Original Article

Sociological Methodology
1–61

DOI: 10.1177/0081175020922879
http://sm.sagepub.com

ESTIMATING CONTEXTUAL EFFECTS FROM EGO NETWORK DATA

Jeffrey A. Smith*
G. Robin Gauthier*

Abstract

Network concepts are often used to characterize the features of a social context. For example, past work has asked if individuals in more socially cohesive neighborhoods have better mental health outcomes. Despite the ubiquity of use, it is relatively rare for contextual studies to use the methods of network analysis. This is the case, in part, because network data are difficult to collect, requiring information on all ties between all actors. In this article the authors ask whether it is possible to avoid such heavy data collection while still retaining the best features of a contextual-network study. The basic idea is to apply network sampling to the problem of contextual models, in which one uses sampled ego network data to infer the network features of each context and then uses the inferred network features as second-level predictors in a hierarchical linear model. The authors test the validity of this idea in the case of network cohesion. Using two complete data sets as a test, the authors find that ego network data are sufficient to capture the relationship between cohesion and important outcomes, such as attachment and deviance. The hope, going forward, is that researchers will find it easier to incorporate holistic network measures into traditional regression models.

*University of Nebraska–Lincoln, Lincoln, NE, USA

Corresponding Author:
Jeffrey A. Smith, University of Nebraska–Lincoln, Department of Sociology, 711 Oldfather Hall, Lincoln, NE 68588-0324, USA.
Email: jsmith@unl.edu

Keywords

ego networks, network sampling, hierarchical linear models, cohesion, adolescents, exponential random graph models

1. INTRODUCTION

How are individuals affected by their social environment? This is a core question in the discipline, motivating hundreds of studies across a diverse set of subfields, from deviance in schools to mental health in neighborhoods (Berkman et al. 2000; Sharkey and Faber 2014; Wray, Colen, and Pescosolido 2011). Many studies in this tradition use network concepts to characterize the features of the social context. For example, past work has asked if individuals in more cohesive neighborhoods have better mental health outcomes, with fewer suicide attempts and lower chances of depression (Ivory et al. 2011; Maimon and Kuhl 2008). Despite the common use of network ideas, contextual studies rarely use the methods and measures of network analysis (even with recent calls to do so; see Entwisle 2007; Pescosolido 2006).

Contextual-network approaches are uncommon, in part, because network data are difficult to collect (Krivitsky and Morris 2017; McCormick and Zheng 2015). Network data offer a rich means of measuring contextual features, but this comes at a cost, as there are practical challenges to collecting network data in multiple contexts. Traditionally, network measures require census information on the population of interest. This means having information on all the actors and all the ties between actors. In an ideal case, a researcher would interview every actor in every context, using the network data (e.g., friendships between students) to measure the global network features of each school, town, and so on in the study (Entwisle et al. 2007; McFarland et al. 2014). Such data can be difficult to collect, however, especially when there is a large number of contexts or a large number of actors in each context (Hipp and Perrin 2006; Smith 2015).

A more typical approach is to sample. For example, a researcher may interview a subset of students from each school in the study. This eases the data collection burden but makes it difficult to measure global network properties, which are based on the interdependences between actors (Smith and Burow 2018). It is telling that most sample-based studies use attitudinal data (e.g., "Do you feel this is a safe neighborhood?"), rather than network data, to infer the social features of a context (Hipp and Perrin 2006).

In this article we ask a practical, important question: is it possible to accurately estimate a contextual model using network features (such as density or transitivity) while avoiding the data burden of traditional network studies? The basic idea is to apply network sampling techniques (Handcock and Gile 2010; Smith 2012) to the estimation of typical regression models, where outcomes such as health or mental health serve as the dependent variable (McPherson and Smith 2019). Network sampling offers a bridge between survey methods and network analysis, as a researcher uses sampled data to infer the global features of a network (Frank 1978; Krivitsky, Handcock, and Morris 2011; Smith 2015). A researcher could thus collect sampled data across contexts, use the sampled data to infer the network features of each context, and then use the inferred network features as second-level predictors in a hierarchical linear model (HLM) (Raudenbush and Bryk 2002). Here, we assume that the data are collected via simple random sampling without replacement.

The potential payoff from network sampling is large, as contextual-network studies would be much easier to undertake. A researcher could avoid collecting census information in every context, yet still use a network measure as a contextual-level predictor. The reduced data burden would also make it easier for contextual-network studies to move beyond institutionally bounded populations. It is an open question, however, whether sampled network data can be used to accurately estimate a contextual model. Past work has focused on the network features themselves (e.g., can we use sampled network data to infer network distance?), leaving the more difficult problem of HLM estimation unexplored (Frank 1978; Granovetter 1976; Smith 2012).[1] For example, Smith (2015) mentioned HLMs as the ideal application for network sampling but did not actually test this idea, suggesting that future research should take up the problem.

Here, we directly apply network sampling techniques to the problem of contextual models. We ask if HLM results based on sampled network data can mimic the results using complete network data. Substantively, we focus on the effect of network cohesion (which we define in Section 4, "Analytic Setup: Testing the Method") on individual-level outcomes, a common topic for contextual studies (e.g., Gottfredson and DiPietro 2010; Legewie and DiPrete 2012). We offer two tests. The first case uses empirical data on adolescents in schools (using National Longitudinal Study of Adolescent to Adult Health [Add Health] data).

We model two individual-level outcomes, school attachment and behavioral problems in school, as a function of network cohesion, measured as a school-level network feature. Do students in more cohesive schools have better outcomes (higher attachment, fewer behavioral problems), net of individual-level characteristics? And can this be answered just using sampled ego network data? The second case uses simulated data as the basis for the test, in which the relationship between network cohesion and the outcomes of interest are known from the start.

The larger hope is that any study with multiple contexts will be able to incorporate a holistic network measure into the analysis, thus connecting the strengths of a network approach (a rich depiction of the social features of a context) with the strengths of traditional survey methods (wide coverage of the population and ease of data collection). We begin with a discussion of network sampling. We then discuss the analytic test of the approach, before presenting two sets of results.

2. BACKGROUND ON NETWORK SAMPLING

Network sampling has a long tradition, stemming from the important early work of Frank (1971, 1978) and Granovetter (1976). The goal is to solve a fundamental problem in network studies: how can a researcher take sampled network data and still undertake a proper analysis of the network structure, which is based on the pattern of relations between all actors? Network samples can take many forms. Most of the early work on network sampling used subgraph samples, where all of the ties between a randomly selected subset of actors are recorded. Unfortunately, subgraph samples on large networks tend to yield little information (as few ties between sampled actors are recorded), so recent work has used alternative sampling schemes, such as ego network sampling, scale-up methods, or snowball sampling designs (e.g., Burt 1984; Feehan and Salganik 2016; Thompson and Frank 2000).

With a snowball sample, for example, a set of initial seeds enumerate their social contacts, or alters, who are brought into the study; these recruits then enumerate their alters who are also brought into the study, and the process is repeated until a sufficient sample is collected (for debates about the use of snowball sampling, see Handcock and Gile 2011). Past work on snowball sampling generally falls into two traditions, one using formula-based estimators and the second using simulation-based approaches to inference (for examples of formula-

based approaches, see Illenberger and Flötteröd 2012; Verdery et al. 2017). Most simulation approaches draw on the exponential random graph model (ERGM) framework. The basic idea is to estimate a model predicting the presence or absence of a tie on the basis of the snowball sample (Handcock and Gile 2010; Pattison et al. 2013; Stivala et al. 2016). The estimated parameters of this model are then used to simulate complete networks, thus inferring the rest of the network (Koskinen, Robins, and Pattison 2010; Rolls and Robins 2017).

Snowball sampling is useful in many settings, but it can be difficult to implement, making it less than ideal for the problem at hand: estimating contextual network models from sampled data. Snowball sampling designs require that a researcher identify, find, and interview the people named by the respondent as a social connection (alternatively, respondents can be tasked with bringing in their associates for the study). This is a difficult undertaking in a single setting, let alone across many settings, which would be necessary in a study of contextual network effects. Additionally, a snowball sampling design is not easily incorporated into existing surveys.

Ego network data, in contrast, are a natural fit for a contextual, network sampling approach and serve as the focus of this article. Ego network data are based on a random sample of independent cases, where each respondent answers questions about their personal, or local, network. The data collection burden is comparatively low (Marsden 2011). All information comes from the respondents themselves, as the named network alters are not identified and not interviewed. Ego network data are thus easy to collect and easily incorporated into a multicontext study. A researcher would sample within each context of interest, interviewing a subset of people in each school, neighborhood, and so on. Importantly, ego network data are easily added to existing surveys, meaning a researcher could turn traditional contextual studies (e.g., on schools) into a network study by adding a few questions to the base survey.

Figure 1 plots three example ego networks. Ego network surveys begin by asking respondents to list their alters, or the individuals they are socially connected to (Burt 1984; Marsden 1990; Smith, McPherson, and Smith-Lovin 2014). A study may ask about friendship, social support, and so on. This offers information on the degree (i.e., number of ties) for each actor. In Figure 1, the first respondent names three friends, the second names two, and the third names zero. In a contextual study, a researcher may have respondents list all their alters; respondents would

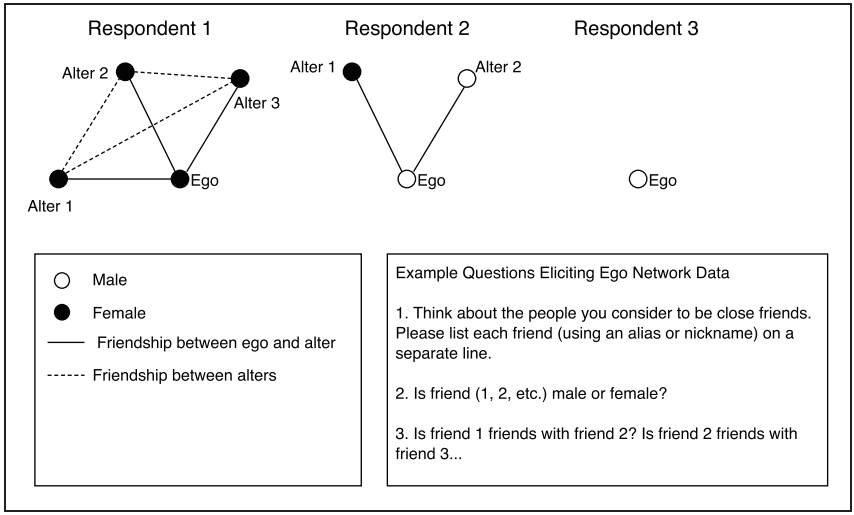

Figure 1. Ego network data from three hypothetical respondents.

then be asked which alters reside in the context (e.g., school, neighborhood) of interest.

Ego network data will also include demographic information about the respondents and named alters. A researcher may ask respondents about their age, education, and gender and also ask them to report on the age, education, and gender of each alter (Burt 1984). Figure 1 demonstrates this in the case of gender. The first respondent is female and reports that all three of the named alters are female, suggesting something about the level of homophily in the network (the tendency for similar actors to interact at high rates).

Finally, ego network surveys often ask respondents to report on ties between alters. For example, the first respondent in Figure 1 reports that alters 1 and 2 are friends, alters 1 and 3 are friends, and alters 2 and 3 are friends. The alter-alter ties capture the tendency for friends to also be friends with one another, or transitive closure. Note that no identifying information on alters is collected. This means the alter-alter tie information cannot be used to map particular ties (or edges) in the network.

The vast majority of inferential work on ego network sampling has been at the local level, measuring items like mean degree (how many alters were named on average?) or homophily (how similar are ego and the named alters in terms of gender, race, and so on?) (e.g., McPherson,

Smith-Lovin, and Brashears 2006). Other, more global, features of a network are difficult to capture from ego network data, as there is no way of connecting one sampled ego to another (McPherson and Smith 2019) (i.e., this is more than a simple missing data problem [Smith, Moody, and Morgan 2017]).

Past work has consequently used simulation as an analytic approach, in a similar way to the literature on snowball sampling (Morris and Kretzschmar 2000; Morris et al. 2009). The basic idea is to take information from the sampled data, such as number of named partners or similarity between ego and alter, and generate full networks that are consistent with that local information (Krivitsky et al. 2011; Smith 2012). For example, Krivitsky and Morris (2017) simulated complete sexual networks from ego network data on the basis of degree, homophily on race (and sex), and degree differences by race and sex. As is typical with simulations based on ego network data, the model does not include any term for transitive closure (i.e., did their sexual partners share other partners?).

Smith (2012) offered a related but distinct approach to simulating full networks from ego network data. His approach fully incorporated alter-alter tie data into the simulation, thus conditioning the generated networks on the pattern of transitive closure found in the ego network data. This is an important development in network inference, as models that do not incorporate transitive closure tend to yield unrealistic networks (i.e., excluding sexual networks, in which closure is relatively weak), with features that do not approximate the true network well. Empirically, past work shows that the approach offers better estimates than models based just on degree and homophily (Smith 2012).

Gjoka, Smith, and Butts (2014) offered a different approach to a similar problem, deriving unbiased estimators for clique size on the basis of independently sampled ego network data (see also Anderson, Butts, and Carley [1999], who discuss estimation of density from ego network data). These formula-based approaches are simple to implement but limited to a narrow set of statistics for which formulas are possible to derive, as opposed to a simulation approach, in which any statistic can be calculated on the generated networks.

In this article, we draw on Smith's (2012, 2015) simulation approach to infer network properties from sampled data across multiple contexts, which are then used as predictors in an HLM. We also consider formula-based estimators where appropriate. We offer a short

background section on the simulation approach before describing the test case of interest.

3. BACKGROUND ON INFERENTIAL APPROACH

The simulation approach described by Smith (2012) has three main steps: first, summarize the key features of the sampled ego networks; second, simulate a full network consistent with the sampled information using an ERGM; and third, map out the path structure of the generated network, using that to measure network properties of interest. We offer a brief overview of the approach here, but see Smith (2015) for a more detailed discussion (and see Appendix B for a discussion of scope conditions). We begin with a short background on ERGMs.

3.1. *ERGMs*

ERGMs are statistical models used to test hypotheses about network structure/formation (Hunter et al. 2008; Wasserman and Pattison 1996). Define $\boldsymbol{y}$ as the observed graph and $\boldsymbol{Y}$ as a random graph on N, where each possible tie, ij, is a random variable. The exponential random graph models the $Pr(\boldsymbol{Y} = \boldsymbol{y})$. The "independent variables" are counts of local structural features in the network (Robins et al. 2007), such as number of ties. We can write the model as

$$P(\boldsymbol{Y}=\boldsymbol{y}) = \frac{exp\left(\theta^T g(\boldsymbol{y})\right)}{\kappa(\theta)}, \tag{1}$$

where $g(\boldsymbol{y})$ is a vector of network statistics, θ is a vector of parameters, and $\kappa(\theta)$ is a normalizing constant.

ERGMs are typically used to test hypotheses about network features, but it is also possible to generate, or simulate, networks from a specified model (e.g., Morris et al. 2009; Robins, Pattison, and Woolcock 2005). The coefficients reflect the effect of different local processes on the probability of a tie existing; those coefficients can then be used to predict the presence/absence of a tie between actors in a synthetic network. This is how we will use ERGMs here.

3.2. *Key Steps in Simulating Networks using ERGMs*

The method begins by summarizing the information available from the sampled ego networks. It is crucial to extract as much information as

possible, because the networks are generated based on the sampled data. The degree distribution is estimated first. This is taken directly from the sampled data: the proportion in the sample with 0, 1, 2, 3, and so on alters is used as the estimate of the degree distribution. We then summarize the information from the alter-alter tie data, showing the tendency toward transitive closure. Smith (2012) proposed a novel characterization of the alter-alter tie data, whereby the data are used to construct a distribution of ego network configurations. Each respondent is characterized as a distinct structural type, on the basis of the size of the ego network and the pattern of ties between alters. See Appendix Figure A1 for an example ego network configuration distribution.

The next main step is to set up the simulation itself. A network of size N (based on the population of interest) is first seeded with the correct degree distribution, estimated from the sample. Each node in the generated network is then seeded with demographic characteristics from the sampled data. Nodes in the generated network are matched with sampled respondents with the same degree; each selected node is assigned all of the demographic characteristics of that respondent, such as gender, age, and education. Such a matching process ensures that demographic groups with higher degree in the ego network data will also have higher degree in the generated network, thus constraining the network on the basis of differential degree.

The method then estimates the initial ERGM coefficients. The terms in the ERGM specify which local features are used to construct the full network. The terms in the model will reflect all the information available from the ego network data, specifically in terms of homophily and the ego network configuration distribution (differential degree and the degree distribution are already used to construct the base network). The model will include homophily terms for every demographic variable available in the ego network survey. For continuous variables, such as age and education, this takes the form of an absolute difference coefficient, reflecting the absolute difference between respondents and their alters on that dimension. For categorical variables, such as race and gender, the model includes a mixing matrix showing the number of ties going between different demographic groups. These coefficients can be estimated within the ERGM framework or using case control logistic regression (see Smith et al. 2014).

The model will also include a term for the geometrically weighted edgewise shared partner (GWESP) distribution (Snijders et al. 2006).

GWESP captures the tendency for actors (who are themselves socially connected) to have shared partners, or common associates (i.e., if *i* and *j* are friends and both are friends with *k* and *l*, that edge would have two shared partners). The GWESP statistic is a geometrically weighted mean of the shared partner distribution and is based on two items: the actual distribution of shared partners and a scaling parameter. The scaling parameter captures the relative weight put on the second, third, fourth, and so on, shared partner when performing the summation. When the scaling parameter is zero, for example, the term puts all the weight on the initial shared partner and simply sums up the number of edges for which there is at least one shared partner in common.

Substantively, GWESP captures higher order transitivity in the network (i.e., given that actors *i* and *j* are friends, do actors *i* and *j* have many friends in common?), or the tendency for small groups to emerge. Within the simulation, GWESP is included to capture the ego network configuration distribution. Because it parameterizes ties between alters, GWESP mirrors many of the structural features of ego networks, making it an appropriate choice to reproduce them. For example, within an ego network, ego's friends may be friends with one another; this is analogous to seeing many shared partners in the network. Unlike with the homophily coefficient, the coefficient for GWESP cannot be easily set prior to the simulation (it is not possible to solve for the value on GWESP that will yield a network with the correct ego network configuration distribution). The GWESP coefficient is thus set at an initial value and is updated during the simulation as the method searches for the best fitting network. Note that the degree distribution and differential degree of the initially seeded network are maintained throughout the simulation.

The framework then takes the initial model (coefficients, terms, and constraints) and simulates a network. The homophily rates in the simulated network are then compared with the empirical data to ensure that they match. The homophily coefficients are updated if this is not the case. The basic idea is to compare the level of homophily in the generated network with that observed in the ego network data (along each dimension available in the ego network data) and update the model accordingly. Coefficients are increased if there are too few ties between groups and decreased if there are too many. See Smith (2012) for technical details on how to adjust the homophily coefficients.

The next step is to evaluate the simulated network by how well it matches the ego network configuration distribution seen in the sample. Ego networks are first drawn from the simulated network and placed into a structural type (see Appendix Figure A1). The distribution of ego network configurations from the simulated network is then compared with the distribution in the sample using a chi-square value. The chi-square value is written as

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{2}$$

where O_i is the observed frequency in the simulated network, E_i is the frequency found in the sample, and n is the total number of possible ego network configurations. A large chi-square value indicates a poor fit, as the distribution in the simulated network deviates from the distribution in the sample.

The model is then updated to find a better fitting network, defined as a network that better matches the true ego network configuration distribution (conditioned on the other features found in the ego network data). The ego network configuration distribution is used as a benchmark to judge the simulated networks. The question then becomes, what coefficient will yield a network with the lowest chi-square value? The framework finds the best fitting network by (1) simulating networks with different values for GWESP, (2) calculating the chi-square value for each network, and (3) searching for a better fitting GWESP coefficient, given the results. This process is repeated until no better fitting network can be found. See Smith (2012, 2015) for more technical details.

The end result of the search process is a network with the same properties as the sampled ego network data. The simulated network will have the same degree distribution, differential degree, homophily, and ego network configuration distribution as in the sample. A researcher can then take the generated network and calculate statistics of interest; in this case to be used as predictors in an HLM.

4. ANALYTIC SETUP: TESTING THE METHOD

The main question of this article is whether it is possible to use sampled network data to estimate traditional HLMs: can a researcher test contextual theories using sample-based estimates of network features? The

empirical test of a network sampling approach focuses on network cohesion, although we do consider other contextual-network measures in the Appendix (i.e., average betweenness, average closeness, transitivity, and proportion isolated). We focus on network cohesion as an empirical example because there is a long tradition in the social sciences of using cohesion as a contextual-level predictor (e.g., Browning and Cagney 2002; Sampson, Raudenbaush, and Earls 1997). Social cohesion is typically understood as a multidimensional concept, with emotional and relational components that are thought to mutually reinforce each other, leading to lower rates of psychological distress, deviant behavior, and the like (Friedkin 2004; Moody and White 2003). Here, we focus on the relational component of cohesion, which can be parameterized naturally using network methods (Bearman 1991).

We use two global measures of network cohesion, density and bicomponent size (defined below). Both measures have often been used by past work as network measures of cohesion (e.g., Moody 2004; Tulin, Pollet, and Lehmann-Willenbrock 2018). Importantly, both are defined at the contextual level (or network level), reflecting a feature of the context itself. The hope is that a researcher could collect sampled data across a number of contexts, use the sampled data to infer density or bicomponent size in each context, and then use the inferred network measures as predictors in an HLM, thus greatly reducing the data collection burden while still retaining the best feature of a network approach.

We test the validity of a sampling approach by comparing the true HLM coefficients with the same coefficients estimated from the sampled ego network data. The baseline estimates serve as the gold standard: the researcher knows the true values of density and bicomponent size for each context and can use those values within the HLM. To make the comparisons equivalent, both the true models and the ego network models are estimated on the basis of the same samples. The only difference is that in the true models, the values for density and bicomponent size are known, whereas in the ego network models, density and bicomponent size must be inferred from the sample itself. This makes it possible to isolate the bias in the estimates (i.e., only bias that arises from using an inferred measure of cohesion).

We present two tests of the approach, one based on an empirical data set and one based on a simulated data set. The empirical data set has the advantage of representing actual conditions (in terms of networks and outcomes), and the simulated case has the advantage of being full

controlled, with known parameters. We focus primarily on the empirical example, describing the test and results in full before moving to the simulated case.

There are seven steps to testing a network sampling approach: (1) select a test data set, (2) specify an HLM to be used in the test, (3) take samples from the complete data set, (4) estimate the true HLM using known values of density and bicomponent size, (5) estimate the values for density and bicomponent size from the sampled ego network data, (6) estimate an HLM using the inferred values of density and bicomponent size (from step 5), and (7) compare the true HLM coefficients with the ego network–based HLM coefficients.

4.1. *Select Test Case*

In step 1, we select a data set for the analytic test. We use Add Health for the first case study because it has a nested structure (students within schools), complete network information for each context, and appropriate outcomes of interest. Add Health is a nationally representative survey of middle schools and high schools, both public and private (Harris et al. 2009). We use the wave I, in-school data. Saturating each school in the survey, all students in the school (or as many as possible) were asked a series of questions about health, behavior, and social connections. The network information is based on friendship ties. Students were asked to list up to five male and five female friends, and these nominations are used to construct the complete, known network for each school. The complete networks are used to calculate the true values of density and bicomponent size. Each network is assumed to be symmetric. This symmetrizing is done using a "weak" rule, whereby a tie exists if either *i* or *j* nominates each other. We restrict the analysis to schools with more than 400 students, which corresponds to the 80 largest schools in the data set.[2]

4.2. *Specify HLMs*

In step 2, we specify the HLMs of interest, used to test a network sampling approach. This entails specifying the dependent variables, the main predictors at the contextual level, and the control variables at the individual level. Students are nested within schools, with schools serving as the contextual (or second) level in the models.

4.2.1. *Dependent Variables.* We conduct two tests of the approach, using two different dependent variables. The first outcome of interest is attachment to school, an individual-level variable capturing the perceived attachment of each student to his or her school. Attachment is a scale variable, constructed from three different questions. Students responded to the following prompts: "I feel close to people at this school," "I feel like I am part of this school," and "I am happy to be at this school." For each question, students could respond on a scale ranging 1 to 5, with 1 = "strongly agree" and 5 = "strongly disagree." We take the mean over the three answers to form the attachment scale. This variable is then reverse-coded, so higher values correspond to feeling more attached to the school.

The second dependent variable captures behavioral problems. We use three ordinal variables to construct a scale, again, based on the means over the three variables. The three questions of interest are "Since school started this year, how often have you had trouble" (1) "getting along with your teachers?" (2) "paying attention in school?" and (3) "getting your homework done?" Responses range from 0 to 4, with 0 meaning "never" and 4 meaning "every day." Values are coded such that higher values equate to having more trouble.

4.2.2. *Contextual-Level Predictors.* The key predictor of interest in the HLM is network cohesion, defined as a contextual-level variable. The two measures are density and bicomponent size. Density is defined as the number of ties (e.g., friendship, social support) relative to the number of possible ties in a network of that size (Wasserman and Faust 1994). Density can be written as

$$Dens = \frac{\sum\sum \boldsymbol{y}}{N*(N-1)}, \tag{3}$$

where $\boldsymbol{y}$ is the observed network and N is the total number of people in the network. As a measure of cohesion, lower density suggests lower cohesion in the network. It is important to remember, however, that the number of possible ties increases nonlinearly as network size increases; this has important implications for understanding density in large networks (Anderson et al. 1999; Mayhew and Levinger 1976).

Bicomponent size is defined as the largest set of actors connected by at least two independent paths, so that removing a single actor leaves the entire set connected (Moody and White 2003). The basic idea is that

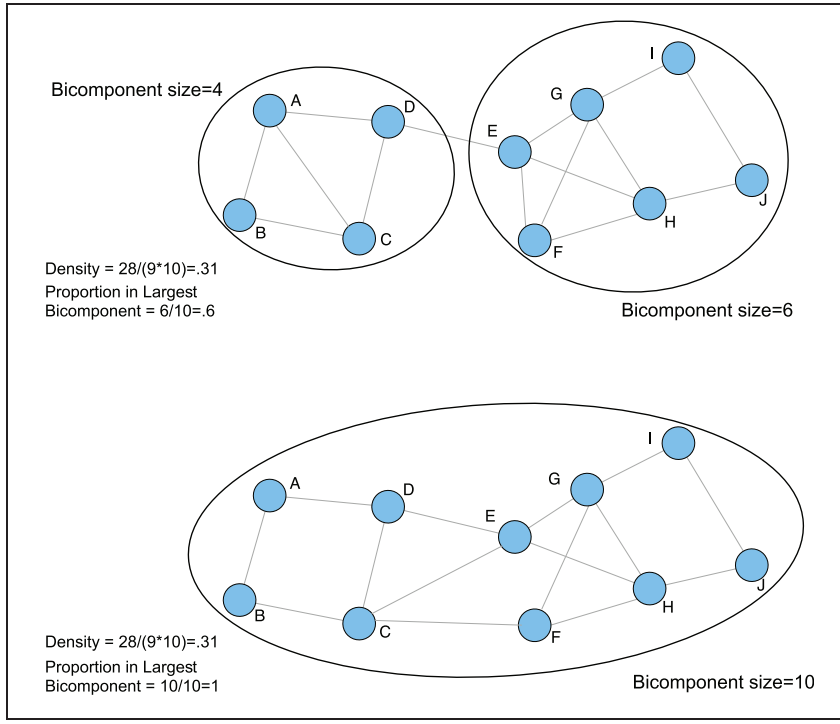

Figure 2. Measuring cohesion using bicomponent size.

a network that is difficult to tear apart is socially cohesive, where higher values for bicomponent size suggest a more cohesive network. Figure 2 explores some of the key properties of bicomponent size. The figure consists of two networks with the same density but different proportions in the largest bicomponent (for a more detailed discussion, see Moody and White 2003). It is clear that the network in the top half effectively splits into two groups and is disconnected by removing a single actor. The bottom half tells a different story. The network has the same density but the pattern of ties is different: our two loosely connected groups are now much more integrated. This would be missed by looking at density alone but is captured by bicomponent size, which is based on the number (and type) of paths connecting the actors. The proportion in the largest bicomponent ranges from .6 to 1, yet density is the same in the two plots.

Density and bicomponent size (defined as the proportion of people in the largest bicomponent) are measured for each school in the data set. We run separate models, first with density as the main predictor and then bicomponent size as the main predictor.

4.2.3. *Control Variables.* The model includes a number of individual-level controls. We include control variables to see if the estimates for the cohesion coefficient are robust to different model specifications. We include variables for gender (male or female), race (black, Asian, Hispanic, white, or other), and social isolation (1 if the person has one or no social ties and 0 otherwise).

4.2.4. *Models.* We run four models. In each case, we predict attachment or behavioral problems as a function of contextual-level network cohesion and a set of control variables. The first model includes only the measure of cohesion, either density or bicomponent size. The second model adds demographic controls to the first model, and the third model adds social isolation to the first model. The final model includes all predictors—cohesion, race, gender, and social isolation:

$$Y_{ij} = b0_j + b1(Asian) + b2(Black) + b3(Hispanic) + b4(Other) + b5(Female) + b6(Isolated) + \varepsilon_{ij},$$

$$b0_j = a00 + a01\left(Cohesion_j\right) + u_{0j},$$

where $a01$ is the coefficient of interest, capturing the fixed effects interaction of cohesion on the intercept (allowed to vary across schools). These four models are estimated for each dependent variable (attachment and behavioral problems) and measure of cohesion (density and bicomponent size).

4.3. *Sampling Setup*

In step 3, we take random samples from the complete Add Health data set. We assume that an independent sample is drawn from each of the 80 Add Health schools in the analysis. We begin by taking a hypothetical survey for each school. It is hypothetical in the sense that we are not actually interviewing any respondents; all information comes from the data. We are simply mimicking, as closely as possible, what a survey on this population would look like. We assume that the survey includes an

ego network component, as well as more general questions about attachment, behavioral problems, and demographics.

For the ego network portion, we assume that the following information is collected: number of alters (with no cap on number reported); respondent and alter characteristics, including race, grade, sex, and club membership (e.g., sports teams or band); and reports on ties between alters. To mimic a realistic survey, we assume that respondents are only asked to report on the first five alters named (in terms of alter characteristics and alter-alter ties).[3] Because this is not an actual survey, the five alters are randomly selected from the full set of alters for each respondent (this is only necessary for respondents with more than five friends).

The analysis is repeated for three different sample sizes: 15 percent, 25 percent, and 35 percent. The sample size refers to the percentage of each school sampled. Note that a 15 percent sample need not yield an absolutely large number of respondents if the school is small. This process is repeated 100 times for each sample size. We take 100 samples to capture the variability in estimates.

4.4. *Estimate HLMs using True Values for Density and Bicomponent Size*

In step 4, we estimate the true, baseline HLMs. The models specified in step 2 are estimated using the samples from step 3 and the true values of density and bicomponent size. The true models thus represent the results one should get with the sample in question. The true values for density and bicomponent size are calculated using the complete Add Health network data for each school (i.e., not the sampled data). There will be sample-to-sample variation in cases in the regression, but the contextual-level variables are assumed to be known and fixed and thus do not vary sample to sample. We estimate separate models predicting attachment and behavioral problems for each sample.

4.5. *Using Sampled Ego Network Data to Estimate Density and Bicomponent Size*

In step 5, we use the sampled ego network data (from step 3) to infer density and bicomponent size for each school/sample. Density is based solely on the volume of ties in the network, which makes the inferential process

much easier than with bicomponent size. Although a researcher cannot simply apply equation (3) (because all the information from the full matrix $\boldsymbol{y}$ is effectively missing), it is possible to directly use sampled ego network data to estimate global network density, without recourse to any complicated computational methods (see Anderson et al. 1999; Marsden 1990). The basic idea is to start with two inputs: first, the list of alters for each respondent and, second, the total size of the network, defined as N (assumed to be known). The first step is to calculate average degree for the sampled respondents, defined as the mean number of alters listed. The second step is to divide that by $(N-1)$, yielding an estimate for density. Formally, define sample average degree as $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, where n is the number of sampled respondents, and d_i is the degree of person i, equal to the number of alters listed. The total volume of ties in the network can then be written as $v = \bar{d} * N$. The sample-based density is thus estimated by

$$\frac{v}{(N*(N-1))} = \frac{\bar{d}}{(N-1)}. \tag{4}$$

Bicomponent size is more difficult to estimate from sampled data, because it captures the pattern of ties between actors more directly (see Figure 2). Ego network data are local, offering pieces of the network that cannot be connected. Ego network data thus cannot be used to directly map out the path structure in the network, yet this is exactly what the measure of bicomponent size is based on. Here we must rely on the full simulation-inferential approach of Smith (2012, 2015). A researcher would take the sampled data in each context, simulate a network consistent with the sampled information, and measure bicomponent size on the generated networks.

4.6. *Estimate HLMs using Ego Network–Based Estimates of Density and Bicomponent Size*

In step 6, we estimate HLMs using the ego network–based estimates of cohesion. The analysis is exactly the same as in step 4, except the values for density and bicomponent size are inferred from the sampled ego network data (estimated in step 5). Thus, the data used to estimate the HLMs (for each sample) are exactly the same as in the true models, with only the estimates for density or bicomponent size varying.

4.7. *Compare True Models with Ego Network–Based Models*

In step 7, we compare coefficients from the ego network models (step 6) with coefficients from the true models (step 4). In both cases, there are 100 × 2 × 2 × 4 × 3 estimates, as there are 100 samples, two dependent variables, two measures of cohesion, four models, and three sample rates. The question is how closely the ego network–based coefficients approximate the true coefficients.

5. RESULTS

5.1. *Comparing True Network Values with Statistics Inferred from Ego Network Data*

We begin the results section by comparing the true values for density and bicomponent size with the estimated values inferred from the ego network samples. The true values are calculated from the complete network data for each school. We begin with network statistics before moving to HLMs (in the next section), as HLMs are dependent on the inferred measures of bicomponent size and density. It is important to check the estimation of bicomponent size and density before moving to the harder test, where those estimates are used as predictors in a regression.

Figure 3 presents the results. The top row shows results for density, and the bottom row shows results for bicomponent size. Each subplot captures a different sampling rate, running from 15 percent to 35 percent. Results are presented as a series of boxplots, with one boxplot for each network in the study (within a subplot). The boxplots capture the distribution of the estimated values for density or bicomponent size for that school and sample rate. The boxplots are aligned so that each school is placed on the *x*-axis at the corresponding true value for density or bicomponent size. We also added a reference line, showing where the boxplots should be centered if the estimator is unbiased.

Looking at the top row, the results suggest that density can be effectively estimated using sampled network data. The boxplots are centered at the true value for density in every case. The variability of the estimates increases somewhat in more dense schools (as the range of degree is higher, leading to more variability sample to sample) but is overall

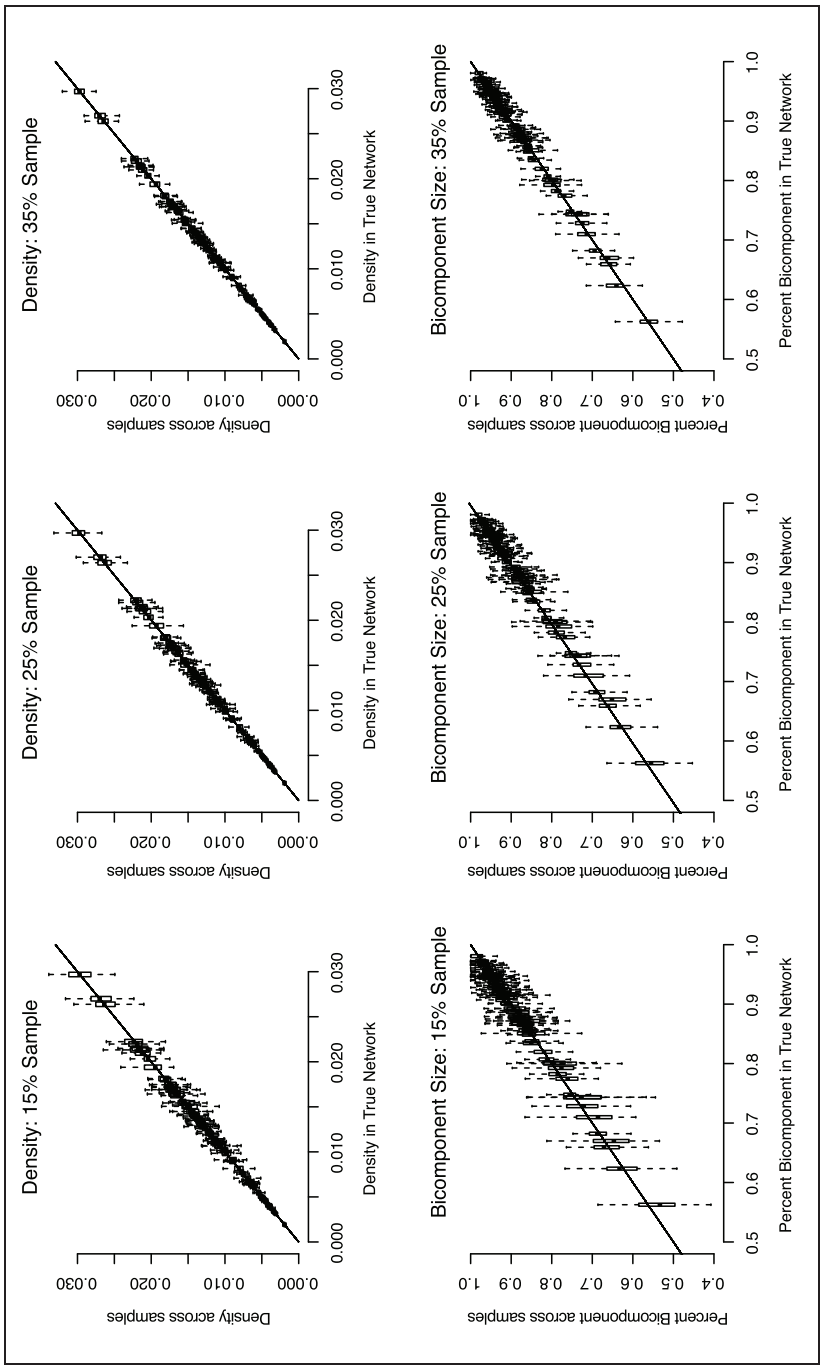

Figure 3. Estimates of density and bicomponent size by true values for Add Health networks.

relatively modest, with sample-to-sample estimates clustered tightly around the true value.

The bottom row presents results for bicomponent size. The results are again quite good, with the inferred values of bicomponent size centered at the true values. The correlation between the mean estimate (for each school, across samples) and the true value is .998 for all three sample rates. Here, however, we see that the variance of the estimates can be high, especially at the lower sampling rates. For example, for the 15 percent sample, the average standard error (across all schools) is .029, with some schools having standard errors above .05. Thus, the results show that bicomponent size can, on average, be accurately captured from sampled ego network data, but any given sample may offer an estimate that deviates substantially from the true value.

The initial results are encouraging. Still, it is an open question if the estimates for density and bicomponent size are good enough to serve as second-level predictors in an HLM. As Figure 3 shows, estimates for density or bicomponent size may be above or below the true value in any given sample. It is these imperfect measures (due to sampling variability) that a researcher must use as inputs into the HLM. With these results in mind, we now test a network sampling approach in the context of HLMs.

5.2. *True HLMs: Attachment to School*

We begin the HLM results section by looking at the true HLMs predicting attachment to school. The true models are estimated using the actual, known values for density and bicomponent size. Table 1 reports HLM coefficients for the 35 percent sample. There are four models, with varying combinations of control variables. In each case, we focus on the estimate for network cohesion, either density or bicomponent size. We report the mean estimate for density and bicomponent size over the 100 samples, as well as the 95 percent error bounds (such that 95 percent of the values fall in that interval). The table reports two sets of results for each model, one for bicomponent size and one for density.

In general, individuals in more cohesive schools report higher levels of attachment. This is clearest in model 1, in which no controls are included in the model. Looking at the density results, 95 percent of estimates fall between 12.70 and 17.00, with a mean value of 14.98. Even in model 4, with controls for gender, race, and social isolation, density

Table 1. Hierarchical Linear Model Results for School Attachment, 35 Percent Sample: Using Empirical Measure of Cohesion, Bicomponent Size, or Density

Variables	Model 1a	Model 2a	Model 3a	Model 4a	Model 1b	Model 2b	Model 3b	Model 4b
Intercept	2.411[a]	2.533[a]	2.487[a]	2.609[a]	2.56[a]	2.669[a]	2.605[a]	2.782[a]
	(2.388, 2.434)	(2.503, 2.565)	(2.465, 2.511)	(2.579, 2.641)	(2.549, 2.572)	(2.650, 2.689)	(2.593, 2.617)	(2.699, 2.739)
Contextual-level variables								
Density	14.976[a]	13.559[a]	11.802[a]	10.681[a]				
	(12.701, 17.002)	(11.222, 15.627)	(9.392, 13.881)	(8.28, 12.861)				
Bicomponent					.654[a]	.381[a]	.253[a]	.007
					(.510, .810)	(.232, .537)	(.110, .409)	(–.137, .162)
Individual-level variables								
Asian		–.069[a]		–.059[a]		–.073[a]		–.064[a]
		(–.115, –.023)		(–.104, –.011)		(–.117, –.026)		(–.109, –.017)
Black		–.195[a]		–.175[a]		–.194[a]		–.177[a]
		(–.234, –.155)		(–.213, –.134)		(–.233, –.154)		(–.215, –.136)
Hispanic		–.101[a]		–.085[a]		–.104[a]		–.091[a]
		(–.138, –.067)		(–.122, –.051)		(–.14, –.071)		(–.128, –.057)
Other		–.202[a]		–.19[a]		–.202[a]		–.191[a]
		(–.243, –.166)		(–.230, –.155)		(–.243, –.166)		(–.231, –.156)
Female		–.073[a]		–.094[a]		–.073[a]		–.094[a]
		(–.094, –.053)		(–.115, –.073)		(–.094, –.053)		(–.115, –.073)
Isolated			–.457[a]	–.459[a]			–.458[a]	–.461[a]
			(–.494, –.416)	(–.496, –.418)			(–.495, –.417)	(–.499, –.42)
N	20,458	20,458	20,458	20,458	20,458	20,458	20,458	20,458

[a]Denotes coefficient for which the 95 percent interval does not contain zero.

remains a strong predictor of school attachment. The mean coefficient is 10.68, and 95 percent of values fall between 8.28 and 12.86. From model 1 to model 4, the effect of density decreases by only 28 percent. This suggests that a dense network facilitates and supports individual attachment to a school. It is difficult to ignore that one is part of a larger group (here the school) if most people are friends with one another.

The results for bicomponent size are very different: here the effect of cohesion is highly dependent on the controls included in the model. In model 1, 95 percent of coefficients for bicomponent size fall between .51 and .81. The effect for bicomponent size clearly decreases, however, when individual-level controls are added to the model. Model 2 includes controls for race and gender, and the effect of bicomponent size is reduced by roughly 40 percent, going from .65 to .38; compare this with the 10 percent drop seen with density. The coefficient for bicomponent size in the full model (model 4) approaches zero, with a mean coefficient of .007 and 95 percent of estimates falling between –.14 and .16. The results for bicomponent size are thus weaker than with density, suggesting robustness (as opposed to whether everyone knows everyone else) does not uniformly increase attachment for the whole school, net of individual-level controls.

5.3. *Ego Network Results: Attachment to School*

The ego network results for the same models are presented in Figures 4 and 5 and Tables A1 and A2. The question is whether an ego network approach will yield the same estimates and conclusions as in the baseline models. Figure 4 presents results for the density coefficient, and Figure 5 presents results for the bicomponent coefficient. The results are presented as boxplots, capturing the distribution of the cohesion coefficient (either density or bicomponent size) across the 100 samples. The figures are organized by model and sampling rate, with models 1 to 4 on the columns and the three sampling rates on the rows. There are two boxplots for each subplot: the *true coefficients*, for which bicomponent size and density are known for each school, and the *ego network–based coefficients*, for which bicomponent size and density are inferred from the sampled ego network data. In the ideal case, boxplots from the ego network approach will be close to boxplots from the true models. The Appendix tables present summary statistics.

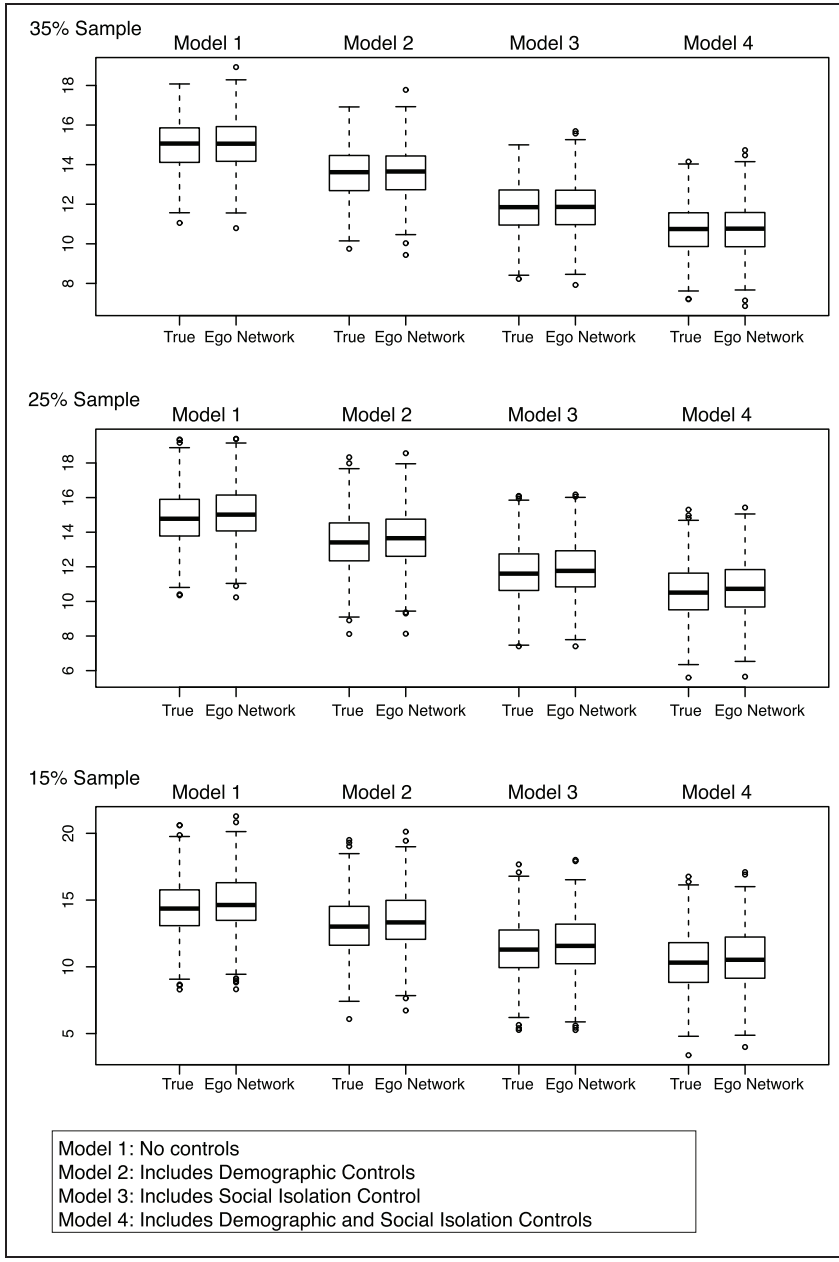

Figure 4. Boxplots of density coefficients for attachment to school models.
Note: Model 1 includes no controls. Model 2 includes demographic controls. Model 3 includes social isolation control. Model 4 includes demographic and social isolation controls.

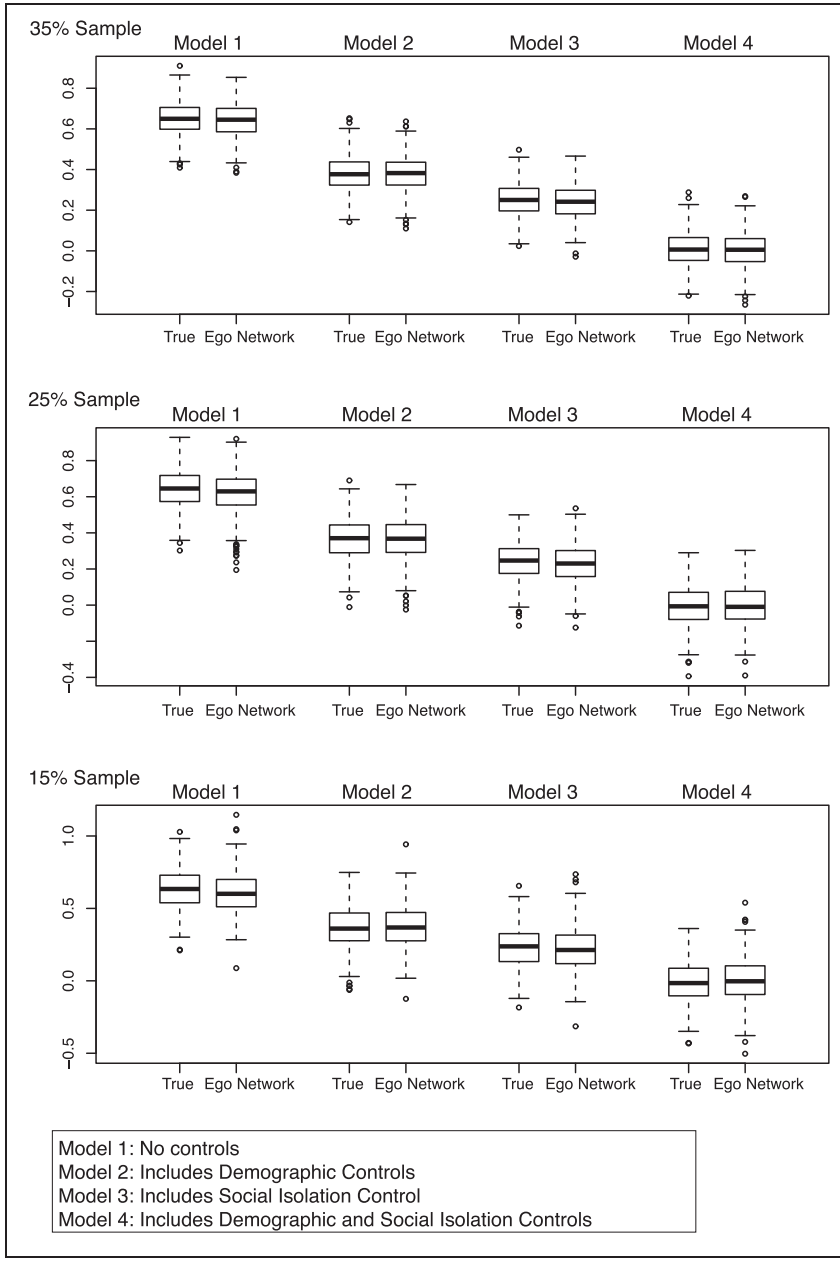

Figure 5. Boxplots of bicomponent coefficients for attachment to school models.

Note: Model 1 includes no controls. Model 2 includes demographic controls. Model 3 includes social isolation control. Model 4 includes demographic and social isolation controls.

Figure 4 presents results for density. The ego network results are quite good across models and sampling rates. Looking at model 1 (no individual-level controls), the ego network–based estimates closely approximate the true coefficients for density, with the ego network boxplots nearly identical to the true boxplots. For example, in the 35 percent sample, the true mean estimate is 14.976, and 95 percent of estimates fall between 12.70 and 17.00. In the ego network analysis, the mean coefficient is 14.99, and 95 percent of estimates fall between 12.60 and 17.16 (see Table A1). Model 2 controls for gender and race, but this does not greatly affect the results. For example, the mean estimate for the 35 percent sample is 13.56 using the true measure of density. The ego network–based estimate is 13.59, a difference of less than 1 percent. Looking at Figure 4, results for model 3 (controlling for isolation) and model 4 (all controls included) are similar: estimates using ego network data are close to estimates using the true values for density. Overall, moving from model 1 to model 4, the density coefficient is reduced by 28 percent in the ego network models (using the 35 percent sample: 14.999 to 10.701), the exact same value as in the true HLMs. As in the true models, the ego network–based results suggest density is a significant predictor of attachment, even controlling for demographics and social isolation (95 percent of the coefficients fall between 8.35 and 12.88 in model 4). Thus, the ego network models yield the same conclusions as the true models, in which density is measured on the complete, known networks. The 15 percent and 25 percent results are similar but have higher variance and are (slightly) less accurate.

Figure 5 and Table A2 present results using bicomponent size as the measure of network cohesion. These models provide a more difficult test, as bicomponent size is harder to estimate from sampled data than density. Results in Figure 5 are quite good, despite the difficulty of the test.[4] Across all models, the coefficients based on the ego network data are very similar to estimates using the true values of bicomponent size: the ego network boxplots are very close to the true boxplots. For example, in model 1, the true coefficient for bicomponent size is, on average, .65 for the 35 percent sample, with 95 percent of coefficients falling between .51 and .81. Using ego network data, the mean estimate is .64 (1.7 percent different from the true estimate), with 95 percent of estimates falling between .48 and .80 (see Table A2). The story is similar for models 2 to 4. For example, controlling for isolation in model 2, the true mean coefficient is .381 (using the 35 percent sample), and the

mean coefficient for the ego network analysis is .379. In the true models, 95 percent of estimates fall between .232 and .537; compare this with .199 and .545 in the ego network models. The 15 percent and 25 percent samples have similar results, although the estimates are predictably more variable as sample size decreases.

The ego network models also capture the reduction in the bicomponent coefficient across models. Looking at the 35 percent sample, the bicomponent coefficient is reduced by 99 percent from model 1 to model 4 (from .654 to .007) in the true models and 99.5 percent in the ego network models (from .643 to .003). More generally, the ego network models capture the null results in model 4, which controls for social isolation, gender, and race/ethnicity. For the ego network analysis, 95 percent of the bicomponent coefficients fall between –.17 and .17 (using the 35 percent sample). The true interval is –.14 to .16. Thus, the ego network models correctly show that the effect of bicomponent size is largely explained via demographic variables and social isolation (zero is contained in the 95 percent interval).

5.4. *True HLMs: Behavioral Problems in School*

Table 2 presents baseline HLM results for behavioral problems in school. The models are estimated as before, using the known, true values for bicomponent size and density. The only difference is that the outcome of interest is now behavioral problems in school, where higher values correspond to worse outcomes.

Results in Table 2 suggest that more cohesive schools have fewer behavioral problems, with (typically) negative coefficients for density and bicomponent size. Thus, more cohesive schools promote attachment, and they reduce the propensity for students to have behavioral problems. The effect for density however, is quite weak, with no significant coefficient in models 1 to 4. The results are more consistent for bicomponent size, for which a strong negative relationship holds across all four models, even controlling for demographics and social isolation. For example, 95 percent of coefficients fall between −.65 and −.26 in model 4 for the 35 percent sample. The results reflect the fact that a structurally robust network is conducive to promoting shared norms, and thus reducing behavioral problems in school.

Table 2. Hierarchical Linear Model Results for Behavioral Problems, 35 Percent Sample: Using Empirical Measure of Cohesion, Bicomponent Size, or Density

Variables	Model 1a	Model 2a	Model 3a	Model 4a	Model 1b	Model 2b	Model 3b	Model 4b
Intercept	1.623[a]	1.574[a]	1.600[a]	1.558[a]	1.605[a]	1.599[a]	1.593[a]	1.589[a]
	(1.594, 1.651)	(1.543, 1.609)	(1.571, 1.63)	(1.527, 1.593)	(1.592, 1.619)	(1.577, 1.619)	(1.579, 1.607)	(1.566, 1.61)
Contextual-level variables								
Density	–1.849	2.023	–.916	2.65				
	(–4.995, 1.023)	(–1.119, 4.847)	(–3.97, 2.063)	(–.453, 5.436)				
Bicomponent					–.945[a]	–.531[a]	–.84[a]	–.458[a]
					(–1.131, –.75)	(–.73, –.331)	(–1.021, –.648)	(–.652, –.262)
Individual-level variables								
Asian		.147[a]		.145[a]		.138[a]		.136[a]
		(.079, .207)		(.076, .204)		(.069, .199)		(.067, .197)
Black		.206[a]		.201[a]		.195[a]		.192[a]
		(.162, .242)		(.159, .238)		(.152, .234)		(.149, .232)
Hispanic		.222[a]		.218[a]		.21[a]		.208[a]
		(.183, .262)		(.179, .258)		(.172, .25)		(.168, .248)
Other		.188[a]		.185[a]		.184[a]		.182[a]
		(.146, .236)		(.144, .232)		(.142, .231)		(.14, .228)
Female		–.164[a]		–.159[a]		–.163[a]		–.159[a]
		(–.19, –.14)		(–.186, –.136)		(–.189, –.139)		(–.186, –.136)
Isolated			.135[a]	.101[a]			.125[a]	.094[a]
			(.093, .177)	(.058, .141)			(.083, .168)	(.052, .134)
N	20,458	20,458	20,458	20,458	20,458	20,458	20,458	20,458

[a]Denotes coefficient for which the 95 percent interval does not contain zero.

5.5. *Ego Network Results: Behavioral Problems in School*

The ego network–based results are presented in Figures 6 and 7 and Tables A3 and A4. As with attachment, the ego network–based results for density are quite good. Looking at the boxplots in Figure 6, the ego network–based coefficients are close to the coefficients based on the true values of density. For example, for model 1 in the 15 percent sample, the median coefficient is −1.82 in the true model and −1.85 in the ego network model, a difference of 1.6 percent. Similarly, 95 percent of coefficients in the true model fall between −6.42 and 2.82, and 95 percent of coefficients in the ego network model fall between −6.60 and 2.93. A researcher would thus arrive at the right conclusion, that density is not a significant predictor of behavioral problems, just using the ego network data. The 25 percent and 35 percent samples offer substantively similar results (see Table A3).

The bicomponent results paint a more complicated picture. In model 1, the ego network–based estimates for the bicomponent coefficient clearly improve as sample size increases. In the 15 percent sample, the mean coefficient in the true model is −.94, and the mean coefficient in the ego network model is −.87, a difference of about 8 percent. The percentage difference (between the true and ego network estimates) decreases to 4 percent in the 25 percent sample and to 1 percent in the 35 percent sample. Results are similar for model 3, in which the ego network coefficients converge with the true coefficients in the 25 percent and 35 percent samples. For example, in the 35 percent sample, the mean coefficient is −.84 in the true model and −.85 in the ego network model, a difference of about 1 percent. Models 2 and 4 provide more consistent results across sampling rates. In model 2, the mean coefficient for the 15 percent sample is −.51 in the true model and −.49 in the ego network model. In the 35 percent sample, the true coefficient is −.531 and the ego network estimate is −.552 (a difference of about 5 percent in the 15 percent sample and 4 percent in the 35 percent sample).

As with school attachment, the ego network models offer the same substantive conclusions as the true models. For example, consider model 4 for the 35 percent sample. Using the ego network data, 95 percent of density coefficients fall between −.56 and 5.52, and 95 percent of bicomponent coefficients fall between −.69 and −.28. This correctly

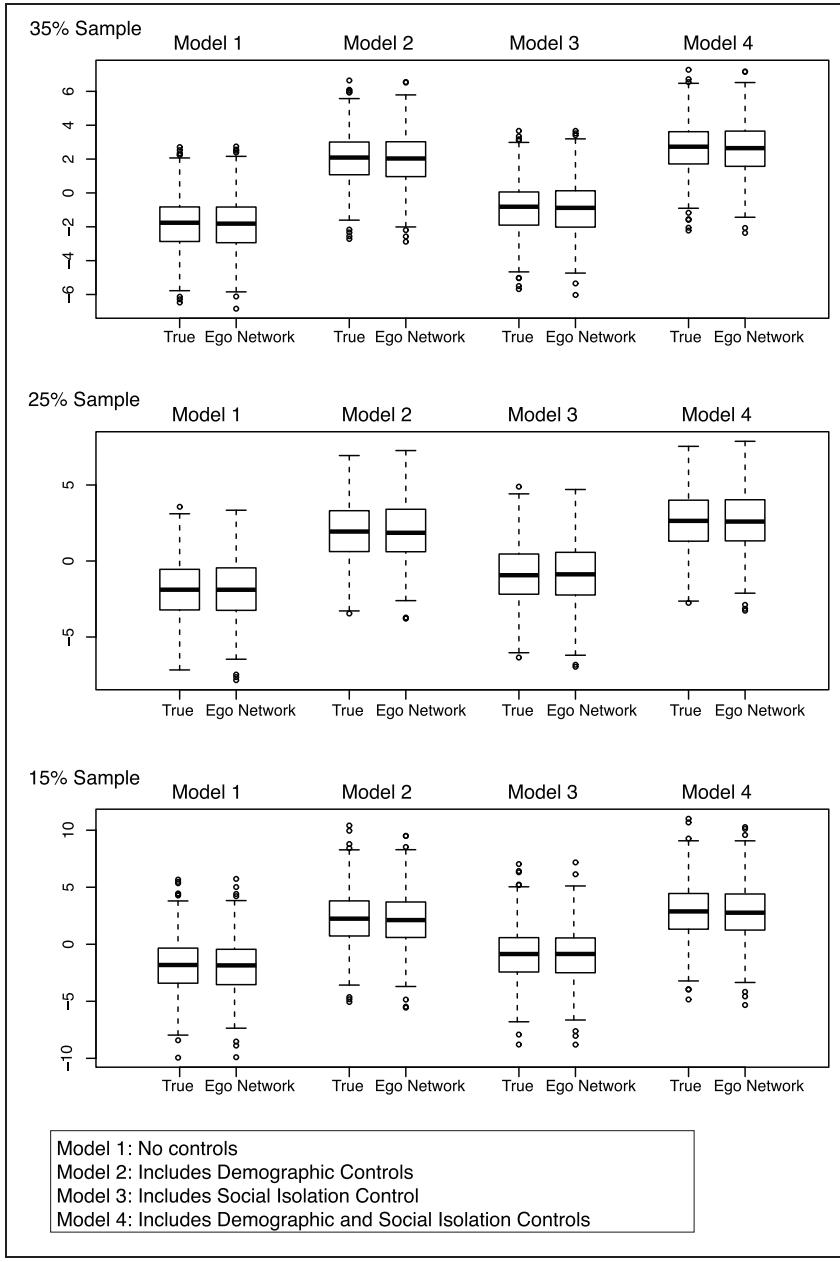

Figure 6. Boxplots of density coefficients for behavioral problems models. *Note:* Model 1 includes no controls. Model 2 includes demographic controls. Model 3 includes social isolation control. Model 4 includes demographic and social isolation controls.

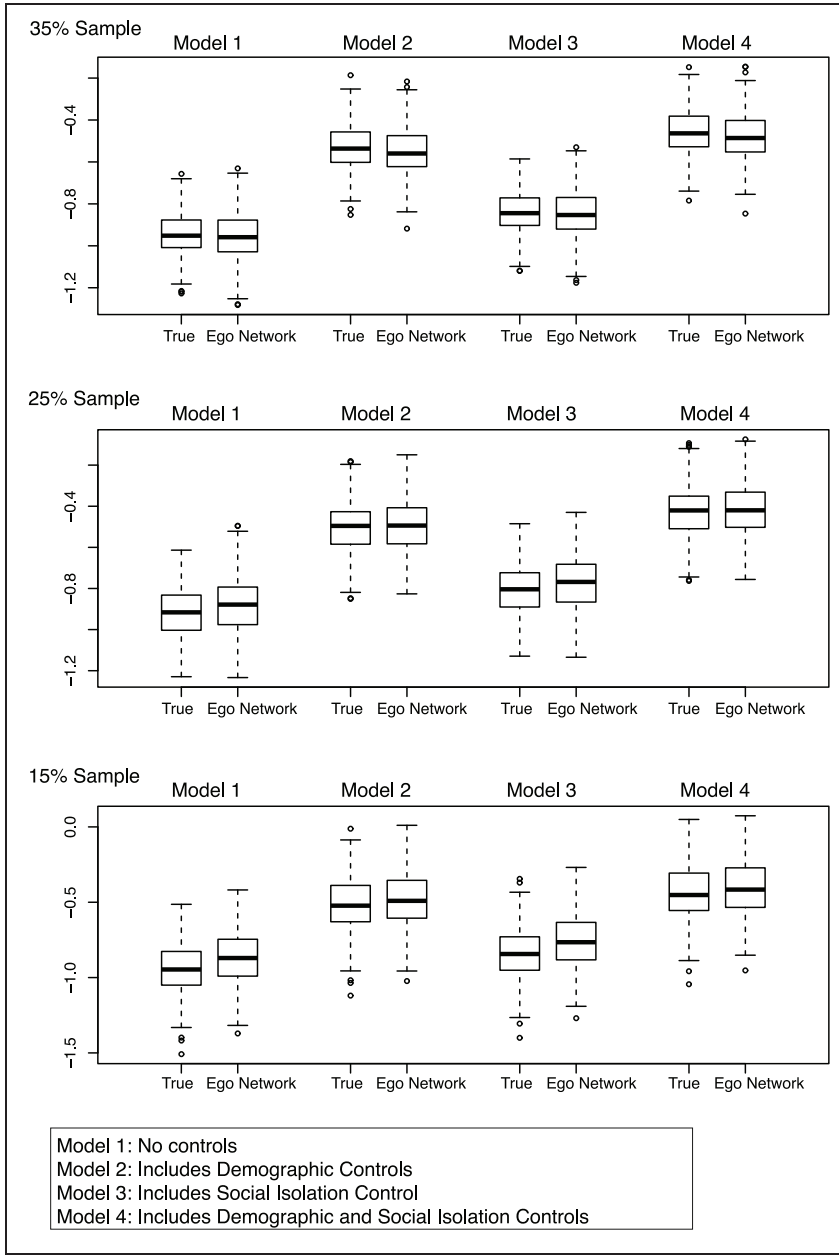

Figure 7. Boxplots of bicomponent coefficients for behavioral problems models.

Note: Model 1 includes no controls. Model 2 includes demographic controls. Model 3 includes social isolation control. Model 4 includes demographic and social isolation controls.

suggests that bicomponent size, but not density, reduces the propensity for students to have behavioral problems, net of other controls.

In summary, the ego network estimates effectively mirror the true estimates, albeit with some inconsistent results across sampling rates. It is instructive to consider why these results, although generally accurate, are more inconsistent than for school attachment. The difference lies in the presence of absence of strong outliers in the data. A small number of schools have very high levels of behavioral problems compared with the rest of the Add Health schools (i.e., more than 1.5 times the interquartile range). These outlier schools have a potentially disproportionate effect on the coefficients for second-level predictors, in particular bicomponent size. There are no parallel outliers in the school attachment analysis. Thus, with behavioral problems, the estimation is more difficult at low sampling rates. A few outlier schools have the potential to drive the estimation, and bicomponent size for those schools is measured imperfectly (each sample will yield a different value for bicomponent size). It is telling that estimation improves as sample size increases, making the measure of bicomponent size more consistent across samples.

6. ROBUSTNESS CHECKS: TESTING THE APPROACH UNDER DIFFERENT EGO NETWORK CONDITIONS

A network sampling approach, although promising, must contend with the practical limitations of ego network data. We have thus far assumed that a researcher could collect information on the number of alters, alters' characteristics, and ties between alters. Such data collection may not always be possible, however. For example, alter-alter tie data can be burdensome to collect. Surveys will thus often restrict the ego network questionnaire, including questions for number of alters and alter characteristics, but not alter-alter ties. Similarly, we have assumed that ego network data are measured without error, yet this may be an optimistic assumption (Almquist 2012). Ego network surveys ask respondents to report secondhand on the relationships between their alters. Respondents may not always know if their named alters are friends, however, and an uncertain respondent may be forced to guess if a tie exists, creating bias. Thus, even when alter-alter ties can be collected, there may still be problems with the data.

We explore these issues in Appendix C, in which we replicate the analysis under different conditions surrounding alter-alter tie data. In

the first analysis, we assume that no information on alter-alter ties is collected. The analysis is exactly the same as in the main text, but here the researcher must infer bicomponent size using ego network data that include only degree and homophily information (there is information only on the number of alters and their characteristics). In the second analysis, we assume that alter-alter ties are available, but there is measurement error in the data. We take the true ego network data (data under perfect reporting) and construct scenarios in which 15 percent of ties are reported with error. These error-filled data become the input to the simulation approach.

Results for the 35 percent sample are presented in Appendix Tables C1 and C2. We present results only for bicomponent size because density does not depend on the alter-alter tie data. Looking at Table C1 (attachment is the dependent variable), we see that the results are good, on the whole, using the limited ego network data, although not as accurate as with the complete ego network data. For example, the true coefficient for bicomponent size in model 1 is .654; the estimate using the complete ego network data is .643, and it is .684 with no alter-alter data and .634 with the measurement error data. Results are similar for the other models. Overall, the results are largely encouraging, with estimates suggesting that a researcher can "get away" with collecting imperfect ego network data and still correctly estimate an HLM using sampled network data, even if the estimates are not as good as with full information. See Appendix C for the full results.

7. ROBUSTNESS CHECKS: OTHER CONTEXTUAL-NETWORK MEASURES

Appendix D presents another set of supplementary results. Here we run the same basic analysis, estimating HLMs using inferred network features, but we use an alternative set of network measures. Instead of focusing on density and bicomponent size, we include results for average betweenness, average closeness, transitivity, and proportion isolated. Table D1 presents the results for attachment and behavioral problems. To simplify the discussion, we focus just on the results for the 35 percent sample and we only include results for model 1 (no controls).

Overall, the results are quite good for attachment, with estimates based on the ego network data approximating the true estimates. For example, for transitivity, the true mean coefficient is .42, and the ego

network–based estimate is .41. Or, for proportion isolated, 95 percent of ego network estimates fall between −1.79 and −1.22; compare this with the true sampling distribution, in which 95 percent of values fall between −1.75 and −1.26. We see similar results for average betweenness, with a true mean coefficient of −3.77E-5 and an inferred estimate of −3.83E-5. The results are not as strong for average closeness, for which the ego-based estimate is −1.19 and the true estimate is −1.04 (a difference of 14 percent). Closeness is notoriously difficult to measure with incomplete data, and thus inference in the HLM context is challenging (Smith et al. 2017). We see similar results for behavioral problems, although the ego network–based estimates fare even worse for average closeness, possibly because of outliers in the data (the transitivity coefficient is also estimated poorly here). The mean true coefficient for average closeness is −1.133 and the ego network estimate is −.771 (a difference of more than 30 percent). The results thus suggest an HLM network sampling approach can, potentially, work beyond measures of cohesion, but some networks measures (and outcomes) will be more conducive than others.

8. ROBUSTNESS CHECKS: TESTING THE APPROACH USING SIMULATED DATA

We offer one more test of a contextual network sampling approach. This test uses the same setup as with the Add Health analysis but uses data constructed with known properties, rather than being based on empirical data. The basic idea is to test a network sampling approach in a case in which the networks and outcomes of interest can be fully controlled. There is thus no measurement error in the data used for the test, and the coefficients of interest are known from the start. We can systematically vary the network features, as well as the relationship between the measures of cohesion and the dependent variables. We consider only measures of cohesion here. The test offers an important robustness check, seeing if a network sampling approach will work on a completely different case, one with different networks, outcomes, and models.

Appendix E presents the methodological details and results, but we briefly describe the results here. The test is based on 36 constructed networks with systematically different features (in terms of density and bicomponent size). The test is based on two outcomes, constructed to have known relationships with our measures of cohesion. The first

outcome is constructed so bicomponent size, but not density, is related to the outcome of interest. The second outcome has the inverse pattern, with only density related to the outcome of interest. The rest of the test is analogous to that used above, with three sample rates (15 percent, 25 percent, and 35 percent), 100 samples per sample rate, and similar assumptions about the ego network data.

Results are presented in Appendix Tables E1 and E2, and they offer strong support for a contextual, network sampling approach. A researcher with only ego network data would arrive at the correct estimates for density and bicomponent size. With the first outcome, the positive, strong effect for bicomponent size is approximated quite well, and the null effect for density is also correctly captured. For example, for the 25 percent sample, the mean estimate is 1.009, and the true coefficient is 1.00. With the second outcome, a network sampling approach correctly picks up the opposite effects, with density, but not bicomponent size, affecting the outcome of interest. For example, the true coefficient for density is 50, and the ego network–based estimates have a mean of 49.272 for the 25 percent sample. One drawback to using ego network data is that the coefficients have relatively high variance, higher than if we had known the true values of bicomponent size and density. Thus, in cases in which the true effect of density or bicomponent size is very strong, measuring those network properties imperfectly (each sample will yield a slightly different estimate) yields coefficients that are measured uncertainly. Still, even with higher variance, the estimated coefficients from the ego network data offer a viable means of doing contextual network models. See Appendix E for the full set of results.

9. CONCLUSION

Network data are a natural fit for contextual models. Global network measures offer a rich picture of a social context, showing how micro-level interactions cohere into a larger whole (Butts 2008; Robins et al. 2005). A researcher could collect network data in each context of interest, measure global network features like cohesion, and use the network measures to predict health, mental health, deviance, and so on. The drawback of a network approach is that the data collection burden (at least traditionally) is quite high, as one would need to collect census data in every context in the study. In this article we consider an

alternative approach, one in which the researcher estimates contextual-network models but is able to avoid the heavy data collection toll. The basic idea is to combine HLMs with network sampling: one uses sampled ego network data to infer the network features of each context and then uses the inferred network features as second-level predictors in an HLM (Raudenbush and Bryk 2002; Smith 2012).

We test the validity of this idea using two complete data sets. The main test uses empirical data from Add Health. We examine the relationship between two measures of network cohesion, density and bicomponent size, and two individual-level outcomes, school attachment and behavioral problems in school. The results are encouraging. Across all models, it is possible to approximate the true coefficients for density and bicomponent size just using the ego network data. Importantly, the substantive conclusions based on ego network data are exactly the same as in the baseline models, using the true values for density and bicomponent size. Our second test uses simulated data, and we find similar results, offering additional supporting evidence.

Overall, the results suggest that contextual-network models can be estimated using sampled data, thus reducing the data burden for the researcher considerably. The implications are clear: any study with sampled individuals can become a network study, in which individual outcomes, behaviors, and interactions are placed within a larger relational context.

A contextual, network sampling approach is not without limitations, however. For example, some network measures are easier to estimate from sampled data than others. Network measures that capture features of path-based connectivity are generally more difficult to estimate than measures that are independent of the path structure. With our cohesion measures, for example, we found that estimates around bicomponent size (dependent on the path structure) are more variable than estimates for density, which do not depend on the path structure. The results for average betweenness and average closeness are also instructive. Both are path-based network measures, but the results are better for betweenness than for closeness. This is the case because average betweenness is based on the number of shortest paths between vertices, whereas average closeness is based on the length of the shortest paths. It is easier to capture general properties about the path structure (e.g., reachability between actors) than more specific properties of those paths (e.g., distance between actors). Overall, the method is most easily applied in

cases in which the network measure does not depend on the path structure (density, proportion isolated). The next best case is for network measures that depend on the general features of the path structure (bicomponent size, average betweenness). The most difficult case is where network measures depend on the specific path lengths between actors (average closeness).

The method is also only appropriate for certain types of networks. The networks of interest must be undirected, as ego networks do not capture asymmetric relationships very well. Similarly, the networks of interest must capture a strong relationship, where it is difficult to maintain a large number of ties, because the inferential approach can have problems when the degree distribution is badly skewed, with one or two actors capturing a disproportionately large number of ties (as a random sample may not always capture these important actors) (Smith 2015). In such cases, researchers may have to consider alternative sampling schemes, such as a two-step snowball sample, in which the ego's alters are interviewed (offering more information than the simple ego network data).

Similarly, the method for generating whole networks from ego network samples is based on a simulation approach and is limited by factors that make computation expensive, such as the number of actors in the network, the number of ties between them, and the transitivity in the network. Large, dense, transitive networks are more difficult to simulate than small, sparse, and nonclustered ones. These factors combine in complex ways to determine the practical application of the approach. A large, sparse network (e.g., 50,000 nodes) with low transitivity may be practical to infer, whereas a smaller network (e.g., 10,000 nodes) that is very dense and very transitive may prove prohibitive (in terms of run time). Practical experience suggests it is possible to infer networks up to about 75,000 nodes, but the most likely applications will be on much smaller networks, especially in the case of HLMs, for which multiple networks must be inferred.

There are fewer limitations when it comes to the dependent variables that are appropriate for the approach. For example, we were able to estimate the HLMs using ego network data even in cases in which the relationship between cohesion and the outcome was weak (e.g., the case of bicomponent size and attachment). The main limitation is with outliers, for which the models are more difficult to estimate when there are

strong outliers on the dependent variable (some contexts have very high or low values on the outcome of interest). Additionally, our results are restricted to the case of continuous variables, and the approach may not fare as well with categorical outcomes (such as binary variables), for which the variance on the dependent variable is constrained.

As a final limitation, measurement error in the ego network data may cause estimation problems (Alwin 2007; Feld and Carter 2002). We considered the case of misreporting in alter-alter ties, but there are other possible sources of error that may affect the results. For example, past work has found that respondents will sometimes report fewer alters than they actually have (Marin 2004). Such underreporting will distort the number of alters listed per respondent, which is a key input into the density and bicomponent size calculations. Alternatively, respondents may name an alter as more of an aspirational tie than an actual one (i.e., someone they wished were their friend), adding an alter to the list who should not actually be included. The HLM estimates (ultimately the item of interest here) will only be affected, however, if the bias is (1) quite high, (2) stronger in some contexts than others, and (3) correlated with the outcome of interest. There is no *a priori* reason to believe that such conditions are likely. Still, it will be important for future work to consider the problem of alter misreporting more closely.

In this article we focused on network cohesion across schools, but an HLM network sampling approach is general and is appropriate for any research setting concerned with the effect of social context on individual outcomes. Methodologically, future work could test the approach using other contextual-level network measures (e.g., centralization or modularity) to see if the approach can accurately reproduce the effect of these features on important outcomes of interest (see Appendix D for an initial test on betweenness, closeness, transitivity, and proportion isolated). More substantively, ego network sampling could be applied to such diverse topics as deviance in neighborhoods, mental health in organizations, and health behaviors in schools. Moving forward, the hope is that researchers will find it easier to blend traditional survey methods with a network approach, thus maintaining the coverage and convenience of a sample without sacrificing the holistic, relational feel of a network study.

APPENDIX A

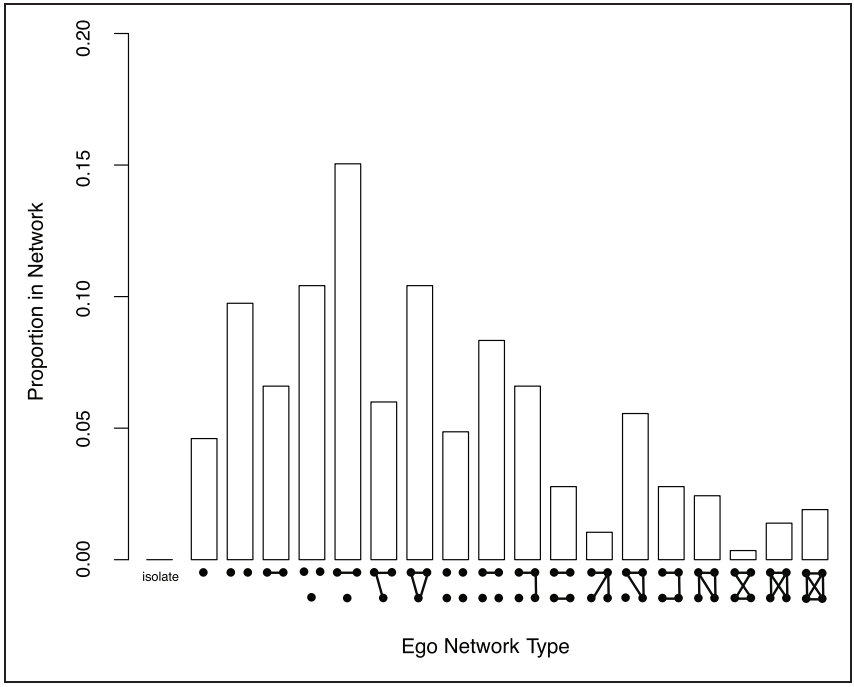

Figure A1. Example ego network configuration.
Note: This figure is a based on a hypothetical ego network configuration distribution. The proportions refer to the proportions of each ego network type in a hypothetical network. Ego is not included in the ego network types. The figure includes only ego network types of size four or less to make the figure legible.

Table A1. Summary of True and Estimated Sampling Distribution for Density Coefficient, Attachment to School Models

		15% Sample		25% Sample		35% Sample	
		True	Ego Network	True	Ego Network	True	Ego Network
Model 1[b]	Mean	14.434	14.775	14.842	15.054	14.976	14.999
	Median	14.364	14.631	14.779	15.016	15.066	15.058
	S.D.	2.139	2.158	1.598	1.577	1.200	1.214
	95% CI[a]	(10.24, 18.711)	(10.426, 18.855)	(11.563, 18.217)	(11.939, 18.384)	(12.701, 17.002)	(12.602, 17.163)
Model 2[c]	Mean	13.105	13.48	13.445	13.676	13.559	13.592
	Median	13.014	13.327	13.41	13.654	13.619	13.653
	S.D.	2.189	2.198	1.671	1.659	1.215	1.225
	95% CI[a]	(9.017, 17.495)	(9.219, 17.768)	(10.055, 17.055)	(10.313, 17.237)	(11.222, 15.627)	(11.293, 15.671)
Model 3[d]	Mean	11.328	11.614	11.674	11.852	11.802	11.816
	Median	11.291	11.568	11.607	11.766	11.852	11.864
	S.D.	2.135	2.17	1.601	1.58	1.207	1.224
	95% CI[a]	(7.151, 15.435)	(7.151, 15.835)	(8.301, 14.965)	(8.707, 15.19)	(9.392, 13.881)	(9.421, 14.057)
Model 4[e]	Mean	10.333	10.643	10.574	10.766	10.681	10.701
	Median	10.314	10.528	10.51	10.728	10.748	10.766
	S.D.	2.192	2.215	1.667	1.655	1.219	1.235
	95% CI[a]	(6.302, 14.509)	(6.296, 14.937)	(7.07, 14.007)	(7.405, 14.096)	(8.28, 12.861)	(8.353, 12.885)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of the estimates fall within that interval.

[b]No controls.

[c]Includes demographic controls.

[d]Includes social isolation control.

[e]Includes demographic and social isolation controls.

Table A2. Summary of True and Estimated Sampling Distribution for Bicomponent Coefficient, Attachment to School Models

		15% Sample		25% Sample		35% Sample	
		True	Ego Network	True	Ego Network	True	Ego Network
Model 1[b]	Mean	.629	.605	.645	.625	.654	.643
	Median	.634	.601	.646	.63	.65	.645
	S.D.	.139	.138	.102	.115	.08	.083
	95% CI[a]	(.347, .88)	(.348, .867)	(.455, .844)	(.367, .848)	(.51, .81)	(.483, .801)
Model 2[c]	Mean	.362	.373	.371	.368	.381	.379
	Median	.361	.368	.371	.368	.377	.383
	S.D.	.147	.146	.113	.121	.082	.087
	95% CI[a]	(.071, .63)	(.085, .656)	(.162, .598)	(.12, .605)	(.232, .537)	(.199, .545)
Model 3[d]	Mean	.228	.215	.243	.231	.253	.241
	Median	.238	.213	.247	.23	.25	.242
	S.D.	.141	.143	.103	.107	.079	.084
	95% CI[a]	(–.045, .488)	(–.056, .472)	(.038, .45)	(.03, .44)	(.11, .409)	(.078, .398)
Model 4[e]	Mean	–.013	.004	–.005	–.002	.007	.003
	Median	–.016	–.003	–.007	–.01	.006	.005
	S.D.	.149	.15	.111	.113	.082	.087
	95% CI[a]	(–.322, .265)	(–.29, .282)	(–.218, .219)	(–.222, .223)	(–.137, .162)	(–.169, .168)

Note: CI = confidence interval.
[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.
[b]No controls.
[c]Includes demographic controls.
[d]Includes social isolation control.
[e]Includes demographic and social isolation controls.

Table A3. Summary of True and Estimated Sampling Distribution for Density Coefficient, Behavioral Problems Models

		15% Sample		25% Sample		35% Sample	
		True	Ego Network	True	Ego Network	True	Ego Network
Model 1[b]	Mean	–1.877	–1.941	–1.856	–1.863	–1.849	–1.874
	Median	–1.818	–1.852	–1.888	–1.89	–1.76	–1.814
	S.D.	2.382	2.443	1.894	1.974	1.516	1.534
	95% CI[a]	(–6.423, 2.822)	(–6.599, 2.933)	(–5.573, 1.709)	(–5.818, 1.948)	(–4.995, 1.023)	(–4.962, 1.063)
Model 2[c]	Mean	2.264	2.128	2.013	1.97	2.023	1.973
	Median	2.243	2.12	1.938	1.856	2.09	2.035
	S.D.	2.428	2.472	1.912	1.964	1.514	1.527
	95% CI[a]	(–2.528, 7.188)	(–2.595, 7.118)	(–1.655, 5.73)	(–2.03, 5.625)	(–1.119, 4.847)	(–1.1, 4.784)
Model 3[d]	Mean	–.904	–.948	–.864	–.856	–.916	–.937
	Median	–.857	–.851	–.935	–.877	–.815	–.881
	S.D.	2.39	2.447	1.904	1.978	1.508	1.526
	95% CI[a]	(–5.432, 3.952)	(–5.829, 3.874)	(–4.702, 2.743)	(–4.722, 2.835)	(–3.97, 2.063)	(–4.079, 2.046)
Model 4[e]	Mean	2.894	2.774	2.68	2.648	2.650	2.603
	Median	2.879	2.766	2.639	2.595	2.731	2.646
	S.D.	2.436	2.478	1.919	1.968	1.509	1.521
	95% CI[a]	(–1.989, 7.815)	(–1.987, 7.811)	(–1.041, 6.404)	(–1.408, 6.211)	(–.453, 5.436)	(–.565, 5.521)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.

[b]No controls.

[c]Includes demographic controls.

[d]Includes social isolation control.

[e]Includes demographic and social isolation controls.

Table A4. Summary of True and Estimated Sampling Distribution for Bicomponent Coefficient, Behavioral Problems Models

		15% Sample		25% Sample		35% Sample	
		True	Ego Network	True	Ego Network	True	Ego Network
Model 1[b]	Mean	–.943	–.868	–.917	–.881	–.945	–.954
	Median	–.946	–.87	–.916	–.878	–.952	–.959
	S.D.	.163	.17	.116	.126	.096	.106
	95% CI[a]	(–1.249, –.64)	(–1.185, –.553)	(–1.13, –.704)	(–1.117, –.643)	(–1.131, –.75)	(–1.16, –.75)
Model 2[c]	Mean	–.511	–.485	–.505	–.496	–.531	–.552
	Median	–.523	–.491	–.496	–.494	–.536	–.56
	S.D.	.174	.173	.117	.123	.103	.11
	95% CI[a]	(–.825, –.185)	(–.793, –.158)	(–.73, –.296)	(–.741, –.278)	(–.73, –.331)	(–.761, –.357)
Model 3[d]	Mean	–.836	–.764	–.806	–.772	–.84	–.849
	Median	–.844	–.765	–.804	–.768	–.844	–.853
	S.D.	.166	.17	.118	.127	.096	.107
	95% CI[a]	(–1.143, –.526)	(–1.072, –.46)	(–1.031, –.581)	(–1.008, –.529)	(–1.021, –.648)	(–1.051, –.645)
Model 4[e]	Mean	–.437	–.411	–.427	–.419	–.458	–.479
	Median	–.452	–.416	–.42	–.42	–.464	–.486
	S.D.	.176	.174	.119	.124	.103	.11
	95% CI[a]	(–.756, –.103)	(–.721, –.071)	(–.666, –.205)	(–.672, –.19)	(–.652, –.262)	(–.694, –.278)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.

[b]No controls.

[c]Includes demographic controls.

[d]Includes social isolation control.

[e]Includes demographic and social isolation controls.

APPENDIX B: SCOPE CONDITIONS FOR SIMULATION APPROACH TO NETWORK INFERENCE

Network sampling holds great promise but is not without its limitations, and it is appropriate only when certain scope conditions are met. For example, the population of interest must have a known sampling frame, with properties conducive to traditional sampling techniques. Similarly, an ego network approach is appropriate only for certain types of social relationships. The relationship of interest must be reasonably strong (e.g., close friends, discuss important matters), as weaker ties would potentially create a long list of alters. One may know hundreds of people, for example. Unfortunately, it can be difficult to report on a long list of alters within a survey. Similarly, past work shows that estimates for bicomponent size can be biased when the skew of the degree distribution is severe (i.e., a few actors have a very large number of ties), which is more likely when the relationship of interest is weak (Smith 2015). This is the case because the high-degree nodes have a disproportionate effect on the network structure but are not any more or less likely to be sampled in a random sample of the population. The relationship of interest must also be symmetric, generating an undirected network (so that if *i* is friends with *j, j* is also friends with *i*). Ego network data are only from the point of view of the respondent, making it difficult to capture asymmetric relationships. Finally, the researcher must be able to estimate the size of each context.

APPENDIX C

This appendix describes a supplementary analysis in which we consider different practical problems with using ego network data. We consider two issues. For the first analysis, we look at practical problems surrounding data collection. Researchers are often unable to collect full network information; in particular, many ego network surveys do not include information on alter-alter tie data, as this can be difficult to collect. We reconsider our analysis in light of such limited ego network data. For the second analysis, we consider the problem of measurement error. Here, we assume that alter-alter tie data are collected but the data are imperfect, suffering from inaccuracies in respondent reporting.

Limited Ego Network Data: No Alter-Alter Tie Data

For the main analysis, we assume that the ego network survey includes information on degree (number of alters), alter characteristics, and ties between

alters. Here, we repeat the analysis under more restrictive data assumptions, in which no information on alter-alter ties is available. The question is whether a researcher can use this more limited survey and still accurately estimate an HLM using sampled network data. It would be unsurprising if the results are worse than before (i.e., including alter-alter ties in the ego network survey), but if the differences are slight, then it may be a worthwhile trade-off, given the comparative ease of data collection.

The analysis is largely the same as before. The key difference is that the input ego network data only include information on number of alters and their characteristics. We still infer bicomponent size by simulating full networks on the basis of the ego network data. The simulation, however, is conditioned only on the degree distribution and homophily—the information that can be inferred from the limited ego network survey. The simulated networks are thus not conditioned on any information about local closure (or shared partners), which would normally be inferred from the alter-alter ties. The analysis is otherwise the same, with the same models, sample rates, and so on as in the main analyses. The results for density are not presented, as the estimates are identical to the previous analysis, not being dependent on alter-alter ties.

Tables C1 and C2 present the results. The tables compare the true results (using the known values of density and bicomponent size), the full ego network results (for which alter-alter tie information is available and perfectly reported), and the limited ego network results (column 3). We present results for the 35 percent sample. We begin with the attachment results in Table C1. The estimates are good on the whole, even though the simulated networks have transitivity levels (is a friend of a friend also a friend?) that are no higher than by chance (i.e., the level of transitivity that can be induced by homophily alone, given the degree distribution). In model 2, the 35 percent sample yields a true median coefficient of .377, and the estimate from the limited ego network data is .397. This is worse, but not dramatically so, than with the complete ego network data, for which the median estimate is .383. The other models offer the same basic story: the estimates using the full ego network data are more accurate than the limited models, but the estimates are closer than one might think. For example, for model 1, the true coefficient is .65 for the 35 percent sample (using the median of the sampling distribution). The limited ego network data yield a median estimate of .682 and a 95 percent interval of (.517, .846), showing the positive effect of cohesion on attachment. The median for the full ego network model is .645. More generally, the limited ego network results offer the same substantive findings as the true models. In model 4, the 95 percent interval is (–.174, .191) for the limited ego network models, correctly capturing

Table C1. Summary of Estimates for Bicomponent Coefficient, Attachment to School Models: Robustness Checks

		35% Sample			
		True	Ego Network (Full Information)	Ego Network (No Alter-Alter Tie Data)	Ego Network (Measurement Error in Alter-Alter Tie Data)
Model 1[b]	Mean	.654	.643	.684	.634
	Median	.65	.645	.682	.626
	S.D.	.08	.083	.087	.087
	95% CI[a]	(.51, .81)	(.483, .801)	(.517, .847)	(.464, .839)
Model 2[c]	Mean	.381	.379	.399	.363
	Median	.377	.383	.397	.358
	S.D.	.082	.087	.092	.092
	95% CI[a]	(.232, .537)	(.199, .545)	(.223, .575)	(.206, .552)
Model 3[d]	Mean	.253	.241	.262	.234
	Median	.25	.242	.259	.232
	S.D.	.079	.084	.088	.087
	95% CI[a]	(.11, .409)	(.078, .398)	(.099, .435)	(.042, .43)
Model 4[e]	Mean	.007	.003	.005	.011
	Median	.006	.005	.005	.015
	S.D.	.082	.087	.093	.092
	95% CI[a]	(–.137, .162)	(–.169, .168)	(–.174, .191)	(–.154, .189)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.

[b]No controls.

[c]Includes demographic controls.

[d]Includes social isolation control.

[e]Includes demographic and social isolation controls.

the weak effect of bicomponent size on attachment when all controls are included in the model.

Table C2 presents results for behavioral problems. The results parallel the attachment results: the full ego network models are not dramatically better than the limited models (with no alter-alter tie information). For model 1 in the 35 percent sample, the median coefficient is −1.008 for the limited ego network model and −.959 for the full ego network model (the true value is −.952). The estimates are even closer in model 2, with median coefficients of −.589 and −.56 for the full and limited ego network models (the true value is −.536). For model 3, the 95 percent interval is (−1.051, −.645) in the full ego network model, (−1.106, −.679) in the limited ego network model, and (−1.021, −.648) in the true model. The true median is −.844; the full ego network estimate is −.853 and the limited estimate is −.897. More generally, we again see that the limited model offers the same basic conclusion as using the true measures of bicomponent size. In all four models, the 95 percent interval does not include 0, implying (correctly) that bicomponent size consistently reduces the level of behavioral problems in a school.

Measurement Error in Alter-Alter Tie Data

The second analysis considers the problem of measurement error. Here we assume the data include all the information from the main analysis: number of alters, alter characteristics, and ties between alters. Unlike the main analysis, however, we assume the alter-alter tie data are measured imperfectly. Past work suggests there may be bias in ego network data, particularly for alter-alter tie reports (Almquist 2012; Feld and Carter 2002). Respondents must report secondhand about the relationships between their alters, but they may not always know if their alters are close, and they may be forced to guess if a tie exists, creating error in the data. The question is how badly the HLM estimates are biased when there is error in alter-alter tie reports (which are used to infer the network properties of interest).

The error generation process takes the following form. We assume that respondents are asked to report their alter-alter ties but that 15 percent of those reports are guesses; that is, the respondent is unsure if a tie actually exists. We simulate guessing by taking a random draw from a binomial distribution, with probability set to .5. We set the alter-alter tie to 0 or 1 depending on the random draw for that tie (for a similar procedure, see Smith and Faris 2015). Thus, 15 percent of alter-alter ties are just guesses, with no actual relationship to the true

Table C2. Summary of Estimates for Bicomponent Coefficient, Behavioral Problems Models: Robustness Checks

		35% Sample			
		True	Ego Network (Full Information)	Ego Network (No Alter-Alter Tie Data)	Ego Network (Measurement Error in Alter-Alter Tie Data)
Model 1[b]	Mean	–.945	–.954	–1.002	–.930
	Median	–.952	–.959	–1.008	–.936
	S.D.	.096	.106	.112	.105
	95% CI[a]	(–1.131, –.75)	(–1.16, –.75)	(–1.213, –.784)	(–1.131, –.745)
Model 2[c]	Mean	–.531	–.552	–.578	–.555
	Median	–.536	–.56	–.589	–.563
	S.D.	.103	.11	.116	.108
	95% CI[a]	(–.73, –.331)	(–.761, –.357)	(–.8, –.37)	(–.76, –.338)
Model 3[d]	Mean	–.84	–.849	–.891	–.826
	Median	–.844	–.853	–.897	–.830
	S.D.	.096	.107	.112	.105
	95% CI[a]	(–1.021, –.648)	(–1.051, –.645)	(–1.106, –.679)	(–1.03, –.625)
Model 4[e]	Mean	–.458	–.479	–.501	–.483
	Median	–.464	–.486	–.507	–.488
	S.D.	.103	.11	.116	.109
	95% CI[a]	(–.652, –.262)	(–.694, –.278)	(–.725, –.287)	(–.694, –.28)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.

[b]No controls.

[c]Includes demographic controls.

[d]Includes social isolation control.

[e]Includes demographic and social isolation controls.

data. The error-filled ego network data are then used as input to the simulation approach. The rest of the analysis follows the analysis above.

The results for the measurement error analysis are presented in column 4 of Tables C1 and C2. The results follow a predictable pattern: the estimates are clearly affected by measurement error but generally fare better than having no alter-alter tie data at all. For example, looking at the attachment results for model 1, the true mean coefficient is .654, the limited ego network mean coefficient (with no alter-alter ties) is .684, and the measurement error mean (with alter-alter ties, but measured imperfectly) is .634. The mean coefficient with full ego network information and no measurement error is .643. We see similar results for the other models, although the limited information estimates are not always worse than the measurement error estimates. For example, for model 2, 95 percent of estimates fall between .232 and 0.537 in the true model. The analogous values are (.206, .552) for the measurement error results and (.223, .575) for the limited survey results.

The behavioral problems results tell a similar story. For model 3, for example, the true mean coefficient is –.84, the full ego network information estimate is –.849, the limited ego network estimate is –.891, and the measurement error estimate is –.826. Or, looking at model 4, the true median coefficient is –.464, and the measurement error median is –.485; compare this with the limited ego network estimates, for which the median is –.507.

Overall, this robustness check suggests that a researcher with imperfect ego network data can still estimate contextual-level network models well, although not as accurately as in the ideal case of full information and accurate reporting. The results also suggest that having no alter-alter tie data is worse than having alter-alter tie data that contain measurement error. Thus, at least in this case, it is better to have error-filled data than no data at all.

APPENDIX D

This appendix describes a supplementary analysis in which we test a network sampling approach on alternative measures for the contextual-network property. The analysis is exactly the same as in the main text (using the Add Health data), but here we consider four different measures of network structure, or four different contextual-level predictors to use in the HLM. The question is whether a network sampling approach can work beyond measures of cohesion. We consider average betweenness, average closeness, transitivity, and proportion isolated. Average betweenness is measured by calculating betweenness centrality for every actor in the network (either the true or the inferred) and taking the

Table D1. Summary of True and Estimated Sampling Distribution for Coefficient on Alternative Network Measures: 35 Percent Sample Using Add Health Data

		Average Betweenness		Average Closeness		Transitivity		Proportion Isolated	
		True	Ego Network	True	Ego Network	True	Ego Network	True	Ego Network
Attachment[b]	Mean	–3.77E-05	–3.83E-05	1.035	1.189	.42	.41	–1.48	–1.418
	Median	–3.74E-05	–3.81E-05	1.026	1.18	.418	.436	–1.461	–1.405
	S.D.	5.97E-06	6.54E-06	.128	.148	.169	.26	.147	.18
	95% CI[a]	(–4.95E-5, –2.75E-05)	(–5.11E-5, –2.70E-05)	(.797, 1.278)	(.924, 1.508)	(.062, .74)	(–.068, .913)	(–1.751, –1.255)	(–1.787, –1.122)
Behavioral problems[b]	Mean	–3.37E-05	–3.72E-05	–1.133	–.771	–.33	–.436	1.623	1.566
	Median	–3.29E-05	–3.68E-05	–1.128	–.768	–.31	–.427	1.611	1.541
	S.D.	6.82E-06	7.46E-06	.171	.172	.203	.237	.23	.262
	95% CI[a]	(–4.81e-5, –2.09e-5)	(–5.28e-5, 2.37e-5)	(–1.472, –.79)	(–1.149, –.451)	(–.758, .026)	(–.961, –.073)	(1.176, 2.082)	(1.124, 2.143)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.

[b]Includes only network feature of interest in the model.

mean over those values; betweenness centrality captures the number of shortest paths going through a specific vertex. Average closeness does the analogous calculation for closeness, on the basis of the inverse of the mean geodesic distance between a given node and all other nodes. Transitivity calculates the proportion of two paths (from *i* to *j* to *k*) that also have a direct path (from *i* to *k*). Proportion isolated is measured as the number of students in the school with no social connections divided by the number of students in the school. For this supplementary analysis, we consider only model 1 (no controls) for the 35 percent sample. This simplifies the discussion while still offering evidence on measures beyond cohesion. Table D1 shows results for attachment and behavioral problems. See the main text for a discussion of the results.

APPENDIX E

This appendix describes a supplementary analysis in which we replicate our analysis using a simulated data set. The analysis follows the same basic form as with the Add Health data, but here we construct the networks and data used in the test to have desired, known properties.

Methods

The first step is to generate the networks used in the test. We vary three key inputs when generating the networks: proportion isolated, mean degree, and level of transitivity (capturing how often a tie between *i* and *k* exists when there is a tie between *i*-*j* and *j*-*k*). There are 3 levels of isolation (0 percent, 5 percent, and 10 percent) $\times$ 3 levels of mean degree (4, 6, and 8) $\times$ 4 levels of transitivity (.04, .10, .16, and .22), yielding 36 networks. We vary these three inputs as a means of generating networks with systemically different global features, particularly in terms of percentage bicomponent size and density. Importantly, many networks with the same density differ on bicomponent size, and many networks with the same bicomponent size differ on density. Density and percentage bicomponent are correlated at only .40 across networks. This makes the test a difficult one, as simply getting the number of edges correct would be insufficient to properly estimate bicomponent size, and thus the effect of bicomponent size on other outcomes (this is also true of the test using Add Health data, for which the correlation between density and bicomponent size is similarly moderate). All networks range between 1,200 and 2,000 nodes. Nodes in the generated network are seeded with attributes (i.e., gender, race, and

education) to be consistent with the U.S. adult population (based on General Social Survey data).

The second step is to specify the HLM for the test. Here we use two constructed variables, with known relationships with bicomponent size and density. The first is based on a substantive case in which the outcome of interest is dependent on bicomponent size but not density. The variable is constructed as

$$Y_i = 1 * percentbicomp + 0 * density + \varepsilon_i,$$

where ε_i is a draw from a normal distribution with mean 0 and standard deviation .15. The true coefficient for percent bicomponent is thus 1. We denote this variable as "bicomponent correlated variable." The second variable is based on a case where the outcome of interest is solely dependent on density:

$$Y_i = 0 * percentbicomp + 50 * density + \varepsilon_i.$$

The true coefficient for density is thus 50 (it must be set at a higher value than with bicomponent size given the scale of the variable). We denote this variable as "density correlated variable." In short, each measure of cohesion is related to only one of the constructed variables, making it easy to isolate the effect of bicomponent size or density on the outcome of interest.

These two variables become the dependent variable in the HLMs used to test the approach. As before, the HLMs are specified with actors nested in networks, with a random intercept included. For the first outcome ("bicomponent correlated"), the main predictor in the model is bicomponent size. No other variables are necessary as the outcome is only a function of bicomponent size. We also show results when density is included as a control. The second outcome ("density correlated") has the opposite structure, where density is included as the main predictor of interest and bicomponent size is included as a control with a known effect of zero.

The rest of the test is analogous to the first case study, in which the question is whether the sampled data are sufficient to estimate the true coefficients. For each network, we take 100 samples at three sampling rates (15 percent, 25 percent, and 35 percent). The sampled actors are used to infer the values for bicomponent size and density. The inferential process is the same as above, with similar assumptions about the ego network data available to the researcher. We then estimate the HLMs on the sampled data, once using the inferred values of bicomponent size and density, and once using the true, known values for bicomponent size and density. The results are presented as a comparison between the inferred estimates and the true values.

Results

Tables E1 and E2 present the main results. The tables compare the true estimates to the ego network-inferred estimates, analogous to Tables A1 to A4. In this case the true coefficients for bicomponent size and density are known for each dependent variable. It is thus unnecessary to include a separate table and discussion for the true results. Table E1 shows results for the bicomponent correlated variable. The table reports the estimates on the bicomponent size coefficient for two models. Model 1 includes only bicomponent size as a predictor, and model 2 includes bicomponent size and density. Table E2 shows results for the density correlated variable and estimates for the density coefficient (one model just including density as a predictor and another including density and bicomponent size).

We begin with the bicomponent correlated variable. Looking at model 1 in Table E1 (just bicomponent size in the model), we see that the ego network–based estimates closely approximate the true coefficient for bicomponent size, 1. For the 25 percent sample, for example, the mean estimate (over all samples) is 1.009; 95 percent of estimates fall between .937 and 1.083 for the ego network–based estimate. The analogous interval is (.983, 1.02) using the true values for bicomponent size. This suggests the ego network–based models approximate the true coefficients, but the variance is much higher across samples than when using the true bicomponent size. The 35 percent sample offers somewhat more certain estimates (the standard deviation decreases from .036 to .032), but we still see similar estimates, with 95 percent of coefficients falling between .964 and 1.081. The results are also similar for the 15 percent sample, with a mean estimate of .973, although the bias is a bit higher here (2.7 percent). Model 2 includes density as a predictor. Here the estimates for bicomponent size are slightly inflated compared with model 1 (we are including two moderately correlated terms that are both measured imprecisely), but the results are still quite good, with a mean coefficient of 1.021 for the 25 percent sample. The model also accurately captures the effect for density on the dependent variable (not reported in the table). Density should itself have no effect, and this is what the results suggest: 95 percent of density coefficients fall between −9.262 and 6.918, a wide range that includes both positive and negative effects. This suggests density is estimated with relatively high uncertainty, but a researcher would still correctly conclude that density was unrelated to the outcome variable.

Table E2 reports analogous results for the density correlated variable. Here we focus on the coefficient for the density measure of cohesion. The true

Table E1. Summary of True and Estimated Sampling Distribution for Coefficient on Bicomponent Size, Bicomponent Correlated Variable: Simulated Test Case

		15% Sample		25% Sample		35% Sample	
		True	Ego Network	True	Ego Network	True	Ego Network
Model 1[b]	Mean	1.004	.973	1.002	1.009	1.001	1.024
	Median	1.004	.97	1.002	1.008	1.002	1.022
	S.D.	.012	.045	.009	.036	.008	.032
	95% CI[a]	(.98, 1.029)	(.889, 1.061)	(.983, 1.02)	(.937, 1.083)	(.987, 1.016)	(.964, 1.081)
Model 2[c]	Mean	1.005	.976	1.001	1.021	1.001	1.033
	Median	1.004	.976	1.001	1.021	1.002	1.035
	S.D.	.013	.051	.01	.043	.008	.042
	95% CI[a]	(.98, 1.029)	(.88, 1.081)	(.982, 1.021)	(.935, 1.106)	(.985, 1.017)	(.942, 1.111)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.

[b]Includes only bicomponent size as a predictor.

[c]Includes bicomponent size and density as predictors.

Table E2. Summary of True and Estimated Sampling Distribution for Coefficient on Density, Density Correlated Variable: Simulated Test Case

		15% Sample		25% Sample		35% Sample	
		True	Ego Network	True	Ego Network	True	Ego Network
Model 1[b]	Mean	50.337	49.207	50.084	49.272	50.095	49.685
	Median	50.298	49.2	50.034	49.224	50.124	49.694
	S.D.	1.294	2.065	.839	1.426	.685	1.207
	95% CI[a]	(47.662, 52.9)	(45.256, 53.246)	(48.525, 51.736)	(46.75, 52.535)	(48.693, 51.444)	(47.475, 52.161)
Model 2[c]	Mean	50.238	49.484	50.036	49.406	50.074	49.755
	Median	50.269	49.334	49.989	49.358	50.076	49.696
	S.D.	1.354	2.47	.936	1.757	.746	1.465
	95% CI[a]	(47.559, 52.9)	(44.712, 54.36)	(48.378, 52.035)	(46.164, 53.409)	(48.581, 51.45)	(47.234, 52.798)

Note: CI = confidence interval.

[a]Reports the 95% quantile of the sampling distribution, such that 95% of estimates fall within that interval.

[b]Includes only bicomponent size as a predictor.

[c]Includes bicomponent size and density as predictors.

coefficient should be 50, and we see in model 1 that the ego network–based coefficients have a mean of 49.272 (for the 25 percent sample), a difference of 1.5 percent compared with the true value; 95 percent of estimates fall between 46.75 and 52.535. Again, the coefficient for the main predictor (here density) increases slightly when we add controls for the other measure of cohesion (bicomponent size). The mean coefficient increases from 49.272 to 49.406 for the 25 percent sample. The results are similar for the 15 percent and 35 percent samples, with mean estimates for the density coefficient of 49.484 and 49.755 (for model 2). The ego network–based model also correctly captures the null effect of bicomponent size on the outcome of interest, with 95 percent of estimates falling between –.045 and .034 (looking at the 25 percent sample).

Acknowledgments

We would like to thank Sela Harcey and Julia McQuillian for helpful comments on earlier versions of this article. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (grant P20 GM130461) and the Rural Drug Addiction Research Center at the University of Nebraska–Lincoln. We would also like to thank the HAAS faculty award program at the University of Nebraska–Lincoln for providing financial support during the writing of this article.

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (grant P20 GM130461) and the Rural Drug Addiction Research Center at the University of Nebraska–Lincoln. We would also like to thank the HAAS faculty award program at the University of Nebraska–Lincoln for providing financial support during the writing of this article.

Data Note

This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

ORCID iD

Jeffrey A. Smith https://orcid.org/0000-0003-1847-1858

Notes

1. HLM estimation is a more difficult problem because a researcher must rely on estimates from a single sample, taken over all contexts, to act as predictors in the model, rather than use the mean network estimate over many samples (almost always the item of interest in past work). Past results on network features are thus no guarantee that an HLM can be properly estimated.
2. We use schools larger than 400 because schools smaller than 400 are unlikely candidates for a sampling strategy. In a school of 200, for example, a researcher's best option is to collect full network data. Sampling would yield too few absolute cases to make inferences (i.e., only 20 people in a 10 percent sample), and complete network data are easily collected under such circumstances.
3. Ego network surveys often cap the number of alters respondents report on to limit respondent fatigue (see Burt 1984).
4. Estimates for bicomponent size are accurate in most cases. On average, the (absolute) difference between the true value and the sample estimate is less than 2 percent.

References

Almquist, Zack W. 2012. "Random Errors in Egocentric Networks." *Social Networks* 34(4):493–505.

Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*, Vol. 547. Hoboken, NJ: John Wiley.

Anderson, Brigham S., Carter Butts, and Kathleen Carley. 1999. "The Interaction of Size and Density with Graph-Level Indices." *Social Networks* 21(3):239–67.

Bearman, Peter S. 1991. "The Social Structure of Suicide." *Sociological Forum* 6(3): 501–24.

Berkman, Lisa F., Thomas Glass, Ian Brissette, and Teresa E. Seeman. 2000. "From Social Integration to Health: Durkheim in the New Millennium." *Social Science & Medicine* 51(6):843–57.

Browning, Christopher R., and Kathleen A. Cagney. 2002. "Neighborhood Structural Disadvantage, Collective Efficacy, and Self-Rated Physical Health in an Urban Setting." *Journal of Health and Social Behavior* 43(4):383–99.

Burt, Ronald S. 1984. "Network Items and the General Social Survey." *Social Networks* 6(4):293–339.

Butts, Carter T. 2008. "Social Network Analysis: A Methodological Introduction." *Asian Journal of Social Psychology* 11(1):13–41.

Entwisle, Barbara. 2007. "Putting People into Place." *Demography* 44(4):687–703.

Entwisle, Barbara, Katherine Faust, Ronald R. Rindfuss, and Toshiko Kaneda. 2007. "Networks and Contexts: Variation in the Structure of Social Ties." *American Journal of Sociology* 112(5):1495–533.

Feehan, Dennis M., and Matthew J. Salganik. 2016. "Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations." *Sociological Methodology* 46(1):153–86.

Feld, Scott L., and William C. Carter. 2002. "Detecting Measurement Bias in Respondent Reports of Personal Networks." *Social Networks* 24(4):365–83.

Frank, Ove. 1971. "Statistical Inference in Graphs." PhD dissertation, Stockholm University, Stockholm, Sweden.

Frank, Ove. 1978. "Sampling and Estimation in Large Social Networks." *Social Networks* 1(1):91–101.

Friedkin, Noah E. 2004. "Social Cohesion." *Annual Review of Sociology* 30:409–25.

Gjoka, Minas, Emily Smith, and Carter Butts. 2014. "Estimating Clique Composition and Size Distributions from Sampled Network Data." Pp. 837–42 in *2014 IEEE Conference on Computer Communications Workshops* (INFOCOM WKSHPS). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Gottfredson, Denise C., and M. DiPietro Stephanie. 2010. "School Size, Social Capital, and Student Victimization." *Sociology of Education* 84(1):69–89.

Granovetter, Mark. 1976. "Network Sampling: Some First Steps." *American Journal of Sociology* 81(6):1287–1303.

Handcock, Mark S., and Krista J. Gile. 2010. "Modeling Social Networks from Sampled Data." *Annals of the Applied Statistics* 4:5–25.

Handcock, Mark S., and Krista J. Gile. 2011. "Comment: On the Concept of Snowball Sampling." *Sociological Methodology* 41(1):367–71.

Harris, K. M., C. T. Halpern, E. Whitse, J. Hussey, J. Tabor, P. Entzel, and J. R. Udry. 2009. "The National Longitudinal Study of Adolescent to Adult Health: Research Design." Retrieved May 1, 2020. http://www.cpc.unc.edu/projects/addhealth/design.

Hipp, John R., and Andrew Perrin. 2006. "Nested Loyalties: Local Networks' Effects on Neighbourhood and Community Cohesion." *Urban Studies* 43(13):2503–23.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steve M. Goodreau, and Martina Morris. 2008. "Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24(3):1–29.

Illenberger, Johannes, and Gunnar Flötteröd. 2012. "Estimating Network Properties from Snowball Sampled Data." *Social Networks* 34(4):701–11.

Ivory, Vivienne C., Sunny C. Collings, Tony Blakely, and Kevin Dew. 2011. "When Does Neighbourhood Matter? Multilevel Relationships between Neighbourhood Social Fragmentation and Mental Health." *Social Science & Medicine* 72(12): 1993–2002.

Koskinen, Johan H., Garry L. Robins, and Philippa E. Pattison. 2010. "Analysing Exponential Random Graph (P-Star) Models with Missing Data Using Bayesian Data Augmentation." *Statistical Methodology* 7(3):366–84.

Krivitsky, Pavel N., Mark S. Handcock, and Martina Morris. 2011. "Adjusting for Network Size and Composition Effects in Exponential-Family Random Graph Models." *Statistical Methodology* 8(4):319–39.

Krivitsky, Pavel N., and Martina Morris. 2017. "Inference for Social Network Models from Egocentrically Sampled Data, with Application to Understanding Persistent Racial Disparities in HIV Prevalence in the US." *Annals of Applied Statistics* 11(1): 427–55.

Legewie, Joscha, and Thomas A. DiPrete. 2012. "School Context and the Gender Gap in Educational Achievement." *American Sociological Review* 77(3):463–85.

Maimon, David, and Danielle C. Kuhl. 2008. "Social Control and Youth Suicidality: Situating Durkheim's Ideas in a Multilevel Framework." *American Sociological Review* 73(6):921–43.

Marin, Alexandra. 2004. "Are Respondents More Likely to List Alters with Certain Characteristics? Implications for Name Generator Data." *Social Networks* 26(4): 289–307.

Marsden, Peter V. 1990. "Network Data and Measurement." *Annual Review of Sociology* 16(1):435–63.

Marsden, Peter V. 2011. "Survey Methods for Network Data." Pp. 370–88 in *The Sage Handbook of Social Network Analysis*, edited by J. Scott and P. J. Carrington. London: Sage Ltd.

Mayhew, Bruce H., and Roger L. Levinger. 1976. "Size and the Density of Interaction in Human Aggregates." *American Journal of Sociology* 82(1):86–110.

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.

McFarland, Daniel A., James Moody, David Diehl, Jeffrey A. Smith, and Reuben J. Thomas. 2014. "Network Ecology and Adolescent Social Structure." *American Sociological Review* 79(6):1088–1121.

McPherson, Miller, Lynn Smith-Lovin, and Matthew Brashears. 2006. "Social Isolation in America: Changes in Core Discussion Networks over Two Decades." *American Sociological Review* 71(3):353–75.

McPherson, Miller, and Jeffrey A. Smith. 2019. "Network Effects in Blau Space: Imputing Social Context from Survey Data." *Socius*. Retrieved April 30, 2020. https://doi.org/10.1177/2378023119868591.

Moody, James. 2004. "The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999." *American Sociological Review* 69(2): 213–38.

Moody, James, and Douglas R. White. 2003. "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups." *American Sociological Review* 68(1): 103–27.

Morris, Martina, and Mirjam Kretzschmar. 2000. "A Micro-simulation Study of the Effect of Concurrent Partnerships on HIV Spread in Uganda." *Mathematical Population Studies* 8(2):109–33.

Morris, Martina, Anne E. Kurth, Deven T. Hamilton, James Moody, and Steve Wakefield. 2009. "Concurrent Partnerships and HIV Prevalence Disparities by Race: Linking Science and Public Health Practice." *American Journal of Public Health* 99(6):1023–31.

Pattison, Philippa E., Garry L. Robins, Tom A. B. Snijders, and Peng Wang. 2013. "Conditional Estimation of Exponential Random Graph Models from Snowball Sampling Designs." *Journal of Mathematical Psychology* 57(6):284–96.

Pescosolido, Bernice A. 2006. "Of Pride and Prejudice: The Role of Sociology and Social Networks in Integrating the Health Sciences." *Journal of Health and Social Behavior* 47(3):189–208.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Vol. 1. Thousand Oaks, CA: Sage.

Robins, Garry, Philippa Pattison, and Jodie Woolcock. 2005. "Small and Other Worlds: Global Network Structures from Local Processes." *American Journal of Sociology* 110(4):894–936.

Robins, Garry, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. 2007. "Recent Developments in Exponential Random Graph (P*) Models for Social Networks." *Social Networks* 29(2):192–215.

Rolls, David A., and Garry Robins. 2017. "Minimum Distance Estimators of Population Size from Snowball Samples Using Conditional Estimation and Scaling of Exponential Random Graph Models." *Computational Statistics & Data Analysis* 116:32–48.

Sampson, Robert J., Stephen W. Raudenbush, and Felton Earls. 1997. "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy." *Science* 277(5328): 918.

Sharkey, Patrick, and Jacob W. Faber. 2014. "Where, When, Why, and for Whom Do Residential Contexts Matter? Moving Away from the Dichotomous Understanding of Neighborhood Effects." *Annual Review of Sociology* 40(1):559–79.

Smith, Jeffrey A. 2012. "Macrostructure from Microstructure: Generating Whole Systems from Ego Networks." *Sociological Methodology* 42(1):155–205.

Smith, Jeffrey A. 2015. "Global Network Inference from Ego Network Samples: Testing a Simulation Approach." *Journal of Mathematical Sociology* 39(2):125–62.

Smith, Jeffrey A., and Jessica Burow. 2018. "Using Ego Network Data to Inform Agent-Based Models of Diffusion." *Sociological Methods & Research*. doi:10.1177/0049124118769100.

Smith, Jeffrey A., and Robert Faris. 2015. "Movement without Mobility: Adolescent Status Hierarchies and the Contextual Limits of Cumulative Advantage." *Social Networks* 40:139–53.

Smith, Jeffrey A., Miller McPherson, and Lynn Smith-Lovin. 2014. "Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004." *American Sociological Review* 79(3):432–56.

Smith, Jeffrey A., James Moody, and Jonathan H. Morgan. 2017. "Network Sampling Coverage II: The Effect of Non-random Missing Data on Network Measurement." *Social Networks* 48:78–99.

Snijders, Tom A. B., Philippa Pattison, Garry L. Robins, and Mark S. Handcock. 2006. "New Specifications for Exponential Random Graph Models." *Sociological Methodology* 36:99–153.

Stivala, Alex D., Johan H. Koskinen, David A. Rolls, Peng Wang, and Garry L. Robins. 2016. "Snowball Sampling for Estimating Exponential Random Graph Models for Large Networks." *Social Networks* 47:167–88.

Thompson, Steven K., and Ove Frank. 2000. "Model-Based Estimation with Link-Tracing Sampling Designs." *Survey Methodology* 26:87–98.

Tulin, Marina, Thomas V. Pollet, and Nale Lehmann-Willenbrock. 2018. "Perceived Group Cohesion versus Actual Social Structure: A Study Using Social Network Analysis of Egocentric Facebook Networks." *Social Science Research* 74:161–75.

Verdery, Ashton M., Jacob C. Fisher, Nalyn Siripong, Kahina Abdesselam, and Shawn Bauldry. 2017. "New Survey Questions and Estimators for Network Clustering with Respondent-Driven Sampling Data." *Sociological Methodology* 47(1):274–306.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Wasserman, Stanley, and Philippa Pattison. 1996. "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and P*." *Psychometrika* 61:401–25.

Wray, Matt, Cynthia Colen, and Bernice Pescosolido. 2011. "The Sociology of Suicide." *Annual Review of Sociology* 37(1):505–28.

Author Biographies

Jeffrey A. Smith is an associate professor in the Department of Sociology at the University of Nebraska–Lincoln. His research explores methodological and substantive problems related to networks, individuals, and the broader social context. He has done methodological work on network sampling and missing data, developing new techniques to make inference about network properties from incomplete data. He is currently working on a National Institutes of Health–funded project that will study the risk factors associated with rural drug use, looking at the social network dynamics and behavioral contexts that contribute to overdose risk and increased risk for addiction.

G. Robin Gauthier is an assistant professor in the Department of Sociology at the University of Nebraska–Lincoln. She has studied relational patterns in adolescent friendship, with a substantive focus on how the structure of peer networks combine with local gender norms to shape adolescent identity, behavior, and well-being. In her most recent work, she is exploring patterns of conflict in family networks with applications for emotional closeness and social support.