Spring 2015

Global Network Inference from Ego Network Samples: Testing a Simulation Approach

Jeffrey A. Smith
University of Nebraska–Lincoln, jsmith77@unl.edu

Published in *The Journal of Mathematical Sociology* 39:2 (2015), pp. 125-162;
doi: 10.1080/0022250X.2014.994621

Published online March 31, 2015

Global Network Inference from Ego Network Samples: Testing a Simulation Approach

Jeffrey A. Smith

Department of Sociology, University of Nebraska–Lincoln, Lincoln, Nebraska, USA

Abstract
Network sampling poses a radical idea: that it is possible to measure global network structure without the full population coverage assumed in most network studies. Network sampling is only useful, however, if a researcher can produce accurate global network estimates. This article explores the practicality of making network inference, focusing on the approach introduced in Smith (2012). The method uses sampled ego network data and simulation techniques to make inference about the global features of the true, unknown network. The validity check here includes more difficult scenarios than previous tests, including those that go beyond the initial scope conditions of the method. I examine networks with a skewed degree distribution and surveys that limit the number of social ties a respondent can list. For each network/survey combination, I take a random ego network sample, run the simulation method, and compare the results to the true values (using measures of connectivity and cohesion). I also test the method on local measures of network structure. The results, on the whole, are encouraging. The method produces good estimates even in cases where the degree distribution is skewed and the survey is strongly restricted. I also find that is it better to not truncate the survey if possible. If the survey must be restricted, the researcher would do well to infer the missing data, rather than use the raw data naively.

Keywords: exponential random graph model (ERGM), simulation, social networks

1. Introduction

Network studies have traditionally been restricted to relatively small, bounded settings with full information on the population of interest. Network structure is easily measured under these ideal conditions, opening up questions about connectivity (i.e., disease/information spread), group divisions and cohesion (Moody & White, 2003; Shwed & Bearman, 2010; Mucha, Richardson, Macon, Porter, & Onnela, 2010; Adams, Moody, & Morris, 2013). In many cases, however, it is impractical to collect full network data on the population of interest (Frank, 1971; Koskinen, Robins, Wang, & Pattison, 2013). The network may be too large and the resources available too small to interview everyone, while electronic information on social ties may be unavailable. This is especially true of comparative network studies. A study on social integration, for example, would need to measure network cohesion across many neighborhoods, villages, or schools (Sampson & Raudenbush, 1997; Browning, Cagney, & Iveniuk, 2012; Verdery, Entwisle, Faust, & Rindfuss, 2012). Traditionally, a researcher would have to collect full network data in every context, a tall order by any survey method standard.

A small but growing literature on network sampling has raised the possibility of making inference on network structure without the full coverage assumed in most network studies (Frank, 1978; Handcock & Gile, 2010; Koskinen, Robins, & Pattison, 2010; Krivitsky, Handcock, & Morris, 2011). Using sampled network data eases the burden of data collection considerably but also raises difficult inference problems. Global network measures depend on all of the ties among individuals while sampled data, by definition, only provide sampled bits of the network. As a solution to this inferential problem, recent work has offered simulation techniques as a means of making inference from sampled network data (Lee, 2008; Morris, Kurth, Hamilton, Moody, & Wakefield, 2009; Smith, 2012; Merli et al., 2015).

This article examines the practical application of simulation to network inference. I focus on the method proposed in Smith (2012). The method uses ego network data to make inference about global network structure. Individuals are first randomly sampled from the population; they then answer questions about themselves (e.g., demographic information) and the people they are socially connected to, such as friends or confidants (Marsden, 1987; Smith, McPherson, & Smith-Lovin, 2014). Respondents also report on the ties between their named alters. The method uses this data to simulate networks consistent with the sampled, local information. For example, the survey yields an estimate of the degree distribution, or the number of ties per person, and the simulation is conditioned on this information. Finally, the method calculates the global statistics of interest on the generated networks, such as average distance, modularity, or cohesion. The approach has great potential as independently sampled data are easy to collect and can be incorporated into existing surveys.

There is also empirical reason to be optimistic, as the method performed quite well in a series of validity tests (Smith, 2012). These tests, while employing a large number of networks, adhered strongly to the initial scope conditions of the method. The original tests were limited to networks based on strong tie relationships, where individuals have a relatively small number of social ties and can easily enumerate them within an ego network survey. Based on this strong tie assumption, the tests were limited to low-degree networks and assumed full information about respondents' social contacts.

The goal of this article is to apply the approach to a wider set of contexts, even if those contexts fall outside the bounds initially specified by the method. For example, ego network sampling may yield biased estimates when the network has a skewed degree distribution so that a few individuals have a disproportionally large number of ties (Barabasi & Albert, 1999; Gould, 2002). High degree individuals have a large impact on network structure but are not any more likely to be sampled (as it is a random sample of the population). Thus, the sampling scheme may miss important nodes in highly skewed networks, leading to potential biases.

Additionally, surveys will often restrict the amount of information that is collected. A typical survey will limit the number of social contacts a respondent can name. This is done for practical reasons, for example, to reduce respondent fatigue and the cost of the survey (Burt, 1984). This will result in a truncated degree distribution, however. Even if one is lucky

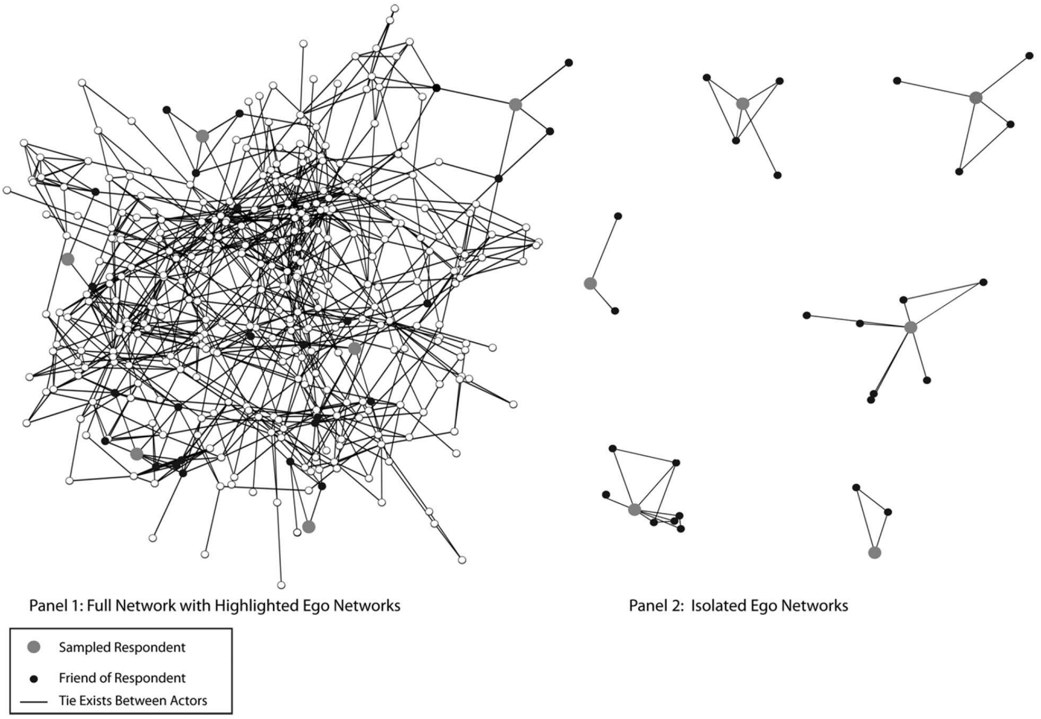
Panel 1: Full Network with Highlighted Ego Networks
Panel 2: Isolated Ego Networks
Sampled Respondent
Friend of Respondent
Tie Exists Between Actors

Figure 1. Example network and ego network sample.

enough to sample a high degree node, one would still have an inaccurate measure of their degree—as there is only information up to the truncated amount (say, 10 ties).

Here, I drop the simplifying assumptions of previous tests and show how degree skew and degree truncation affect the bias in the estimates. Is the method appropriate for all levels of degree skew and survey truncation? More generally, what should a researcher do when faced with less than ideal conditions? I begin the article with a background section on ego network sampling and the simulation approach. I then discuss the experimental setup used to test the method, before moving to the results.

2. Ego Network Sampling

Figure 1 summarizes the problem of network sampling. Panel 1 in Figure 1 plots a typical (hypothetical) network structure. We can assume that the network represents friendships among adolescents in a school. A researcher would normally collect information on all nodes, here adolescents, and all ties between nodes in the network. This information can be used to characterize the topology of the network. For example, the paths between nodes are easy to enumerate when the network is complete. Substantively, the path structure yields a measure of diffusion potential, where lower average distance and higher levels of connectivity equals higher probabilities of global diffusion (Watts, 2002; Centola & Macy, 2007).

But what if it is impractical to collect information on every individual in the network? In such cases, it is necessary to make inference from a sample. There are many ways to sample a network, such as snowball and subgraph approaches (Thompson & Frank, 2000; Goodman, 2011), but I focus on the simplest possible option, the ego network sample (Marsden, 1988).

In Panel 1, I have highlighted a hypothetical ego network sample from the full network. The grey nodes represent the randomly sampled respondents from the set of all individuals in the network. The white nodes represent the nonsampled respondents, and the black nodes are the friends of the respondents. The black nodes (here the friends of the respondents) are not interviewed, but we receive information about them indirectly through the respondent's reports on them. For example, we would know if the friends of the respondents are themselves friends (Louch, 2000). The sample thus offers information on the grey nodes and the black nodes, leaving independent bits of the network. These sampled substructures are plotted separately in Panel 2. The sampled parts of the network cannot be connected, as ego network surveys do not collect identifying information on the named friends (i.e., the black nodes).

Ego network sampling poses more than a missing data problem (Kossinets, 2006; Borgatti, Carley, & Krackhardt, 2006; Smith & Moody, 2013). It is a problem of statistical inference, where almost all of the network information must be "filled in." The question is how we can take information on the respondents and their friends, or these disconnected, representative pieces of the network, and make inference about the entire graph (as one cannot simply trace out the paths between nodes anymore).

2.1. Background on Simulation Approach

Smith (2012) provides a simulation solution to the problem of network inference. The method takes independently sampled ego network data, like that in Figure 1, and makes inference about the entire network structure.[1] The simulation approach rests on a simple premise: that one should generate networks consistent with the local information found in the sampled data. A simulated network consistent with the local information should have similar macro features as the real network. The method ultimately works well because it utilizes so much of the information embedded in the network survey.

As with most surveys, an ego network sample will provide information about the demographic characteristics of the respondents, or the sampled grey nodes in our example network. Individuals may be asked about their gender, race/ethnicity, education, age, and so forth. This information is useful in the simulation because it yields the demographic composition of the network.

An ego network survey will also ask respondents to name their alters, or those individuals to whom they are socially connected. The alters are defined as friends in Figure 1. The

[1] It is important to note that the method is only appropriate for well defined populations with a sampling frame. The population of interest is thus assumed to be nonhidden (i.e., not female sex workers or drug injectors), and the size of the population is assumed to be known. The method also assumes that the relationship of interest is symmetric, so that if *i* nominates *j*, then *j* nominates *i*.

node on the top right of Panel 2 would list four alters, or have degree 4, while the node on the bottom right would list two alters, or have degree 2. This yields an estimate of the degree distribution, equal to the number of alters per respondent. The alter list also offers information on differential degree, or the mean degree by demographic group. This can be inferred from the data because there is information on degree and the demographic characteristics of the sampled respondents. The highly educated may have more ties than those with less education (e.g., Lizardo, 2006; McPherson, Smith-Lovin, & Brashears, 2006).

Ego network samples also provide information on the demographic characteristics of the alters. For example, for the top-right respondent, a researcher may ask about the race, gender, and age of the four friends. Respondents are often asked to report on only a subset of their alters. A respondent may list 15 friends but only be asked about the demographic characteristics of the first five (e.g., Marsden, 1987). This is done to limit respondent burden.[2] The alter demographic data are important for the simulation because they offer information on homophily, or the tendency for demographically similar individuals to be socially connected (McPherson, Smith-Lovin, & Cook, 2001; Smith et al., 2014).[3] The data show if respondents and their alters have the same race/gender/age/etc.; we can thus ask if there is strong homophily along religious lines, for example (Cheadle & Schwadel, 2012). The data also capture the pattern of ties among demographic groups (Rosenfeld, 2008). Thus, we can ask if ties between Whites and Asians are more likely than ties between Whites and Blacks (Qian, 2002; Qian & Lichter, 2007).

Finally, ego network data provide information on the ties between alters. Respondents are asked about the ties that exist between alter 1 and 2, 1 and 3, 2 and 3, and so on. Again, the surveys are often limited to curtail respondent burden. A respondent may be asked about a subset of all ties that exist between alters (i.e., the ties that exist among the first five friends). For our top-right respondent, we would know that two of their friends are tied together; we would also know that these two friends are not tied to the other two friends, who are not tied to each other. The alter–alter tie data offer information about the local structural patterns in the network. Are individuals tied to the respondent also tied to each other, so a friend of a friend (the respondent) is also a friend (Goodreau, Kitts, & Morris, 2009)?

One of the contributions of Smith (2012) was its unique characterization of the alter–alter tie data. Past work has generally relied on density (the number of ties amongst alters divided by the total number of possible ties) to measure the structure of ego networks (Fischer, 1982; Mardsen, 1987; although see Louch, 2000, for an exception).[4] Unfortunately, this does not offer a precise enough measure for the purposes of the simulation: many networks with the same local density have very different global structures (Smith, 2012). As

[2] It can be quite tedious to describe the demographic characteristics of many alters along many demographic dimensions.

[3] One can estimate the strength of homophily as one knows the characteristics of the respondents and the respondents' alters.

[4] This is largely because ego network data provide biased estimates for many typical triadic measures; such as global transitivity, defined as the proportion of two-step paths where there is also a one-step path (Soffer & Vazquez, 2005; Bansal, Khandelwal, & Meyers, 2009). Thus, for our top-right respondent, there is one tie out of a possible six.

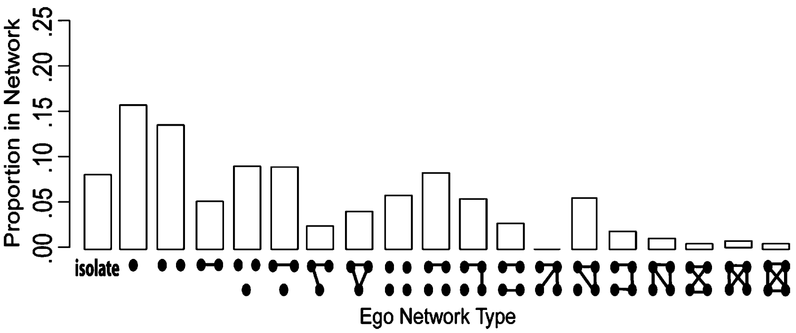

Figure 2. Example ego network configuration distribution. *Note*: This figure is a based on a hypothetical ego network configuration distribution. Ego is not included in the ego network types. I only include ego network types of size four or less to make the figure legible.

the networks are generated from the local information, it is important to have a discerning measure of local structure.

Smith (2012) offered an alternative measure of ego network structure, one that encapsulates all of the information available from the sampled data. The basic idea is to form a distribution of ego network configurations from the alter–alter tie data (see Holland & Leinhardt, 1976, and Middendorf, Wiggins, & Honig, 2005, for similar intuition). Figure 2 plots a hypothetical ego network configuration distribution. The histogram is limited to four alters for the sake of space considerations, but the actual distribution is not limited by size. Each respondent is placed into a distinct structural type, based on the size of the ego network and the pattern of ties between alters. For example, our top-right respondent would fall into the 10th configuration from the left (four alters with one tie between them), while the top-left respondent would fall into the 11th configuration.

Formally, let Y_p be a square matrix of dimensions $m \times m$, consisting of the alters in the ego network of respondent p. Define $Y_{pij} = 1$ if a tie exists between alter i and j. The ego network configuration can then be defined by the unique combination of:

$$\begin{cases} 1.\ Size_p = m \\ 2.\ (d_i)_p = Y_{pi+}, \text{ where } d_i \text{ is the degree of alter } i \\ 3.\ T_p = \sum_t Y_{pij} * Y_{pjk} * Y_{pik}, \text{ where } t \text{ is set of all triads in } Y_p \end{cases}$$

A distribution of ego network configurations offers a more discriminating measure because it fully captures the structural features of the ego networks. Ego networks of the same size and density can exhibit very different structural *patterns*, but this is obscured using traditional measures. For example, there are four distinct configurations with four alters and three ties between alters. Substantively, the forces that constrain the real network, such as transitive closure (where a friend of a friend is a friend) will also constrain the ego network configurations, making it more likely that the simulated network will reflect the properties of the true network (as the simulated networks are conditioned on the ego network configurations).

The simulation generates full networks that are consistent with the information extracted from the ego networks: the degree distribution, differential degree, homophily, and the ego network configuration distribution. A network consistent with this local information should have similar global features as the true network—as the simulated networks are heavily constrained by the empirical data. The method draws on two models to simulate the networks, exponential random graph models (ERGMs) and case control logistic regression. I briefly discuss both models before moving to the details of the approach.

2.2. ERGM

ERGMs offer a statistical approach to modeling network data. The model form is quite general, and can be used to test hypotheses about network structure/formation (Holland & Leinhardt, 1981; Frank & Strauss, 1986; Wasserman & Pattison, 1996; Snijders, Pattison, Robins, & Handcock, 2006; Handcock, Goodreau, Hunter, Butts, & Morris, 2008). Formally, we can define a network, Y_{ij}, over the set of nodes N ($N = 1, 2, \ldots n$), such that $Y_{ij} = 1$ if a tie exists and 0 otherwise. Let **y** be the observed networks. **Y** is then a random graph on N, where each possible tie, ij, is a random variable. An ERGM will model the Pr(**Y** = **y**). The "independent variables" are counts of local structural features in the network (Robins, Snijders, Wang, Handcock, & Pattison, 2007; Goodreau et al., 2009), such as the volume of ties, homophily (e.g., the number of ties that match on race), or transitivity (e.g., the number of transitive triads). The model can be written as:

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\exp\left(\theta^{T} g(y)\right)}{\kappa(\theta)}$$

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

Typically, ERGMs are used to test hypotheses about the formation of the network, but it is also possible to simulate networks from an underlying model. The coefficients capture the effect of different local processes on the formation of the network. We can then use those coefficients to predict the presence of ties between individuals in a constructed network.

Prior to a simulation, a researcher must specify the model terms and coefficients used to generate the network. The model should be specified with the end goal in mind. In this case, the goal is to simulate networks consistent with the local network information. The model terms should reflect the information available from the ego network sample, while the coefficients should generate networks with the right local features.

2.3. Case Control Logistic Regression

The second model, case control logistic regression, is used to estimate the initial coefficients for the ERGM (Smith et al., 2014). The case control model is specifically used to estimate the coefficients for the homophily terms. The model is also used to adjust the homophily coefficients during the simulation (see Merli et al., 2015, for an example). Case control models are often used in medical research to study rare conditions that are difficult to capture through random sampling (Breslow & Day, 1980). The models compare the cases, a set of individuals with the "disease" (e.g., cancer), to the controls, a set of individuals without the "disease." A case control model is ideal for ego network data because the sample captures the rare event of interest, the social relationships between actors. The cases are all respondent–alter pairs, or those pairs with a known social relationship (the "disease" of interest). The controls represent a random sample of pairs that do not have a known social relationship. This is formed by randomly pairing the sampled respondents together, capturing random mixing in the population.

The case control model compares the cases to the controls on some behavior or condition of interest (e.g., smoking). Here, the condition of interest is the sociodemographic distance between individuals in a pair. For numerical variables, like age, this is measured as the absolute distance between *i* and *j*. For categorical variables, like race or religion, distance is measured as a matching term (are *i* and *j* the same race?) or a set of dummy terms, describing the rate of contact between all categories (e.g., what is the rate of social contact between Whites and Blacks, Whites and Asians, and so forth?). The sociodemographic distance between the respondents and alters, or the cases, is compared to the sociodemographic distance between randomly paired respondents, or the controls. The model thus compares the sociodemographic distance observed in the data to that expected under random mixing in the population (Smith et al., 2014). The model is a simple logistic regression, where the 1s are the respondent–alter pairs and the 0s are the random respondent pairings. Formally

$$\ln\left(\frac{p(Y)}{1 - p(Y)}\right) = \theta\mathbf{D}$$

where Y_{ij} is the presence or absence of a tie, D_{ij} is the sociodemographic distance between *i* and *j* for each dyad, and θ is the vector of coefficients.

The case control model is useful because of its flexible form: The controls are constructed independently from the cases, making it easy to alter the comparison. This makes the initial estimation straightforward, and, more importantly, makes it possible to update the coefficients during the simulation itself.

3. Methods Overview

The simulation approach draws on both models to generate networks consistent with the sampled information. I describe the approach here in some detail but see Smith (2012) for a complete discussion.

The method has three basic parts: first, summarizing the sampled data prior to the simulation; second, setting up the simulation; and third, simulating the full networks. The first part gathers information from the sampled ego networks, while the second and third parts generate full networks that are consistent with the local, sampled information. I assume, when describing the method, that the survey has collected demographic information on both the respondents and alters. I also assume that there is information on the number of alters and the ties between alters.

The method begins by calculating the degree distribution from the sampled data. The degree distribution is taken directly from the sampled data itself. Thus, the proportion of respondents with 0, 1, 2, etc., alters is used as the estimate of the degree distribution. This is a quite simple approach, but it is also possible to use more complicated, model-based methods when estimating the degree distribution (e.g., Zhang, Kolaczyk, & Spencer, 2014). Thus, one could estimate the generating function of the degree distribution, take draws from that model, and use that as the degree distribution for the network. This would serve to smooth out the distribution from the sample. In the case of a truncated survey design (e.g., where only 10 names are collected), it will be particularly important to use a model-based approach—as the sample degree distribution will not offer a good approximation of the true degree distribution.

As a second step, the method calculates the ego network configuration distribution from the sampled data (see formula and discussion above).

The next series of steps sets up the simulation. Here, the method first generates a network of size *N* with the correct degree distribution, already calculated from the ego network data.[5] The size of the population is assumed to be known.[6] The method then assigns demographic characteristics to the nodes in the simulated network. A sampled respondent is first selected at random; a node from the seeded network is then selected with the same degree as the sampled respondent. The selected node is assigned all of the characteristics of the selected respondent, such as race, gender and education. This maintains differential degree in the simulated network, as demographic groups with high average degree in the sample will also have high average degree in the network (as characteristics are assigned to nodes based on degree).[7] The initially simulated network will thus have the right size, degree distribution, demographic composition, and differential degree (where everything but size is based on the sampled data).

The method then specifies an ERG formula to simulate the full network from. The ERG formula specifies which local features are used to form the full network of interest. The terms in the model should reflect all of the local information available from the ego network

[5] This initial simulation can be done within an ERGM framework or using a stub-based algorithm (Newman, Strogatz, & Watts, 2001; Viger, Latapy, & Wang, 2005).

[6] See Pattison, Robins, Snijders, and Wang (2013) for approaches that do not require the size of the network to be known.

[7] It is important to note that the number of people in the simulated network is larger than the number of sampled respondents. This means that a sampled respondent may be seeded multiple times in the simulated network. This is unlikely to cause problems, however, as the network ties are probabilistically determined; thus, nodes in the simulated network with the exact same set of characteristics need not be tied together. Or, there is no definitional reason that a node seeded multiple times will have to be tied to herself. Any nodes with similar characteristics will have a high probability of being tied together. More substantively, it may be the case that many people have the same race and grade in a school; in which case the simulation does not deviate far from the empirical setting. Future work could, however, explicitly deal with this duplication of nodes by modeling how the characteristics go together, rather than simply drawing them from the data itself.

sample: differential degree, homophily, and the ego network configuration distribution. There are two steps here: first, determining which terms should be included; second, calculating the initial coefficients for those terms.

The model will include homophily terms for each demographic dimension available in the data. The homophily terms take the form of an absolute difference if they are continuous, like age. Thus, we would be interested in the absolute age difference between respondents and their alters. The method uses a mixing matrix to capture homophily if the terms are categorical, like race or gender. The mixing matrix reflects the number of social ties between each category. For example, the terms for race may include Black–Black, Black–Hispanic, White–White, and so forth, capturing the number of social ties between each racial group.

The model also includes a term for geometrically weighted edgewise shared partner (GWESP) distribution. GWESP counts the number of shared partners that *i* and *j* have (assuming *i* and *j* are themselves tied), or the number of common associates. GWESP substantively captures higher order transitivity in the network, or the tendency for local clusters to emerge. The GWESP term is designed to capture the ego network configuration distribution. GWESP is an appropriate choice as it mirrors many of the structural features of the ego networks. For example, the shared partner distribution in an ego network is the same as the degree distribution of the alters (from the point of view of the respondent), a key component in defining the ego network configurations.

The method then sets the initial coefficients for each term. The homophily coefficients are estimated using case control logistic regression (Smith et al., 2014). The GWESP coefficient, in contrast, cannot be easily set prior to the simulation; this is the case as it is not possible to analytically solve for the value that will yield the right ego network configuration distribution. Rather, GWESP is set at an initial value and is updated during the simulation as the method looks for a better fitting network.[8] The degree distribution and differential degree (specified while seeding the network) are also maintained throughout the simulation.

The framework takes the initial ERGM (coefficients, terms, and constraints) and simulates a network. The simulated network is then checked to make sure the homophily rates are correct. The homophily coefficients (and network) are updated if any discrepancies are found. The simulated networks may have incorrect homophily rates because the initial case control model only includes homophily terms. The initial homophily coefficients are thus unconditioned. The simulated networks, however, have a nonzero GWESP coefficient, making the homophily coefficients biased when simulating the networks.

Formally, the case control model is used to update the coefficients, comparing the true rates of homophily to those in the simulated network and adjusting accordingly. Coefficients are decreased if there are too many ties among groups and increased if there are too few. The method takes the tied dyads from the simulated network and the respondent–alter dyads from the sampled data and forms a combined dataset. This dataset includes the demographic characteristics of each person in the dyad. The dataset also includes a 0/1 variable: equal to 0 if the dyad comes from the simulated network and 1 if the dyad comes from

[8] The method calculates a starting value by estimating a dyadic independent ERGM on the ego networks. It is also possible to use other terms than GWESP.

the empirical sample. The method then runs a logistic regression, predicting the 1s as a function of the sociodemographic distance between individuals in the dyads. Substantively, the logistic regression compares the sociodemographic distance from the empirical sample (i.e., between respondents and alters) to the sociodemographic distance between tied individuals in the simulated network. For example, for categorical variables, a positive coefficient means that the simulated network has too few ties between those groups. The estimated coefficients are then added to the original homophily coefficients. This adjusts the original homophily coefficients up or down, depending on the bias in the simulated network. The network is then updated based on these new coefficients. (See Smith, 2012, for more formal details.)

The simulated network is then evaluated on how well it captures the ego network configuration distribution from the sampled data. The ego network configurations found in the simulated network are compared to the observed distribution in the sample using a chi-square value.[9] Specifically, the chi-square calculation has two parts: first, calculate the ego network configuration distribution from the simulated network, and second, compare the ego network configuration distribution from the simulated network to the distribution from the sampled data (calculated prior to the simulation). We can write the chi-square value as

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

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. The chi-square value will be large, and the fit poor, if is there is a large difference between the true distribution and the distribution from the simulated network.

The coefficient on GWESP is then updated to find a better fitting network, if possible. A better fitting network will have ego network configurations that match the true distribution more closely, condition on the other local features in the sampled data. As the simulation looks for a better fitting network, it is necessary to compare the homophily rates in the simulated network to the rates in the sampled data.

The ego network configuration distribution thus serves as the benchmark by which to judge the simulated networks. The question is what coefficient on GWESP will yield a network with the lowest chi-square value. The minimization process begins by simulating a sample of networks with different values for GWESP; the values are set above and below the initial value for GWESP. The method then adjusts the simulated networks for any homophily bias and calculates the chi-square value, comparing the true ego network configuration distribution to the distribution in the simulated network. The coefficients on GWESP and the chi-square values are then used to fit an OLS regression. The chi-square values are regressed on linear and quadratic terms of the GWESP coefficients. Formally

$$\chi_i^2 = \beta_0 + \beta_1 (G_i) + \beta_2 (G_i)^2$$

[9] A good fit means that the ego network configurations in the simulated network are found at the same rate as in the sampled data.

where X_i^2 is the chi-square value for network i, and G_i is the coefficient on GWESP for network i. The regression coefficients are then used as inputs into an optimization routine; the method uses the Nelder-Mead algorithm to find the GWESP coefficient that yields the lowest chi-square value. The coefficient that minimizes the equation is then used as the starting point for the next iteration of the simulation. This process is repeated until it is impossible to improve the fit by changing the GWESP coefficient and updating the homophily coefficients. In general, the simulation rests on a kind of approximated likelihood ratio test: The coefficients are updated to find a more likely network, where a network is considered more likely if the ego network configuration distribution more closely matches the true distribution from the sample.

In the end, the simulation will yield a network with the same local properties as the sampled ego network data: It will match the degree distribution, differential degree, homophily and the ego network configuration distribution. The researcher can then calculate the global statistics of interest on the generated network.

4. Analytical Setup: Testing the Method

The key question is how closely the method reproduces the true network. A test of the simulation approach has four steps: first, select a known, complete network as a test case; second, take an ego network sample from the known network; third, use the simulation approach to estimate network properties of interest; and fourth, compare those estimates to the true values on the known network.

I use this general setup to examine two validity threats to the approach: skewed degree distributions and truncated survey designs. A degree distribution is skewed when a few nodes have very high degree relative to the rest of the population. A network with a skewed degree distribution may be difficult to simulate accurately. A random sample of the population could miss the high degree nodes (because they are not any more likely to be sampled than anyone else), but these actors are disproportionally important for network structure. Those with high degree are connected to many people and are likely to connect disparate parts of the network. A researcher is likely to underestimate the global connectivity of the network if the sample misses these important hubs.

Similarly, survey designs will often limit the number of alters a respondent can name. A person with 15 friends may only be able to list five or 10. This is done to limit respondent burden, but it also makes it hard to infer the true degree distribution. Even if a researcher was fortunate and sampled a high degree node, they would not discover this information if the survey was limited to a small number of alters. This will be especially consequential for networks based on weaker relationships, where an individual can maintain a large number of social ties.

I incorporate varying levels of degree skew and survey truncation into the larger framework for testing the method. Degree skew and truncation are the only variables allowed to vary during the analysis (along with the researcher's response to these problems). I divide the discussion into two parts. I first describe aspects of the test that are held fixed throughout the analysis. I divide this discussion into three subsections: the known network; the sample from the known network; and the global network properties of interest. I then describe the experimental setup itself, describing how degree skew and survey truncation are allowed to vary. I also discuss what steps the researcher can take to limit the level of bias.

4.1. Known, Complete Network

I use networks simulated from an underlying ERGM as the basis for the test.[10] I generate networks from a known model (as opposed to using an empirical network) because simulated networks can be fully controlled, making it easier to represent the different features in the experimental setup (Borgatti et al., 2006). The generated networks are 500 nodes and the model includes terms for transitive closure and homophily on grade and race. The composition of the generated networks and the coefficients in the ERGM are modeled after an empirical Add Health network (e.g., Haynie, 2001; Moody, 2001; Schaefer, Kornienko, & Fox, 2011). The network is also assumed to be symmetric. The degree distribution is allowed to vary by experimental condition (see below), while all other features are held constant throughout the analysis. The networks thus have empirically realistic features but are easily manipulated to satisfy the test conditions. I will assume, for ease of exposition, that the generated networks come from a school setting, representing adolescent friendships.

4.2. Sampling Setup

I assume that a 20% ego network sample is taken from the full, known networks. The *hypothetical* survey has 100 "respondents" and is assumed to collect the following information: the number of alters per respondent, the race and grade of the respondent, the respondent's report on the race and grade of the alters, and the respondent's report on the ties between alters. The survey is hypothetical in the sense that no respondents are actually interviewed and all information on the sampled nodes is taken from the true network.[11] For both the alter–alter ties and the demographic information, the respondent "reports" on only five alters. In this way, the hypothetical survey mimics real data collection limitations—where respondent fatigue is a concern. Since this is not an actual survey, the five alters are randomly selected from the set of all alters for that node (acting as the five alters they chose to report on).

4.3. Network Properties of Interest

I include five global measures in the test: component size, defined as the largest set of nodes connected by at least one path; bicomponent size, defined as the largest set of nodes connected by at least two independent paths (Moody & White, 2003); mean distance, defined as the length of the shortest path between nodes (on average); and reachability, defined as the proportion of nodes reachable X steps out into the network (on average). I include two measures of reachability, one going three steps out into the network and the other going five steps out into the network. I also include results for the full distance distribution, defined as the proportion of dyads that are 0, 1, 2, etc., distance apart. These represent typical

[10] Note that the simulations here are not part of the inferential process, but are rather used to generate networks to make inference about.

[11] Specifically, because this is a hypothetical survey, we do not have the respondent's report on alter race, gender and the alter–alter ties. I thus use information from the actual network as a (perhaps idealistic) proxy of what a respondent would report for the characteristics of their alters. Similarly, the number of alters is their degree from the true network.

measures of global connectivity and diffusion potential in a network. Bicomponent size offers a measure of cohesion, showing how robust the network is to disconnection.

I also include a set of more local network measures, as many researchers will be interested in local network structure. The local measures include: the mean, standard deviation and skew of the degree distribution, density, the number of two-paths, transitivity, the proportion of open triads, the proportion of closed triads, and the geometrically weighted dyad shared partner (GWDSP) distribution. The degree distribution is measured as the proportion of people with degree 0, 1, 2, and so forth. The mean, standard deviation and skew measures are summaries of the full degree distribution. Density is defined as the number of observed ties in the network divided by the number of possible ties. A two-path exists between i and j if there exists a tie between $i - k$ and $k - j$; the summary measure is the count of two-paths in the network. Transitivity is the proportion of two-step paths where there is also a one-step path. The proportions of unclosed and closed triads summarize the triad census: an unclosed triad is defined as a triad with two ties (as the networks are undirected), while a closed triad is defined as a triad with three ties. The dyad shared partner distribution is defined by the number of nodes that i and j are both connected to, or the number of shared friends. This leads to a distribution of shared partners (at the level of the dyad), which is summarized as a geometrically weighted summation. I set alpha to 1 when calculating the summation.

5. Experimental Setup

Three variables are manipulated in the experimental setup: the baseline networks, ranging from low degree skew to high degree skew; the survey design, ranging from no truncation to strong truncation; and the researcher's response to truncation, ranging from nothing to inferring the missing data. There are three conditions for each experimental variable, leaving a 3 × 3 design and nine data points overall. The experiment captures the negative effect of degree skew on the estimates; it also recognizes that the rate of degradation is conditioned on the available survey information and the researcher's response to the survey conditions.

5.1. Varying the Networks of Interest

The three test networks range from low degree skew to high degree skew. Figure 3 plots the degree distributions for the three baseline networks. The top-degree nodes range from 26 in the low skew network to 47 in the medium skew network to 75 in the high skew network. Similarly, the skew in the degree distribution increases as we move down the plot. The bulk of the degree values are still between 0 and 20 in the medium and high skew networks, but the tails are much longer. One or two people now collect a much larger proportion of the total number of ties.

The method was initially designed with strong tie relationships in mind. Relationships based on trust and emotional support, such as close friendships or confidants, require large time/energy investments; this limits the number of ties an individual can maintain. Networks based on strong tie relationships are appropriate for the method precisely because it is difficult to maintain many strong ties. The degree distribution will have a short tail, and it is relatively easy to list of all one's social contacts—minimizing problems of fatigue and the

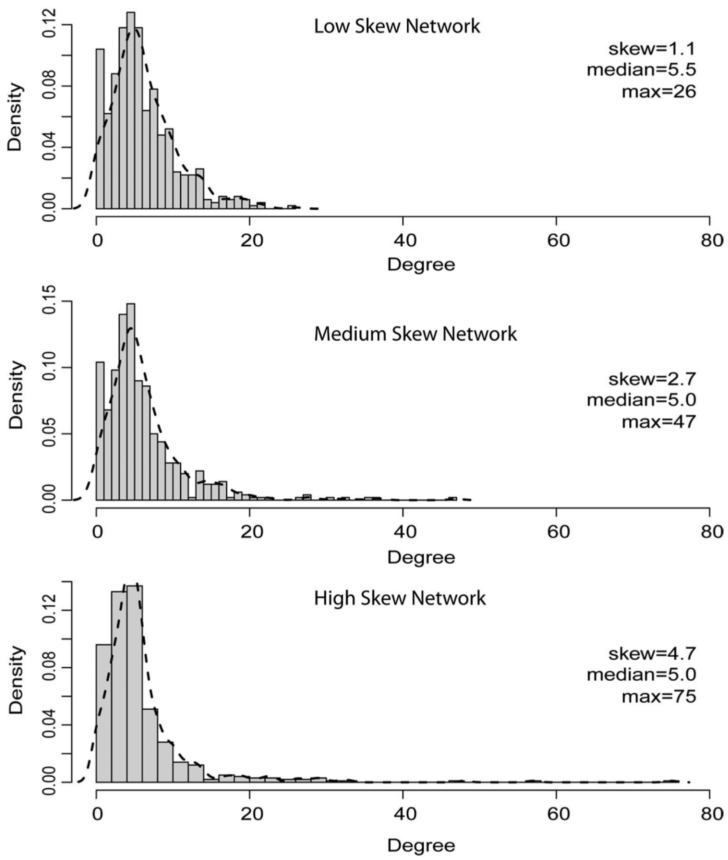

Figure 3. Degree distribution for test networks: Low to high skew.

necessity of truncating the survey. This makes the high skew network a particularly difficult test. The high skew network includes individuals with 75 ties, and represents networks based on weak social relationships, such as acquaintanceship. The question is whether the method will work even in situations where individuals maintain a large number of ties.

I have plotted the three networks in Figure 4. The nodes are colored according to grade and sized according to degree; larger nodes have higher degree. It is clear in all of the networks that the system is organized along grade levels. Grade 12 is separated from Grade 11 and Grade 10, which form a loosely connected group of their own. It is also clear that the top degree in each network increases as move from the low skew network to the high skew network. In the high skew network, the highest degree node is found in Grade 12. This node plays a large role in connecting the Grade 12 students to each other. The high skew network exhibits lower average distance and higher reachability than the low/medium skew networks.[12]

[12] Note that even though the top degree is larger in the high skew network, homophily is still quite important in organizing the social structure of the school.

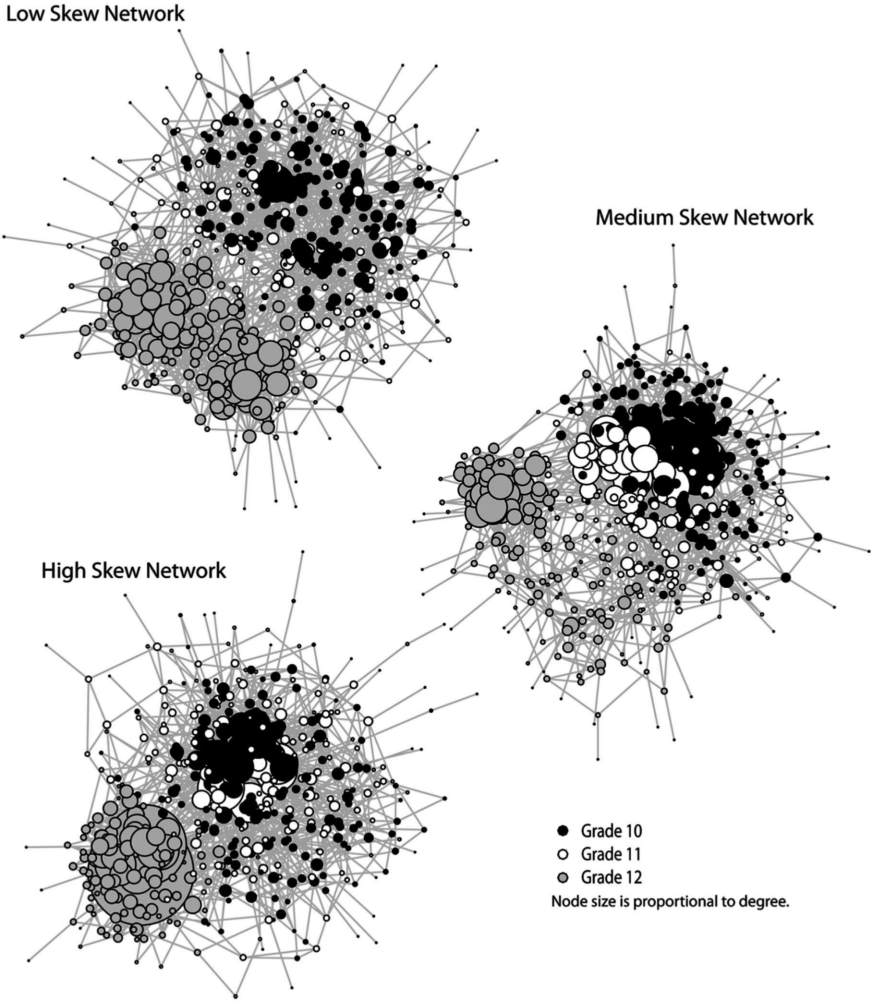

Figure 4. Networks used to test simulation method.

5.2. Varying the Survey Design

Given the networks of interest, I also vary the amount of information assumed to be collected in the ego network survey. The experimental design includes three survey types: strong truncation, medium truncation and no truncation. The surveys vary how many friends the respondents are allowed to name—up to 10 friends (strong truncation), up to 25 friends (medium truncation), or no limit (no truncation). For example, if a sampled node has 30 friends in the full network, we would know about the first 10 under strong truncation, the first 25 under medium truncation, and all 30 under no truncation.[13] In the simulation

[13] Even in the low skew network there could be differences across survey conditions as many individuals have more than 10 friends. Across all survey conditions, I assume that alter demographic information is only recorded for five alters.

experiment, a node with degree 30 would appear in the sample as a node with degree 10, 25, or 30, depending on the level of truncation.

5.3. Varying the Researcher's Response to Survey Truncation

The final variable in the experimental design is the researcher's response to the survey. Surveys will often truncate the alter listing to limit respondent burden. A researcher may react in a number of ways to this (potential) validity threat. The experiment varies the researcher's response to reflect different courses of action.

As a baseline, the researcher could do nothing, taking the data as is. Here, if the survey truncates the alter listing at 10, then the maximum degree recorded for the respondents will be 10—even if they really have 15 ties. In the simulation, a sampled node with degree 15 would be recorded as degree 10. Thus, there will be a mass of people at the truncated value.

Alternatively, the researcher could fill out the truncated degree distribution. The basic idea is to fit a model to the known, truncated data and then use that model to impute the truncated portion. Specifically, the researcher assumes that the true degree distribution follows a negative binomial distribution.[14] They then estimate the parameters of the full distribution, inferring the mean and size parameter from the truncated data.[15] They then use this model to impute the degree of all individuals who report the truncated amount.[16] The researcher takes a draw from the inferred model and assigns the respondent a value greater than or equal to the truncated value (say 13 if the survey is truncated at 10 and the respondent lists up to 10 alters). As the "researcher" in the experiment, I apply this approach to all sampled nodes with degree greater than or equal to the truncated amount. Thus, nodes with degree less than the truncated amount are treated as before, with the value taken directly from the sampled data.

Finally, a researcher could use categorical responses to deal with the truncated survey (e.g., Handcock & Jones, 2004). Here, respondents who list the maximum number of friends (or, in the simulation, have higher degree than the truncated value) are asked an additional question. They are asked to estimate how many friends they actually have. They are given the following categorical options: 11–25, 26–40, 41–60, 61–80, and 81–100+. The categorical options will vary depending on the level of truncation. For example, there is no 11–25 option if the survey is truncated at 25. The researcher then takes a draw from a Poisson distribution with mean set at the mean of the categorical response.[17] Since this is not an actual

[14] Past work has shown that the process of gaining and losing ties will often yield a Poisson distribution for the total number of ties (McPherson, 2009); a negative binomial distribution offers a more flexible form for fitting the data, but is still theoretically close to a Poisson distribution, making it an ideal option.

[15] The best parameters will generate the observed degree distribution, once we collapse all of the values above the truncated value (say 10) into the truncated value. Thus, the model should generate a distribution with the right proportions in each value, including the right proportion above the truncated value.

[16] The simulated value is not allowed to fall below the truncated amount. In this way, respondents cannot have values below the number of alters they listed.

[17] The simulated values are restricted to the range of the categorical response.

survey, and respondents are not actually answering the question, I assume that the "respondent" accurately records their degree. If they really have 30 ties in the full network, then I record 26–40 as their answer. I (playing the role of the researcher) would then take a draw from a Poisson distribution with mean of 32.

The experimental setup is thus a 3 × 3 × 3, and we can think of the test being repeated for each level of truncation with three networks and three types of responses.[18] The test includes 100 iterations per condition to account for variation in the estimates. For each iteration, I take a completely independent ego network sample and repeat the process again. The key question is how much the estimates are affected by increasing degree skew and truncation. It is also important to see if the researcher can limit the level of bias. Is it better to impute the degree distribution or to do nothing under a truncated survey? Or is it worth asking one more question and employing categorical responses?

6. Results: Global Network Measures

I begin the results section by focusing on measures of global network structure. The first set of results focus on the effect of degree skew and truncation on the validity of the estimates. I initially ignore the role of the researcher in the estimation process. The initial section thus presents the worst case scenario: how would the results look if one did nothing in response to the truncation found in the surveys.

6.1. Results Part 1: Researcher Does Nothing in Response to Degree Truncation

Figure 5 presents a snapshot view of the results. The figure is organized by global measure of interest. There is a separate set of boxplots for component size, bicomponent size, distance, three-step reachability, and five-step reachability. The x-axis in each subplot corresponds to truncation level, moving from strong truncation (at 10) to medium truncation (at 25) to no truncation. Each boxplot captures the difference between the true values and the estimates taken from the simulation. The boxplot values are positive if the simulation overestimates the true value. The boxplots are also divided by network of interest, ranging from low skew to high skew. Boxplots corresponding to the same network have the same color (e.g., white = low skew). I have placed the true value for each network/measure underneath the boxplots as a point of reference. The estimates for each network/truncation level are good if they are centered around 0 and have low variance.

I begin with component and bicomponent size. Here the results are encouraging. The method successfully produces good estimates, and this is true for every network and every level of truncation. It is not the case, as we might expect, that the estimates deteriorate as truncation increases. It is even possible to estimate the size of the largest component and bicomponent when the network is highly skewed and degree is truncated at 10. Thus, a researcher interested in component or bicomponent size could make inference using sampled data even under particularly difficult conditions. (Note

[18] Although it is important to note that under no truncation the researcher response is not varied, as there is nothing to try and "fill in."

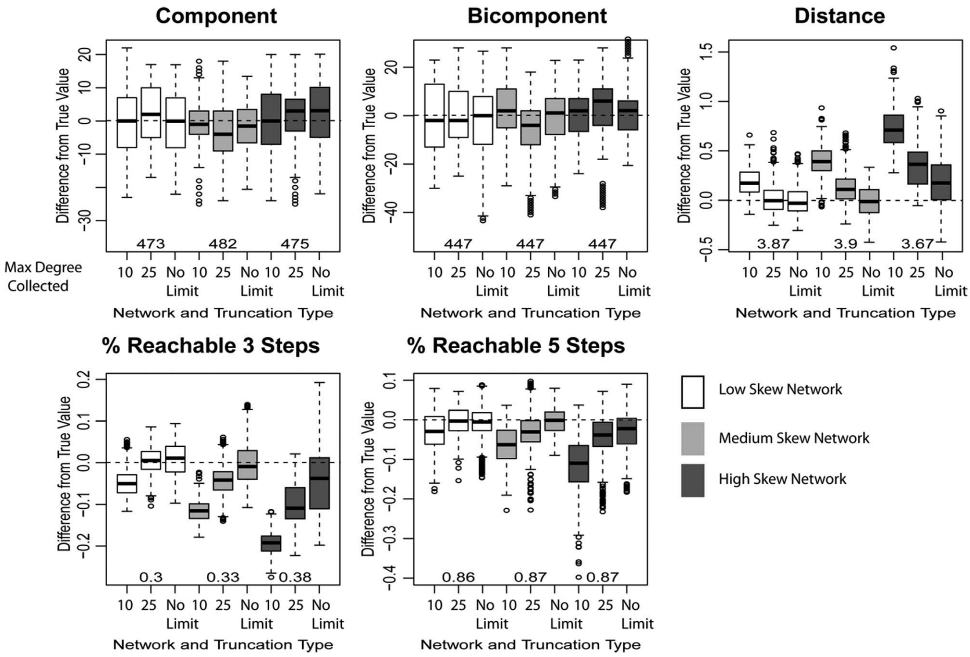

Figure 5. Estimates for global measures by network and truncation type: No researcher response.

that there are also analytical solutions showing if there is a giant component in the network; Grannis, 2010.)

The distance results are more heterogeneous, dependent on both the level of skew in the network and the level of truncation in the survey. Beginning with the low skew (white) network, there is little bias under no truncation and medium truncation (close to no truncation for the low skew network). There is, however, some bias when the survey is strongly truncated, allowing only 10 alters to be named. Thus, as long as degree is not strongly truncated, it is possible to produce accurate estimates for distance on the low skew network. Even when the survey is strongly truncated, the bias is under 5%. The medium skew network offers a similar story, with severe bias only when the degree distribution is truncated at 10, and essentially no bias with no truncation. We begin to see systematic bias in the high skew network. The true distance is 3.67 while the mean estimate is 4.40 under strong truncation (see Table A3 in the Appendix). The bias clearly decreases as the level of truncation decreases: The mean estimate is 4.01 when truncating at 25 and 3.84 with no truncation. Yet there is still residual bias even when the researcher has full information on the sample. Under no truncation, the interquartile range (IQR) is (3.67, 4.03), just barely including the true value. Figure A1 in the Appendix presents the full distance distribution. The method once again reproduces the true distance between nodes when there is no truncation, but overestimates distance under strong truncation. For example, in the medium skew network under strong truncation, the simulated networks underestimate the number of nodes that are separated by two or three paths and overestimate the number separated by four or five paths—but this is not the case when there is no or weak truncation in the survey.

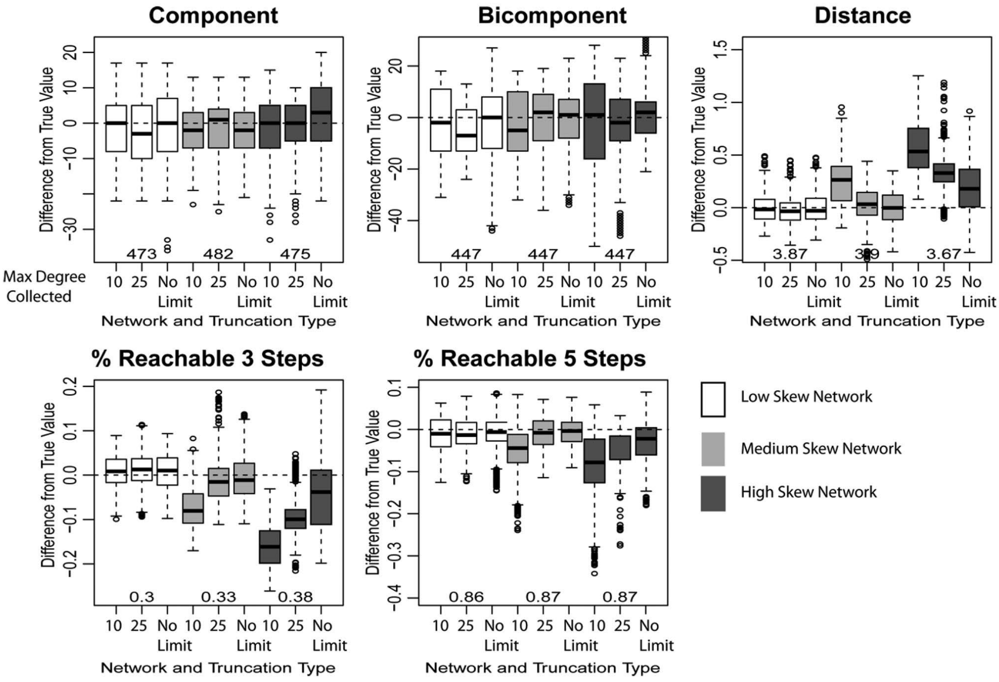

Figure 6. Estimates for global measures by network and truncation type: Simulate truncated tails.

Reachability offers similar results. There is noticeable bias for the low skew network only when the survey is strongly truncated (limited to 10 alters); the high skew network, in contrast, has some bias and large variance even when there is no limit on the number of alters a respondent can name. Similarly, there is little bias in the medium skew network when there is no truncation but significant bias when the survey is truncated at 10. The method generally underestimates the true value. For example, the true value in the medium skew network is .327 while the mean estimate is .286 (under medium truncation).

Five-step reachability yields much lower bias than three-step reachability and is easier to estimate on the whole. It is even possible to estimate five-step reachability on the high and medium skew networks, as long as the survey is not truncated at 10. The low skew network is estimated quite well at all levels of truncation.

In short, it is clear that some measures/networks are easier to estimate than others (e.g., bicomponent size versus distance; low skew versus high skew networks) while truncated surveys lead to worse estimates. Encouragingly, a researcher can accurately estimate many global measures even if they do nothing in response to survey limitations (i.e., the low and medium skew networks when the survey is not strongly truncated).

6.2. Results Part 2: Filling in the Truncated Degree Distribution

Figure 6 presents the results for the imputed tail analysis. The figure has the same form as Figure 5, but now the researcher imputes the truncated portion of the degree distribution.

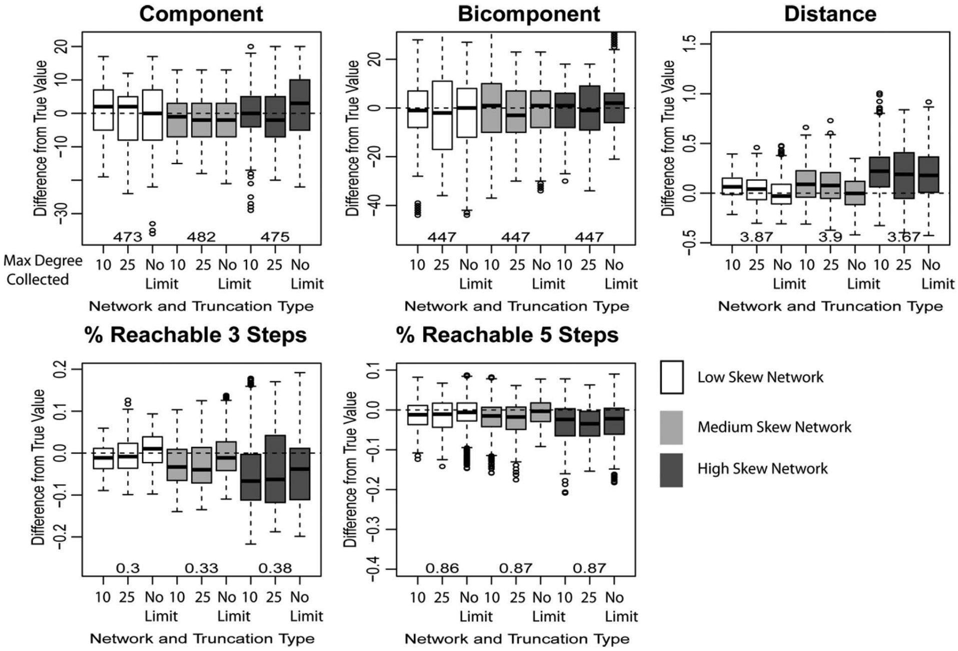

Figure 7. Estimates for global measures by network and truncation type: Use categorical responses.

As before, the method produces excellent estimates for component and bicomponent size across all networks and survey designs.

The results for distance are quite different than in Figure 5, where there was no response from the researcher. Here, there is no bias when estimating distance in the low skew network, even when the survey is truncated at 10. The bias is also reduced in the medium skew network, although there is still some bias under strong truncation. The results for the high skew network show little improvement, however. The full distance distribution (see Figure A2 in the Appendix) offers an analogous story. There is only bias for the medium skew network under strong truncation (again, underestimating the number of shorter paths in the network), and none for the low skew network.

The results are similar for three-step reachability: there is no bias for the low skew network across all survey designs, while the medium skew network shows lower bias than before (for both strong and medium truncation). The five-step reachability results are very similar across the no response and imputed tails figures.

Overall, filling in the degree distribution improves the estimates in every case but the high skew network. It is now possible to estimate the medium skew network for all survey designs expect the strong truncation case, while there is no bias for the low skew network, independent of survey design.

6.3. Results Part 3: Using Categorical Responses

Figure 7 presents the results for the categorical response analysis. Here, the researcher asks the respondents to estimate their number of alters (if above the truncated number); they then use that information to fill in the truncated portion of the degree distribution.

Once again, component size and bicomponent size are accurately estimated for all networks and survey designs. The distance results, however, largely favor the categorical response approach—both compared with doing nothing and imputing the tails of the distribution (although this does not hold for the medium skew network at medium truncation). Using categorical responses, one can accurately estimate distance for the low skew and medium skew networks under any level of truncation. Similarly, the bias for the high skew network is much reduced under the categorical response approach. For example, the true distance in the high skew network is 3.67; the IQR is (3.91, 4.10) when imputing the tails and (3.61, 4.07) when using categorical responses (under medium truncation). The story is similar for three-step reachability. It is possible to estimate reachability well for the medium skew network at all levels of truncation, although the best estimates still come when there is no truncation. There is essentially no bias for five-step reachability, and this holds for all networks and all truncation levels.

Tables A1–A5 in the Appendix focus more explicitly on the role of the researcher. The tables are organized around measures, with one table per global statistic of interest. Within the tables, the columns correspond to different researcher responses: no researcher response; impute truncated tails; use categorical responses. The rows capture the level of skew in the network and the level of truncation in the survey.

Beginning with distance (Table A3 in the Appendix), it is clear that the categorical response approach offers the lowest bias when there is strong truncation (at 10) and the network has medium or high skew. For example, the true distance score in the medium skew network is 3.90. Under strong truncation, the mean estimate is 4.29 when doing nothing, 4.15 when imputing the tails, and 4.00 when using categorical responses. This general pattern also holds for reachability. There is thus good reason to use categorical responses when the survey has a strongly truncated degree design. Similarly, the high skew network is almost uniformly handled best by the categorical response approach.

It is not the case, however, that categorical responses will always offer the best option. For example, in the medium skew network under medium truncation (truncated at 25), filling out the degree distribution is generally preferable to using categorical responses. Similarly, the low skew network is handled equally well by both approaches (and both are better than doing nothing). In general, the categorical response approach is most appropriate when the information about the network is limited: so that truncation is strong or the skew of the network is high. Either active approach (i.e., imputing the tails or using categorical responses) is better than doing nothing, which is only appropriate when the network is not skewed and there is no survey truncation.

7. Results: Local Network Measures

The global network results are encouraging, but it is important to see if the same trends hold for the local measures. Collectively, the local measures capture: (a) the properties of the degree distribution (or the number of ties per person), (b) the level of local closure (is a friend of a friend a friend?), and (c) the level of higher order clustering (i.e., how many

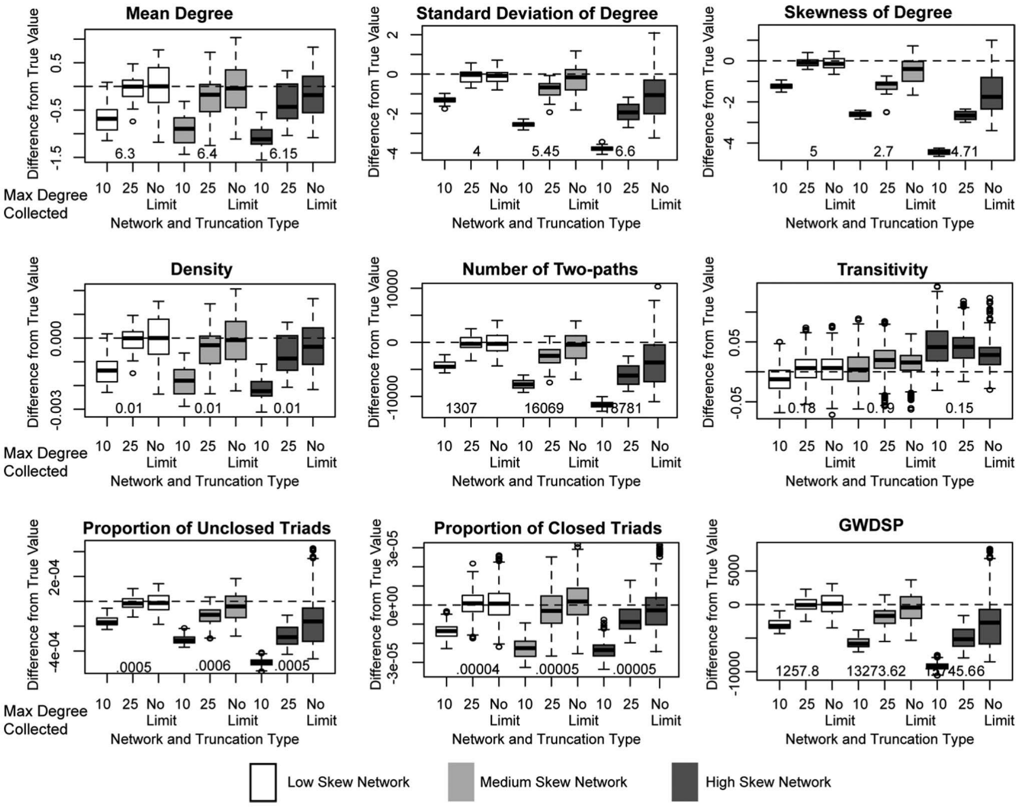

Figure 8. Estimates for local measures by network and truncation type: No researcher response.

friends does *i* and *j* have in common?). Measures of cohesion and connectivity exhibit a regularity that the local measures often lack (as they are more robust to small changes), making this a more difficult test of the method.

Figures 8 to 10 present the results for the local network measures (see also Tables A6–A13 in the Appendix.) The top line of Figures 8 to 10 presents the results for the mean, standard deviation and skew of the degree distribution. It is clear from the figure that mean degree is accurately estimated in the low and medium skew networks when there is no truncation. The estimates for mean degree are only slightly biased downwards when truncation is weak (truncated at 25). Under strong truncation, it is important for the researcher to take some action. For example, for the low skew network, under strong truncation, the mean bias is about 3% under the actual value based on categorical responses but 10% if the researcher does nothing. The results are similar for the high skew network, although the bias is higher at every level of truncation. Note that the density plot is a scaled version of the mean degree plot, and offers identical results.

The bias is higher for the standard deviation and skew of the degree distribution: there is little bias in the low and medium skew networks under no truncation, but some bias when the survey is truncated, even when imputing the tails or using categorical responses. The bias is particularly noticeable for the medium skew network. It is even harder to estimate

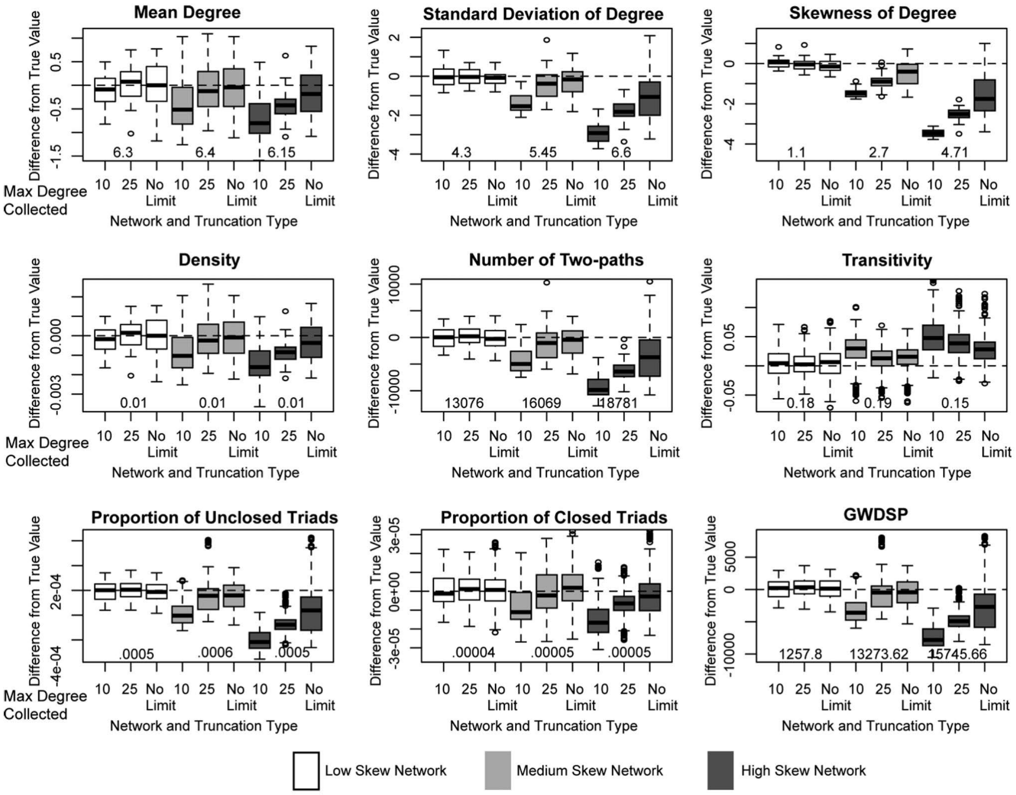

Figure 9. Estimates for local measures by network and truncation type: Simulate truncated tails.

the skew and standard deviation in the high skew network. This is true at all levels of truncation and for all researcher responses (although categorical responses fares the best of the three). For example, the bias in standard deviation is upwards of 25% under weak truncation. The results suggests three things: first, that the sample is missing the high degree nodes in the high skew network; second, that the method is underestimating the degree of the truncated cases in the medium skew network; and third, that the high degree nodes have a large impact on the shape of the degree distribution.

The remaining measures capture local clustering and reachability in the network.

The first measure, number of two-paths, is estimated quite well in the low skew network, especially when simulating the tails of the distribution (with bias under 1%). The worst estimate is the case of strong survey truncation and no researcher response. Similarly, the medium skew network is estimated well under weak truncation (when simulating the tails of the distribution) and no truncation (bias under 5%). The largest bias clearly resides with the high skew network, where the number of two-paths is underestimated across the board. This is the case as high degree nodes create many two-paths; missing those high degree nodes means underestimating the extent of local reachability.

In a similar manner, the proportion of unclosed triads is estimated well in the low and medium skew networks and underestimated in the high skew network. For example, in the

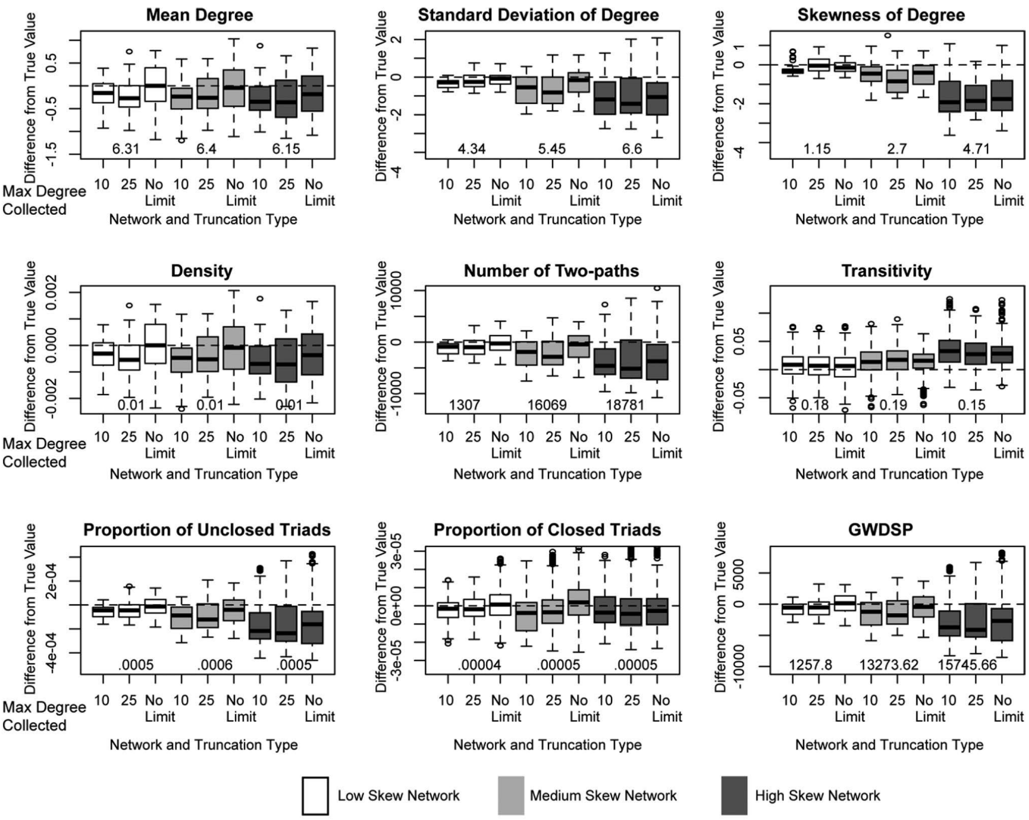

Figure 10. Estimates for local measures by network and truncation type: Use categorical responses.

low skew network, the bias under no truncation is 2%. The closed triad estimates are even better, with little bias for the low and medium skew networks under weak or no truncation. Under strong truncation, the estimates are still quite good for the low skew network, save in the case of no researcher response. The bias is also much lower for the high skew network; under no truncation, the bias is 5% for closed triads but 20% for unclosed triads. The unclosed triads are more sensitive to missing high degree nodes because high degree nodes often connect people who are themselves not connected (just as we saw with two-paths).

The final plot presents the results for the geometrically weighted dyad shared partner distribution. The measure captures higher-order clustering in the network, or the tendency for individuals to have more than one friend in common. The high skew network, once again, offers the worst estimates, with bias of 16% under no truncation. The method performs quite well otherwise, even though ego network data do not directly capture these wider connections. There is very little bias for the low skew network, and this holds for all levels of truncation (under categorical responses or simulating tails). Similarly, the method estimates GWDSP well for the medium skew network under no truncation. Under strong truncation, the categorical response approach performs best, with bias around 11%; under weak truncation, simulating the tails performs best, with bias around 5%.

8. Conclusions

Network sampling holds great promise for network scholars and public policy practitioners. The benefits of sampling are clear: a researcher no longer needs a census to analyze the network structure of a system (Handcock & Gile, 2010). Network sampling can be quite difficult, however, and we are still trying to understand when accurate inference is possible and when it is not (Smith & Moody, 2013). Ego network data pose an extreme example of the network sampling tradeoff. Ego network data are easy to collect for the very reason inference is hard: the sample only collects bits of information about the respondents and their social connections. The question is how far one can push ego network data and simulation techniques in making inference about global network properties—such as cohesion, diffusion potential or group structure.

Past work has demonstrated the validity of simulation techniques on a restricted range of networks. The simulation approach of Smith (2012) yields accurate estimates for networks without a skewed degree distribution and surveys that do not truncate (or limit) the number of alters named. The test here explored a wider range of circumstances: I systematically varied the level of degree skew in the network and the level of truncation in the survey when evaluating the method. I also asked if the researcher could limit the level of bias.

The low skew network represents the case closest to previous tests. Here, the top-degree node has less than 25 ties. Low skew networks characterize many networks of interest, including networks based on close friendship, confidants, and social support (Wellman & Wortley, 1990; Smith et al., 2014). The results suggest that one can accurately make global network inference in low skew networks under all survey conditions. The researcher must, however, treat the data in some way (using categorical responses or imputing the tails) when the survey is strongly truncated. Similarly, there is little bias for the local clustering measures. This holds for all truncation levels, as long as the researcher does something to deal with the truncation in the survey (with imputing the tails offering the better option).

The results for the medium skew network are similarly encouraging. The medium skew network has a high degree around 50, representing networks of weak friendship or regular affiliation (Moody, 2004). The model produces accurate global network estimates when there is no degree truncation. When there is degree truncation, the results clearly depend on the response of the researcher. The categorical response approach offers accurate estimates across all survey designs, while imputation is a viable option for all but the strongest level of truncation. It is, however, quite difficult to produce unbiased estimates when the researcher does nothing in response to survey truncation (although one can still estimate component and bicomponent size without bias). The method also performs well for more local network measures, although the bias is generally higher than with the global measures. The method is particularly appropriate for local clustering measures when there is no/weak truncation, but less appropriate if the survey is strongly truncated.

Taken together, simulation based inference is appropriate for many empirical settings of interest—ranging from the close ties of friendship and cohabitation to the wider ties of regular affiliation and association.

High skew networks reflect considerably weaker relationships, and represent the least appropriate case for ego network data.[19] Component and bicomponent size show little bias across survey/response conditions, but other network measures are harder to estimate. The method produces biased estimates for distance and reachability (5% for distance and 10% for reachability) even when there is no truncation in the survey. The ego network samples often miss the high degree nodes, leading to inaccurate estimates. This is particularly clear when looking at the local network measures: where it is difficult to estimate the properties of the degree distribution in the high skew network, even under no truncation. This also makes it difficult to capture local clustering measures in the high skew network, like number of two-paths or proportion of unclosed triads.

What are the larger implications of the results? First, it is always better to collect full degree information if possible. The low skew and medium skew networks had little or no bias when there was no survey truncation. The estimates were always worse (or at best equal) when the survey was truncated.

Second, if the survey must be truncated, the researcher should infer the truncated portion of the data. Across the board, the imputed tails and categorical response results were better than using the truncated degree distribution from the sample. Often the results were just as good as if the survey had not been truncated at all. The categorical response approach is particularly appropriate when the network is highly skewed or the survey is strongly truncated. The researcher only adds one question to the survey but receives information on the high degree nodes.

And third, the researcher may need to collect additional data if they believe the network has a skewed degree distribution. The level of bias in Tables A1–A13 in the Appendix may be acceptable to a researcher, given their particular research question; but a study requiring a higher level of accuracy would need to consider alternative kinds of networks or sampling schemes. For example, it may be necessary to collect a larger sample, thus increasing the probability of sampling the high degree nodes. Alternatively, one could embed a two-step sample within the larger ego network survey. Thus, for a subset of the initial sample (say 20%), the researcher would contact the reported alters of the respondents. This second step would capture most of the high degree nodes (as individuals with high degree are disproportionately named as alters), and should improve the estimates. This will be particularly important in trying to capture local network measures.

Overall, the results are encouraging. It is possible to accurately estimate local and global network features, like GWDSP, component size or reachability, from independently sampled network data. The difficulty of the task clearly increases as the skew of the degree distribution increases and the survey becomes more truncated, but in most cases the researcher can make appropriate adjustments and produce accurate network estimates.

The results, while encouraging, rest on a number of assumptions, and it is important for future work to consider those assumptions more carefully. For example, I assume that respondents accurately report the number of alters they have. In reality, respondents may list fewer (or more) social contacts than actually exist (Marin, 2004). One way of capturing

[19] The high degree nodes approach 75 ties, well beyond the number found in strong tie networks for which the method was initially designed.

this uncertainty is to add a disturbance term to the reported degree; this would produce bounds on the estimated network statistics (based on individuals under or over estimating their degree). More generally, a researcher concerned about the degree data could model the degree distribution, rather than use the sampled data as is. Of course, a more robust name generator would also improve the reported data and make it less necessary to smooth out the distribution (Marin & Hampton, 2007).

In a similar manner, truncating the survey may lead to artificial spikes in the degree distribution: if the survey asks for five people, then respondents may be inclined to "find" five friends, instead of listing the actual number of associates. It is, however, fairly straightforward to avoid such spikes. The key is not revealing how many alters the respondent can list; by probing for more alters, but not telling them when they will stop, individuals are not cognitively biased towards a certain number of alters.

I also assume that respondents accurately report on the characteristics of their alters. This is likely true in some but not all cases. Respondents may report demographic characteristics, such as age or gender, accurately but have more difficulty with behavioral/cultural characteristics, such as political attitudes. Similarly, I also assume that respondents can accurately report on the ties between alters. The reported ties may, however, be biased towards transitive relations, as individuals cognitively try to maintain social balance in their personal network (Krackhardt & Kilduff, 1999). This is likely to be exasperated in content-specific relations (do *i* and *j* talk about politics?), where the respondent may be unaware if a tie exists, and minimized in more general relations (do *i* and *j* know each other?), where it is easier to observe an existing tie.

A researcher concerned with the validity of the self-reports could take two paths. First, they could collect more data, going out one more step in the network and interviewing the alters of the original respondents; this would provide first hand reports of the alter characteristics and social connections. Second, the researcher could induce error into the reported alter–alter ties and alter characteristics; they would then rerun the analysis, using those simulations to put bounds on the estimates. This would capture some of the uncertainty in the estimates, but not force the researcher to collect more data.

Moving forward, the hope is that ego network sampling and simulation will become a general option for network scholars. The long-term agenda is to make comparative network studies more feasible. A researcher would first collect sampled network data in multiple contexts as opposed to census data in one context (as is typically done). They would then infer the network structure in each setting, using those estimates to measure contextual variation in cohesion or connectivity. Cohesion/connectivity could then be used as a contextual level predictor of health, suicidality and the like (Bearman & Moody, 2004; Wray, Colen, & Pescosolido, 2011). In this way, more studies could incorporate social structure into their analyses, with data limitations no longer standing in the way.

Acknowledgments — The author would like to thank Miller McPherson, James Moody, Peter Mucha, Robin Gauthier, and the network working group at Duke for comments on earlier drafts and presentations of this work.

References

Adams, J., Moody, J., & Morris, M. (2013). Sex, drugs, and race: How behaviors differentially contribute to the sexually transmitted infection risk network structure. *American Journal of Public Health*, *103*, 322–329.

Bansal, S., Khandelwal, S., & Meyers, L. (2009). Exploring biological network structure with clustered random networks. *BMC Bioinformatics*, *10*, 405.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Bearman, P. S., & Moody, J. (2004). Suicide and friendships among American adolescents. *American Journal of Public Health*, *94*, 89–95.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). Robustness of centrality measures under conditions of imperfect data. *Social Networks*, *28*, 124–136.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Vol. 1. The analysis of case-control studies*. Lyon, France: IARC Scientific Publications.

Browning, C. R., Cagney, K. A., & Iveniuk, J. (2012). Neighborhood stressors and cardiovascular health: Crime and C-reactive protein in Dallas, USA. *Social Science & Medicine*, *75*, 1271–1279.

Burt, R. S. (1984). Network items and the general social survey. *Social Networks*, *6*, 293–339.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*, 702–734.

Cheadle, J. E., & Schwadel, P. (2012). The ‘friendship dynamics of religion,’ or the ‘religious dynamics of friendship’? A social network analysis of adolescents who attend small schools. *Social Science Research*, *41*, 1198–1212.

Fischer, C. (1982). *To dwell among friends: Personal networks in town and city*. Chicago, IL: The University of Chicago Press.

Frank, O. (1971). *Statistical inference in graphs* (Unpublished doctoral dissertation). Stockholm University, Sweden.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, *1*, 91–101.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, *81*, 832–842.

Goodman, L. A. (2011). Comment: On respondent-driven sampling and snowball sampling. In Hard-to-reach populations and snowball sampling not in hard-to-reach populations. *Sociological Methodology*, *41*, 347–353.

Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography*, *46*, 103–125.

Gould, R. V. (2002). The origins of status hierarchies: A formal theory and empirical test. *American Journal of Sociology*, *107*, 1143–1178.

Grannis, R. (2010). Six degrees of “who cares?” *American Journal of Sociology*, *115*, 991–1017.

Handcock, M., Goodreau, S. M., Hunter, D. R., Butts, C. T., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, *24*, nihpa54860.

Handcock, M. S., & Gile, K. J. (2010). Modeling social networks from sampled data. *Annals of the Applied Statistics*, *4*, 5–25.

Handcock, M. S., & Jones, J. H. (2004). Likelihood-based inference for stochastic models of sexual network formation. *Theoretical Population Biology*, *65*, 413–422.

Haynie, D. L. (2001). Delinquent peers revisited: Does network structure matter? *American Journal of Sociology*, *106*, 1013–1057.

Holland, P. W., & Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, *7*, 1–45.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, *76*, 33–51.

Koskinen, J. H., Robins, G. L., & Pattison, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, *7*, 366–384.

Koskinen, J. H., Robins, G. L., Wang, P., & Pattison, P. E. (2013). Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Social Networks*, *35*, 514–527.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, *28*, 247–268.

Krackhardt, D., & Kilduff, M. (1999). Whether close or far: Social distance effects on perceived balance in friendship networks. *Journal of Personality and Social Psychology*, *76*, 770–782.

Krivitsky, P. N., Handcock, M. S., & Morris, M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, *8*, 319–339.

Lee, J.-S. (2008). *Inferring adolescent social networks using partial ego-network substance use data*. (Unpublished thesis). Carnegie Mellon University, Pittsburgh, PA.

Lizardo, O. (2006). How cultural tastes shape personal networks. *American Sociological Review*, *71*, 778–807.

Louch, H. (2000). Personal network integration: Transitivity and homophily in strong-tie relations. *Social Networks*, *22*, 45–64.

Marin, A. (2004). Are respondents more likely to list alters with certain characteristics?: Implications for name generator data. *Social Networks*, *26*, 289–307.

Marin, A., & Hampton, K. N. (2007). Simplifying the personal network name generator. *Field Methods*, *19*, 163–193.

Marsden, P. V. (1987). Core discussion networks of Americans. *American Sociological Review*, *52*, 122–131.

Marsden, P. V. (1988). Homogeneity in confiding relations. *Social Networks*, *10*, 57–76.

McPherson, M. (2009). A baseline dynamic model for ego networks. *American Behavioral Scientist*. Advance online publication. doi: 10.1177/0002764209331530

McPherson, M., Smith-Lovin, L., & Brashears, M. (2006). Social isolation in America: Changes in core discussion networks over two decades. *American Sociological Review*, *71*, 353–375.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Merli, M. G., Moody, J., Smith, J., Li, J., Weir, S., & Chen, X. (2015). Challenges to recruiting population representative samples of female sex workers in China using respondent driven sampling. *Social Science & Medicine*, *125*, 79–93. doi: 10.1016/j.socscimed.2014.04.022

Middendorf, M. E. Z., Wiggins, C. H., & Honig, B. H. (2005). Inferring, network mechanisms: The Drosophila melanogaster protein interaction, and network. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 3192–3197.

Moody, J. (2001). Race, school integration, and friendship segregation in America. *American Journal of Sociology*, *107*, 679–716.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Journal of Sociology*, *69*, 213–238.

Moody, J., & White, D. R. (2003). Structural cohesion and embeddedness. *American Sociological Review*, *68*,103–127.

Morris, M., Kurth, A. E., Hamilton, D. T., Moody, J., & Wakefield, S. (2009). Concurrent partnerships and HIV prevalence disparities by race: Linking science and public health practice. *American Journal of Public Health*, *99*, 1023–1031.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*, 876–878.

Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, *64*, 026118.

Pattison, P. E., Robins, G. L., Snijders, T. A. B., & Wang, P. (2013). Conditional estimation of exponential random graph models from snowball sampling designs. *Journal of Mathematical Psychology*, *57*, 284–296.

Qian, Z. (2002). Race and social distance: Intermarriage with non-Latino whites. *Race & Society*, *5*, 33–47.

Qian, Z., & Lichter, D. T. (2007). Social boundaries and marital assimilation: Interpreting trends in racial and ethnic intermarriage. *American Sociological Review*, *72*, 68–94.

Robins, G., Snijders, T., Wang, P., Handcock, M., & Pattison, P. (2007). Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, *29*, 192–215.

Rosenfeld, M. J. (2008). Racial, educational and religious endogamy in the United States: A comparative historical perspective. *Social Forces*, *87*, 1–31.

Sampson, R. J., & Raudenbush, S. W. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *277*, 918–924.

Schaefer, D. R., Kornienko, O., & Fox, A. M. (2011). Misery does not love company: Network selection mechanisms and depression homophily. *American Sociological Review*, *76*, 764–785.

Shwed, U., & Bearman, P. S. (2010). The temporal structure of scientific consensus formation. *American Sociological Review*, *75*, 817–840.

Smith, J. A. (2012). Macrostructure from microstructure: Generating whole systems from ego networks. *Sociological Methodology*, *42*, 155–205.

Smith, J. A., McPherson, M., & Smith-Lovin, L. (2014). Social distance in the United States: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004. *American Sociological Review*, *79*, 432–456.

Smith, J. A., & Moody, J. (2013). Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks*, *35*, 652–668.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*, 99–153.

Soffer, S., & Vazquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E*, *71*, 057101.

Thompson, S. K., & Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, *26*, 87–98.

Verdery, A. M., Entwisle, B., Faust, K., & Rindfuss, R. R. (2012). Social and spatial networks: Kinship distance and dwelling unit proximity in rural Thailand. *Social Networks*, *34*, 112–127.

Viger, F., Latapy, M., & Wang, L. (2005). Efficient and simple generation of random simple connected graphs with prescribed degree sequence computing and combinatorics. In L. Wang (Ed.), *Lecture notes in computer science* (Vol. 3595, pp. 440–449). Berlin: Springer.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika*, *61*, 401–425.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 5766–5771.

Wellman, B., & Wortley, S. (1990). Different strokes from different folks: Community ties and social support. *American Journal of Sociology*, *96*, 558–588.

Wray, M., Colen, C., & Pescosolido, B. (2011). The sociology of suicide. *Annual Review of Sociology*, *37*, 505–528.

Zhang, Y., Kolaczyk, E. D., & pencer, B. D. (2014). *Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks*. Online at http://arxiv.org/abs/1305.4977

Appendix: Supplementary Tables and Figures

Table A1. Bias Table, Component Size

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	473	472.378	.001	471.130	.004	473.477	.001
Medium	10	482	481.432	.001	480.317	.003	481.136	.002
High	10	475	474.797	.000	474.230	.002	474.805	.000
Low	25	473	474.901	.004	471.130	.005	471.414	.003
Medium	25	482	478.64	.007	480.317	.001	479.828	.005
High	25	475	476.548	.003	474.230	.001	474.454	.001
Low	None	473	471.434	.003	471.434	.003	471.434	.003
Medium	None	482	480.159	.004	480.159	.004	480.159	.004
High	None	475	476.228	.003	476.228	.003	476.228	.003

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A2. Bias Table, Bicomponent Size

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	447	446.544	.001	445.312	.004	446.301	.002
Medium	10	447	449.358	.005	444.889	.005	447.595	.001
High	10	447	447.587	.001	445.444	.003	447.163	.000
Low	25	447	447.663	.001	442.354	.010	444.419	.006
Medium	25	447	442.419	.010	447.068	.000	444.809	.005
High	25	447	450.133	.007	443.181	.009	445.599	.003
Low	None	447	445.204	.004	445.204	.004	445.204	.004
Medium	None	447	446.709	.001	446.709	.001	446.709	.001
High	None	447	448.840	.004	448.840	.004	448.840	.004

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A3. Bias Table, Distance

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	3.875	4.054	.046	3.869	.002	3.941	.017
Medium	10	3.899	4.293	.101	4.149	.064	4.001	.026
High	10	3.666	4.395	.199	4.228	.154	3.878	.058
Low	25	3.875	3.886	.003	3.845	.008	3.916	.011
Medium	25	3.899	4.018	.031	3.940	.011	3.977	.020
High	25	3.666	4.012	.095	3.998	.091	3.851	.051
Low	None	3.875	3.866	.002	3.866	.002	3.866	.002
Medium	None	3.899	3.899	.000	3.899	.000	3.899	.000
High	None	3.666	3.844	.049	3.844	.049	3.844	.049

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A4. Bias Table, % Reachable, Three Steps

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	.297	.248	.165	.305	.025	.286	.039
Medium	10	.327	.213	.349	.254	.224	.297	.092
High	10	.381	.186	.512	.224	.412	.330	.134
Low	25	.297	.303	.020	.310	.041	.291	.020
Medium	25	.327	.286	.127	.314	.042	.303	.075
High	25	.381	.281	.262	.286	.250	.338	.113
Low	None	.297	.305	.027	.305	.027	.305	.027
Medium	None	.327	.322	.016	.322	.016	.322	.016
High	None	.381	.343	.102	.343	.102	.343	.102

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A5. Bias Table, % Reachable, Five Steps

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	.857	.829	.032	.844	.015	.844	.015
Medium	10	.867	.803	.074	.820	.054	.847	.023
High	10	.875	.760	.131	.791	.096	.841	.039
Low	25	.857	.856	.001	.846	.013	.840	.019
Medium	25	.867	.838	.034	.858	.010	.846	.025
High	25	.875	.834	.046	.827	.055	.842	.037
Low	None	.857	.846	.012	.846	.012	.846	.012
Medium	None	.867	.862	.006	.862	.006	.862	.006
High	None	.875	.850	.028	.850	.028	.850	.028

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A6. Bias Table, Mean Degree

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	6.312	5.669	.102	6.211	.016	6.117	.031
Medium	10	6.404	5.460	.147	5.969	.068	6.111	.046
High	10	6.152	5.070	.176	5.449	.114	5.879	.044
Low	25	6.312	6.289	.004	6.310	.0003	6.086	.036
Medium	25	6.404	6.172	.036	6.319	.013	6.225	.028
High	25	6.152	5.812	.055	5.740	.067	5.863	.047
Low	None	6.312	6.301	.002	6.301	.002	6.301	.002
Medium	None	6.404	6.341	.010	6.341	.010	6.341	.010
High	None	6.152	5.939	.035	5.939	.035	5.939	.035

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A7. Bias Table, Standard Deviation of Degree

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	4.337	3.005	.307	4.369	.007	3.966	.086
Medium	10	5.452	2.906	.467	4.055	.256	4.796	.120
High	10	6.597	2.819	.573	3.686	.441	5.541	.160
Low	25	4.337	4.235	.024	4.331	.001	4.155	.042
Medium	25	5.452	4.684	.141	5.040	.076	4.858	.109
High	25	6.597	4.69	.289	4.829	.268	5.593	.152
Low	None	4.337	4.237	.023	4.237	.023	4.237	.023
Medium	None	5.452	5.165	.053	5.165	.053	5.165	.053
High	None	6.597	5.586	.153	5.586	.153	5.586	.153

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A8. Bias Table, Skewness of Degree

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	1.147	–0.082	1.071	1.186	.035	.927	.191
Medium	10	2.700	0.091	.966	1.225	.546	2.278	.156
High	10	4.714	0.26	.945	1.247	.735	3.076	.347
Low	25	1.147	1.059	.076	1.107	.035	1.158	.010
Medium	25	2.700	1.521	.437	1.840	.319	1.916	.290
High	25	4.714	2.032	.569	2.191	.535	3.018	.360
Low	None	1.147	1.102	.040	1.102	.040	1.102	.040
Medium	None	2.700	2.229	.174	2.229	.174	2.229	.174
High	None	4.714	3.178	.326	3.178	.326	3.178	.326

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A9. Bias Table, Number of Two Paths

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	13076	8899.049	.319	12962.522	.009	11793.245	.098
Medium	10	16069	8221.248	.488	11646.036	.275	13714.447	.147
High	10	18781	7159.127	.619	9590.866	.489	15205.758	.190
Low	25	13076	12834.591	.018	13136.986	.005	12123.673	.073
Medium	25	16069	13537.751	.158	14952.708	.069	14239.320	.114
High	25	18781	12567.329	.331	12717.645	.323	15350.508	.183
Low	None	13076	12903.668	.013	12903.668	.013	12903.668	.013
Medium	None	16069	15308.785	.047	15308.785	.047	15308.785	.047
High	None	18781	15528.971	.173	15528.971	.173	15528.971	.173

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A10. Bias Table, Transitivity

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	.178	.166	.066	.181	.019	.186	.042
Medium	10	.185	.191	.029	.213	.149	.200	.077
High	10	.153	.198	.291	.204	.328	.188	.225
Low	25	.178	.184	.035	.182	.020	.185	.037
Medium	25	.185	.206	.110	.197	.064	.202	.089
High	25	.153	.195	.271	.191	.248	.181	.181
Low	None	.178	.182	.024	.182	.024	.182	.024
Medium	None	.185	.198	.070	.198	.070	.198	.070
High	None	.153	.181	.180	.181	.180	.181	.180

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A11. Bias Table, Proportion of Unclosed Triads

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	.00051	.00036	.309	.00051	.014	.00046	.107
Medium	10	.00063	.00032	.492	.00044	.300	.00053	.161
High	10	.00077	.00027	.639	.00037	.520	.00060	.220
Low	25	.00052	.00051	.027	.00052	.001	.00048	.081
Medium	25	.00063	.00052	.180	.00058	.084	.00055	.131
High	25	.00077	.00049	.362	.00050	.353	.00061	.205
Low	None	.00052	.00051	.019	.00051	.019	.00051	.019
Medium	None	.00063	.00059	.063	.00059	.063	.00059	.063
High	None	.00077	.00062	.197	.00062	.196	.00062	.196

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A12. Bias Table, Proportion of Closed Triads

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	3.75E-5	2.37E-5	.366	3.81E-5	.017	3.53E-5	.057
Medium	10	4.80E-5	2.53E-5	.472	3.99E-5	.167	4.41E-5	.080
High	10	4.63E-5	2.27E-5	.509	3.15E-5	.320	4.50E-5	.028
Low	25	3.75E-5	3.82E-5	.019	3.86E-5	.031	3.61E-5	.036
Medium	25	4.80E-5	4.52E-5	.058	4.77E-5	.006	4.61E-5	.039
High	25	4.63E-5	3.91E-5	.156	3.91E-5	.155	4.37E-5	.056
Low	None	3.75E-5	3.81E-5	.0156	3.81E-5	.0156	3.81E-05	.0156
Medium	None	4.80E-5	4.90E-5	.021	4.90E-5	.021	4.90E-05	.021
High	None	4.63E-5	4.42E-5	.045	4.42E-5	.045	4.42E-05	.045

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

Table A13. Bias Table, Geometrically Weighted Dyad Shared Partner Distribution

			No researcher response		Impute truncated tails		Use categorical responses	
Network skew	Truncation	True value	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]	Mean estimate	Relative bias[a]
Low	10	11257.844	8313.961	.261	11311.449	.005	10480.317	.069
Medium	10	13273.617	7576.140	.429	10013.939	.246	11739.427	.116
High	10	15745.660	6587.430	.582	8383.076	.468	12838.329	.185
Low	25	11257.844	11268.902	.001	11519.546	.023	10718.811	.048
Medium	25	13273.617	11442.981	.138	12584.982	.052	12068.162	.091
High	25	15745.660	10767.273	.316	10950.629	.305	13044.604	.172
Low	None	11257.844	11371.809	.010	11371.809	.010	11371.809	.010
Medium	None	13273.617	12891.368	.029	12891.368	.029	12891.368	.029
High	None	15745.660	13221.088	.160	13221.088	.160	13221.088	.160

The values for mean estimate, SE, bias, and relative bias are calculated over 100 independent samples, where each sample yields one estimate of the network measure.

a. Relative Bias = | (*E*(estimates) – True Value) / True Value |

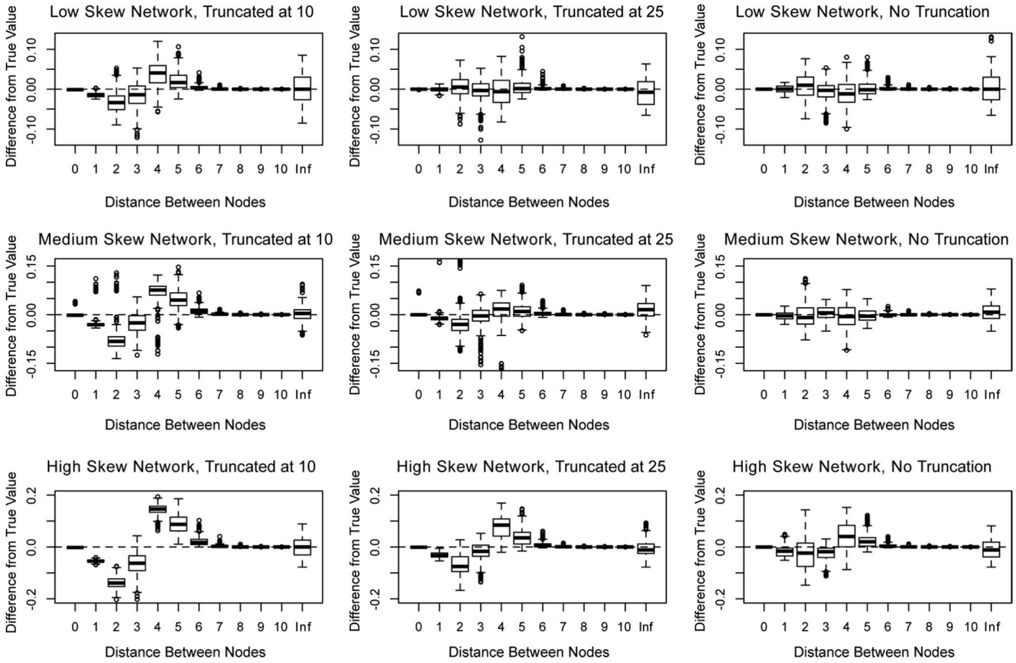

Figure A1. Estimates for distance distribution by network and truncation type: No researcher response.

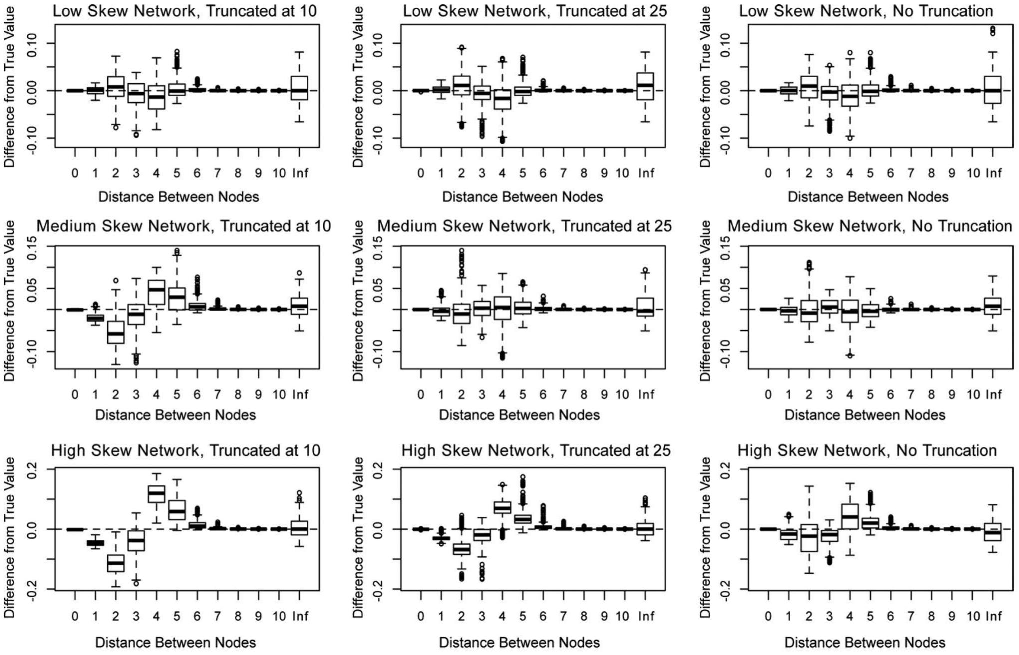

Figure A2. Estimates for distance distribution by network and truncation type: Simulate truncated tails.

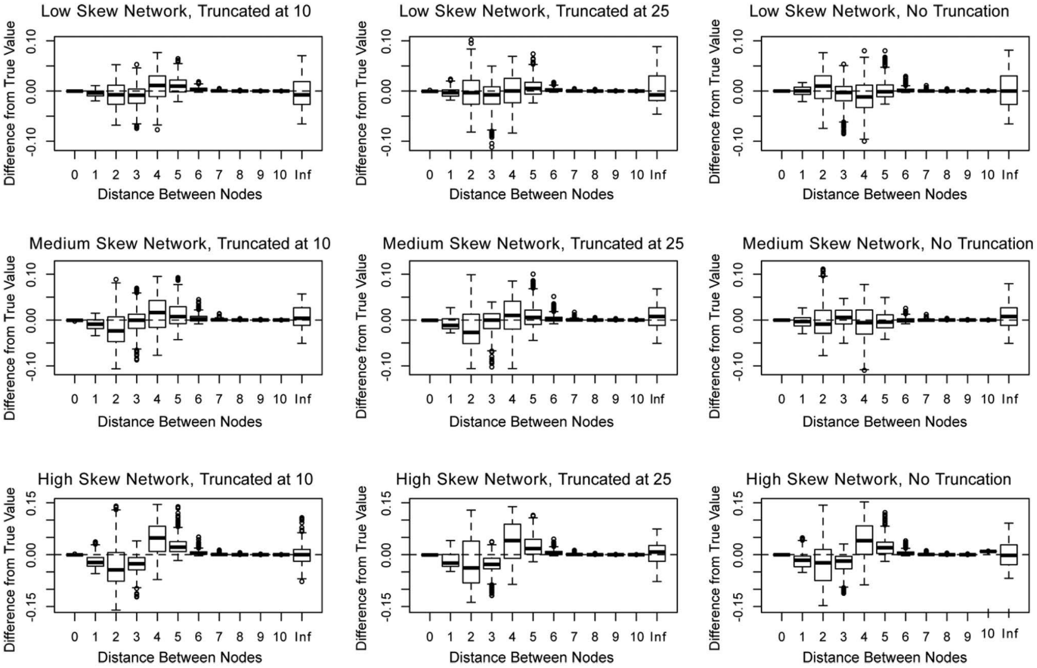

Figure A3. Estimates for distance distribution by network and truncation type: Use categorical responses.