4-2018

Using Ego Network Data to Inform Agent-based Models of Diffusion

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

Jessica Burow
The Hartford

Published in *Sociological Methods & Research* (2018), 46p.
doi 10.1177/0049124118769100

Published April 24, 2018.

Using Ego Network Data to Inform Agent-based Models of Diffusion

Jeffrey A. Smith[1] and Jessica Burow[2]

1 Department of Sociology, University of Nebraska–Lincoln, Lincoln, NE, USA
2 Hartford Financial Service Group Inc., Hartford, CT, USA

Corresponding author — Jeffrey A. Smith, Department of Sociology, University of Nebraska–Lincoln, 711 Oldfather Hall, Lincoln, NE 68588, USA. *email* jsmith77@unl.edu

Abstract

Agent-based modeling holds great potential as an analytical tool. Agent-based models (ABMs) are, however, also vulnerable to critique, as they often employ stylized social worlds, with little connection to the actual environment in question. Given these concerns, there has been a recent call to more fully incorporate empirical data into ABMs. This article falls in this tradition, exploring the benefits of using sampled ego network data in ABMs of cultural diffusion. Thus, instead of relying on full network data, which can be difficult and costly to collect, or no empirical network data, which is convenient but not empirically grounded, we offer a middle-ground, one combining ABMs with recent work on network sampling. The main question is whether this approach is effective. We provide a test of the approach using six complete networks; the test also includes a range of diffusion models (where actors follow different rules of adoption). For each network, we take a random ego network sample and use that sample to infer the full network structure. We then run a diffusion model through the known, complete networks, as well as the inferred networks, and compare the results. The results, on the whole, are quite strong: Across all analyses, the diffusion curves based on the sampled data are very similar to the curves based on the true, complete network. This suggests that ego network sampling can, in fact, offer a practical means of incorporating empirical data into an agent-based model.

Keywords: ego networks, network sampling, agent-based models, diffusion

Agent-based modeling holds great promise as an analytical tool in the social sciences (Macy and Flache 2009; Sterman 2006). Agent-based models (ABMs) rely on simulation as a means of analysis, offering an alternative to traditional statistical techniques (Macy and Willer 2002; Railsback and Volker 2011). A researcher specifies a virtual world, where actors are seeded with certain characteristics and set to interact based on a system of behavioral rules; macrolevel outcomes ultimately emerge out of these individual-level interactions (De Marchi and Page 2014; Hedström and Bearman 2009; Miller and Page 2007). Agent-based modeling is a useful tool for a number of reasons (see Axtel 2000). First, ABMs offer a social laboratory of sorts, where key conditions are allowed to vary, but all else can be held constant, allowing the researcher to pinpoint plausible causal mechanisms and generate testable hypotheses (Hedström and Ylikoski 2010; Manzo 2007).[1] Second, ABMs encourage analytical clarity, as the researcher must be explicit about their assumptions and theoretical model (Manzo 2007). And third, ABMs capture aspects of social life that are difficult to represent in traditional statistical models, such as the (nonlinear) relationship between microlevel interactions and emergent collective outcomes (Bonabeau 2002; Mabry et al. 2008).

The cost of an ABM is that one must typically make do with a stylized hypothetical world, often quite divorced from actual social conditions (Boero and Squazzoni 2005; Hedström and Manzo 2015). Such models are tractable and useful for specifying a theoretical model, but the results are vulnerable to critique, especially in fields with a heavy empirical bent (Richiardi et al. 2006). Recent work has responded to this concern by pushing for more empirically grounded models, models that wed simulation approaches with empirical data on the population and setting of interest (Bruch and Axtell 2015; Hedström and Manzo 2015). Empirical data can be used to inform the simulation itself or as a validation tool (Windrum, Fagiolo, and Moneta 2007). In either case, the simulations are constrained by the empirical evidence, making it more likely that the theoretical conclusions are valid and apply to the population/setting of interest (e.g., see Liu and Bearman 2015; Manzo 2013; Verdery 2015).

This article falls in this tradition, exploring the feasibility and returns of using sampled network data in ABMs (see also Rolfe 2014). Specifically, we explore empirically grounded ABMs in the context of diffusion models (the ABM) and ego network data (the empirical

data). Past work has used simulation models to characterize the diffusion, or spread, of cultural items (new products, innovations, ideas, etc.) through social networks (e.g., Centola 2015; Gondal 2015). Such simulations are based on a virtual world of interacting actors who follow a set of prespecified rules and are thus ABMs. Actors will, for example, adopt or drop a behavior probabilistically, based on the behavioral traits of their social partners, as well as the rules surrounding adoption.

The problem with these studies, from an empiricist point of view, is that the network structure used in the simulation is typically not based on empirical data. The network used may then be criticized as unrealistic or arbitrary. As a response to such concerns, one would ideally use the actual network structure that corresponds to the population of interest (e.g., Adams and Schaefer 2016; Wang et al. 2017). Such data can be very difficult to collect, however, as this requires information on all actors and all ties between actors (i.e., a full census of the population; Krivitsky and Morris 2015; Smith 2012).[2] These data may not be available and may not be cost-effective to collect, particularly when the point is to seed a realistic network in an ABM.

We consider the potential of a middle ground: where one uses independently sampled ego network data to infer a realistic full network, which is then employed in an ABM. Thus, instead of relying on full network data (most accurate but difficult to collect) or no empirical network data (least realistic but most convenient), the approach combines ABMs with recent work on network sampling (Luke and Stamatakis 2012).[3] Ego network data, in particular, are an ideal choice because they are so easy to collect. Individuals are randomly sampled from the population of interest, answering questions about themselves (such as demographic information) and the people they are socially connected to, or their network alters (such as friends or confidants; Marsden 1987; Smith, McPherson, and Smith-Lovin 2014). The data collection burden is low, as the data are based on a sample (rather than a census) and independent respondents (thus, it is unnecessary to identify and interview the named social contacts).

The "middle-ground" approach proposed here thus has the advantage of employing widely available, easy to collect data, while, potentially, still yielding a realistic network structure by which to seed an ABM. The validity of such an approach is dependent on getting realistic looking networks from independently sampled data. Traditionally,

this has been a very difficult task. Global network structure, by definition, depends on all of the ties between all of the actors; sampled data, in contrast, only provide bits, or pieces, of the network. Here, we apply the framework of Smith (2012), which uses exponential random graph models (ERGMs) to make global network inference from ego network data.

The question is whether a sampling approach will be effective in the context of ABMs of diffusion. Past work has tested a sampling approach on the features of the network (e.g., can we use ego network data to estimate transitivity?) but has not considered diffusion processes directly (Smith 2015). Such tests miss the behavioral, stochastic component at the heart of an ABM, where different adoption rules (i.e., under what conditions will an actor adopt a new product?) *combine* with the network structure to yield different diffusion outcomes. It is thus an open question if a diffusion simulation based on sampled data can yield the same insights as if one had the full network (see also Rolfe 2014).

We begin this article with a background section on ABMs of cultural diffusion, before turning to sections on ego network sampling and the inferential approach. We then discuss the test of the method before moving to the results.

ABMs of Cultural Diffusion

Past work on ABMs and diffusion has typically been applied to problems of disease spread or cultural transmission (e.g., Keeling and Eames 2005; Kitts 2006; Rocha, Liljeros, and Holme 2011). The basic question is how an outcome of interest (HIV, a new product) moves through a population via a social network, where an "infected" case can pass on the item to someone they are connected to, for example, through sexual contact. Here, we focus on the case of cultural transmission, such as adopting a new product. We focus on culture as a motivating example as much of the theoretical work using ABMs in sociology has focused on such outcomes (e.g., Baldassari and Bearman 2007; Mark 1998; Mäs and Flache 2013). Additionally, cultural models of diffusion rarely employ empirical data (compared to models of disease spread where this much more common; e.g., Merli et al. 2015; Morris et al. 2009), making the test particularly

appropriate. Practically, this means focusing on networks and diffusion processes that are appropriate for cultural, but not necessarily disease, transmission.

Past work on cultural diffusion has, itself, employed a wide range of networks and diffusion processes. A network sampling approach will be useful in different ways to different approaches, and we split the discussion into different sections based on the type of simulation.

Models Employing a Known Network Structure to Explore Cultural Change

One tradition in the agent-based modeling literature begins with a social network of desired, or known, properties and uses that network as the basis for the diffusion model. Here, the network is used to explore cultural change in the population. The network is typically held constant as actors adopt or drop cultural items across time based on interactions with other actors. The focus is thus on interpersonal processes, like influence and distancing, that shape the distribution of cultural items across the simulated population, given the network structure. For example, Centola and Macy (2007) use small-world networks to explore diffusion processes in the context of complex contagion. Small-world networks have certain signature features, with shortcuts between well-defined groups, or high clustering but short overall path length (Watts 1999). Centola and Macy (2007) explore the diffusion potential in such networks, comparing simple contagion (where a product could be passed on with only one friend already adopting) to complex contagion— where the product of interest is viewed as risky or uncertain, so that actors need more than one of their friends, or multiple signals, to take on the item of interest before they will consider doing so.[4]

More recent work by DellaPosta, Shi, and Macy (2015) uses a similar approach to explore cultural consumption across social space. They use a similar network structure but a somewhat more elaborate diffusion model, one that draws on McPherson's (1983, 2004) ecological model. They build a simulation where actors are positively influenced (so convergence) by those close in social space and negatively influenced (so divergence) by those who are distant in social space. The simulation proceeds by allowing the behaviors to evolve while holding the network structure fixed. The basic idea is that such processes

can lead to amplified correlations between demographic characteristics and cultural items (see also Mark 2003).

A network sampling approach is directly useful for models that employ known network structures to explore cultural change. These models depend on having a realistic network structure in the simulation, making a sampling approach directly applicable. A researcher could collect an ego network sample, infer the full network, and use that network as the basis for the diffusion simulation. The advantage of such an approach is that the researcher no longer has to come up with and justify the network used in the diffusion simulation; one could simply take the inferred network, based on the actual data, and use that as a realistic network structure within the simulation. Arguments over the chosen network (e.g., does it have realistic features?) would be cut off from the start.

Models Generating Networks From Microrules

Other work in the ABM tradition allows the networks to emerge within the simulation itself, rather than using a given network as a starting point (e.g., Carley 1991; Centola et al. 2007). Here, the researcher specifies the underlying mechanisms that make a tie more or less likely to form between two actors (e.g., sharing some demographic or cultural characteristic), varying the strength of those mechanisms in the course of the analysis. The network structures that emerge do not have predetermined structural properties, as in studies using an a priori network structure, but are dependent on the range of behavioral rules specified in the simulation. For example, DiMaggio and Garip (2011) use simulation to explore network externalities. The model is designed for cases where the adoption of a new product or innovation, such as the Internet, depends on the number of people in ones' immediate social network that has already adopted (thus a local threshold model of adoption; see also Granovetter 1978). The simulation varies two main parameters: the nature of this threshold effect and the strength of homophily on demographic dimensions like race and education (controlling the tendency for similar individuals to interact).[5] Centola (2015) uses a similar setup to explore the effect of homophily on diffusion in the context of complex contagion (see also Gondal 2015).

A network sampling approach can be used to inform generative models in a number of ways. First, a researcher could use the ego

network data to inform the simulation itself. The ego network data are thus used to inform the rules of interaction employed in the simulation. In this case, the inferred network (from the ego network data) serves to constrain the simulation, showing the range of input parameters that are empirically realistic. Second, the inferred network could be used to judge the output of the simulation. Here, the inferred network is used as a gold standard, showing which generated networks are consistent with the empirical data. The inferred network is not used in the simulation itself but is rather used after the fact as a check. The question is what set of microrules could have yielded the inferred network structure and cultural diffusion curves.[6]

Coevolution of Network Ties and Behavioral/Cultural Items

Many ABMs are specifically concerned with the coevolution of network ties and behavioral (or cultural) items (Mark 1998; Baldassari and Bearman 2007). Here, actors change their behavior while simultaneously adding and dropping ties. Much of the literature on network/behavioral coevolution falls in the stochastic actor-based model (SABM) tradition (Snijders, Van de But, and Steglich 2010). SABMs use simulation to estimate parameters on complete, longitudinal network data (Steglich, Snijders, and Pearson 2010). The model takes the observed network at time T and $T + 1$ and asks what processes (e.g., selection/influence) could account for shifts in network ties and behaviors (e.g., alcohol use) over time (e.g., Schaefer, Haas, Bishop 2012). SABMs are primarily used to estimate parameters, but it possible to employ these models as a general simulation framework. The researcher would use an empirically grounded model to explore the coevolution of network ties and behaviors under different theoretical conditions (Schaefer, Adams, and Haas 2013; Wang et al. 2017). For example, Adams and Schaefer (2016) demonstrate the effect of increasing/decreasing peer influence on the level of smoking across schools (using complete longitudinal network data to parameterize the base model).

Our test of a network sampling approach does not, ultimately, fall in this tradition despite the potential use of ego network data within SABMs and similar models.[7] Instead, we focus on simulations where the network structure is held fixed and only the item of interest (i.e., a new product) is allowed to shift over time. We focus on this "static" test for two reasons. First, using a static network makes it easier to

assess a network sampling approach. In a dynamic network model, the network updates over time based on the specified model; the network may wander far from the original, inferred network (reflecting, in part, the parameters of the ABM itself), making it hard to directly assess a sampling approach. Second, and more substantively, a fixed network structure is often used in simulations (e.g., DellaPosta et al. 2015). For example, a fixed network makes it easier to see how behavioral rules combine with network structure to affect global outcomes; for example, a small-world network may be conducive to quick diffusion under simple but not complex, contagion (Centola and Macy 2007; Rolfe 2014). This opens up questions of how different contexts facilitate or hinder diffusion and why we might expect differential adoption across subgroups (DiMaggio and Garip 2011). By using a fixed network, we can test if a sampling approach can be applied to such questions, reflecting the joint effect of network structure and adoption behavior on diffusion.

Ego Network Sampling

This article explores the payoff of using sampled network data to inform ABMs of diffusion. Sampled ego network data are used to infer complete network structures, which are then used within an ABM. We test this approach by running diffusion models through inferred networks (based on sampled data), asking how close those simulations are to the analogous simulation on the full, true network. If the sample-based simulations are close to the true simulations, then a researcher could plausibly seed an ABM using a bit of sampled data—rather than trying to collect full census data or using no empirical network at all.

A sample-based approach only makes sense, of course, if it is possible to make inference about the network structure from sampled data. We present a short background section on network sampling before presenting a test of the approach.

Figure 1 outlines the basics of ego network sampling. Panel 1 in Figure 1 plots a typical, complete network structure. This is the ideal case. A researcher has collected information on all nodes, or actors, and all ties between nodes in the network. This information is sufficient to map out the paths between nodes in the network as well as

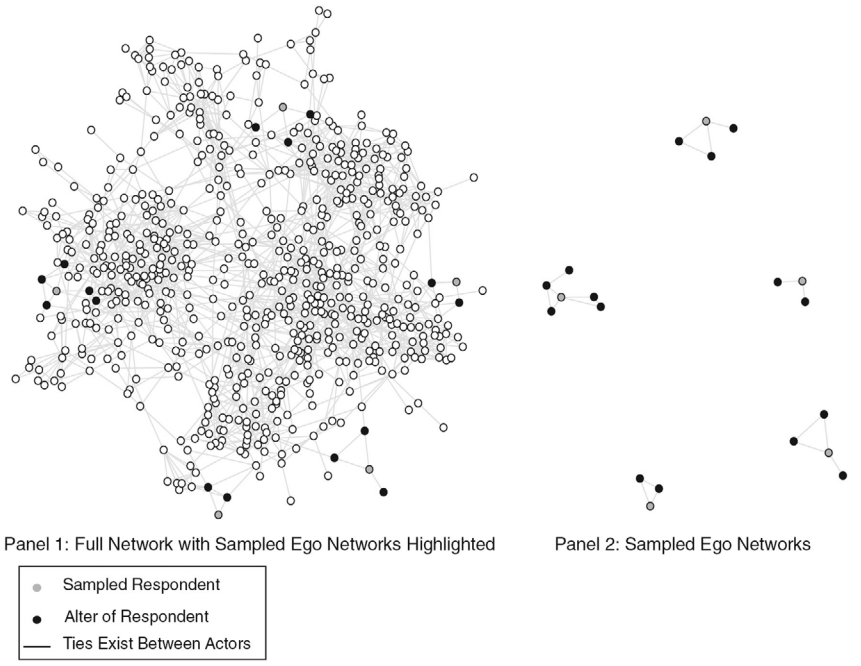

Figure 1. Example of an ego network sample from full network. (Panel 1) Full network with sampled ego networks highlighted. (Panel 2) Sampled Ego Networks.

to measure other global network features of interest. A complete network could also be used within an ABM, for example, serving as the network structure in a simulation exploring cultural diffusion.

Many times, however, it will be impractical to collect complete network data. The network may be too large or the resources too scant to interview everyone in the network. In such cases, one must make inference from a sample. There are a number of ways of sampling a network and thus reducing the data collection burden of the researcher (Handcock and Gile 2010; Thompson and Frank 2000). We focus on the simplest option, an ego network sample, as ego network data are (a) easy to collect data and (b) surprisingly rich, offering useful information that can be used to infer the global features of the network (Smith 2012).

Ego network data are based on a random sample of respondents, where each respondent reports on their local social network. Panel 1 highlights a hypothetical ego network sample from the full network. The gray nodes, our egos, represent the randomly sampled respondents (a subset of all individuals in the network). The white nodes are the nonrespondents, while the black nodes are the named alters

(e.g., friends) of the respondents. The black nodes are themselves not interviewed, but the researcher will receive information about them indirectly via the respondent's reports on each alter. We may know, for example, if the named alters know each other or not. An ego network sample will thus provide information on the gray nodes and the black nodes, offering independent pieces of the full network. These independent pieces are plotted in panel 2. Note that the survey does not collect identifying information on the named alters (i.e., the black nodes). This means that the sampled pieces of the network cannot be connected. It also means that the ego network data cannot be used to map particular edges in the network. We may know if ego has a dense personal network, but we cannot tell which nonsampled respondents are friends.

Ego network sampling thus poses a difficult inferential problem, as all of the network information must be "filled in" (as opposed to a simple missing data problem; Kossinets 2006; Smith and Moody 2013). The question is how to take information on the respondents and the named alters, or the sampled pieces of the network, and make inference about the structure of the entire network, here, for the purpose of incorporating a realistic network structure into an ABM.

Background on Inferential Approach

We draw on the work of Smith (2012), which provides a simulation solution to the problem of network inference. The basic idea is to take independently sampled ego network data, extract as much information as possible, and use that information to make inference about the full network structure. The simulation produces a set of networks that are consistent with all of the local information found in the sampled data. Networks that are consistent with the local information are likely to have similar global features as the real network. The approach is ultimately effective because it utilizes so much of the information found in the ego network survey.

An ego network survey will, most simply, provide data on the demographic characteristics of the respondents (gender, race/ethnicity, education, etc.). This makes it possible to generate networks with the correct demographic composition. An ego network survey will, more importantly, ask the respondents to list their alters, or those

individuals they are socially connected to (e.g., friends, discussion partners, sources of social support). For example, in Figure 1, the respondents in the middle row would list five and two alters, respectively. This information offers an estimate of the degree distribution, describing the number of alters per respondent. The data also offer information on differential degree, showing the mean degree by demographic groups (putting together the information on degree and the demographic characteristics of the respondents).

Ego network data also provide information on the named alters. Ego network surveys will typically ask the respondents to describe the demographic characteristics of the alters such as gender, age, or education. This alter demographic information can be paired with the respondent demographic information to measure homophily, or the tendency for demographically similar individuals to be socially connected (McPherson, Smith-Lovin, and Cook 2001; Smith et al. 2014). One can ask if the respondents and alters share the same gender, age, education, and so on. Similarly, the data capture the mixing between demographic groups, showing the frequency of social ties between each group. For example, are ties between college graduates and high school graduates more/less likely than ties between college graduates and PhD holders?

Ego network surveys also ask respondents to report on the ties between alters. Respondents will report on the existence of a tie between alters 1 and 2, 1 and 3, and so on (often limited to a small number of alters, say five, to limit respondent burden). The alter–alter tie data capture the local structural tendencies surrounding the focal respondents. Do individuals tied to ego also know each other, so that a friend of a friend (ego) is also a friend (Goodreau, Kitts, and Morris 2009)? Smith (2012) offered a novel characterization of the alter–alter tie data. Most work relies on density (the number of ties in the ego network divided by the number possible) to measure ego network structure (Fischer 1982; Mardsen 1987). Ego network density does not, however, offer a precise enough measure for the simulation: Many networks with the same (mean) local density have very different global network features (Smith 2012). Ultimately, the generated networks are based on the sampled information, making it important that the measure of local network structure is sufficiently discerning.

The measure by Smith (2012) uses the alter–alter tie data to construct a distribution of ego network configurations. Figure 2 plots the

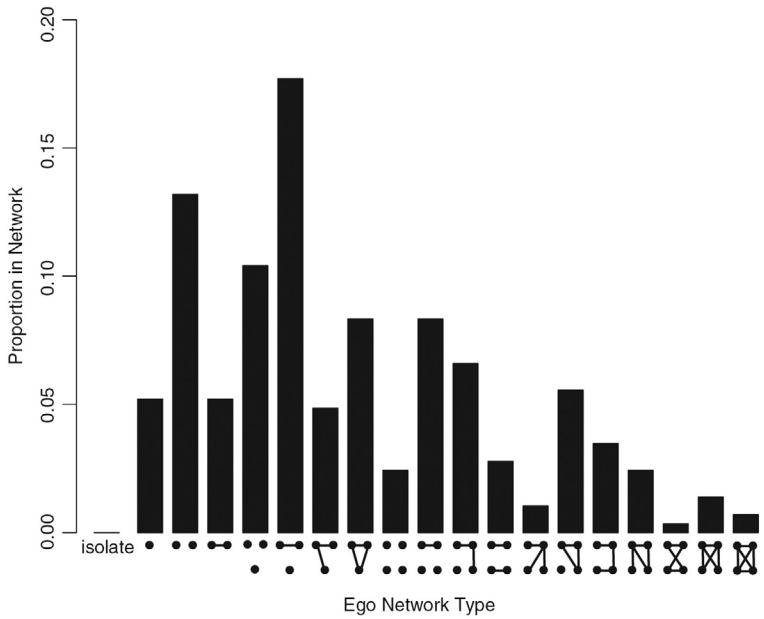

Figure 2. Example of ego network configurations. This figure is based on a hypothetical ego network configuration distribution. Ego is not included in the ego network types. We only include ego network types of size 4 or less to make the figure legible.

ego network configuration distribution for the network plotted in Figure 1; note that the plot is limited to four alters for the sake of space, but the actual distribution is not limited to four alters. The histogram presents the ego network configurations on the x-axis and the proportion in the network on the y-axis. Each respondent is categorized as a distinct configuration based on the size of the ego network and the pattern of ties between alters. For example, the top ego in Figure 1 would fall into the fourth configuration from the left (as they have three alters with no ties between them). A distributional approach offers a more discriminating measure than density because it captures the full pattern of ties within the ego network. Ego networks of the same size and density can have different structural patterns, but this is missed using traditional summary measures. See Smith (2012) for technical details on how to place each ego into a structural type.

The simulation approach constructs full networks that are consistent with each piece of information extracted from the ego network sample: the degree distribution, differential degree, homophily, and the ego network configuration distribution. The generated networks are thus heavily constrained by the empirical data, making it

more likely to have the same global features as the true network. The method draws on ERGM to simulate the networks. We briefly discuss ERGMs before turning to the approach itself.

ERGMs

ERGMs are statistical models used to test hypotheses about network structure/ formation (Hunter et al. 2008; Wasserman and Pattison 1996). Formally, define a network, Y_{ij}, over the set of nodes N (N = 1, 2, . . . , n), where Y_{ij} = 1 if a tie exists and 0 otherwise. Define y as the observed network. Y is then a random graph on N, where each possible tie, ij, is a random variable. The ERG models the Pr($\boldsymbol{Y} = \boldsymbol{y}$). The "independent variables" are counts of local structural features in the network (Goodreau et al. 2009; Robins et al. 2007), such as number of ties and homophily (e.g., the number of ties that match on gender). The model can be written as:

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

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

ERGMs are generally used to test hypotheses about the formation of a network, but it is also possible (and increasingly common) to simulate networks based on a specified model (e.g., Morris et al. 2009; Robins, Pattison, and Woolcock 2005). The coefficients reflect the effect of different local processes on tie formation. These coefficients can then be used to predict the presence/absence of a tie between actors in a generated network. A researcher must specify two items prior to the simulation: The model terms and coefficients used to generate the network. Here, the goal is to construct networks that are consistent with the local information found in the ego network sample. The model is thus specified with the ego network data in mind. The model terms are based on the information available from the survey, while the coefficients should generate networks with the same local features as observed in the ego networks.

Simulation Approach

The simulation approach uses ERGM to generate networks consistent with the sampled data. We describe the approach in general terms here, but see Smith's (2012, 2015) study for technical details.

The simulation begins by constructing a network of size N (the size of the population of interest) with the correct degree distribution. The degree distribution is estimated from the sampled data. Demographic characteristics are then assigned to the nodes in the generated network. The assigning of characteristics is done to be consistent with the empirical data. Nodes from the simulated network are matched to sampled respondents with the same degree; each selected node is assigned the characteristics of that respondent. This ensures that demographic groups with higher degree in the sample also have higher degree in the simulated network (thus making sure that differential degree is correct in the generated networks).

The next step is to estimate the initial ERGM coefficients. The model includes homophily terms for every demographic dimension available in the empirical data (specifically the full mixing matrix for each dimension). The homophily coefficients are estimated using case control logistic regression (Smith et al. 2014). The degree distribution and differential degree (incorporated while seeding the initial network) are maintained throughout the simulation as well. The model will also include a term for geometrically weighted edgewise-shared partner (GWESP) distribution. GWESP captures the distribution of shared partners (if i is tied to j, how many common friends do i and j have?), capturing higher order transitivity in the network, or the tendency for local clusters to emerge. The coefficient for GWESP is set at an initial value and updated during the simulation as the method searches for the best-fitting network.[8]

The framework takes the ERGM coefficients, terms, and constraints and simulates an initial network using the seeded network as a starting point. This network is evaluated on how well it captures the ego network configuration distribution seen in the sampled data. The method uses a w^2 value to compare the ego network configurations found in the simulated network to the distribution seen in the sample. A small w^2 value suggests that the ego network configurations found in the simulated network are found at the same rates as in the sample. The initial coefficient on GWESP is then updated to find a

better fitting network, where a better fitting network means having an ego network configuration distribution that more closely matches the true distribution (conditioned on the other local features in the sampled data). Note that the homophily coefficients must be updated as the simulation looks for a better fitting network (see Smith 2012). The homophily rates in the simulated network are compared to the rates in the sampled data to ensure that there are no discrepancies as GWESP changes. Formally, a case control model is used to update the coefficients, comparing the true rate of homophily in the sample to the rate in the simulated network and adjusting accordingly. The whole process is repeated until it is not possible to improve the fit by updating GWESP and/or the homophily coefficients.

As a final step, the framework simulates networks from the best coefficients found during the search process. The framework keeps all generated networks with a w^2 statistic that is below a certain threshold (i.e., within 30 points of the minimum w^2 value based on the best-fitting network). The final product is thus a set of candidate networks that are consistent with the local, sampled data. The networks will match on the degree distribution, differential degree, homophily, and the ego network configuration distribution. A researcher may then use all of the candidate networks in their analysis (i.e., run ABMs through each generated network). Alternatively, for simplicity, they may limit their analysis to a single network; for example, selecting the best-fitting network among the set of candidate networks. The selected network can then be used to measure global network features of interest or, as in this article, to employ as a realistic network in an ABM.

Analytical Setup: Testing the Method

The core question of this article is whether this approach will be useful for ABMs, so that a diffusion simulation based on sampled data will yield the same insights as if one had the full network. If so, then it would be worthwhile to collect ego network data, infer the full network structure, and use that to condition an ABM. There are six basic steps to testing the validity of a sampling approach: (1) select a known, complete network as a test case; (2) run an ABM of diffusion through the known, complete network; (3) take an ego network sample from the complete network; (4) generate a full network

consistent with the ego network data; (5) run an ABM (same as in step 2) through the inferred network from step 4; and (6) compare the diffusion estimates from the sampled data (step 5) to the estimates from the known, complete network (step 2).

Selecting Known, Complete Networks

The first step in the test is to gather a set of networks to act as the complete, known networks. It is necessary to have complete networks in order to assess the validity of the approach. The complete networks serve as a baseline to judge if the sample-based diffusion results are valid. The exact networks chosen are not especially crucial, as the inputs to the test (the sample) come from the network of interest, and the test is thus self-contained, that is, we only care if the method can replicate a given network from a sample *on that population*. It is nonetheless important to have networks with a variety of features, as it is important to see if the method is inappropriate for certain settings. We thus use synthetic, generated networks as our test cases. Generated networks are ideal as they can be fully controlled, yielding the desired properties for each network. This makes it easier to compare the results across networks and to pinpoint where the method does, and does not, work. The networks include 700 nodes and are based on a symmetric relationship; for concreteness, we can assume that a tie is defined by friendship. The networks are based on an empirical school setting, as the network follows the size and demographic distribution of one Add Health network (e.g., McFarland et al. 2014). The networks have the same race and grade distribution as in the empirical data.

We generate four different networks. Each network has the same basic composition but different network features. We specifically vary two network features known to be important for diffusion processes: density and transitive closure (Moody and Benton 2016). Density captures the total number of edges in the network relative to the number possible. Transitivity captures the proportion of two steps that also include a direct tie; or, is a friend of a friend also a friend? Each network contains 700 nodes with the same distribution of race and grade but with different levels of density and transitivity. In the case of simple contagion (so one infected alter is enough to pass on the “disease”), denser networks with low transitivity should have faster global diffusion. This is the case as there are more nonredundant ties, or more

edges reaching out to individuals in socially distant groups (and more technically, there is a higher number of independent paths; Moody and White 2003). We include a variety of networks to test if the approach can differentiate networks with different features; that is, can we tell if the network in question has high/low diffusion potential just using sampled data? More generally, the variation in network structure makes it easier to judge the validity of the sampling approach, as we can see how the approach fares under different conditions.

We include two levels of density and two levels of transitivity: high density, low transitivity; high density, high transitivity; low density, low transitivity; and low density, high transitivity. The four networks are plotted in Figure 3. The first network, high density, low transitivity, has density of .015 and transitivity of .012. This amounts to a random network with a mean degree of 10.26. In Figure 3, it is clear that the network is dense (as there a large number of edges) and has little group structure. Moving to the right, the high density, high transitivity

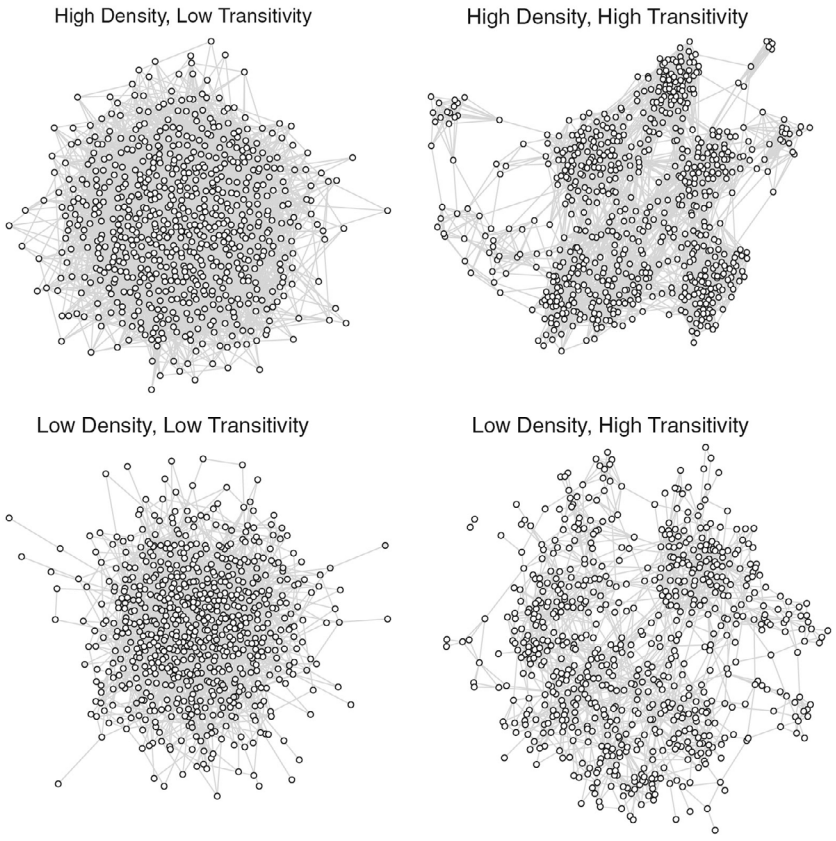

Figure 3. Networks used in testing a sampling approach.

network has the same mean degree (10.26) but a much higher level of transitivity. The rate of transitive closure is .306, so that 30 percent of all possible triangles are closed. At a more aggregate level, the high level of transitivity yields strong group divisions. The bottom two networks have the same basic form, but the density decreases from .015 to .007, with mean degree of 5.13. The low density, low transitivity network has transitivity of .012, while the low density, high transitivity network has transitivity of .307.

Thus, we have four networks, varying systematically by density and transitivity. The high-density networks have the same density, the high-transitivity networks have the same transitivity, and so on. This facilitates comparison, as the networks share everything in common except the level of density or transitivity.

The generated networks thus have the advantage of offering a controlled set of comparisons. There are, however, potential drawbacks to using synthetic networks: Most crucially, the networks may be unrealistic, offering a too simple "toy" test of the approach. With this concern in mind, we have also included the original, raw Add Health network in the set of test networks. Here, the network features are not set a priori but are determined solely by the nomination data in the original survey.

We must also recognize that the Add Health–based networks represent a test of the approach on relatively small networks, only 700 nodes, while many studies want to explore the properties of larger networks (where the returns to sampling are largest). We thus present an additional analysis on a much larger network, the Colorado Springs drug-user network, size = 5,492. The results are presented in Online Appendix B and offer an important supplement to the main analysis.

ABM of Cultural Diffusion

Step 2 runs an ABM of diffusion through the complete networks from step 1. This serves as the baseline by which to judge the sample-based results. The model of diffusion is based on a simple premise that there exists a new product or innovation and that this product can be passed on through social ties (i.e., individuals introduce new products to their friends). The actors in our ABM are the 700 students in the seeded network. The network structure, and thus the interaction partners of our actors, is held fixed throughout the simulation. We assume a

simple contagion process, so that the probability of "infection," or taking up the new product, is based solely on the probability of direct transmission from actor *i* to actor *j* (Keeling and Eames 2005). Thus, the product can be adopted even if an actor only has one friend who has adopted. We also assume that it is possible to drop the product once adopted. Actors that have "recovered," or dropped the product, are susceptible to adopting again in the future.

The simulation begins with one randomly selected seed, designated as the first adopter in T1 (time period 1). In T2 (the second time period), our first adopter interacts with their immediate neighbors. The neighbors are introduced to the new product within this interaction, with each neighbor adopting the new product with probability *p*, set at different values in different analyses. We run three different analyses, setting the probability of adoption at .1, .2, and .3. There are two goals here: first, to make sure that the results are robust to model specification; and second, to see if a sampling approach can capture the substantive changes in diffusion that result from shifts in adoption behavior. More formally, if *i* is tied to *j* and *i* has already adopted the product, then *j* takes up the product with probability *p* (.1, .2, or .3). Once the product is adopted, an actor has .2 probability of "recovering" or no longer using the product of interest. Every actor who has already adopted considers dropping the product after every time period. The analysis is run over 30 time periods. Each time period allows new individuals to adopt (if they are connected to someone already adopting the innovation) and current adopters to drop the product. The results from the simulation are summarized as the proportion adopting after each time period, or a cumulative distribution of adopters. The whole process is repeated 1,000 times, and we summarize the results as the mean over the 1,000 runs. For step 2, the diffusion model is run through the five known, complete networks.

Sampling Setup

The third step in the analysis takes ego network samples from the true, complete networks. This serves as input for step 4, where we make inference about the full network structure from sampled data. We assume that a 25 percent ego network sample is taken for each network. Our hypothetical survey thus has 175 respondents (for each network). The survey is hypothetical in the sense that no

respondents are actually interviewed and all information on the sampled actors is taken from the true network. We assume that the following information is "collected," mimicking a typical ego network survey: the number of alters for each respondent, the race and grade of each respondent, the race and grade of each named alter, and reports on the ties between the named alters. The alter–alter tie information and the alter characteristics are restricted to only five alters in order to mimic actual surveys where time and fatigue are often a problem (e.g., Burt 1984). As this is not an actual survey, the five alters are randomly selected from the set of all alters for that "respondent" (acting as the five alters they chose to report on). There is no limit, however, on the number of named alters. This process is repeated 100 times for each network. Each time through we take a new sample from the original network, making it possible to assess variability due to sampling.

Using Sampled Ego Network Data to Infer Full Networks

Step 4 takes the sampled ego network data (for each network/sample) and uses the simulation approach by Smith (2012) to make inference about the full network. This means generating networks consistent with the local properties found in the sampled data. In this case, we assume that the researcher only keeps one generated network (per sample) to be used in the ABM.

Running Agent-based Diffusion Model on Inferred Networks

Step 5 takes the best-fitting networks from step 4 (one for each network/ sample) and runs the same agent-based diffusion model as in step 2. The parameters and setup are exactly the same. The only thing that is different is the network structure used within the simulation. In step 2, the simulation uses the complete, known networks. In step 5, the simulation uses the networks inferred from the sampled data. Again, we run the simulation 1,000 times for each network, summarizing the results as the proportion adopted after each time period. We again take the mean over the 1,000 iterations. Note that there are $100 \times 3 \times 5$ estimates, as there are 100 samples, three adoption probabilities, and five networks in the analysis.

Compare Diffusion Estimates From Sample to Diffusion Estimates From Known Network

The final step in the analysis compares the sample-based diffusion estimates (step 5) to the population-based diffusion estimates (step 2). We compare the distribution of adopters based on the sampled data to the distribution of adopters based on the complete networks. The distribution of adopters is measured as the proportion of adopters after each time period.

Results

We begin the results section by looking at the true diffusion curves. The results are based on the five known, complete networks and represent the standard by which the sampling-based results will be judged. In particular, it is important to note the differences across networks and adoption rates, so we can determine if the sampling approach is working as intended. Figure 4 plots the empirical diffusion curves. There are three subplots, corresponding to the three adoption probabilities used in the diffusion simulations. The *y*-axis plots the proportion of people in the network that have adopted the new product (using the mean over the 1,000 iterations).[9] The *x*-axis captures the time periods in the simulation.

Figure 4 clearly shows the effect of network structure on diffusion processes (Moody and Benton 2016). Looking within subplots, it is clear that the network with high density and low transitivity (the green line) has the fastest diffusion rates through the network. This is the case as the diffusion simulation is based on simple contagion. For example, looking at the third subplot (where adoption is equal to .3), we can see that the proportion adopting the innovation increases quite rapidly, with a sharp increase between periods 5 and 7. The network reaches saturation (so that the proportion adopting does not increase after that period) by the eighth time period, with a saturation proportion around .8. The high-density, high-transitivity network (the blue line) has the same basic shape, but the diffusion curve is considerably less steep and saturation happens later. The high-density, high-transitivity network has slower diffusion as the network has stronger

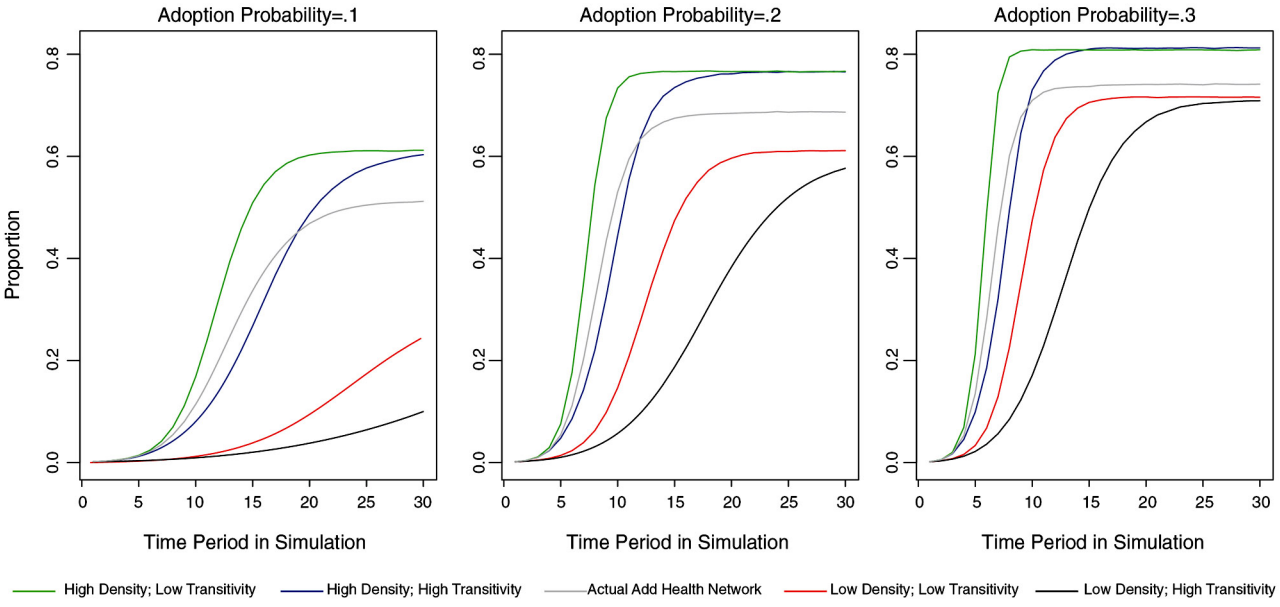

Figure 4. True diffusion curves by adoption probability.

group divisions, meaning it takes longer for a new product to exit a group once enmeshed (as most ties go within the group). Diffusion still happens quite fast, however, as the network is so dense that an absolutely high number of ties connect distant groups, facilitating global diffusion.

The low-density networks have curves even further to the right of the plot, indicating slower rates of adoption. The low-density, low-transitivity network (the red line) can be directly compared to the high-density, low-transitivity network (the green line). Both are random networks with low levels of transitivity; they only differ in the total number of edges. Looking at the red line (low density, low transitivity) in the third subplot (adoption probability equal to .3), saturation does not occur until time period 15, twice as long as with the high-density, low-transitivity network. The low-density, low-transitivity network also has slower diffusion than the high-density, high-transitivity network. The low-density network has slower diffusion because fewer new adoptions are possible in each time period (due to the low number of social connections). The low-density, high-transitivity network (the black line) offers the extreme case of slow adoption: The overall density is low and strong group divisions make diffusion difficult. The saturation point does not occur until time period 25 (looking at the high adoption subplot).

Finally, the empirical Add Health network (the gray line) falls somewhere in the middle of the four curves. The empirical network has higher density than the low-density networks but lower density than the high-density networks (with mean degree of 8.74). Similarly, the transitivity in the empirical network is in between the values of the low/high-transitivity networks (.189). The diffusion rates are thus in between those of the constructed networks.

Figure 4 also clearly shows the importance of adoption behavior in structuring the diffusion curves. For all networks, the pace of diffusion is faster when the adoption probability is higher. This effect is particularly strong in the low-density networks. For example, for the low-density, low-transitivity network (the red line), the proportion adopting at period 30 is .247 when the adoption probability is .1, .611 when it is .2, and .716 when it is .3. The effect of adoption behavior is similar, but weaker, in the high-density networks. Here, for example, in the high-transitivity network, the proportion adopting at period 30 goes from .603 to .812 as we move from .1 to .3 adoption probability.

Sampling Results

The sample-based results are plotted in Figures 5–7, one figure for each adoption probability (low = .1, medium = .2, high = .3). Each figure has ten lines, with two lines for each network: a solid line corresponding to the diffusion curves based on the known, complete network; and a dotted line corresponding to the diffusion curves based on the sampled data. Curves from the same network (true and sample based) have the same color (e.g., green is the high-density, low-transitivity network). For each figure, we compare the solid and dotted lines, showing the difference between the true and estimated values (the true lines are the same as in Figure 4). Note that the sample-based lines correspond to the means over all 100 samples. We consider sampling variability below.

The results in Figures 5–7 are quite encouraging: The estimated values are, on average, very close to the true diffusion curves. Across all figures, the dotted lines very closely approximate the true diffusion curves and this is true for each network. Note also that the fit for the empirical Add Health network (the gray line) is quite good, on par with the synthetic networks. We can thus be confident that the results are not dependent on using synthetic networks. The results are

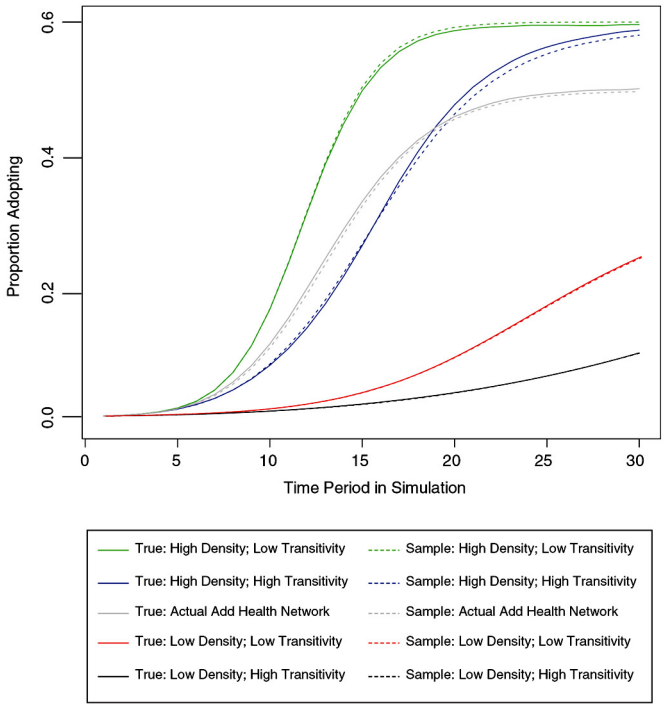

Figure 5. Comparing diffusion curves from true networks to sampled-based estimates (low-adoption probability).

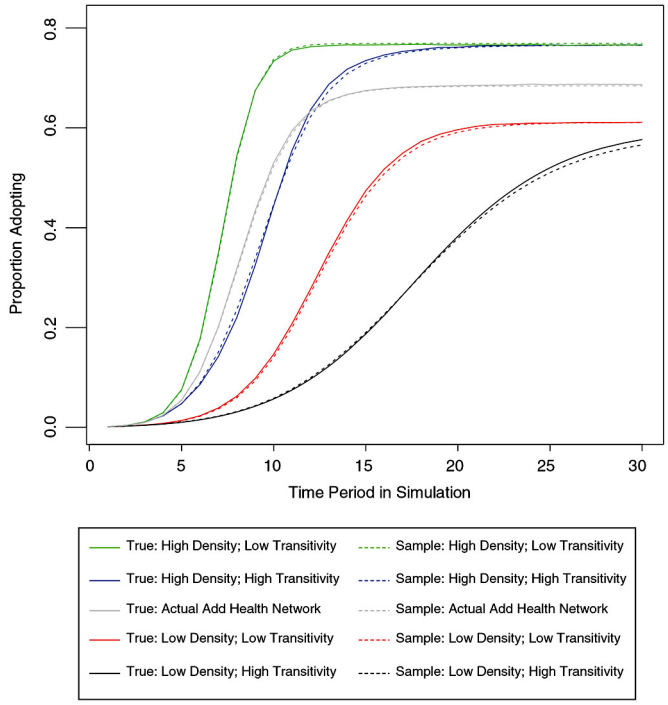

Figure 6. Comparing diffusion curves from true networks to sampled-based estimates (medium-adoption probability).

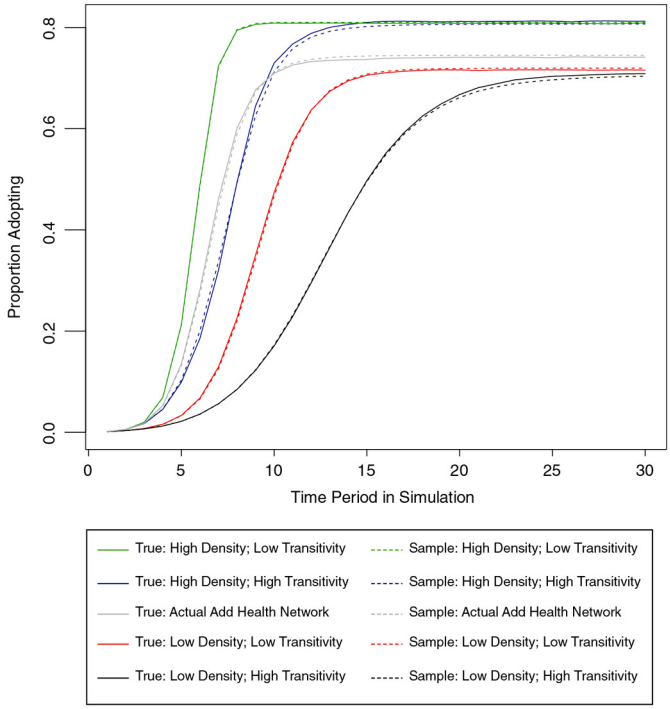

Figure 7. Comparing diffusion curves from true networks to sampled-based estimates (high-adoption probability).

also consistent across adoption probabilities: The sample-based diffusion curves closely approximate the true values as the probabilities of adoption increase. This suggests that sampled network data can be used to explore the effect of adoption behavior and network structure on diffusion. For example, just based on the ego network data, it is clear that adoption behavior has the greatest effect on diffusion in the low-density networks. It is also clear that diffusion is faster in high-density, low-transitivity networks. This initial snapshot is important, suggesting the potential payoff from a sampling approach. The results suggest that a researcher could run a diffusion simulation using sampled data and return the same results as if they had used the full network. The researcher would know, just from the sample, what kind of social world they are investigating: whether it is a world of quick diffusion and saturation or a world of slower diffusion and incomplete adoption.

We present more formal results in Figures 8–11. There is one figure for each network. The three subplots present the true proportion adopting (dots), the mean sample estimate (dotted line), and the 95 percent error bounds (solid lines), such that 95 percent of the sample

estimates fall within that bound. We also include an online appendix figure (Figure A1) that presents the relative bias for each adoption probability and network. Relative bias is defined as:

$$\frac{E(\text{estimates}) - \text{true value}}{\text{true value}}$$

We start with the high-density, low-transitivity network. It is clear from Figure 8 that the sample-based results closely approximate the true diffusion curves (i.e., the black dots fall close to the sample estimates) and do so with relatively little variation. The bias is quite small across all adoption probabilities. For example, the bias is under 1 percent for every time period when the adoption probability is set at .3. Similarly, the mean level of bias (over all 30 time periods) is .006 when the rate is set at .2. This suggests, on average, that the sample-based estimates will yield the same proportion of adopters

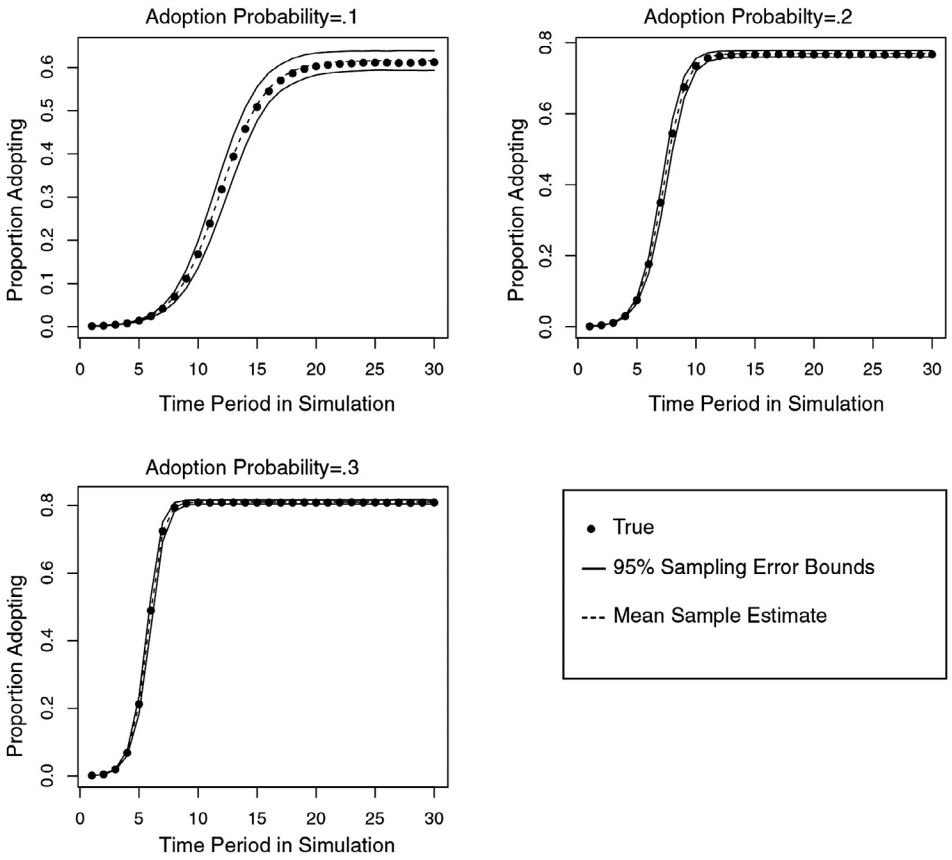

Figure 8. High-density, low-transitivity diffusion curve results.

as in the full network. The variation is also quite modest. The highest variation occurs in those periods that constitute the upward slope of the diffusion curve—those in-between moments between very low adoption and saturation. It is not surprising that the variance is highest here. Large changes occur in this transition phase, meaning sample-to-sample variation in the takeoff point (when the adoptions rates start to increase rapidly) leads to variation in the proportion adopting. The variation is relatively small even for this transition phase, however. For example, with high adoption (.3), time period 6 has the highest variance across samples with a standard error of .025. The true value is .489, while 95 percent of the sample estimates fall between .44 and .53. The results are similar for the other adoption probabilities, for example, the highest standard error is .026 when the adoption probability is set to 1.

The high-density, high-transitivity network offers a more difficult test and is presented in Figure 9. The bias here is still quite low, with

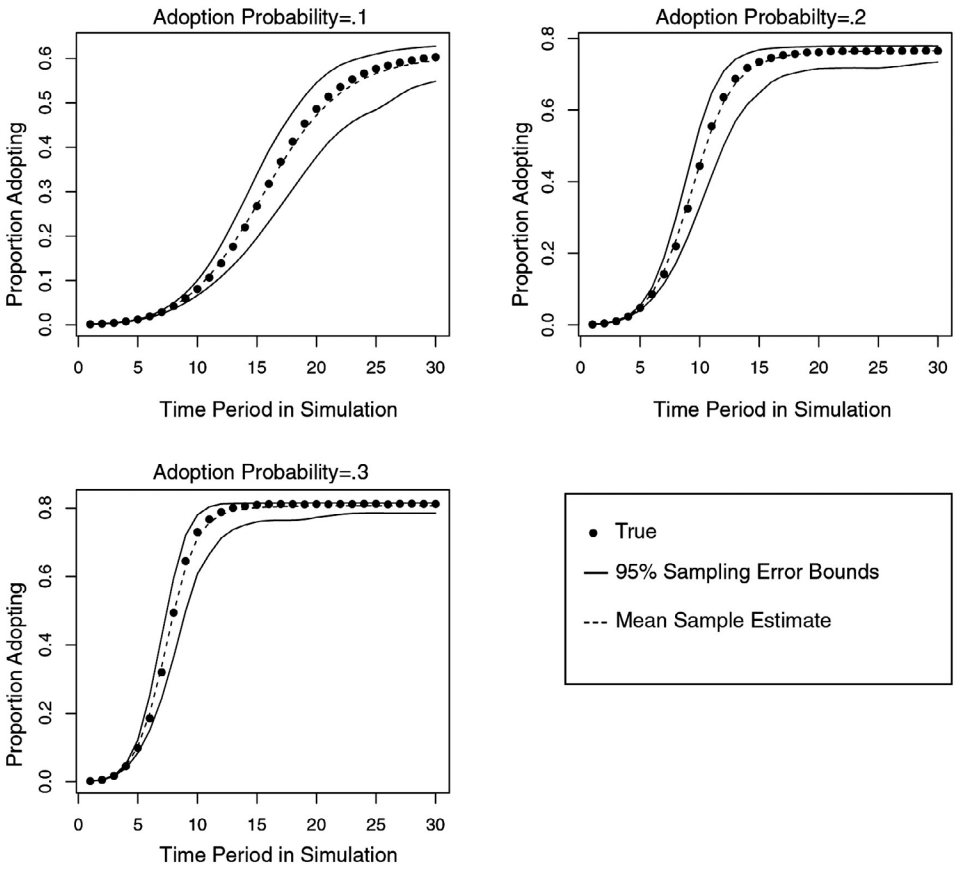

Figure 9. High-density, high-transitivity diffusion curve results.

the median bias over the 30 time periods under 1 percent for the high adoption analysis (adoption probability = .3). For example, in period 4, the true proportion of adopters is .045, while the mean value over the samples is .046, a bias under 1 percent. The bias is, however, higher than with the high-density, low-transitivity network. The low-transitivity network has bias under 1 percent for all time periods; this is not the case for the high-transitivity network, where seven periods have bias exceeding 1 percent (with a mean of .036 for those seven periods). Similarly, the median bias (over the 30 periods) for the low adoption analysis is .018, higher than with the low-transitivity network (only .008 median bias). The high-transitivity network yields higher bias for two reasons: First, the underlying network structure is harder to capture from sampled data; and second, the diffusion curve itself spends more days in a transition phase (where the adoption rate starts to pick up quickly), and this region is the most difficult to accurately estimate. The variance is also considerably higher than under the low-transitivity case, with standard errors ranging from .00001 to .063 in the high adoption case. The highest standard error under the low-transitivity case is only .025. We see similar results in the low and medium adoption analysis, with median standard errors (over the 30 days) of .019 and .028. The analogous values in the low-transitivity network are .005 and .012. The variance and bias are, however, low enough that the estimates provided by a network sampling approach still offer an excellent approximation of the true network, particularly when the goal is to seed a realistic looking network for an ABM. For example, for period 12, the true proportion who adopt in the full network is .788 in the high adoption analysis, while the mean in the samples is .781, a bias under 1 percent. 95 percent of the estimates fall between .71 and .81.[10] For the low adoption analysis, 95 percent of the estimates fall between .11 and .18 (with a true value of .144) for period 12, sufficient to show the effect of adoption behavior on diffusion.

How does a sampling approach do when density is lower? Figure 10 presents the low-density, low-transitivity results. Like the high-density, low-transitivity network, the bias is low. For the high adoption case, the bias is under 3 percent for every single time period and under 1 percent for most (with a median under 1 percent). For example, for period 11, the proportion adopting from the true, complete network is .573, while the mean over the samples is .568, a bias under 1

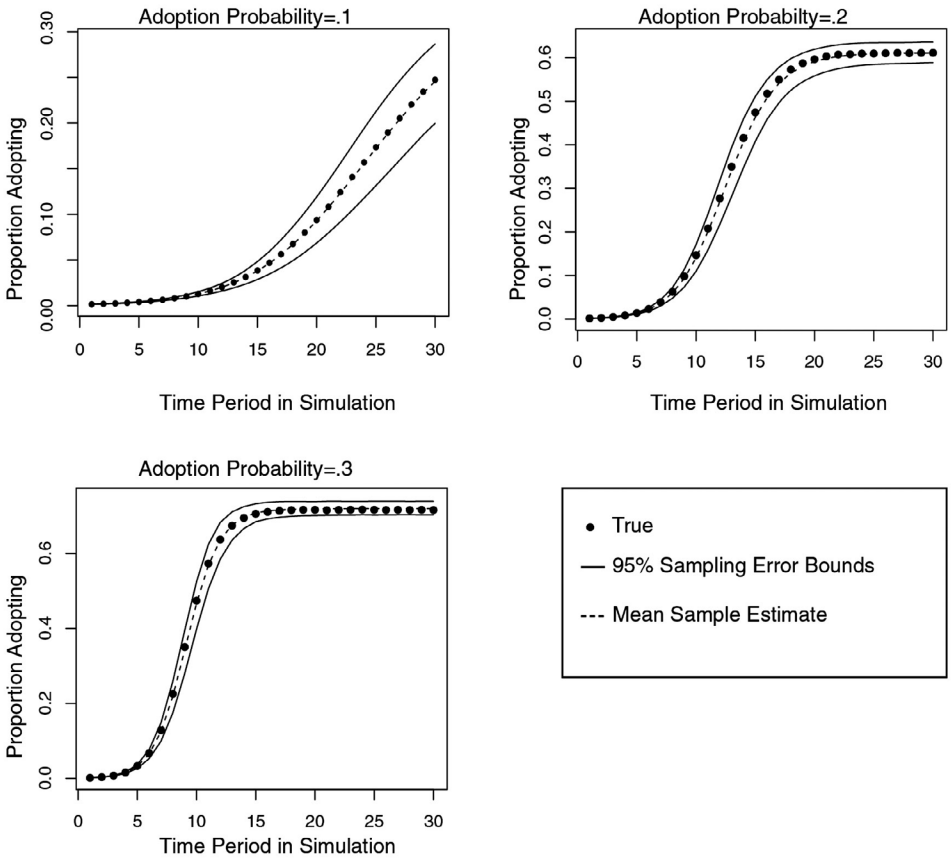

Figure 10. Low-density, low-transitivity diffusion curve results.

percent. The bias is similarly low when the probability of adoption is low. The mean bias over the 30 periods is under 1 percent. For example, at period 20, the true value is .094, while 95 percent of the sample-based estimates fall between .068 and .119 (with a mean of .093). These estimates are sufficient to show the effect of adoption behavior on diffusion: Under medium adoption, 95 percent of the estimates fall between .556 and .623 for period 20 and the true value is .596.

Figure 11 presents the results for the low-density, high-transitivity network. The results are largely consistent with the other networks. The median bias is .011, .019, and .008 in the low, medium, and high adoption probability analyses. For example, for the medium adoption analysis, the true proportion adopted is .187 in period 15, while the mean over the samples is .190, a bias of 1.5 percent. The standard errors are, however, often higher than in the previous networks. For the medium adoption analysis, there is a median standard error of .032

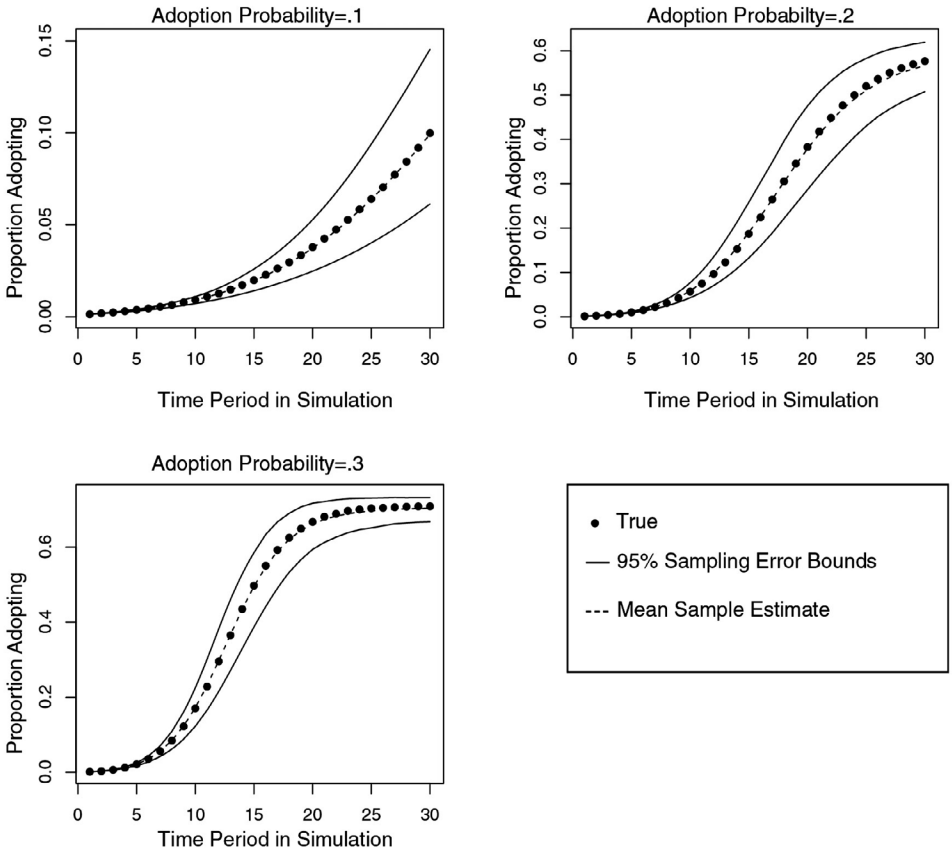

Figure 11. Low-density, high-transitivity diffusion curve results.

over all time periods compared to .013 in the low-density, low-transitivity network or .019 in the high-density, high-transitivity network. For example, for period 15, 95 percent of the sample values fall between .133 and .258. The high adoption estimates are also uncertain, with 95 percent of the values falling between .38 and .58 for the same period (note this is still sufficient to show the effect of adoption behavior on diffusion). The low-density, high-transitivity network has higher variance, in part, because it converges slower to a saturation point, where the variance across samples is higher.

Exploring Assumptions Through Additional Analyses

The results presented thus far are encouraging. A sampled-based approach can effectively mimic the diffusion curves from the true network, making it possible to capture the effect of network structure

and adoption behavior on diffusion. The results are, however, based on a constrained test of the approach. For example, the analysis has thus far been restricted to a simple diffusion model and has not considered alternative diffusion processes. Similarly, the analysis has assumed that the ego network data are measured without error. We examine each of these issues in turn, rerunning the analysis to examine the consequence of each assumption.

Alternative Diffusion Model: Homophily Based

We begin this supplemental section by replicating the analysis with a different diffusion model. The analysis thus far has followed simple diffusion, where the probability of adoption is the same across all *i*–*j* pairs, as long as they are socially connected. A more complicated behavioral model could relax this assumption, allowing the probability of adoption to vary across pairs of tied actors. For example, adoption may be more likely to occur when two people share a characteristic. Thus, individuals are more likely to mimic the behavior of friends who are similar to themselves (Centola 2011).[11] In this way, homophily (the tendency for similar actors to interact) has a dual effect on diffusion, where it shapes the network itself (creating social divisions based on demographic characteristics) as well as the adoption probability (Centola 2015; Salathé and Khandelwal 2011). The question is whether a sampling approach can capture the diffusion curves in this more complicated scenario.

Our homophily-based diffusion model extends the original model by allowing the probability of adoption to vary, depending on whether *i* and *j* share a key characteristic. We define the key characteristic as being in the same grade in school. Formally, we rerun the analysis as before, setting the probability of in-group adoption at .2 and out-group adoption at .1. If *i* is tied to *j*, and *j* has adopted the product, then *i* adopts with probability .1 if *j* is not in the same grade, and .2 if *j* is in the same grade. The networks are the same as in the main analysis.

Figure 12 presents the results for our homophily diffusion model. The results are presented as before, with simple plots describing the diffusion curves for each network (taking the mean over the samples). There are two subplots. The left-hand plots the simple diffusion results, where the adoption probability is the same within grade as across grade (set at .1). The right-hand plots the homophily results,

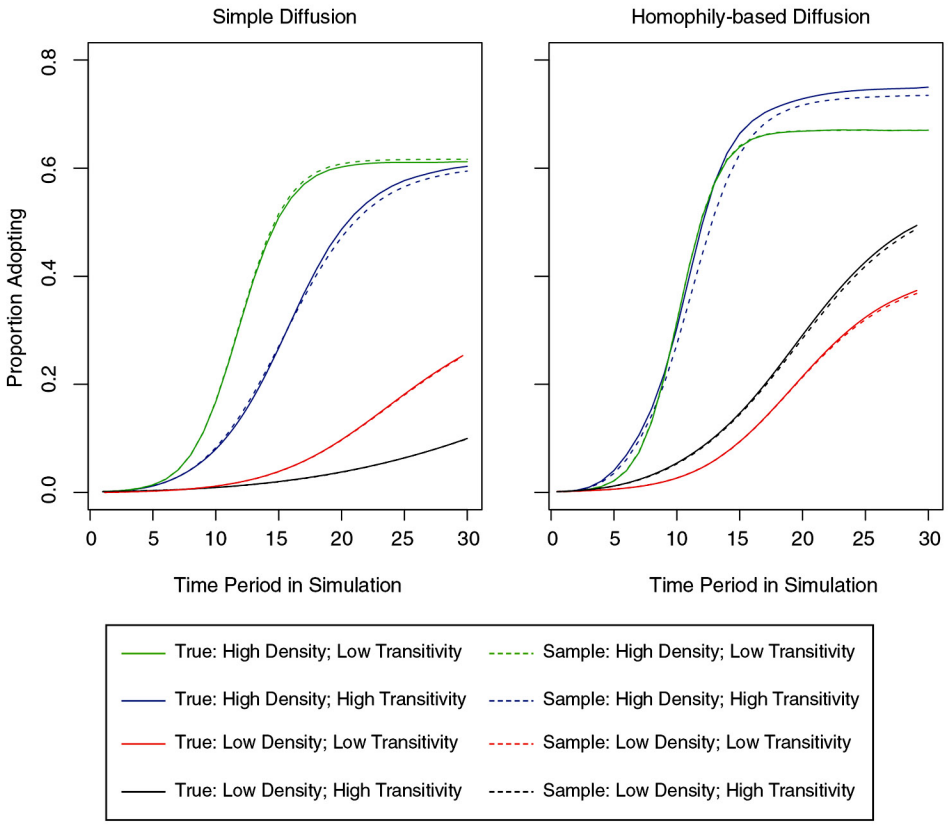

Figure 12. Comparing simple diffusion with homophily-based diffusion.

where the adoption probability within grade is double the adoption rate across grades. The simple diffusion results serve as a baseline for the homophily results.

Figure 12 clearly shows the effect of adoption behavior on diffusion. The true diffusion curves (the solid lines) show very different patterns across the two subplots. For example, under simple diffusion, the low-density, low-transitivity network has a steeper diffusion curve than the low-density, high-transitivity network (so diffusion happens faster). The exact opposite happens in the homophily simulation. This is the case as the high-transitivity network has higher levels of homophily on grade (so individuals who are friends tend to be in the same grade) than the low-transitivity network. Under simple diffusion, homophily on grade creates barriers to diffusion, and networks that approach random mixing will have the fastest rate of diffusion. In the homophily simulation, the social divisions on grade actually facilitate diffusion (as within-grade

adoption is more likely), creating an incubator for the product to spread widely (Centola 2011).

The dotted lines in Figure 12 represent the sampled-based diffusion curves under our two scenarios. Looking at the homophily results, the sampled-based lines closely approximate the true curves. For example, the bias in the homophily analysis is the same as with simple diffusion for the two low-transitivity networks. The bias is also low in the two high-transitivity networks, although slightly higher than in the simple diffusion case. For example, with the high-density, high-transitivity network, the median bias (over the 30 time periods) is .017 in the simple diffusion case but .034 in the homophily-based analysis. Similarly, in the low-density, high-transitivity network (the black line), we see a median bias of .01 in the simple diffusion case but .018 in the homophily analysis.

More generally, the sample-based analysis offers the same conclusions as the true ABM based on the known networks. For example, just using the sample data, it is clear that the low-density, low-transitivity network has faster diffusion than the low-density, high-transitivity network under simple diffusion but not homophily-based diffusion. Thus, the effect of network structure on diffusion depends crucially on the assumed behavior of adoption, and this is captured just using the sampled data.

Measurement Error

The second additional analysis looks at the problem of measurement error. Thus far, the analysis has assumed that the input ego network data are collected without error, where respondents accurately report on their alters as well as the connections between those alters. Past work has raised doubts about this assumption, discussing the possible sources of bias in ego network data (Brewer et al. 2005; Feld and Carter 2002). There may, for example, be bias in the alter–alter ties (Almquist 2012). Respondents report secondhand on relationships between alters. Respondents may, however, not always know if their friends, for example, are friends themselves. An uncertain respondent may simply guess if a tie exists or, alternatively, may include/exclude ties to make their network cognitively consistent (e.g., adding ties to minimize intransitive relations; Krackhardt and Kilduff 1999).

The question is how well a network sampling approach fares in the face of measurement error. Here, we rerun the main analysis but induce error into the "reported" ego network data. We specifically focus on the alter–alter ties, as they may be particularly prone to misreporting.

The error generation process takes a simple form. We assume that respondents report on all of their alter–alter ties, but that some of those reports are actually guesses, where the respondent is unsure if a tie exists or not. To simulate guessing, we take a draw from a binomial distribution with probability set to .5 (i.e., flipping a fair coin), setting the alter–alter tie to 0 or 1 depending on the simulated draw (see Smith and Faris 2015, for a similar procedure). We run this measurement error analysis under three levels of error: .05 (low), .15 (medium), and .25 (high). Thus, under high error, 25 percent of the alter–alter ties (randomly selected) are guesses, with no necessary relationship to the actual network. This error-filled ego network data serve as input into the simulation. The rest of the analysis is the same as before, with simple diffusion and our four synthetic networks. We set the adoption probability to .3 to simplify the discussion. The results are presented in Figure 13 as a set of relative bias plots (analogous to Online Appendix Figure A1).

The results suggest that measurement error does, in fact, negatively impact the sampled-based estimates of diffusion. Looking at Figure 13, the relative bias increases as the level of measurement error increases. The bias is, however, generally modest, even when 25 percent of the alter–alter ties are guesses. For example, the median bias (over the 30 days) in the low-density, low-transitivity network is .005 under no measurement error, .025 under .05 error, .039 under .15 error, and .084 under .25 error. Similarly, for the high-density, high-transitivity network, the median bias goes from .008 under no measurement error to .031 under .25 measurement error. Specific estimates can, however, have much higher bias under conditions of high error. For example, with the low-density, low-transitivity network, the bias for period 10 is .015 with no measurement error but .171 with .25 measurement error. Practically, the results suggest that a sampling approach can be used to inform ABMs under conditions of imperfect respondent reporting; one must, however, be wary of possible bias if very high levels of error are suspected.

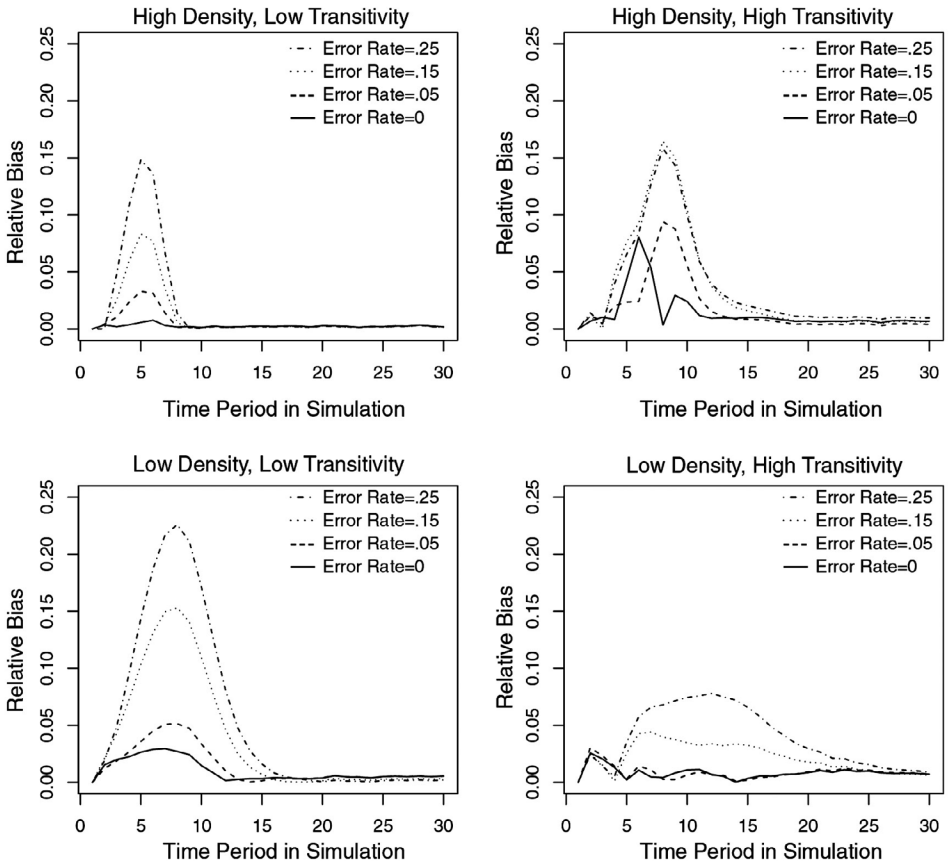

Figure 13. Effect of measurement error in alter–alter ties on diffusion estimates.

Conclusion

Agent-based modeling holds great promise as a theoretical and analytical tool for the social sciences. Simulation models make it easier to specify theories that are based on systems of interdependent, interconnected actors (Macy and Willer 2002). Agent-based modeling is, nonetheless, open to certain criticisms, particularly in disciplines with a heavy empirical focus (Bruch and Atwell 2015). For those of a purely empirical bent, a simulation can be dismissed as "mere" theory or simply a demonstration of an idea, but not a result. An ABM not rooted in real-world data and conditions is particularly vulnerable to such critiques. Given these concerns, there has been a recent call to more fully incorporate empirical data into ABMs, making the virtual world that actors inhabit mirror the actual conditions of the

social environment (e.g., Bruch and Atwell 2015; Hedström and Manzo 2015). This article falls into this tradition, asking about the payoff of incorporating ego network data into ABMs of diffusion. Is it worthwhile to collect a small ego network sample, use the data to infer a complete network, and then use the inferred network as a basis for ABMs of cultural transmission? We test this idea by creating an ABM of diffusion, where actors are exposed to new cultural items via their friends. We run different diffusion processes through the true networks and the inferred networks (based on the sampled data) and compare the results.

The results are, on the whole, encouraging. Across all test networks, the diffusion curves based on the sampled data are very similar to the curves based on the true, complete network. The bias in the estimates (for the proportion adopting at a given time point) is generally under 1 percent and almost always under 3 percent. The variance of the estimates is also generally modest, meaning that sample to sample, there are similar results. Importantly, the sample estimates yield the same substantive conclusions as the true, known networks. For example, the low-density, low-transitivity network has faster diffusion than the low-density, high-transitivity network under simple diffusion but not homophily-based diffusion. The sample estimates alone are sufficient to make such conclusions and, more generally, to capture the effect of network structure and adoption behavior on diffusion.

It is important to recognize that not every estimate has ideal properties, despite the generally encouraging results. For example, the estimates can have high standard errors, particularly for the low-density, high-transitivity network. A researcher concerned about the variance of the diffusion estimates could, most directly, collect more data, thus reducing the uncertainty in the network structure. Alternatively, they could tweak the ABM to reduce the uncertainty in the simulation; for example, increasing the number of initial seeds or varying the adoption probability. Additionally, it is important to note that the specific takeoff point (or timing) of the diffusion process can be difficult to capture when there is measurement error in the input data; the final level of adoption is more consistently measured across all sampling/measurement conditions.

Overall, the results suggest that ego network sampling can offer a practical means of incorporating empirical data into an ABM (Rolfe

2014). The data are easy to collect and widely available (as the data are based on independently sampled cases) but still yield excellent approximations of the true diffusion curves. A researcher collecting sampled ego network data and the full census (all actors and all ties between actors) would arrive at very similar results; here, concerning the rates of adoption of a new cultural product. Thus, a researcher would not have to collect full network data to employ a realistic network in their ABM.

There are a number of ways that ego network data could be incorporated into an ABM, depending on the substantive and theoretical setup used in the analysis. In each case, the goal is to limit the number of variables in the simulation that may feel arbitrary from a critical point of view.

First, and most directly, ego network data could be used in much the same way as in this article: A researcher could collect an ego network sample, infer the full network, and use that network as the basis for a diffusion simulation. Here, the researcher avoids having to come up with (and justify) the network structure over which the diffusion model is run (i.e., small-world network, power law, etc.; Flache and Macy 2011; Hamil and Gilbert 2010; Rahmandad and Sterman 2008). Such a strategy is most appropriate for projects interested in stratification/diffusion outcomes where the network structure can assumed to be (more or less) fixed, while the behavioral model is allowed to vary. The sampled data would be collected on the particular population(s) of interest. If it is not feasible to collect sampled data, it may still be possible to use data from a similar population, given the wide availability of ego network data.

Second, ego network data could be used to explore the joint dynamics of culture and structure (Adams and Schaefer 2016; Centola et al. 2007; Wang et al. 2017). Simulations that allow both network ties and cultural consumption to update over time (i.e., network ties may be added or broken through time while a product may be adopted or dropped) could use a network inferred from ego network data as a starting point for the model. This would condition the ABM on a realistic network, thus holding fixed a very large variable in the model (i.e., the starting point in the simulation). This will make it easier to demonstrate how endogenous network processes affect the outcome of interest.[12]

Third, ego network data could be used in purely generative models, where the network structure emerges within the simulation, based on the interactions of the seeded actors (e.g., Baldassarri and Bearman 2007; Centola 2015; Gondal 2015). With purely generative models, there is no a priori network substrate; rather, a set of local rules determine the nature of the interactions and these interactions determine the features of the macronetwork structure. An ego network sampling approach could be used in two ways to inform generative ABMs. First, a researcher could use the ego network data to inform the simulation itself. Here, the researcher would use the ego network data to inform the rules (or range of rules) dictating interactions in the simulation (i.e., taboos on certain interactions; tendency for transitive closure, etc.). Second, the inferred network structure could be used solely to judge the output of the simulation, showing which generated networks, or region of networks, are consistent with the empirical data. Thus, the inferred network is not used in the simulation itself but is rather used after the fact as a check (e.g., Schreiber and Carley 2013). This would show what set of microrules could have given way to the observed network structure and/or diffusion curves.

The approach, while promising, rests on a number of assumptions. For example, we assume that measurement error in the ego network data is not severe enough to badly bias the estimates of diffusion. Our own results suggest that measurement error in the alter–alter ties becomes problematic only at very high levels of misreporting. Future work could consider other kinds of measurement error as well as tactics to minimize bias (e.g., Marin and Hampton 2007). For example, respondents given an ego network survey may list fewer alters than they actually have (Marin 2004). Such "forgetting" will distort the number of alters listed and thus the inferred degree distribution. Similarly, many studies truncate the number of alters one can list, but this too can lead to a distortion in the degree distribution (Smith 2015).

We also make a number of assumptions about the network employed in the ABM. We assume that the network is static and undirected. Future work could relax these assumptions, testing a sampling approach in a wider range of circumstances. Additionally, we only consider a single sampling scenario, where a researcher samples a large portion of a small network (25 percent of a size 700 network). We examine this last assumption more carefully in the Appendix. We replicate the entire analysis using a much larger network, the Colorado

Springs high-risk network (size = 5,492), and a much smaller sampling rate, 5 percent. The results are, again, encouraging: The results mimic the main findings, with the sampled-based estimates closely approximating the true diffusion curves. See Online Appendix B for the full discussion.

Finally, we assume that the researcher knows the size of the true network. This may not be true in every research setting, however. A researcher that does not know the size of the actual network has three basic options. First, they could generate networks that correspond to a "typical" network of that type; for example, an ABM based on a school would use networks of a moderate size, less than 2,500. Second, a researcher could repeat the analysis under different assumptions about the size of the network, showing how the results do (or do not) change as network size increases (see Della- Posta et al. 2015). Third, the researcher could try and estimate the size of the actual network, for example, using network scale-up methods (Maltiel et al. 2015).[13]

This article has focused on the adoption of new products, or cultural diffusion, but a network sampling approach is useful for any ABM where networks features are important, including models of neighborhood segregation, labor market outcomes, and status inequalities (Bruch 2014; Fountain and Stovel 2014; Manzo and Baldassari 2015). Moving forward, the hope is that more researchers will see the validity of combining network sampling with ABMs and, more generally, will continue to combine empirical data with the controlled, experimental feel of a simulation.

Note — This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. No direct support was received from grant P01-HD31921 for this analysis. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516, USA; http://addhealth@unc.edu

Acknowledgments — The authors would like to thank Robin Gauthier and Jennifer Clarke for their helpful comments on earlier versions of this article. The author would also like to thank the Haas Faculty Award Program at the University Nebraska–Lincoln for providing financial support during the writing of this article. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for their assistance in the original design.

Declaration of Conflicting Interests — The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding — The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Haas Faculty Award Program at the University Nebraska–Lincoln provided financial support during the writing of this article.

Supplemental Material — Supplementary material for this article follows the References.

Notes

1. Although note that multiple individual-level models could be consistent with the same macrolevel results.
2. This is the case as one must be able to trace out the direct and indirect paths between actors.
3. Even if one cannot collect new data, there may be data from a similar population that can be used instead.
4. See also Moody (2009) who considers the timing of relationships in the context of diffusion in small-world networks.
5. The simulation varies the strength of homophily to see how stronger/weaker in-group bias affects inequality in adoption rates.
6. The diffusion curves capture the proportion who adopt across time in a simulation using the inferred network.
7. Ego network data could, potentially, be used to inform an stochastic actor-based model. The researcher would first infer the full network from the cross-sectional sampled data; they would then take that network and estimate an initial set of parameters (for estimating the model with cross-sectional data, which is atypical, see Snijders and Steglich 2015). That model could then be used to simulate a dynamic network process, where both the behaviors and network are allowed to vary over time. Note that the selection/influence parameters could not be estimated from the sampled data and would need to be set purely by the researcher.
8. The method calculates a starting value by estimating a dyadic independent exponential random graph model on the ego networks.
9. Note that it is possible for a product to "die out" in a given simulation. In this case, the proportion adopting is recorded as 0 for all time periods, where no actors currently use the new product. All values of 0 are incorporated into the overall mean calculation across the 1,000 simulations. These die-out simulations serve as potential outliers, but as this is done for both the true network and the sampled-based networks, the comparison is consistent.
10. Note that the empirical Add Health network offers similar results to the high-density networks, with a mean bias (over the 30 time periods) of .01 and a standard error of .019.
11. This may be the case as individuals who share one salient characteristic are likely to share others (e.g., language, class background, and cultural tastes). Such similarities make the friend a clearer reference group, making adoption more likely.
12. Ego network data could also be used to help parameterize the dynamic simulation itself (by showing the tendencies for tie formation).
13. This would necessitate asking an additional question in the survey. Specifically, the survey would need to ask respondents how many people they know in the setting of interest. The research could then use that information to estimate the size of the population.

References

Adams, Jimi and David R. Schaefer. 2016. "How Initial Prevalence Moderates Network-based Smoking Change." *Journal of Health and Social Behavior* 57(1): 22-38. doi: 10.1177/0022146515627848.

Almquist, Zack W. 2012. "Random Errors in Egocentric Networks." *Social Networks* 34(4): 493-505. doi: 10.1016/j.socnet.2012.03.002.

Axtell, R. 2000. "Why Agents? On the Varied Motivations for Agent Computing in the Social Sciences." Working Paper No. 17, The Brookings Institution, Center on Social and Economic Dynamics, Washington, DC.

Baldassarri, Delia and Peter Bearman. 2007. "Dynamics of Political Polarization." *American Sociological Review* 72(5): 784-811.

Boero, Riccardo and Flaminio Squazzoni. 2005. "Does Empirical Embeddedness Matter? Methodological Issues on Agent-based Models for Analytical Social Science." *Journal of Artificial Societies and Social Simulation* 8(4):6.

Bonabeau, Eric. 2002. "Agent-based Modeling: Methods and Techniques for Simulating Human Systems." *Proceedings of the National Academy of Sciences* 99(Suppl 3): 7280-87.

Brewer, Devon D, Giovanni Rinaldi, Andrei Mogoutov, and Thomas W. Valente. 2005. "A Quantitative Review of Associative Patterns in the Recall of Persons." *Journal of Social Structure* 6(1).

Bruch, Elizabeth E. 2014. "How Population Structure Shapes Neighborhood Segregation." *American Journal of Sociology* 119(5): 1221-78. doi: 10.1086/675411.

Bruch, Elizabeth E. and Jon Atwell. 2015. "Agent-based Models in Empirical Social Research." *Sociological Methods & Research* 44(2):186-221. doi: 10.1177/0049124113506405.

Burt, Ronald S. 1984. "Network Items and the General Social Survey." *Social Networks* 6(4): 293-339. doi: 10.1016/0378-8733(84)90007-8.

Carley, Kathleen. 1991. "A Theory of Group Stability." *American Sociological Review* 56(3): 331-54.

Centola, Damon. 2011. "An Experimental Study of Homophily in the Adoption of Health Behavior." *Science* 334(6060): 1269-72.

Centola, Damon. 2015. "The Social Origins of Networks and Diffusion." *American Journal of Sociology* 120(5): 1295-338.

Centola, Damon, Juan Carlos Gonzalez-Avella, Victor M. Eguiluz, and Maxi San Miguel. 2007. "Homophily, Cultural Drift, and the Co-evolution of Cultural Groups." *Journal of Conflict Resolution* 51(6): 905-29.

Centola, Damon and Michael Macy. 2007. "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology* 113(3): 702-34.

De Marchi, Scott and Scott E. Page. 2014. "Agent-based Models." *Annual Review of Political Science* 17:1-20.

DellaPosta, Daniel, Yongren Shi, and Michael Macy. 2015. "Why Do Liberals Drink Lattes?" *American Journal of Sociology* 120(5): 1473-511. doi: 10.1086/681254.

DiMaggio, Paul and Filiz Garip. 2011. "How Network Externalities Can Exacerbate Intergroup Inequality." *American Journal of Sociology* 116(6): 1887-933.

Feld, Scott L. and William C. Carter. 2002. "Detecting Measurement Bias in Respondent Reports of Personal Networks." *Social Networks* 24(4): 365-83.
Fischer, Claude S. 1982. *To Dwell among Friends: Personal Networks in Town and City*. Chicago, IL: University of Chicago Press.
Flache, Andreas and Michael W. Macy. 2011. "Small Worlds and Cultural Polarization." *The Journal of Mathematical Sociology* 35(1-3): 146-76.
Fountain, Christine and Katherine Stovel. 2014. "Turbulent Careers: Social Networks, Employer Hiring Preferences, and Job Instability." Pp. 339-70 in *Analytical Sociology: Actions and Networks*, edited by G. Manzo. Chichester, England: John Wiley.
Gondal, Neha. 2015. "Inequality Preservation through Uneven Diffusion of Cultural Materials across Stratified Groups." *Social Forces* 93(3): 1109-37.
Goodreau, Steven M., James A. Kitts, and Martina. Morris. 2009. "Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks." *Demography* 46(1): 103-25.
Granovetter, Mark. 1978. "Threshold Models of Collective Behavior." *American Journal of Sociology* 83(6): 1420-43.
Hamill, Lynne and Nigel Gilbert. 2010. "Simulating Large Social Networks in Agent-based Models: A Social Circle Model." *Emergence: Complexity and Organization* 12(4): 78.
Handcock, Mark S. and Krista J. Gile. 2010. "Modeling Social Networks from Sampled Data." *Annals of the Applied Statistics* 4(1): 5-25.
Hedström, Peter and Peter Bearman, eds. 2009. *The Oxford Handbook of Analytical Sociology*. Oxford, NY: Oxford University Press.
Hedström, Peter and Gianluca Manzo. 2015. "Recent Trends in Agent-based Computational Research a Brief Introduction." *Sociological Methods & Research* 44(2): 179-85.
Hedström, Peter and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36(1): 49-67. doi: 10.1146/annurev.soc.012809.102632
Hunter, David R., Mark S. Handcock, Carter T. Butts, Steve M. Goodreau, and Martina Morris. 2008. "ergm: A Package to Fit, Simulate and Diagnose Exponential-family Models for Networks." *Journal of Statistical Software* 24(3): 1-29.
Keeling, Matt J. and Ken T. D. Eames. 2005. "Networks and Epidemic Models." *Journal of the Royal Society Interface* 2(4): 295-307.
Kitts, James A. 2006. "Social Influence and the Emergence of Norms amid Ties of Amity and Enmity." *Simulation Modelling Practice and Theory* 14(4): 407-22. doi: 10.1016/j.simpat.2005.09.006.
Kossinets, Gueorgi. 2006. "Effects of Missing Data in Social Networks." *Social Networks* 28(3): 247-68.
Krackhardt, David and Martin Kilduff. 1999. "Whether Close or Far: Social Distance Effects on Perceived Balance in Friendship Networks." *Journal of Personality and Social Psychology* 76(5): 770-82.
Krivitsky, Pavel N. and Martina Morris. 2015. "Inference for Social Network Models from Egocentrically-sampled Data, with Application to Understanding Persistent Racial Disparities in HIV Prevalence in the U.S." University of

Wollongong, National Institute for Applied Statistics Research Australia. Retrieved May 30, 2018; http://niasra.uow.edu.au/publications/UOW190187

Liu, Kayuet and Peter S. Bearman. 2015. "Focal Points, Endogenous Processes, and Exogenous Shocks in the Autism Epidemic." *Sociological Methods & Research* 44(2): 272-305. doi: 10.1177/0049124112460369.

Luke, Douglas A. and Katherine A. Stamatakis. 2012. "Systems Science Methods in Public Health: Dynamics, Networks, and Agents." *Annual Review of Public Health* 33:357-76. doi: 10.1146/annurev-publhealth-031210-101222.

Mabry, Patricia L., Deborah H. Olster, Glen D. Morgan, and David B. Abrams. 2008. "Interdisciplinarity and Systems Science to Improve Population Health: A View from the NIH Office of Behavioral and Social Sciences Research." *American Journal of Preventive Medicine* 35(2 Suppl): S211-S24. doi: 10.1016/j.amepre.2008.05.018

Macy, Michael W. and Andreas Flache. 2009. "Social Dynamics from the Bottom Up: Agent-based Models of Social Interaction." Pp. 245-68 in *Oxford Handbook of Analytical Sociology*, edited by P. H. A. P. Bearman. Oxford, NY: Oxford University Press.

Macy, Michael W. and Robert Willer. 2002. "From Factors to Actors: Computational Sociology and Agent-based Modeling." *Annual Review of Sociology* 28:143-66.

Maltiel, Rachael, Adrian E. Raftery, Tyler H. McCormick, and Aaron J. Baraff. 2015. "Estimating Population Size Using the Network Scale Up Method." *The Annals of Applied Statistics* 9(3): 1247-77. doi: 10.1214/15-AOAS827

Manzo, Gianluca. 2007. "Variables, Mechanisms, and Simulations: Can the Three Methods Be Synthesized?" *Revue Française de Sociologie* 48(5): 35-71.

Manzo, Gianluca. 2013. "Educational Choices and Social Interactions: A Formal Model and a Computational Test." *Comparative Social Research* 30:47-100.

Manzo, Gianluca and Delia Baldassarri. 2015. "Heuristics, Interactions, and Status Hierarchies: An Agent-based Model of Deference Exchange." *Sociological Methods & Research* 44(2): 329-87. doi: 10.1177/0049124114544225

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

Marin, Alexandra and Keith N. Hampton. 2007. "Simplifying the Personal Network Name Generator." *Field Methods* 19(2): 163-93. doi: 10.1177/1525822x06298588

Mark, Noah P. 1998. "Beyond Individual Differences: Social Differentiation from First Principles." *American Sociological Review* 63(3): 309-30.

Mark, Noah P. 2003. "Culture and Competition: Homophily and Distancing Explanations for Cultural Niches." *American Sociological Review* 68(3): 319-45.

Marsden, Peter V. 1987. "Core Discussion Networks of Americans." *American Sociological Review* 52(1): 122-31.

Mäs, Michael and Andreas Flache. 2013. "Differentiation without Distancing. Explaining Bi-polarization of Opinions without Negative Influence." *PLoS One* 8(11): e74516. doi: 10.1371/journal.pone.0074516.

McFarland, Daniel A., James Moody, David Diehl, Jeffrey A. Smith, and Reuben J. Thomas. 2014. "Network Ecology and Adolescent Social Structure." *American Sociological Review* 79(6): 1088-121. doi: 10.1177/0003122414554001

McPherson, Miller J. 1983. "An Ecology of Affiliation." *American Sociological Review* 48(4): 519-32.
McPherson, Miller J. 2004. "A Blau Space Primer: Prolegomenon to an Ecology of Affiliation." *Industrial and Corporate Change* 13(1): 263-80.
McPherson, Miller J., Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415-44.
Merli, Giovanna, James Moody, Joshua Mendelsohn, and Robin Gauthier. 2015. "Sexual Mixing in Shanghai: Are Heterosexual Contact Patterns Compatible with an HIV/AIDS Epidemic?" *Demography* 52(3): 919-42.
Miller, John H. and Scott E. Page. 2007. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton, NJ: Princeton University Press.
Moody, James. 2009. "Network Dynamics." Pp. 447-74 in *Oxford Handbook of Analytical Sociology*, edited by P. Hedström and P. S. Bearman. Oxford, NY: Oxford University Press.
Moody, James and Richard A. Benton. 2016. "Interdependent Effects of Cohesion and Concurrency for Epidemic Potential." *Annals of Epidemiology* 26(4): 241-48. doi: 10.1016/j.annepidem.2016.02.011.
Moody, James and Douglass R. White. 2003. "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups." *American Sociological Review* 68(1): 103-27.
Morris, Martina, Ann E. Kurth, Deven T. Hamilton, James Moody, and Steve Wakefield. 2009. "Concurrent Partnerships and HIV Prevalence Disparities by Race: Linking Science and Public Health Practice." *American Journal of Public Health* 99(6): 1023-31.
Morris, Martina and Richard Rothenberg. 2011. "HIV Transmission Network Metastudy Project: An Archive of Data from Eight Network Studies, 1988-2001." Inter-University Consortium for Political and Social Research (ICPSR) [distributor]. Ann Arbor, MI.
Rahmandad, Hazhir and John Sterman. 2008. "Heterogeneity and Network Structure in the Dynamics of Diffusion: Comparing Agent-based and Differential Equation Models." *Management Science* 54(5): 998-1014.
Railsback, Steven F. and Volker Grimm. 2011. *Agent-based and Individual-based Modeling: A Practical Introduction*. Princeton, NJ: Princeton University Press.
Richiardi, Matteo, Roberto Leombruni, Nicole J. Saam, and Michele Sonnessa. 2006. "A Common Protocol for Agent-based Social Simulation." *Journal of Artificial Societies and Social Simulation* 9(1): 15.
Robins, Garry, Philippa Pattison, and Jodie Woolcock. 2005. "Small and Other Worlds: Global Network Structures from Local Processes." *American Journal of Sociology* 110(4): 894-936. doi: 10.1086/427322.
Robins, Garry, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. 2007. "Recent Developments in Exponential Random Graph (P*) Models for Social Networks." *Social Networks* 29(2): 192-215.
Rocha, Luis E. C, Fredrik Liljeros, and Petter Holme. 2011. "Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts." *PLoS Computational Biology* 7(3): e1001109.

Rolfe, Meredith. 2014. "Social Networks and Agent-based Modelling." Pp. 233-60 in *Analytical Sociology*. John Wiley & Sons.

Salathé, Marcel and Shashank Khandelwal. 2011. "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control." *PLoS Computational Biology* 7(10): e1002199.

Schaefer, David R, Jimi Adams, and Steven A. Haas. 2013. "Social Networks and Smoking: Exploring the Effects of Peer Influence and Smoker Popularity through Simulations." *Health Education & Behavior* 40(Suppl 1): 24S-32S.

Schaefer, David R, Steven A. Haas, and Nicholas J. Bishop. 2012. "A Dynamic Model of Us Adolescents' Smoking and Friendship Networks." *American Journal of Public Health* 102(6): e12-e18.

Schreiber, Craig and Kathleen M. Carley. 2013. "Validating Agent Interactions in Construct against Empirical Communication Networks Using the Calibrated Grounding Technique." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43(1): 208-14.

Smith, Jeffrey A. 2012. "Macrostructure from Microstructure: Generating Whole Systems from Ego Networks." *Sociological Methodology* 42(1):155-205. doi: 10.1177/0081175012455628

Smith, Jeffrey A. 2015. "Global Network Inference from Ego Network Samples: Testing a Simulation Approach." *The Journal of Mathematical Sociology* 39(2): 125-62. doi: 10.1080/0022250X.2014.994621

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

Smith, Jeffrey A., Miller McPherson, and Lynn Smith-Lovin. 2014. "Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004." *American Sociological Review* 79(3): 432-56. doi: 10.1177/0003122414531776

Smith, Jeffrey A. and James Moody. 2013. "Structural Effects of Network Sampling Coverage I: Nodes Missing at Random." *Social Networks* 35(4):652-68. doi: 10.1016/j.socnet.2013.09.003

Snijders, Tom A. B. and Christian E. G. Steglich. 2015. "Representing Micro–Macro Linkages by Actor-based Dynamic Network Models." *Sociological Methods & Research* 44(2): 222-71.

Snijders, Tom A. B, Gerhard G. Van de Bunt, and Christian E. G. Steglich. 2010. "Introduction to Stochastic Actor-based Models for Network Dynamics." *Social Networks* 32(1): 44-60.

Steglich, Christian, Tom A. B. Snijders, and Michael Pearson. 2010. "Dynamic Networks and Behavior: Separating Selection from Influence." *Sociological Methodology* 40(1): 329-93.

Sterman, John D. 2006. "Learning from Evidence in a Complex World." *American Journal of Public Health* 96(3): 505-14.

Thompson, Steven K. and Ove Frank. 2000. "Model-based Estimation with Linktracing Sampling Designs." *Survey Methodology* 26:87-98.

Verdery, Ashton M. 2015. "Links between Demographic and Kinship Transitions." *Population and Development Review* 41(3): 465-84. doi: 10.1111/j.1728-4457.2015.00068.x

Wang, Cheng, John R. Hipp, Carter T. Butts, Rupa Jose, and Cynthia M. Lakon. 2017. "Peer Influence, Peer Selection and Adolescent Alcohol Use: A Simulation Study Using a Dynamic Network Model of Friendship Ties and Alcohol Use." *Prevention Science* 18(4): 1-12.

Wasserman, Stanley and Philippa Pattison. 1996. "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and P*." *Psychometrika* 61(3): 401-25.

Watts, Duncan J. 1999. "Networks, Dynamics, and the Small-world Phenomenon." *American Journal of Sociology* 105(2): 493-527.

Windrum, Paul, Giorgio Fagiolo, and Alessio Moneta. 2007. "Empirical Validation of Agent-based Models: Alternatives and Prospects." *Journal of Artificial Societies and Social Simulation* 10(2):8.

Author Biographies

Jeffrey A. Smith is an Assistant Professor of Sociology at the University of Nebraska-Lincoln. His research interests fall at the intersection of network analysis, traditional statistical methods and social stratification. He has done methodological work on network sampling and missing data, as well as more substantive work on network processes like homophily and status.

Jessica Burow is a data scientist at The Hartford in Hartford, Connecticut. She received her master's degree in Statistics at the University of Nebraska, Lincoln. Her areas of interest include predictive analytics, statistical network analysis and diffusion.

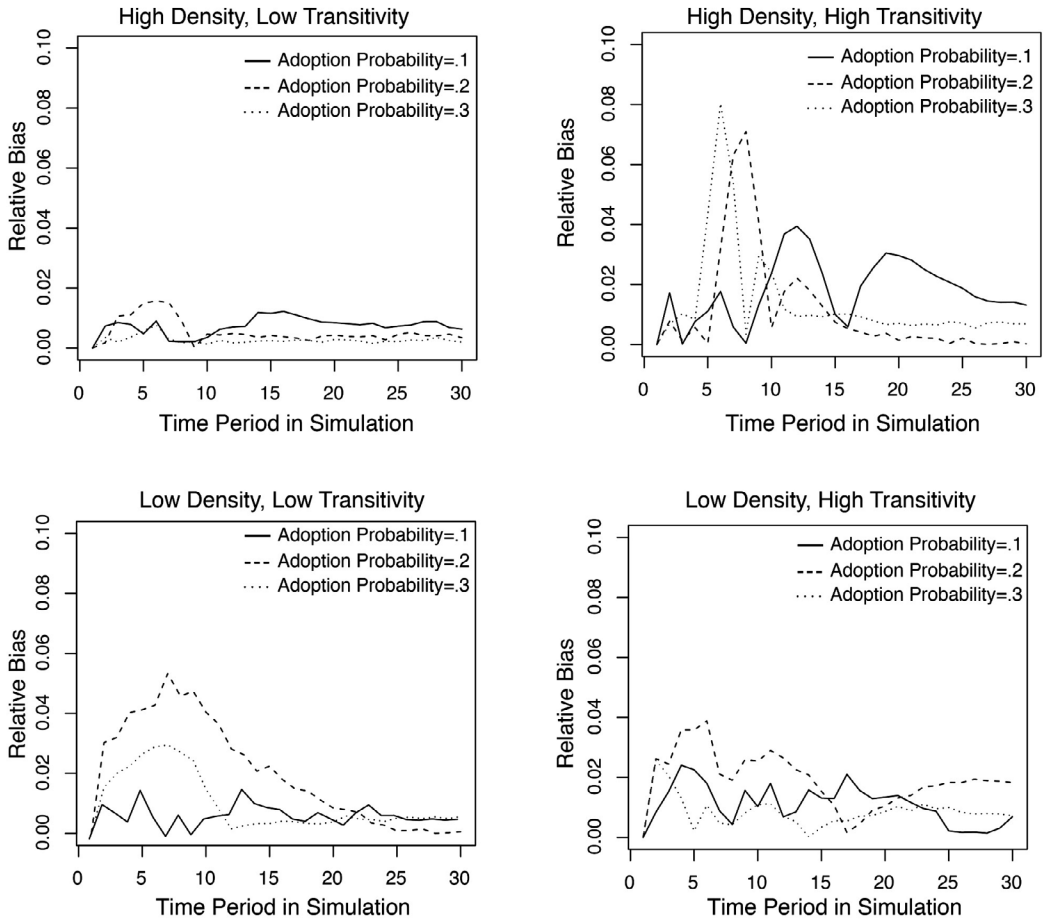

Figure A1. Relative Bias by Network and Adoption Rate

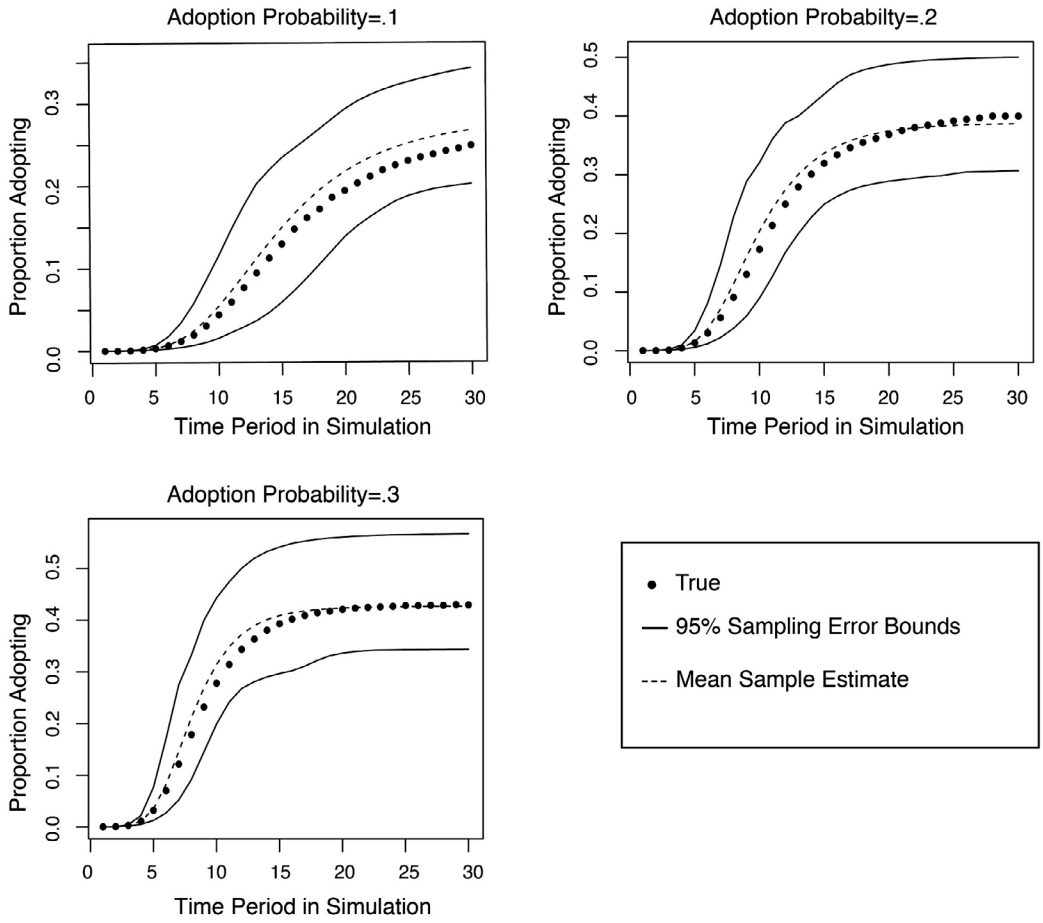

Figure B1. Colorado Springs Diffusion Curve Results

Appendix B

This appendix offers a supplementary test of a sampling approach. In the main analysis, we focus on a single sampling scenario: where a researcher samples a large portion of a small network. We take a 25% sample from a network of size 700. Here, we expand the analysis, replicating the results with a much larger network and a much lower sampling rate. In this way, we can assess the validity of a sampling approach under more difficult conditions.

The network of interest comes from Project 90, the Colorado Springs study of high-risk individuals (see Morris and Rothenberg 2011 for the data source). The population of interest includes at-risk individuals for HIV transmission, including drug injectors and sex workers. Researchers attempted to saturate the population in this city and we treat the network as a full census. The data include social connections based on sex, needle sharing and social ties. The true network includes 5492 nodes and 21644 edges. We base our analysis on a 5% sample of the network.

The test presented here is a difficult one, as we use a larger network and a lower sampling rate than in the main analysis. Additionally, the properties of the network make this a particularly difficult test of a sampling approach. First, the network has high transitivity (.37) and high average degree (7.88), and we have already seen that the bias is higher in such networks. Second, the network has a skewed degree distribution. Inference is harder when the degree distribution is skewed: a few actors have disproportionately high degree, yet they are no more likely to be sampled than any other node. High degree nodes thus have a large impact on network structure, but are often missed in a random sample (see Smith 2015). And third, the network is disconnected, with 20% of the nodes outside the main component (a component is a set of nodes connected by at least one path; the main component is the largest set of nodes connected by at

least one path). The diffusion simulation will be highly variable under such conditions. Global diffusion is possible (albeit not necessary) when the initial seed is in the main component; in contrast, global diffusion is impossible if the initial seed is not in the main component (as they are disconnected from the rest of the network and cannot pass the product beyond their own borders). The results are thus highly dependent on the initial seed, making inference more difficult.

The analysis is the same as before. The agent-based model of diffusion follows a simple contagion process, with three adoption probabilities: .1, .2, and .3. We again take 100 independent samples. We assume that the data collected have the same pattern as in the main text. The only difference is in the demographic characteristics assumed to be collected. Here, the characteristics of interest include: race, gender, employment status, and illicit activity (drug dealer, sex worker, pimp or none).

We present the results below in Figure B1. The results follow the same form as in Figures 8-11. There are three subplots, one for each adoption probability. Each subplot shows the true proportion adopting, the mean estimate and the error bounds. The estimate are, in general, quite good, despite the difficulty of the test. Looking at the high adoption results, the median bias over the 30 time periods is under 2%. The results are on par with the findings in the main text. For example, for period 20, the mean estimate is .424, while the true value is .421 (a relative bias under 1%). The estimates are, as expected, more uncertain than before. The median standard error (over the 30 time periods) is .06, higher than with any network used in the main analysis. For time period 20, 95% of the estimates fall between .34 and .56, a wide range of values. We see similar results with the lower adoption probabilities, although the bias is higher here. The median relative bias (over the 30 time periods) is about .037 for the medium adoption analysis.

For example, for period 20, the true proportion adopting is .368, the mean estimate is .374, and 95% of the estimates fall between .289 and .488.

Overall, the results suggest that it is possible to produce good approximations of the true diffusion curves using a small sample on a large network. The caveat is that the estimates can be quite uncertain, with high variability sample-to-sample. A researcher concerned with the variability of the estimates would have to sample more than 5% of the network.