the ARD response from node i about trait k.

Theorem 2. *Suppose Assumption 1 holds. Then, up to a permutation on the community labels, the community membership vector estimated using Algorithm 1,* $\hat{c}$*, satisfies*

$$\max_{1 \leq i \leq n} \mathbf{1}\{\hat{c}_i \neq c_i^\star\} \xrightarrow{p} 0,$$

as $n \to \infty$*. The estimated number of communities* $\hat{C}$ *as well as* $\hat{P}$*,* $\hat{Q}$*, and* $\hat{\pi}$ *are all consistent as* $n \to \infty$*.*

Latent Space Model

We consider a broad class of latent space models. Broadly speaking, each node has a position in a latent (or unobserved) space, and the closer two nodes are in this space, the more

likely they are to connect. Each node also has a gregariousness parameter, which controls the baseline edge probability for that node (17–19, 24, 25).

We formally define one variant of the latent space generative model, which we study in this work. We draw the gregariousness parameter $\nu_i^\star$ from a distribution F_ν with compact support in $(a, 0)$ for some $a < 0$. We draw traits $t_i^\star \in \{1, \ldots, K\}$ independently with probabilities $\pi_1, \ldots, \pi_K$. Conditioned on traits, we also draw node positions $z_i^\star | t_i^\star = t \sim F_t$, where F_t is some distribution over the latent surface $\mathcal{M}^p(\kappa)$. Here, $\mathcal{M}^p(\kappa)$ is a complete, simply connected Riemannian manifold with constant curvature κ, which means by the Killing–Hopf theorem that it is Euclidean, spherical, or hyperbolic space of dimension p and curvature κ (26). We suppose that F_t is a symmetric distribution over $\mathcal{M}$ and is uniquely determined by its mean μ_t and variance σ_t^2. Some examples of this include the Gaussian distribution over $\mathbb{R}^p$ and the von Mises–Fisher distribution over the p-sphere. In words, the node positions $z_i^\star$ are drawn from a mixture distribution on $\mathcal{M}^p(\kappa)$, with weights determined by $\pi_k = \mathbb{P}(t_i^\star = k)$. Conditioned on these parameters, we draw edges independently with probability

$$\mathbb{P}(g_{ij} = 1 | \nu_i^\star, \nu_j^\star, z_i^\star, z_j^\star) = \exp\{\nu_i^\star + \nu_j^\star - d(z_i^\star, z_j^\star)\}. \quad [3]$$

Again, we suppose that we only have access to ARD about these K traits.

We write $\eta = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$ to refer to the "global" parameters. To build our estimators $\hat{\nu}_i, \hat{z}_i$, and $\hat{\eta}$, we proceed in two steps: a) estimate the global parameters and b) use them as a plug-in to estimate the node parameters. Our proof is based mainly on the following calculation. Consider the marginal probability of a connection between person i and group k, P_{ik}. The form of P_{ik} comes from integrating across all individuals in group k in Eq. **3**, which is consistent with the information in ARD since no individual connections are observed. Further, following refs. 8 and 17, we can model $y_{ik} \mid \nu_i^\star, z_i^\star \sim \text{Binomial}(n_k, P_{ik})$, where P_{ik} is the probability that i connects to a member of group k (the explicit form is derived in *SI Appendix* and is a function of $\nu_i^\star$ and $z_i^\star$) and n_k is the number of nodes in group k.

In step (a), we derive the estimators for the global parameters, $\hat{\eta}$, in *SI Appendix*, section S.1 but provide the intuition here. If we consider the probability of an arbitrary link between two members of the same group k, it does not depend on μ_k but only on the variance σ_k^2 and the expected shift in linking probability due to node effect ν_i. Similarly, if we consider the probability of an arbitrary link across two groups k, k' knowing the variance terms, then this provides information on centers μ_k, μ_k'. We can therefore equate the probability of connection between traits with the observed number of traits and solve for the parameter η. Given estimates of the global parameters η, we now estimate the node locations and fixed effects. Since $E(y_{ik}) = P_{ik}/n_k$, we construct a system of equations by equating the ratio of the marginal probability of connection for person i in group k to that in group k' $(P_{ik}/p_{ik'})$ to the ratio of sample averages $(y_{ik}n_{k'}/y_{ik'}n_k)$, which does not depend on the fixed effect of node i. This allows us to estimate the locations of all nodes, up to a global isometry in the latent space. We then similarly estimate the node fixed effects once we have estimated the node locations and global parameters, by equating y_{ik} and p_{ik}. In summary, we construct Z-estimators of the global parameters, the node locations, and the node fixed effects by constructing 4 systems of equations, which allows us to consistently estimate all of the parameters in the latent space model. Equivalently, one can interpret the moments-based estimators for the location and fixed effects parameters as coming from maximizing a pseudolikelihood, which we describe in *SI Appendix*.

We now state the assumptions for consistency of these estimators. $E_{kk'}[\exp\{-d(z, z')\}]$ denotes the expectation of $\exp\{-d(z, z')\}$, where $z \sim F(\mu_k^\star, \sigma_k^\star)$ is independent of $z' \sim F(\mu_{k'}^\star, \sigma_{k'}^\star)$.

Assumption 2. *For each k, μ_k is in a compact subset of $\mathcal{M}^p(\kappa)$ and σ_k is in a compact subset of $(0, \infty)$.*

Assumption 3. *The node effects $\nu_i^\star \overset{iid}{\sim} H$ satisfy $E\{\exp(\nu_i^\star)\} < \infty$.*

Assumption 4. *The distribution F is a symmetric distribution on $\mathcal{M}^p(\kappa)$ that is completely characterized by its mean and variance and satisfies the following two conditions. The function $z_i \mapsto E_k[\exp\{-d(z_i, z)\}]$ is Lipschitz for every $k \in \{1, \ldots, K\}$ and $z_i \mapsto E_k[\exp\{-d(z_i, z)\}]/E_{k'}[\exp\{-d(z_i, z')\}]$ has a pseudoinverse that is Lipschitz.*

Assumption 5. *Define the function $F_1 : (z_i, \sigma_k, \sigma_{k'}) \mapsto E_k[\exp\{-d(z_i, z)\}]/E_{k'}[\exp\{-d(z_i, z')\}]$. The inverse function F_1^{-1} is continuous in σ, and for every k, k', ℓ, and ℓ', the following two functions are Lipschitz:*

$$\eta \mapsto \frac{E_{kk'}[\exp\{-d(z, z')\}]}{E_{\ell\ell'}[\exp\{-d(z, z')\}]}, \quad \eta \mapsto \frac{E_{kk'}[\{\exp(-d(z, z')\}]^2}{E_{\ell\ell'}[\{\exp(-d(z, z')\}]^2}.$$

Assumptions 4 and 5 ensure that the probabilities from Eq. **3** vary smoothly with changes in the distribution of points on $\mathcal{M}^p(\kappa)$. In *SI Appendix*, we verify that common distributional choices (e.g., Gaussian in Euclidean space or von Mises–Fisher on the hypersphere) satisfy these assumptions and discuss the pseudoinverse defined in the assumptions above. For simplicity, we suppose that $n_k = n/K$ for each trait, so that traits are evenly divided among the nodes, and write $\tilde{n} = n/K$.

Theorem 3. *Suppose Assumptions 2, 3, 4, and 5 hold. The estimators $\hat{z}_i$ and $\hat{\nu}_i$ computed from equating the ARD responses and the marginal probability of connections, as well as $\hat{\eta}$ (defined in SI Appendix), are consistent for $z_i^\star$, $\nu_i^\star$, and $\eta^\star$ as $m, n \to \infty$, up to isometry on $\mathcal{M}^p(\kappa)$ and satisfy*

$$\max_{1 \le i \le m(n)} d_{\mathcal{M}^p(\kappa)}(\hat{z}_i, z_i^\star) \le \sqrt{\frac{3\log(\tilde{n})}{2\tilde{n}}},$$

$$\max_{1 \le i \le m(n)} |\hat{\nu}_i - \nu_i^\star| \le \sqrt{\frac{3\log(\tilde{n})}{2\tilde{n}}},$$

with probability $1 - O(m/\tilde{n}^3)$.

The proof of Theorem 3 and associated simulations are in *SI Appendix*.

A Taxonomy for Estimating Graph Statistics

We assume that data arise from one of three models considered in the previous work (beta-model, stochastic block model, or latent space model) and that ARD allow us to estimate the model parameters $\theta^\star$. We leverage Theorems 1, 2, and 3 and assume throughout the rest of this work that the researcher has access to an estimator $\hat{\theta}_n(\mathbf{y})$ of $\theta^\star$. Here, $\theta^\star$ denotes the true parameters

of one of the three models, and $\hat{\theta}_n(\mathbf{y})$ denotes the estimates of the model parameters from Theorems 1, 2, and 3. We separate our discussion into two cases: 1) The researcher has a single large network with n nodes, and 2) the researcher has many independent networks. We recall for convenience that the user has access to an ARD survey from $m \leq n$ nodes.

Single Large Network. Starting with the first case, assume that the researcher is interested in estimating a network statistic, $S_i(g_n^\star)$ for node i computed on the graph $g_n^\star$. For simplicity, we write this as a function of a single node, though it can easily be extended to functions of multiple nodes. For the purposes of this argument, there is one actual realization of the graph, g_n^*. This is what we would have observed if we had collected information about all actual connections between members of the population, rather than collecting ARD. Importantly, the researcher collecting ARD cannot observe g_n^*. This actual network realization does, however, come from a generative model with parameters that can, by Theorems 1, 2, and 3, be estimated from ARD.

In the following results, we characterize settings where network statics can be consistently estimated using only the $n \times K$ matrix of ARD, $\mathbf{y}$. For simplicity, we set $m = n$, though our results hold when $m < n$ as well, though a researcher would need to sample a sufficiently large fraction of the graph to capture the structure of interest (27). Based on observing ARD, we compute $E\{S_i(g_n) \mid \hat{\theta}(\mathbf{y})\}$, where $\hat{\theta}(\mathbf{y})$ is the estimator from Theorems 1, 2, or 3 using the ARD $\mathbf{y}$. We are interested in the condition when $E\{S_i(g_n) \mid \hat{\theta}_n(\mathbf{y})\}$ is a good estimator of $E\{S_i(g_n) \mid \theta^\star\}$ and therefore of $S_i(g_n^\star)$.

There are two general conditions required to consistently estimate graph parameters from ARD. First, the statistic of interest must be one that is relatively stable between draws from the graph-generating process. This condition is required since our estimators in the previous section concern parameters of the network formation model, but the goal is to estimate a statistic for a particular draw from this generating process, g_n^*. Second, we require that these estimates of generating model parameters are sufficiently precise, and the form of the statistics is such that we can control the variation in the estimated network statistic in the presence of small variance in the estimated model parameters. We formalize these conditions in the following theorem. We use the notation $\theta_{j,n}^\star$ to refer to the jth entry of the vector of true parameter values $\theta^\star \in \mathbb{R}^n$. Finally, let the partial derivative with respect to the ith component be denoted by $\partial_i E\{S_i(g_n) \mid \theta_n\}$.

Theorem 4. *Let $g_n^\star$ denote the graph of interest drawn from a conditional edge-independent graph model with parameters $\theta_1^\star, \ldots, \theta_n^\star$, and let $\hat{\theta}_n$ denote estimates of these parameters. Suppose that*

1. $1/n \sum_j |\hat{\theta}_{j,n} - \theta_{j,n}^\star| \xrightarrow{p} 0$,
2. $|E\{S_i(g_n) \mid \theta^\star\} - S_i(g_n^\star)| \xrightarrow{p} 0$, *and*
3. *the function $\theta_n \mapsto E\{S_i(g_n) \mid \theta_n\}$ is differentiable and*

$$\max_j \sup_{\theta_n} \partial_j E\{S_i(g_n) \mid \theta_n\} \leq C/n$$

for some finite constant $C > 0$.

Then, $|E\{S_i(g_n) \mid \hat{\theta}_n(\mathbf{y})\} - S_i(g_n^\star)| \xrightarrow{p} 0$ as $n \to \infty$.

We provide a proof of Theorem 4 in *SI Appendix*. The proof relies on a Taylor series approximation of the network statistic $E\{S_i(g_n) \mid \hat{\theta}_n(\mathbf{y})\}$. In particular, we require that the approximation term due to the estimation of $\theta_n^\star$ with $\hat{\theta}_n(\mathbf{y})$ disappear as $n \to \infty$. One sufficient condition for this to occur is given in Conditions 1 to 3 of Theorem 4.

Condition 1 of Theorem 4 requires that the average estimation error goes to zero in probability as the graph size grows. The estimators from Theorems 1, 2, and 3 satisfy Condition 1 of Theorem 4 since the average estimation error is always upper-bounded by the maximum estimation error. Thus, Theorem 4 implies that the researcher can use $E\{S_i(g_n) \mid \hat{\theta}_n(\mathbf{y})\}$ to estimate $S_i(g_n^\star)$, provided the network statistic $S_i(g_n^\star)$ satisfies Conditions 2 and 3.

Condition 2 of Theorem 4 requires that $|E\{S_i(g_n) \mid \theta^\star\} - S_i(g_n^\star)| \xrightarrow{p} 0$, which must be true regardless of the estimator used to estimate $\theta^\star$. Many network statistics are an average of terms, such as the clustering coefficient or the centrality coefficient, and so, this condition holds for many statistics of interest. Condition 3 of Theorem 4 requires that changing the graph model parameters slightly does not change the value of $E\{S_i(g_n) \mid \theta_n\}$ too much. For many common network statistics, this condition is true, as we show in Corollary 2.

To clarify when the conditions of Theorem 4 hold and when they fail, we provide several pedagogical examples. Our first example is an obvious failure of the second condition. Specifically, we show that the statistic from a given realization does not converge to its expectation; then, even after more nodes are observed, there is no increasing information, and the mean-squared error of the estimate should not go to zero. Let $p_{ij}(\theta^\star)$ denote the probability that nodes i and j connect.

Corollary 1. *Consider a sequence of distributions of conditional edge-independent graphs $\mathbb{P}(g_n|\theta^\star)$ on n nodes, where $\theta^\star$ is known. Given an (unobserved) graph of interest, g_n^*, and $0 < p_{ij}(\theta^\star) < 1$, then the mean squared error for $E\{S_i(g_n)\} = E(g_{ij})$, the expectation of a draw from the distribution of any single link g_{ij}, is*

$$E[\{E(g_{ij}) - g_{ij}^*\}^2] = p_{ij}(\theta^\star)\{1 - p_{ij}(\theta^\star)\}.$$

When a link exists, the mean squared error is $\{1 - p_{ij}(\theta^\star)\}^2$ and when a link does not, it is $p_{ij}(\theta^\star)^2$. In edge-independent models, node-level exchangeability ensures that $p_{ij}(\theta^\star)$ does not vanish with n, which means that the mean squared error cannot go to zero as $n \to \infty$. However, for graph models in which p_{ij} tends to zero, Condition 2 does hold.

However, for many commonly used and nontrivial network statistics, the conditions of Theorem 4 do hold. By verifying the conditions of Theorem 4, we have the following result.

Corollary 2. *Suppose $g_n^\star$ is drawn from the β-model, stochastic block model, or latent space model, and $\hat{\theta}_n$ is computed from Theorems 1, 2, and 3, respectively. Suppose that the function $\theta_i \mapsto p_{ij}(\theta)$ is differentiable and has a uniformly bounded derivative for any i and j and any parameter vector θ. For the following statistics $S_i(g_n)$, we have that $|E\{S_i(g_n^\star) \mid \hat{\theta}_n(\mathbf{y})\} - S_i(g_n^\star)| \xrightarrow{p} 0$.*

1. *Density (normalized degree): The density of node i is $S_i(g_n) = \sum_j g_{ij}/n$.*
2. *Diffusion centrality (nests eigenvector centrality and Katz–Bonacich centrality): Define $S_i(g_n) = S_i(g_n, q_n, T) = \sum_j \{\sum_{t=1}^T (q_n g_n)^t\}_{ij}$ for some $q_n = C/n$ and any T.*
3. *Clustering: Let $N(i) = \{j : g_{ij} = 1\}$ denote the neighbors of node i. The clustering coefficient is defined as $S_i(g_n) = \sum_{j,k \in N(i)} g_{jk}/(|N(i)||N(i) - 1|)$.*

Diffusion centrality is a more general form which nests eigenvector centrality when $q_n \geq 1/\lambda_1^n$, and because the maximal eigenvalue is on the order of n, this meets our condition. Here, λ_1^n is the largest eigenvalue of the adjacency matrix of g_n. It also nests Katz–Bonacich centrality. In each of these, $T \to \infty$. It also captures a number of other features of finite-sample diffusion processes that have been used particularly in economics (13, 28). These notions each relate to the eigenvectors of the network—objects that are ex ante not obviously captured by the ARD procedure but ex post work since in this model, statistics converge to their expectations.

These results give two practical extreme benchmarks. ARD should not perform well for estimating a realization of any given link in the network. In contrast, it should perform quite well for statistics such as density or eigenvector centrality. Other statistics may fall somewhere in the middle of this spectrum. For example, whether a notion of centrality such as betweenness—which relies on the specifics of the exact realized paths in the network—works well may depend on the specific statistic and network distribution. We explore these predictions empirically in Fig. 1.

Many Independent Networks. Consider the setting where the researcher has R networks each of size n_r, and the networks are over disjoint sets of nodes. This setting occurs in practice in areas such as economics (29–34), psychology (35), and health (36). We use the terminology independent networks to refer to such a collection of networks. For each network r, we observe ARD $n_r \times K$ matrix $\mathbf{y}_r$. We take $n_r = n$ for simplicity, but our results do not require this. Also, we drop the dependence on n in the notation g_r. Every network is generated from a network formation process with true parameter $\theta_r^\star$. In this case of many networks, we consider how well the ARD procedure performs when the researcher wants to learn about network properties, aggregating across the R graphs. This is the case in a large literature (32, 37, 38).

Let $S_r^* = S(g_r^*)$ be a network statistic from the R unobserved graphs generating the ARD. For any given graph from the data generating process, define $S_r = S(g_r)$. For notational simplicity, we consider network-level statistics, but the argument can easily be extended to node, pair, or subset-based statistics. We use the notation $\theta_{i,n,r}^\star$ to denote the ith entry of the vector of parameters $\theta_{n,r}^\star \in \mathbb{R}^n$ for network r. We use similar notation for the estimator $\hat{\theta}_{i,n,r}$.

We consider two regression problems. In the first problem, the goal of the researcher is to estimate the model

$$O_r = \alpha + \beta S_r^* + \epsilon_r \quad r = 1, \ldots, R,$$

where O_r is some socioeconomic outcome of interest and the parameter of interest is β. As before, S_r^* is unobserved because g_r^* is unobserved and the researcher only has ARD, $\mathbf{y}_r$. The researcher instead estimates the expectation of the statistic given using ARD, $\bar{S}_r = E\{S_r \mid \hat{\theta}_{n,r}\}$. The regression becomes

$$O_r = \alpha + \beta \bar{S}_r + u_r. \quad [4]$$

and $\hat{\beta} = \hat{\beta}_{n,R}$ is the ordinary least squares (OLS) estimator of β from Eq. **4**. Critically, $\hat{\beta}$ depends on the size of each network n and the number of networks R.

In the second regression model we consider, the network feature is an outcome that responds to an intervention, T_r:

$$S_r^* = \alpha + \gamma T_r + \epsilon_r.$$

We let $\hat{\gamma}_{n,R}$ denote the OLS estimator of γ from the regression

$$\bar{S}_r = \alpha + \gamma T_r + \epsilon_r. \quad [5]$$

Theorem 5. *Let $\hat{\beta}_{n,R}$ denote the OLS estimate from Eq. **4** and let $\hat{\gamma}_{n,R}$ denote the OLS estimate from Eq. **5**. Suppose that*

1. *the estimators of the parameters for the rth network, denoted by $\hat{\theta}_r(n)$, satisfy*

$$\max_{1 \leq r \leq R} \frac{1}{n} \sum_{i=1}^{n} |\hat{\theta}_{i,n,r} - \theta_{i,n,r}^\star| \xrightarrow{p} 0 \text{ as } n, R \to \infty;$$

2. *the function $\theta_n \mapsto E\{S_r \mid \theta_{r,n}\}$ is differentiable for each network r and each network size n. Suppose also that*

$$\max_{1 \leq r \leq R} \max_j \sup_{\theta_{n,r}} \partial_j E\{S_r \mid \theta_{n,r}\} \leq \frac{C}{n},$$

for some finite constant $C > 0$.

If $E\{\epsilon_r \mid S_r^\star\} = 0$, $E\{S_r \mid \theta_{n,r}\} < \infty$ for any $\theta_{n,r}$, and the design matrix has full rank, then $|\hat{\beta}_{n,R} - \beta| \xrightarrow{p} 0$ and $|\hat{\gamma}_{n,R} - \gamma| \xrightarrow{p} 0$ as $n, R \to \infty$.

The following theorem shows that the three conditions from Theorem 5 hold.

Theorem 6. *Suppose that each network $g_{n,r}^\star$ is known to be drawn from the beta-model, stochastic block model, or latent space model and $\hat{\theta}_n$ is computed from Theorems 1, 2, and 3, respectively. If $S_r^\star$ is the density, centrality, or clustering of a node in network r, as defined in Corollary 2, then Conditions 1 and 2 of Theorem 5 hold if $Rn/\exp(n) \to 0$.*

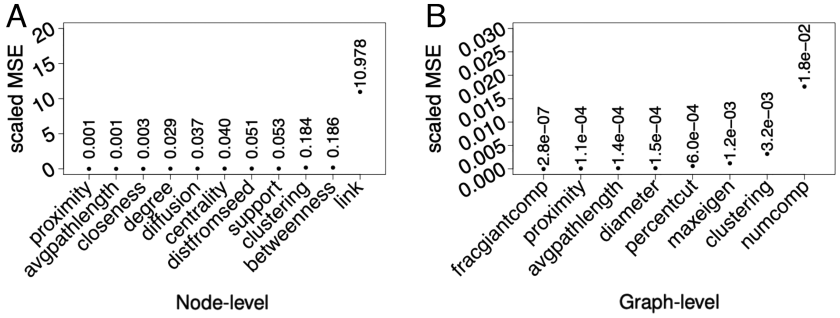

Fig. 1. Scaled mean squared error of node-level (panel *A*) and graph-level (panel *B*) network features. These results corroborate the theoretical intuition we developed. Specifically, we show in Corollary 1 that the mean squared error should be large for a single link and in Corollary 2 that the mean squared error should diminish for (normalized) degree and diffusion at the node level and clustering at the graph level.

In words, Theorem 6 states that a researcher is able to run the regression in Eq. **4** using the estimators in Theorems 1, 2, and 3 to consistently estimate β, the true effect of the network statistics on the observed socioeconomic outcomes.

Take the most extreme example of a single link, where we know that its presence cannot be identified in a single large network. Even if we were interested in a regression of $y_{12,r} = \alpha + \beta g_{12,r} + \epsilon_r$, where whether nodes 1 and 2 are linked affects some outcome variable of interest across all R networks, we can use $E\{g_{12,r} \mid \hat{\theta}(y_r)\}$ in the regression to consistently estimate β. Here, nodes 1 and 2 refer to arbitrarily labeled nodes and can be different across the R networks. In contrast to the single network case, where the mean squared error of the estimate of $g_{12,r}$ does not tend to zero as n grows, here, simply having the conditional expectation is enough to estimate the slope of interest, β. Therefore, with many graphs, the ARD procedure works well under weaker conditions on the network statistics. However, despite the generality of Theorem 5, Condition 2 of Theorem 5 still must hold. Some statistics are more sensitive to the input parameters and thus might not satisfy Condition 2. For example, the number of connected components has a higher mean squared error than the other statistics, which suggests that this statistic might lead to poor OLS estimators in Eqs. **4** and **5**.

Simulation Results

Single Large Graph. We explore the results for a single large graph through simulation exercises on 250 simulated graphs. Each network consists of 250 nodes, similar to the size of villages in ref. 13. We first generate traits for each node and then simulate connections independently using the generating process in Eq. **3**. We then draw a sample of nodes from the graph and construct ARD using traits. Our simulation does not reflect error in the ARD, which may arise if, for example, a person is a member of a group but the respondent does not have this information (e.g., refs. 15, 23, 39, or 40). We then estimate graph statistics using the procedure outlined in ref. 1, which uses a latent space model with positive curvature (8) for the underlying network.

Fig. 1 plots the mean squared errors of our estimation procedure across a range of common network statistics. These mean squared errors reflect uncertainty in estimation of the model parameters and in the underlying network statistics. In order to make the mean squared errors comparable across statistics, we scale by $1/E(S_i)^2$. Fig. 1*A* focuses on node-level statistics. We compile 11 node-level statistics: 1) proximity (average of inverse of shortest paths); 2) average path length; 3) closeness centrality (the average inverse distance from i over all other nodes); 4) degree (the number of links); 5) diffusion centrality (as defined in ref. 13—an actor's ability to diffuse information through all possible paths); 6) eigenvector centrality (the ith entry of the eigenvector corresponding to the maximal eigenvalue of the adjacency matrix for node i); 7) the average distance from a randomly chosen seed (as in a diffusion experiment where the seed has a new technology or piece of information); 8) support (as defined in ref. 41—whether linked nodes ij have some k as a link in common); 9) clustering (the share of a node's links that are themselves linked); 10) betweenness centrality (the share of shortest paths between all pairs j and k that pass through i); and 11) whether link ij exists. The results from the simulation, ordered in terms of scaled mean squared error in the figure, are consistent with the theoretical results. Statistics such as density and centrality take values for each realization that are nearly their expectation, meaning that we can recover the statistics with low mean squared error. For a single link, this is not the case and, correspondingly, the simulations show higher error.

Fig. 1*B* focuses on graph-level statistics. The graph-level statistics are as follows: 1) share of nodes in the giant component; 2) average proximity (average of inverse of shortest paths); 3) average path length; 4) diameter; 5) the share of links across the two groups relative to within the two groups where the cut is taken from the sign of the Fiedler eigenvector (this reflects latent homophily in the graph); 6) maximal eigenvalue; 7) clustering; and 8) number of components. All network statistics, with the exception of the number of components one, have small scaled mean squared error. This reflects the intuition of Corollary 2. ARD recover statistics that converge to their expectations, such as density, and might fail to recover statistics that do not.

We also evaluate our approach using observed, fully elicited graphs. We use data from ref. 13, which consists of completely observed graphs from 75 villages in rural India. In each village, about one-third of respondents were asked ARD questions. Ref. 1 compares statistics estimated with ARD (using estimated formation model parameters) from these graphs with the same statistics calculated using the complete graph. We leverage these results and present a different aspect: How the mean squared error changes as the size of the graph grows. We present results for individual-level statistics from these graphs and compute mean squared error across individuals. Our results using graphs with real-world complexity and properties (e.g., density and community structure) confirm the results from our simulation experiments. These results are presented in *SI Appendix*, Fig. S2.

Many Independent Networks. Multiple independent networks often arise in experiments, so we simulate a setting where we assign graph-level treatment randomly to half of the graphs. Graphs in the control group have expected degree generated from a normal distribution with mean 15 and variance 25, while graphs in the treatment group are generated from a normal density with mean 25 and variance 25. Each graph has 250 nodes. All graphs have a minimum expected degree of 5 and a maximum expected degree of 35. Due to the association between density and treatment, we expect treatment effects on graph-level statistics, such as average path length and diameter. The average sparsity over all graphs is 20/250 = 0.08, which is a value similar to Karnataka data discussed in ref. 13. For individual measures, 50 actors are randomly selected in each network. For links measured between actors, 1,000 pairs are randomly selected in each network. For network-level measures, there is one measure per network, so the regression consists of R data samples, where R is the number of networks.

Fig. 2 shows the simulation exercise with multiple independent networks. We use formation model parameters, $\theta^\star$ from the positive curvature latent space model (8), to get $\bar{S}_{ij,r}$ or $\bar{S}_r$ and include results using estimated model parameters in (*SI Appendix*, Figs. S3, S4, and S5), where we obtain estimates using the procedure in ref. 1. We present results with $R = 200$ ($R = 50, 100, 200$ are in *SI Appendix*). ϵ_r comes from a normal distribution with zero mean, and $\text{var}(\epsilon_r) = \text{var}(S^*_{ij,r})$ to maintain a 0.5 noise to signal ratio.

The first two panels in Fig. 2 show the distribution of the estimate of β in a regression where the network statistic predicts an outcome of interest. The middle line of each boxplot is the median $\hat{\beta}$, and the borders of boxes denote the first and third quartiles. All boxplots have outliers removed. The *Leftmost* panel gives results for individual-level measures, while the center panel gives network-level measures. Among the node-level statistics, we

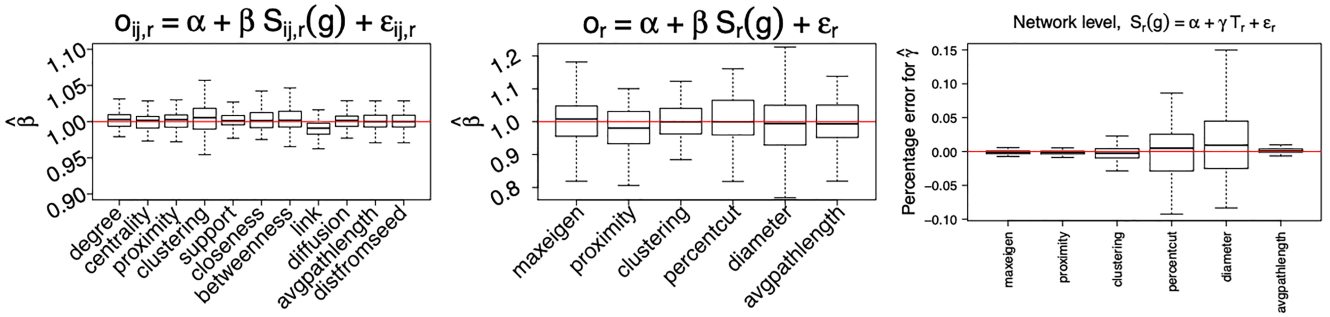

Fig. 2. Boxplots for the simulation experiments with multiple independent networks. In the *Left* figure, we consider a regression where the node-level network statistics determine outcomes on one network. In the *Middle* figure, we consider a regression where network-level statistics determines outcomes on multiple networks. In the *Right* figure, we consider a regression where a treatment determines a network-level statistics. On the *x*-axis, we provide the network statistics used, and the *y*-axis represents the value of the regression coefficients estimators. The red line indicates the true value of the regression coefficients. These results corroborate the theoretical intuition developed in Theorems 4 and 5.

see that all estimated $\hat{\beta}$s are close to the simulation value of one. The individual link measure, though empirically similar, is not centered around the true simulated value. The downward bias is an example of attenuation bias or regression dissolution since there is variability in the network statistic acting as the covariate. The indicator of the presence/absence of a single link is the most variable of the network measures, and thus, bias persists for the link measure when it does not for the others. For graph-level measures, all estimated coefficients are centered around the generated values.

The *Rightmost* panel in Fig. 2 shows results for the case where the network statistic is the outcome and is predicted by another covariate, in this case, treatment status. The percentage error is defined as $(\hat{\gamma}-\gamma)/\gamma$. Percent cut and diameter has large variation of percent errors than the other measures. This is due to the fact that the treatment effect, density differences between treatment and control, has a smaller effect on percent cut and diameter than on other measures. The average percent of variation explained by treatment in S_r for percent cut and diameter is around 0.3, while it is around 0.5 for other measures.

Discussion

Collecting full network data in large networks (e.g., a city) or across many networks (e.g., villages or schools) requires enumerating all egos and alters and therefore can be prohibitively expensive, logistically hard, or face privacy concerns. The use of ARD allows the researchers to overcome these problems by fitting frequently used and rich generative models, which can be then used to estimate socioeconomic quantities and parameters of interest. This can include features of the network but also responses in network structure to interventions as well as how socioeconomic outcomes are affected by network structure.

In this work, we first demonstrated that by using ARD, we are able to consistently estimate parameters in several families of frequently used generative network models, including ones where the number of parameters grows as the graph size grows. Second, we provided a taxonomy to describe when we may expect to estimate socioeconomic features consistently using ARD. Together, our theoretical results and supportive simulations using empirical data present insights into settings where researchers can count on ARD to reliably estimate socioeconomic quantities of interest. This makes the study of socioeconomic networks much more accessible to a wide set of researchers; in our own setting, using ARD delivers the same economic conclusions as the full network data does but at 80% less cost (1).

There are several promising avenues for future work. First, the techniques studied here are likely more relevant for networks of the scale of villages or cities/counties but certainly not necessarily things like large social media networks. It is true that when the number of nodes is very large, one needs many more traits K to exceed the number of latent communities C (since presumably a large C is needed to fit the network well). Note that geography can be included, to some degree, in a reasonably natural way. After all, one can imagine carving out a aset of locations (as set if L regions) and now a "type" K is the subtrait (e.g., caste) crossed with the location. So, $K = T \times L$, and we would use $K > C$ in this way. This is not the only approach, but we leave a complete exposition of this strategy to future work. Second, we demonstrate consistent estimation for edge-independent network models. To do this, we use a strategy based on the insight from Eq. **1**, which requires the fraction of connections between a person and a group, k, concentrates in the sense of a weak law of large numbers as the graph grows. If, for example, the edge probability depended on the graph size (e.g., in an asymptotically sparse model), then we would not have this property. Extending these results to a broader class of models, particularly those that are asymptotically sparse or which have correlated edges, would extend the reach of our work, and we believe that much of the infrastructure we developed around the necessary properties of network statistics would still apply (42–44). Third, a natural question to ask is whether other data collection strategies might be more useful to deliver consistent estimates for quantities that fall outside of the taxonomy of statistics that are estimable with ARD.

Data, Materials, and Software Availability. Previously published data were used for this work https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/U3BIHX.

ACKNOWLEDGMENTS. E.B. and A.G.C were supported by the Alfred P. Sloan Foundation. T.H.M. is supported by the National Institute of Mental Health of the NIH under Award Number DP2MH122405 and by the Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2CHD042828, to the Center for Studies in Demography & Ecology at the University of Washington. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. E.B., T.H.M., and A.G.C. are supported by NSF SES-2215369 from the National Science Foundation.

1. E. Breza, A. G. Chandrasekhar, T. H. McCormick, M. Pan, Using aggregated relational data to feasibly identify network structure without network data. *Am. Econ. Rev.* **110**, 2454–2484 (2020).
2. P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, E. C. Johnsen, Estimation of Seroprevalence, rape, and homelessness in the united states using a social network approach. *Eval. Rev.* **22**, 289–308 (1998).
3. O. Scutelniciuc, Network scale-up method experiences: Republic of Kazakhstan. *Consultation on estimating population sizes through household surveys: Successes and challenges* (New York, NY), 2012.
4. L. Jing, C. Qu, H. Yu, T. Wang, Y. Cui, Estimating the sizes of populations at high risk for HIV: A comparison study. *PloS One* **9**, e95601 (2014).
5. T. A. DiPrete, A. Gelman, T. McCormick, J. Teitler, T. Zheng, Segregation in social networks based on acquaintanceship and trust. *Am. J. Sociol.* **116**, 1234–1283 (2011).
6. D. M. Feehan, M. Mahy, M. J. Salganik, The network survival method for estimating adult mortality: Evidence from a survey experiment in Rwanda. *Demography* **54**, 1503–1528 (2017).
7. M. Leung, Two-step estimation of network-formation models with incomplete information. Working paper (2013).
8. T. H. McCormick, T. Zheng, Latent surface models for networks using aggregated relational data. *J. Am. Stat. Assoc.* **110**, 1684–1695 (2015).
9. V. Boucher, A. Houndetoungan, Estimating peer effects using partial network data. (Centre de recherche sur les risques les enjeux économiques et les politiques, 2020).
10. E. Sadler, Seeding a simple contagion. Available at SSRN 4032812 (2022).
11. S. Lubold, B. Liu, T. H. McCormick, Spectral goodness-of-fit tests for complete and partial network data (2021).
12. A. Chandrasekhar, *Econometrics of Network Formation. The Oxford Handbook Economic Networks* (2016), pp. 303–357.
13. A. Banerjee, A. Chandrasekhar, E. Duflo, M. Jackson, Diffusion of microfinance. *Science* **341**, 1–7 (2013).
14. T. H. McCormick, M. J. Salganik, T. Zheng, How many people do you know? Efficiently estimating personal network size. *J. Am. Stat. Assoc.* **105**, 59–70 (2010).
15. D. M. Feehan, A. Umubyeyi, M. Mahy, W. Hladik, M. J. Salganik, Quantity versus quality: A survey experiment to improve the network scale-up method. *Am. J. Epidemiol.* **183**, 747–757 (2016).
16. P. Orbanz, D. M. Roy, Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 437–461 (2015).
17. P. D. Hoff, A. E. Raftery, M. S. Handcock, Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* **97** (460), 1090–1098 (2002).
18. M. S. Handcock, A. E. Raftery, J. M. Tantrum, Model-based clustering for social networks. *J. R. Stat. Soc.: Ser. A (Stat. Soc.)* **170**, 301–354 (2007).
19. S. Lubold, A. Chandrasekhar, T. McCormick, Identifying latent space geometry of network formation models via analysis of curvature. *J R Stat. Soc. Series B Stat. Methodol.* **85**, 240–292 (2023).
20. S. Chatterjee, P. Diaconis, A. Sly, Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21**, 1400–1435 (2011).
21. B. S. Graham, An econometric model of network formation with degree heterogeneity. *Econometrica* **85**, 1033–1063 (2017).
22. T. H. McCormick, The network scale-up method. *Oxford Handb. Soc. Netw.* **153** (2020).
23. T. Zheng, M. J. Salganik, A. Gelman, How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *J. Am. Stat. Assoc.* **101**, 409–423 (2006).
24. D. Asta, C. R. Shalizi, "Geometric network comparisons" in *Proceedings of the 31st Annual Conference on Uncertainty in AI (UAI)* (2014).
25. C. R. Shalizi, D. Asta, Consistency of maximum likelihood for continuous-space network models. arXiv [Preprint] (2017). http://arxiv.org/abs/1711.02123 (Accessed 1 November 2022).
26. W. Killing, Ueber die Clifford-Klein'schen raumformen. *Math. Ann.* **39**, 257–278 (1891).
27. A. Chandrasekhar, R. Lewis, Econometrics of sampled networks. Stanford Working Paper (2016).
28. A. Banerjee, A. G. Chandrasekhar, E. Duflo, M. O. Jackson, Jackson, Using gossips to spread information: Theory and evidence from two randomized controlled trials. *Rev. Econ. Stud.* **86**, 2453–2490 (2019).
29. M. T. Hansen, M. L. Mors, B. Løvås, Knowledge sharing in organizations: Multiple networks, multiple phases. *Acad. Manag. J.* **48**, 776–793 (2005).
30. S. Heß, D. Jaimovich, M. Schündeln, Development projects and economic networks: Lessons from rural gambia. *Rev. Econ. Stud.* **88**, 1347–1384 (2021).
31. J. Cai, A. D. Janvry, E. Sadoulet, Social networks and the decision to insure. *Am. Econ. J.: Appl. Econ.* **7**, 81–108 (2015).
32. L. Beaman, A. BenYishay, J. Magruder, A. M. Mobarak, Can network theory-based targeting increase technology adoption? *Am. Econ. Rev.* **111**, 1918–1943 (2021).
33. V. Alatas, A. Banerjee, A. G. Chandrasekhar, R. Hanna, B. A. Olken, Network structure and the aggregation of information: Theory and evidence from Indonesia. *Am. Econ. Rev.* **106**, 1663–1704 (2016).
34. L. Beaman, A. Dillon, Diffusion of agricultural information within social networks: Evidence on gender inequalities from Mali. *J. Dev. Econ.* **133**, 147–161 (2018).
35. T. Gu *et al*., Formation mechanism of contributors' self-identity based on social identity in online knowledge communities. *Front. Psychol.* **13**, 1–14 (2022).
36. G. F. Chami, S. E. Ahnert, N. B. Kabatereine, E. M. Tukahebwa, Social network fragmentation and community health. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7425–E7431 (2017).
37. J. Cai, A. De Janvry, E. Sadoulet, Social networks and the decision to insure. *Am. Econ. J.: Appl. Econ.* **7**, 81–108 (2015).
38. E. Breza, *Field Experiments, Social Networks, and Development* (Oxford University Press, Oxford Handbook on the Economics of Networks, Oxford, 2016).
39. P. D. Killworth *et al*., Two interpretations of reports of knowledge of subpopulation sizes. *Soc. Netw.* **25**, 141–160 (2003).
40. S. Ezoe, T. Morooka, T. Noda, M. L. Sabin, S. Koike, Population size estimation of men who have sex with men through the network scale-up method in Japan. *PLoS One* **7**, e31184 (2012).
41. M. O. Jackson, T. R. Rodriguez-Barraquer, X. Tan, Social capital and social quilts: Network patterns of favor exchange. *Am. Econ. Rev.* **102**, 1857–1897 (2012).
42. T. P. Peixoto, Disentangling homophily, community structure, and triadic closure in networks. *Phys. Rev. X* **12**, 011004-1–011004-23 (2022).
43. S. Wasserman, P. Pattison, Logit models and logistic regressions for social networks: I. an introduction to Markov graphs and*p*. *Psychometrika* **61**, 401–425 (1996).
44. T. Snijders, Markov chain Monte Carlo estimation of exponential random graph models. *J. Soc. Struct.* **3**, 240 (2002).