Published in final edited form as:
Soc Networks. 2015 May 1; 41: 56–71. doi:10.1016/j.socnet.2014.12.004.

Research Note: The consequences of different methods for handling missing network data in Stochastic Actor Based Models

John R. Hipp[*], **Cheng Wang**, **Carter T. Butts**, **Rupa Jose**, and **Cynthia M. Lakon**

[*]Department of Criminology, Law and Society and Department of Sociology, University of California, Irvine

Abstract

Although stochastic actor based models (e.g., as implemented in the SIENA software program) are growing in popularity as a technique for estimating longitudinal network data, a relatively understudied issue is the consequence of missing network data for longitudinal analysis. We explore this issue in our research note by utilizing data from four schools in an existing dataset (the AddHealth dataset) over three time points, assessing the substantive consequences of using four different strategies for addressing missing network data. The results indicate that whereas some measures in such models are estimated relatively robustly regardless of the strategy chosen for addressing missing network data, some of the substantive conclusions will differ based on the missing data strategy chosen. These results have important implications for this burgeoning applied research area, implying that researchers should more carefully consider how they address missing data when estimating such models.

Keywords

Smoking; stochastic actor based model; SIENA; peer selection; peer influence substance use behavior

Introduction

For at least the last 30 years, researchers have explored the question of whether networks of relations among individuals have important consequences for behaviors such as substance use, general delinquency, and even obesity (Baerveldt, Volker, and Van Rossem 2008; Christakis and Fowler 2007; Shakya, Christakis, and Fowler 2012). The importance of social networks for behavioral phenomena is well-known for populations throughout the life course, from adolescents in schools (Mouw and Entwisle 2006) to elderly adults in

Address correspondence to John R. Hipp, Department of Criminology, Law and Society, University of California, Irvine, 3311 Social Ecology II, Irvine, CA 92697; john.hipp@UCI.edu.

communities (Rook, August, and Sorkin 2011). A well-known challenge for such integrated network/behavior studies is that the processes driving network tie formation and behavior change (e.g., delinquency) are tightly intertwined. As a consequence, a growing research paradigm utilizes a novel methodological solution to this problem: stochastic actor-based models (popularly implemented via the SIENA (Simulation Investigation for Empirical Network Analysis) toolkit developed by Snijders and colleagues (Snijders, van de Bunt, and Steglich 2010). The growing use of tools like SIENA has allowed researchers to study both influence effects—in which individuals adopt the behavior of their alters—and selection effects—in which individuals form ties with others who engage in similar behaviors—for a variety of behaviors.

As an analytic approach, stochastic actor-based models (SAB) require longitudinal network data, which frequently results in a substantial amount of missing information present (particularly if there are several waves of data). Missing data occurs due to factors such as respondent or item-level non-response, accidental or intentional censoring of responses, or administrative error. Although missing data is a well-known challenge for traditional cross-sectional and longitudinal studies (where missingness can be even more of a problem since a higher proportion of persons are typically missing for one or more waves when compared to a cross-sectional study), the problem is even greater for network studies—and longitudinal network studies in particular. Given a single cross-section from an undirected network from which some fraction of nodes f are observed, the fraction of unobserved edge variables scales as $1\text{-}f^2$; thus, if a researcher lacks information on half of the persons in the network, approximately 75% of the possible edge variables in the network will be missing (Handcock 2002). This problem is exacerbated by the fact that many network properties (e.g., connectivity, betweenness) can be sensitive to the addition or deletion of small numbers of edges (Borgatti and Everett 2006), making network analysis in the presence of missing data particularly treacherous.

Thus, there is reason to believe that missing data may be a serious problem for researchers utilizing longitudinal network data to estimate SAB models. As might be expected given the scope of the problem, various means of dealing with missingness in SAB modeling have been suggested (e.g., Huisman and Steglich 2008). Nonetheless, this issue has received limited attention in the literature and we do not currently know how much impact the choice of missing data strategy used has on the estimated results. The goal of this research note is to describe four possible approaches that are employed in the literature with varying degrees of plausibility vis a vis how we might expect the network to be generated, and to compare how each of these approaches work on real-world data when estimating SAB models. In the process, we hope to shed some light on the question of how much difference the choice of missing data methodology makes for substantive conclusions drawn from SAB models in a practical setting.

The SAB modeling approach

We first briefly describe the SAB modeling approach, as implemented in SIENA. An important challenge for modeling of influence and selection processes is that longitudinal network data is often collected at discrete time points on a time scale that is comparable to

the rate of structural and behavioral change (e.g., multiple months for adolescent friendship networks). Because social ties are created and dissolved in continuous time, ties can form and dissolve between data collection points; likewise, behavior change occurs in continuous time. While these events could be modeled directly (e.g., via a relational event process Butts 2008) if their timing were exactly known, or approximated by discrete-time dynamics if their dynamics were slow relative to the time-scale of measurement (see e.g., Lerner, Indlekofer, Nick, and Brandes 2013), the combination of "fast" dynamics and discrete-time measurement requires that they be modeled as a latent process in continuous time. SIENA accomplishes this objective by simultaneously modeling selection and influence processes with an agent-based simulation model that imputes latent trajectories of structural and behavioral co-evolution consistent with observed data and a hypothesized set of behavioral mechanisms (Baerveldt, Volker, and Van Rossem 2008; Snijders 2001; Snijders, van de Bunt, and Steglich 2010). An important strength of this approach is that it directly links hypothesized behavioral processes with observed social dynamics in a statistically principled way, making it an increasingly popular approach for joint modeling of structural and behavioral evolution within groups (Snijders, van de Bunt, and Steglich 2010).

Strategies to handle missing network data with SAB models

In this section we describe four approaches for handling missing network data when estimating a longitudinal SAB model. These strategies have varying levels of a priori plausibility given the structure of the underlying processes being modeled, but nonetheless have all been used in empirical studies. Table 1 briefly describes various possible missing network data patterns, and how each is addressed by the four strategies.

The first strategy discards the greatest amount of information: this approach *only includes in the analysis individuals who are present at all time points*. Thus, in this strategy anyone who does not appear in the sample in any of the waves is assumed to not be in the network. Of course, we know that these persons are indeed part of the network, and thus this strategy makes a particularly large assumption that discarding said actors from the network will not affect the model estimates. This approach has been used in a number of studies, incorporating various types of individual behaviors (e.g., Agneessens and Wittek 2008; Baerveldt, Volker, and Van Rossem 2008; Burk, Kerr, and Stattin 2008; de Cuyper, Weerman, and Ruiter 2009; Flashman 2012; Light and Dishion 2007; Pearson, Steglich, and Snijders 2006; Schaefer, Haas, and Bishop 2012; Simpkins, Schaefer, Price, and Vest 2013).

The second strategy *discards all persons who were missing in the first wave of the study*, but does not discard persons missing from later waves as long as they were eligible at the first wave. For the later waves, this strategy (as well as strategies 3 and 4) employs a built in feature of SIENA in which network ties at later time points are imputed based on previous observed values with missingness treated as ignorable (Huisman and Snijders 2003; Huisman and Steglich 2008). Although this strategy does not assume that a person missing at any of the later time points is not in the network (as done in strategy 1), it does assume that persons that did not respond to the survey at the first wave are not in the network. This is a somewhat strong assumption that is clearly inaccurate (although it may be a reasonable approximation where e.g. those initially missing are generally peripheral to the group under

study). As with the first strategy, various studies have adopted this strategy when estimating SAB models (e.g., de Klepper, Sleebos, van de Bunt, and Agneessens 2010; Mathys, Burk, and Cillessen 2013; Mundt, Mercken, and Zakletskaia 2012; Shoham, Tong, Lamberson, Auchincloss, Zhang, Dugas, Kaufman, Cooper, and Luke 2012).

The third strategy includes all persons who appeared in the study in any of the waves of the survey. However, *for persons who are missing at the first time point this approach simply imputes null ties to them*. This assumes that persons who did not respond to the survey at the first time point have no social ties at time 1. This is an assumption that seems quite implausible. Furthermore, making this assumption causes missingness at the first time point to act as a de facto proxy for isolation, implying that the persons who did not respond to the survey at the first time point have no social ties during the period. To the extent that this assumption is inaccurate, it will appear that such persons are "gaining" quite a few ties between the first time point and the second time point, when in fact this occurs because of the incorrect assumption that they had no social ties at the first time point. Numerous studies have employed this strategy (e.g., Berger and Dijkstra 2013; Burk, Steglich, and Snijders 2007; Cheadle and Goosby 2012; Dijkstra, Lindenberg, Veenstra, Steglich, Isaacs, Card, and Hodges 2010; Light, Greenan, Rusby, Nies, and Snijders 2013; Logis, Rodkin, Gest, and Ahn 2013; Ojanen, Sijtsema, and Rambaran 2013; Osgood, Ragan, Wallace, Gest, Fienberg, and Moody 2013; Rambaran, Dijkstra, and Stark 2013; Van Workum, Scholte, Cillessen, Lodder, and Giletta 2013; Veenstra and Steglich 2011).

The fourth strategy also includes all persons who appeared in any wave of the study. However, instead of assuming that persons who did not respond to the survey at the first wave have no social ties, *it builds an imputation model based on the characteristics of the sample at the first time point to impute social ties to these persons*. This builds on the insights of Handcock (2002) that the latent missing data framework developed by Rubin (1976) in a non- network context can also be applied to full networks. The strategy begins by fitting a cross- sectional exponential random graph model (ERGM) on the data at wave 1, using the method of Gile and Handcock (2006) to estimate model parameters using the observed-data likelihood. Simulated draws from this fitted model, conditional on the observed data, are then used to impute the states of the edge variables in time 1 that were not observed; the SAB model is then fit to the resulting data set, with the standard SAB imputation scheme being used to handle missing values in subsequent time points (see Appendix 1 for a more complete description).. The uncertainty brought to the resulting estimates by the time 1 imputation process can be accounted for by imputing the network multiple times, and then estimating the SAB model on each of these imputed networks.[1] The general principle of multiple imputation is well-known (Schafer 1997), and the properties have been well-studied in Monte Carlo simulations.

[1]We assessed the variability in model results when estimating the model on 50 different networks for a single school. The results across estimates only had a modest amount of variability. Given that it can take several days to estimate a single model for the larger schools, it was not practical to estimate all models on the multiple imputed networks. Given the similarity of the results over network imputations, the pattern of our singly imputed results is quite similar. Thus, for practical reasons, and because the multiple imputation results are outside the scope of the current study, we focus here on estimates from a single imputation.

In general, imputation strategies (as in methods (3) and (4) above) will only do as well as the imputation model on which they are built. If the imputation model accurately represents the true social process, the imputation-based estimate will often be a close approximation to what would have been obtained had complete data been available. Of course, the researcher will not know the true social process in practice. It is therefore incumbent on the researcher to specify the most plausible imputation model possible. Although the specified network imputation model one proposes is unlikely to be exactly correct, a model that captures basic network properties will typically bias the resulting SAB estimates less than one that produces highly unrepresentative networks. For this reason, strategy (4) has considerable opportunity to outperform strategy (3), in which values of 0 are imputed to the ties of all non-respondents.

In the current study, we utilize these four missing data strategies to estimate SIENA models on six separate schools with three waves of data. We chose two large schools (077 and 058, N = 2104 and N = 1,024, respectively), two mid-sized schools (007 and 008, N = 181 and N = 133, respectively), and two small schools (002 and 126, N = 78 and N = 62, respectively). We estimated the model for each school using the four missing data strategies, and then compared the results. While many such comparisons could be made, our focus here is on differences that would impact the substantive interpretation of the results. We thus compare over missing data strategies: 1) the difference in the size of the coefficients; and 2) the difference in inferential conclusions.

Data and Methods

The data used for analysis comes from three waves of the AddHealth survey (Harris 2009). We used the two large schools, two mid-sized schools, and two small schools from the special oversample data with network information, termed the “saturation sample”. Wave 1 (referred to as the “in-school” sample) was collected from September 1994 - April 1995, wave 2 (the in-home 1 sample) was collected from April-December 1995, and wave 3 (the in-home 2 sample) was collected from April-August 1996. We use AddHealth because that it is one of the few widely available longitudinal network studies, it contains schools of varying sizes allowing studying the effects of missing data in these various contexts, and it is commonly studied in the dynamic network literature (Cheadle and Schwadel 2012; Cheadle and Goosby 2012; de la Haye, Green, Kennedy, Pollard, and Tucker 2013; Flashman 2012; Shoham et al. 2012; Simpkins, Schaefer, Price, and Vest 2013). Nonetheless, the fact that the network data were not collected at a single time point is not ideal and raises additional challenges for dynamic network modeling, which are outside the scope of the present study. We compared the results across these different sized schools. For the two smallest schools, the full model was not able to converge to a proper solution. We therefore estimated models for those schools focusing only on the network dynamics.

Dependent Variables

We examine two main outcome variables: tie choice by adolescents, and smoking behavior. To measure *smoking*, at waves 2 and 3 the respondents were asked “During the past 30 days, on how many days did you smoke cigarettes?” The results are categorized into: 1) “no days”; 2) 1 to 3 days; 3) more than 3 days but less than 22 days; 4) 22 or more days. At

wave 1 a different question was asked: “How often in the last 12 months did R smoke cigarettes?” We re- categorized the results so that they matched the category framing at waves 2 and 3, which include: 1) “never”; 2) “once or twice” to “2 or 3 days a month”; 3) “once or twice a week” to “3 or 5 days a week”; 4) “nearly every day.” *Tie choice* was measured by adolescents selecting up to 5 males and 5 females as ties from a roster of students in the school.

Independent Variables

Truncated friendship identifier—Due to an administrative error within the AddHealth study, certain students were only allowed to nominate one female friend and one male friend during wave 2 and wave 3 of data collection. To correct for this effect within our network model, we account for this limited nomination effect with a variable capturing the change in possible nominations during that period: -1 = going from full to limited nominations, 0 = no change, and +1 = going from limited to full nominations.

We included several measures in the tie choice equation. Rate functions for each of the two periods between waves capture the average number of changes in friendship ties (*friendship rate*). The measure *out-degree* captures the overall propensity to be tied to another adolescent, and *reciprocity* captures the extent to which ties are reciprocated. *Transitive triplets*, the tendency for a focal actor to nominate a friends' friend as their friend, and *three cycles*, the tendency for a friends' friend to nominate the focal actor as a friend, are two effects that assess the presence of triadic closure within a friendship network. *In-degree popularity* is the proclivity to choose popular alters (peers) as friends. *In-in degree assortativity* (square root) is the proclivity to choose friends with similar levels of popularity (Ripley, Snijders, Boda, Vörös, Preciado, and at 2014).

In addition to structural network effects, measures are constructed for the *ego effect*, an *alter effect*, and a *similarity effect* (Burk, Steglich, and Snijders 2007; Steglich, Snijders, and Pearson 2010). For example, we included a measure of the smoking behavior of the ego: this assesses whether higher-level smokers form more ties. The measure capturing the smoking behavior of the alter assesses whether higher-level smokers receive more ties (i.e., are more popular). And the measure of similarity in smoking behavior between ego and alter assesses whether adolescents are more likely to form social ties with others who are similar in smoking behavior (a selection effect). Given that ties are more likely to form among adolescents in the same grade, we included a grade similarity effect. In the largest school, we included a measure of race similarity to capture race homophily (the smaller schools were too racially homogeneous to include such a measure).

In the equation predicting change in smoking behavior over time, we included several measures. Rate functions estimate the average number of changes in smoking behavior (*behavior rate*) between waves. The *linear shape effect* and the *quadratic shape effect*, capture the general tendency to change smoking behavior over time. A measure of in-degree assesses whether adolescents with more ties in the network (more popular) smoke more over time. We assess influence effects with the sum of the Negative Absolute Difference of smoking behavior between ego and all his or her alters averaged by ego's out-degree. This

assesses whether the adolescent's smoking behavior mimics his or her peers' over time (an influence effect).

We account for gender with a measure of *female*. Only the largest school had racial heterogeneity, so we coded an indicator variable for *Black*. A measure of *depression* was constructed as a factor score of 19 items modified from the Center for Epidemiologic Studies Depression Scale (CES-D) all assessing mood in the past week (Radloff 1977). Sample items included “felt depressed,” “felt lonely,” “felt happy,” and “felt life was not worth living.” Responses were coded on a 3-point scale (i.e., 0=“never or rarely”, 1=“sometimes,” 2=“a lot of the time”, and 3=“most of the time or all of the time”) with higher values indicating higher levels of depressed mood. The *home smoking environment* was measured as a sum of parent smoking and household smoking items. To determine parent smoking, we used parents' self-reported smoking behavior (i.e., replies from “Do you smoke?” in parent questionnaire coded 0=no and 1= yes) and adolescent reports of residential parent's smoking (i.e., the average of both parents for each in-home assessment with 0= no parent smokes, 1= at least one parent smokes, 2= both parents smoke). To determine household smoking, we used interviewer's remarks on whether there was “evidence of smoking in the household” (0 = no, 1=yes) and adolescent reports of whether cigarettes were “easily available” at the home (0=no, 1=yes) during each in-home assessment. *Parental support* was measured as the average of two factor scores assessing maternal and paternal emotional support. For both the mother and father, adolescents rated their parents on whether they were “warm and loving” (1= “strongly disagree” to 5 = “strongly agree”), communicated well with each other (1= “strongly disagree” to 5=“strongly agree”), had a “good relationship” (1=“strongly disagree” to 5= “strongly agree”), felt cared for (1= “not at all” to 5= “very much”), felt close (1=not at all” to 5= “very much”), and discussed personal problems together (0= “no”, 1= “yes).

Missing covariate information was imputed in STATA before data was transferred to SIENA. In the Appendix 2 tables, the summary statistics for the six schools are displayed. As can be seen, there is some variability across these schools, as well as across the missing data strategies. For example, in Table A2 the percentage that does not smoke at all at time 1 — according to strategy 3—ranges from 42% in school 058 to 79% in school 126. Notably, these values all are somewhat higher when using strategy 1, implying that the observations dropped from the network using this missing data strategy are systematically different from the rest of the network in this sample. Likewise, heavy smokers at time 1 are 3, 4, 5, 6, and 9 percentage points greater using strategies 3 and 4 compared to strategy 1 in schools 077, 058, 126, 008, and 002, respectively. This highlights that researchers employing these strategies dropping observations missing network data at time 1, or at any time point, encounter the risk of a resultingly biased sample. Assessing whether this is indeed the case is necessary for such an approach. We will assess here the consequences of violating this assumption.

Results

Given that we estimated the model using four different missing data strategies across six networks (for 24 total estimated models), some summarization of results is necessary. We

present the SAB model school-by-school results across the four missing data strategies in Appendix 2 (Tables A4-A9). We focus on a condensed version of the results in Table 2 in which we compare the size of coefficients across the various missing data strategies, averaged over the six networks. We compare missing data strategies 1 through 3 to strategy 4, which uses a principled approach to imputation.[2] For example, Column 1 of Table 2 shows the *average bias* across the six schools when comparing strategy 1 (dropping any cases with missing network data at any time point) to strategy 4. *Average bias* between strategies 1 and 4 is computed as follows: 1) for each school, we compute the difference between the estimated parameter for strategy 1 and that for strategy 4, and divide this by the parameter for strategy 4, and 2) we then compute the average of these values for the six networks. Positive values indicate that that strategy 1 yields larger positive coefficients (or smaller negative coefficients) than strategy 4, on average. Negative values indicate that strategy 4 yields larger positive coefficients (or smaller negative coefficients) than strategy 1, on average. The average bias between strategies 2 or 3 and strategy 4 are computed similarly. Column 2 shows the *average error* when comparing strategy 1 to strategy 4. This is computed as follows: 1) for each school, we compute the absolute value of the difference in the estimated parameters from strategy 1 and strategy 4 divided by the parameter for strategy 4, and 2) we then compute the average of these values for the six networks. Larger positive values indicate greater differences in the coefficients across the two strategies, on average, whereas values closer to zero indicate minimal differences in the coefficients across the two strategies. The average error between strategies 2 or 3 and strategy 4 computed similarly. Columns 3 and 4 display the average bias and average error across all six schools when comparing strategy 2 to strategy 4. And columns 5 and 6 display the average bias and average error across all six schools when comparing strategy 3 to strategy 4.

In comparing the results across all six schools, we distinguish between consistent patterns and those that are unique to specific schools. In the first column of Table 3, we see that the average estimated parameter for the rate of change in friendship is about 50% smaller during the first period for strategy 1 (dropping cases missing any network data at any time point) versus strategy 4, and the parameter during the second period is about 40% less for strategy 1 versus strategy 4. These parameter values are always under-estimated in strategy 1 versus strategy 4 across these six schools, and hence the average bias has the same value as the average error (absolute value of these differences). In columns 3 and 4 we see that these rate parameters are also under-estimated in strategy 2 (dropping any cases missing network data at time 1) relative to strategy 4, although the gap is a little narrower: the rate during period 1 is 35% less and the rate in period 2 is 23% less. The gap is narrower yet when comparing strategy 3 (imputing all missing network at time 1 to be null ties) to strategy 4, although even here we see that the rate in period 1 is 20% less in strategy 3 and the rate in period 2 is 8% less. This general pattern can also be seen in Table 2 showing the results for the largest school: the friendship rate parameter during period 1 increases from 8.4 to 22.9 across the four missing data strategies.

[2]Of course, comparisons between the other strategies can be made implicitly given that these are relative comparisons.

For the out-degree and reciprocity parameters, there is little systematic bias in the estimates for strategies 1 or 2 relative to strategy 4. The parameter estimates across the six schools are 6% to 9% different across the strategies (as seen in the average error columns) but the average bias is less than 5%. These two parameters are slightly under-estimated in strategy 3 compared to strategy 4—about 6% less on average across these schools.

For four of the network structural measures—transitive triplets, three cycles, in-degree popularity, and in-in-degree assortativity (square root)—the parameter values are over-estimated using strategies 1 through 3. For example, comparing strategy 1 to strategy 4 these parameter values are over-estimated on average by 32% for three cycles, 68% for transitive triplets, 91% for in-degree popularity, and 169% for in-in-degree assortativity. The average bias values are less extreme, though still pronounced, when comparing strategy 2 to strategy 4, as they range from 31% to 95%. The average bias is also quite pronounced for strategy 3, with estimated parameters that are 29% to 69% larger on average than those for strategy 4. We also observe that two of these structural network parameters showed relatively consistent differences across the various strategies in each of the schools. First, the absolute value of the in-degree popularity parameter tended to be largest in strategy 1 and smallest in strategy 4. Comparing strategy 1 to strategy 4, this positive parameter was between 46% and 157% larger across these schools. Second, the absolute value of the in-in degree assortativity (square root) parameter tended to be largest in strategy 1 and smallest in strategy 4. Comparing strategy 1 and strategy 4, the negative in-in degree assortativity parameter was two to three times larger in the various schools. These are consistent patterns across these four strategies of treating missing data, and highlight that the way in which missing data is handled is far from a trivial issue.

We also point out that conclusions regarding *statistical significance* of the in-in-degree assortativity parameter vary across these strategies. In one of the large schools, one of the mid- sized schools, and both of the small schools, the statistical conclusion will differ based on the missing data strategy employed. For example, in the largest school (077) in Table 2 this parameter is not statistically significant ($p > .05$) in strategy 4, but is significant using the other strategies. In mid-sized school 008 this parameter attains statistical significance at $p < .05$ using strategy 2, but not under any of the other strategies (see Table A6 in Appendix 2). In small school 002 the parameter is statistically significant when using strategy 1 and 4, but not the other two strategies (Table A7 in the Appendix). And for small school 126 the parameter is statistically significant for strategies 1 and 4, but not the other two (Table A8 in the Appendix). Table 3 displays the number of times a parameter achieves statistical significance (based on $p < .05$) in one technique but not another.

There is also evidence of different results across the missing data strategies for the network covariate measures. For example, the grade similarity effect is upwardly biased 24%, 15%, and 13% for strategies 1, 2, and 3, respectively, on average, compared to strategy 4. In the largest school in Table 2, although race homophily (the race similarity variable) is statistically significant in all strategies, the size of the coefficient is 29% larger when using strategy 1 compared to the other strategies.

A key feature of SAB models is simultaneously exploring influence and selection processes; in the present models smoking behavior is of particular interest. There are some notable differences in the estimated parameter values for the smoking similarity (selection) measure across the different missing data strategies. The average error across the five of the schools comparing strategies 1 and 4 is 23% (we exclude small school 002 which has a very large difference and would skew the results), it is 82% for strategy 2, and 36% for strategy 3. There is minimal evidence of systematic bias across these strategies, suggesting that a researcher cannot be sure if the estimated parameter is larger or smaller compared to what would be obtained using the principled imputation procedure of strategy 4. For example, in one large school (077) the parameter estimates for smoking selection from strategies 2 and 3 are more than twice as large as that for strategy 4 (Table 2). In one mid-sized school (007) this parameter is statistically significant in strategies 1 and 2 but not the other two strategies, but in the other mid- sized school (008) it is statistically significant in strategies 1 and 4 but not the other two strategies. And in one of the small schools (002) the size of the effect differs considerably across these strategies (and is not statistically significant for strategy 3). Thus, the decision on how to handle missing network data can have important implications in terms of estimated selection effects.

The smoking alter parameter captures the popularity of smokers in these networks, and the estimated parameter for this characteristic differs considerably over the various missing data strategies. For example, the average error across schools is 36% comparing strategy 1 to strategy 4 for this measure, it is 18% comparing strategy 2 to 4, and 39% comparing strategy 3 to 4. The parameter estimates across strategies are larger in some schools, but smaller in other schools compared to strategy 4. In the largest school (077) the parameter for alter smoking in strategy 1 is not statistically significant compared to the other three strategies. The same pattern is also detected in the other large school (058). And for mid-sized school 008, only strategy 4 has enough statistical power to conclude that smokers (smoking ego) name fewer ties than others.

Turning to the equation in which smoking behavior is the outcome, it is notable that the effect of smoking similarity—the influence effect—differs considerably over these missing data strategies. When using strategy 1 instead of 4, this parameter is 85% smaller, on average, across these schools. And when using strategy 2 instead of 4, this parameter is 42% smaller, on average, across these schools. Although the conclusions regarding statistical significance for this parameter did not differ across these strategies in these particular schools, the sharp differences in estimated parameters might have considerable consequences for researchers who wish to explore the dynamic implications of these models by perturbing various parameter values or altering the composition of the sample along values of key variables and then simulating the model forward in time.

We also find striking differences in some of the covariate effects for the evolution of smoking behavior between strategies 1 or 2 versus strategies 3 or 4. The effect of depression on smoking behavior is typically inflated in strategy 1 compared to the other strategies: in one- midsized school (007) it has a much larger coefficient and is statistically significant under strategy 1 but not with the other strategies. In the largest school (077), although strategy 1 has the largest coefficient, statistical significance only reaches $p < .10$ given the

greater uncertainty in this strategy (in part due to the reduced sample size). In the largest school, the negative effect of parental support appears twice as strong in strategy 1 compared to the other strategies. The effect of the home smoking environment varies somewhat over these strategies, although not in a systematic way: whereas there is 26% and 21% more average error using strategies 1 or 2 compared to strategies 3 or 4, the bias is low given that the estimates can be either higher or lower. In the largest school (077), the size of the effect for females is almost 50% larger in strategy 1 compared to the other strategies.

Among the baseline parameters for smoking behavior, we note that the rate parameters are somewhat under-estimated in strategies 1 and 2 compared to strategy 4 across these schools. However, the linear and quadratic shape parameters do not differ much over these missing data strategies.

Finally, we assessed whether there were systematic differences in the pattern of biases across missing data strategies depending on the size of the network in this study (small, medium, or large). In general, there was little evidence of systematic differences for these strategies based on the size of the network. There was some suggestive evidence that the selection parameter based on smoking behavior was more strongly biased upwards in smaller networks when using strategies 1 or 2 compared to strategy 4, but the small number of networks in the study precludes making a more confident assessment of such a pattern.

Discussion

The stochastic actor based approach is growing in popularity as a technique for handling longitudinal network data, in part because it is designed to explore influence and selection processes. Despite this growing popularity, there is limited knowledge regarding the consequences of various strategies for handling missing data. Scholars are well aware that missing data can be particularly challenging in longitudinal designs. The consequences of such missingness for inferences regarding social processes may be even more challenging. As a first attempt at assessing the scope of the issue, we have explored the impact of four missing data strategies on inference for SAB model parameters using SIENA on a commonly used dataset. Our findings demonstrate that management of missing data is not a trivial decision, but in fact has serious consequences for parameter estimates and substantive conclusions.

The current study is a first step in demonstrating that missing data can have strong consequences for the obtained results of longitudinal network models. The fact that we found sometimes striking differences when employing different missing data strategies suggests that much more research is needed to explore these issues. Whereas our study had the virtue of studying the consequences of missing network data on an existing dataset that is frequently used in applied research, we are obviously limited in our ability to draw conclusions regarding the “right” approach to take: the fact that different procedures provide different answers implies that at least some are problematic, but a definitive performance assessment requires examination of cases in which the correct answer is known (e.g., via simulation). There is a clear need for Monte Carlo studies that explore the effects of different types of missingness by first drawing data from a known SAB process,

systematically introducing different types of missingness into the synthetic data, and then estimating models on the resulting data sets.

In light of our results, key questions to be addressed by such Monte Carlo studies should include the impact of missingness associated with behavioral or structural characteristics (e.g., smoking behavior or degree), missingness that occurs at different rates over time (e.g., due to declining participation or late enrollment), and missingness in small versus large networks. Such studies should also examine impact of the various analyses strategies discussed here (as well, perhaps, as others), to provide useful guidance on what can be done in practice to minimize the effect of missing data on substantive conclusions. Finally, we suggest that such studies consider not only effects of missing data and analysis strategy on estimated parameter values and significance levels, but also on predictions resulting from the fitted models (e.g., rates of smoking or mean degree). Given the widespread interest in SAB models as tools to inform policy, inaccuracies in predicted behavior patterns are at least as important as inaccuracies in inferred model parameters.

Our study raises broader issues for questions relating to missing data and dynamic network (and network/behavioral) analyses. Although the consequences of different types of missingness are relatively well understood for cross-sectional data, they are less well understood for dynamic network models. For example, although data that is missing completely at random (MCAR) may have the most benign consequences for SAB models, it still is not well understood whether missing data techniques that simply exclude such observations will affect the parameter estimates from SAB models given that this will change the apparent structure of the network (e.g., by changing network size). Even if network data is missing at random (MAR, but *not* MCAR), individuals with missing network data may differ systematically.[3] For example, those with missing network data may systematically differ based on such characteristics as age, their substance use behavior, or their position in the network (e.g., popular actors). In the MAR case, a principled imputation strategy can correct for these biases; even where data is not fully MAR, a strategy that takes at least some of these factors into account will likely yield more accurate results than strategies that simply exclude these cases. Notably, almost none of the existing research using SAB models adopts such a principled strategy for imputing missing network data (at least in the first time point).

Given that a primary goal of researchers employing SAB models is disentangling influence and selection effects, it is particularly notable how different the estimated influence and selection effects could be based on the particular missing data strategy employed. In this study, the size of the estimated parameter for selection based on smoking behavior differed considerably across these missing data strategies. Furthermore, the direction of the bias was uncertain, as the parameter estimates were both over- or under-estimated compared to estimates based on a principled imputation strategy at time 1. Whereas researchers often impute 0's to missing ties at time 1, this approach had an average error of 35% for this

[3]Data is missing at random if its probability of inclusion depends only on observed data and covariates (i.e., not on the values of unobserved quantities); despite the name, MAR does not imply that all variables are equally likely to be observed, nor that this probability is unrelated to other factors.

selection parameter compared to a principled imputation strategy at time 1. Notably, the conclusions regarding statistical significance of this parameter often differed depending on the missing data strategy employed. Given that the statistical significance of this parameter is often of paramount interest, these results highlight the importance of considering seriously how to address missing network data in such longitudinal network designs.

It was also notable that the average smoking influence effect was *underestimated* when using certain missing data strategies compared to a principled imputation strategy at time 1. Thus, dropping cases that are missing on network data at any time point (strategy 1) or at time 1 (strategy 2) can seriously impact estimates of influence effects: in this study these were 85% and 40% smaller parameter estimates on average, respectively. These two missing data strategies will not necessarily always bias such estimates downward—the consequences will depend on which cases are dropped from the analysis based on the pattern of missingness—but the results of this study highlight that the impact can be quite consequential.

Another instance in which biased estimates of structural network parameters might matter is when researchers use the estimates of SAB models for forward simulation of the network based on various perturbations (Schaefer, adams, and Haas 2013). The large differences we detected in the influence and selection parameters could also have considerable effects on such forward simulations. The biased estimates of these network parameters may impact the conclusions drawn from such simulations that could otherwise potentially provide key insights into the possible consequences of various policy manipulations.

Another parameter of much substantive interest to adolescent smoking researchers is the measure of smoking popularity (smoking alter, in our models). There is debate in the literature, as some have suggested that smokers tend to be more popular (Alexander, Piazza, Mekos, and Valente 2001; Valente, Unger, and Johnson 2005), whereas others have suggested that smokers tend to be isolates (Ennett and Bauman 1993). Notably, the estimates of this parameter varied over missing data strategies. Furthermore, the conclusions regarding statistical significance often varied depending on the missing data strategy employed. Given that in each strategy we are estimating the same model on what was initially the same network sample (before observations were dropped in some strategies) these differing conclusions highlight that missing data decisions are not some arcane statistical decision, but rather of crucial importance for substantive conclusions.

We also found that whereas ego and dyadic network measures (e.g., out-degree, reciprocity) were typically less affected by the choice of missing data strategy, higher order network measures were quite strongly impacted. In our study, measures of transitive triplets, 3-cycles, and in-in degree assortativity were upwardly biased anywhere from 30% to 170% on average by alternative missing data strategies compared to a principled imputation strategy at wave 1. The parameter estimates for in-degree (capturing popularity) similarly differed over missing data strategies. Although these network structural parameters are often not of primary focus to SAB modelers—who are more typically focused on the relative effects of influence and selection—these parameters may nonetheless impact the estimated values of

the selection and influence parameters, which we saw indeed often differed over these strategies.

Dropping observations due to missing data generally relies (tacitly or otherwise) on an assumption of MCAR, and the differences we detected across missing data strategies for the covariate estimates in the smoking equation suggest that MCAR does not characterize the missing network data in our sample. The fact that we observed differences in some of the summary statistics for the subset of persons with network information at time 1 compared to the full sample implies that persons with missing network data are systematically different in various ways compared to the entire sample in these networks. The concern of violating the MCAR assumption is why researchers are always wary of using an approach that simply excludes all cases with missing data from the analyses; our results highlight that this issue is equally important for researchers with missing network data. Again, a more principled imputation approach for missing network data is called for in such instances. We also note that omission of nodes from a network alters its size, which may have non-trivial impact on other network properties (see, e.g., Anderson, Butts, and Carley 1999; Butts 2006; Faust 2007).

Conclusion

Although we make no claim that any of these approaches yields the “true” results, nor that these are exhaustive of all possible missing data strategies, it is nonetheless the case that some of these strategies for treating missing data are more defensible compared to others based on our presumptions about how the processes underlying these networks operate. We hope that this study will bring about two developments in the SAB literature. First, we hope that our results will spur additional research exploring the consequences of different types of missing network data using Monte Carlo simulation studies. Simulations would allow isolating the consequences of different types of missing network data for the parameter estimates of SAB models. It would also allow assessing whether the consequences differ based on the size of the network, or based on various characteristics of the network (e.g., density, clustering).

Second, it is our hope that applied researchers will give more consideration to missing network information, and even consider other possible strategies that might be employed. We have focused here on the strategies most commonly utilized in the existing literature. Strategy 1, which discards all persons who are missing at any wave in the study, may be particularly hard to defend given that it assumes that those persons are not part of the network. Given that they are in fact part of the network, this potentially biases the structural network parameters. To the extent that excluded persons are different from those included in the network, this strategy can yield biased results. Strategy 2, which excludes individuals who were not in the sample at wave 1, suffers from a similar strong assumption, albeit somewhat weaker since it includes persons present in the network at subsequent waves. Although strategy 3 has the virtue of including all members of the network present at any of the waves, it makes the rather strong assumption that those who did not report network data at wave 1 in fact had no social ties at that time. This artificially deflates network density at the first time point (potentially in a manner that is conflated with one or more predictors, if

non-response is non-random), and exaggerates the apparent rate of change between the first and second time points (since some ties appearing in time 2 appear to be novel due to the fact that they were taken to be absent at time 1). Strategy 4 attempts to build a model predicting which ties actually exist for these missing persons at the first time step: although we cannot know whether this model of ties is the true one, it does seem likely that it will do a better job of predicting ties than simply assuming that none exist (as is done in strategy 3). Of the four strategies presented here, we recommend this last approach on the grounds of its substantive plausibility, maximal use of available data, ease of robustness testing (via multiple imputation), and capacity for future refinement. Regardless, our findings clearly demonstrate that researchers are well-advised to carefully consider the strategies they use for handling missing data when fitting SAB models, given that the results obtained can differ quite strongly based on this choice.

Acknowledgments

This research is supported in part by NIH grant R21 DA031152-01A1.

Appendix 1: Description of ERGM imputation procedure

Due to missing data, each school has an incomplete adjacency matrix whose *i,j* entries are: 0 if it is known that student *i* did *not* nominate student *j*; 1 if it is known that student i *did* nominate student *j*; and NA if it is not known whether student *i* nominated student *j* as a friend. For each student we have the minimum and maximum number of nominations from him or her to: the set of all male students; the set of all female students; the set of all male off-roster students; and the set of all female off-roster students. These counts are inferred from the invalid and/or off-roster entries in each respondent's male and female nominee lists, and (for non- respondents) from the global male/female out-degree constraint. The edge variables that are coded as NA are the portion of the adjacency matrix that the ERG model will impute during the simulation portion of the process, subject to these group-specific out-degree constraints.

We use a combination of inference and simulation with a model-based procedure within an ERGM-based framework to estimate uncertain edge states associated with missing data. The ERGM approach specifies the model based on structural network measures and a covariate set *X*. We follow Gile and Handcock (2006) in constructing the observed data likelihood for the above model that contains both the missing and non-missing portion. Maximum likelihood inference requires a complex MCMC-based algorithm described by Snijders (2002), Snijders et al. (2006), and Wasserman and Robins (2005), and we employ the implementation of this method in the ergm package (Hunter, Handcock, Butts, Goodreau, and Morris 2008) of the statnet (Goodreau, Handcock, Hunter, Butts, and Morris 2008; Handcock, Hunter, Butts, Goodreau, and Morris 2008) software suite.

To impute the unobserved elements in our respective nomination networks, we must first model each network. Using the above approach, we estimated a model that contained: 1) the edge count statistic (i.e., a homogeneous Bernoulli digraph with support constraints); 2) a mutuality/reciprocity effect; 3) the absolute difference in school grade (de facto age) and gender; 4) homophily effects for those in the same class(es), the same club(s), and the same

sport- team(s); 5) a geometrically weighted edgewise shared partner term (gwesp) fixed at its optimal value (δ_{opt}).[4] For all but one school the models converged within 2,000 iterations.[5]

After estimating the model parameters, we employ conditional ERGM simulation to impute missing edge states. In this case, missing edge variables (NAs) are imputed based on the model estimated with the observed data, with observed edges (1s) and nulls (0s) unaltered and all degree constraints based on the observed data enforced.

Appendix 2

Table A1
Network statistics

School	Net statistics	Strategy 1	Strategy 2	Strategy 3	Strategy 4
077	# of nodes	851	1,674	2,104	2,104
	# of edges at t_1	1,765	3,706	4,585	5,685
	# of edges at t_2	1,237	2,514	4,201	4,201
	# of edges at t_3	1,074	1,469	2,296	2,296
058	# of nodes	479	757	1,024	1,024
	# of edges at t_1	1,854	3,331	4,037	6,063
	# of edges at t_2	1,476	2,437	3,713	3,713
	# of edges at t_3	1,308	1,673	2,484	2,484
007	# of nodes	121	160	181	181
	# of edges at t_1	456	812	853	1,193
	# of edges at t_2	211	338	416	416
	# of edges at t_3	280	375	421	421
008	# of nodes	70	111	133	133
	# of edges at t_1	183	372	413	706
	# of edges at t_2	116	239	320	320
	# of edges at t_3	191	281	359	359
002	# of nodes	48	54	78	78
	# of edges at t_1	129	192	255	367
	# of edges at t_2	62	90	155	155
	# of edges at t_3	87	110	183	183
126	# of nodes	42	46	62	62
	# of edges at t_1	168	209	229	285
	# of edges at t_2	89	105	144	144
	# of edges at t_3	66	74	110	110

[4]The optimal value is determined by estimating a series of models with gwesp fixed from 0 to 3 with 0.1 increment at a step and locating the one with the smallest AIC, BIC, and log-likelihood values.
[5]We raised the iteration limit for the largest school (077) given that the models required between 7,000 to 22,000 iterations before reaching convergence.

Table A2
Smoking level at t_1

School	Smoking level (past 30 days, %)	Strategy 1	Strategy 2	Strategy 3 & 4
077	0 = never	70.04	70.67	68.60
	1 = 1-3days	19.39	17.68	17.54
	2 = 4-21 days	4.70	4.54	4.91
	3 = 22 or more days	5.88	7.11	8.95
058	0 = never	43.01	43.46	42.01
	1 = 1-3days	24.63	22.59	21.31
	2 = 4-21 days	8.98	10.04	9.02
	3 = 22 or more days	23.38	23.91	27.66
007	0 = never	57.02	55.62	55.84
	1 = 1-3days	15.70	16.85	15.74
	2 = 4-21 days	4.96	5.62	5.58
	3 = 22 or more days	22.31	21.91	22.84
008	0 = never	51.43	45.67	44.30
	1 = 1-3days	21.43	18.90	20.13
	2 = 4-21 days	10.00	12.60	12.08
	3 = 22 or more days	17.14	22.83	23.49
002	0 = never	91.67	92.59	75.64
	1 = 1-3days	8.33	7.41	14.10
	2 = 4-21 days	0.00	0.00	1.28
	3 = 22 or more days	0.00	0.00	8.97
126	0 = never	85.71	84.78	79.03
	1 = 1-3days	14.29	13.04	12.90
	2 = 4-21 days	0.00	2.17	3.23
	3 = 22 or more days	0.00	0.00	4.84

Smoking level at t_2				
School	**Smoking level (past 30 days, %)**	**Strategy 1**	**Strategy 2**	**Strategy 3 & 4**
077	0 = never	79.20	79.93	78.28
	1 = 1-3days	6.93	6.99	7.44
	2 = 4-21 days	8.11	6.93	7.07
	3 = 22 or more days	5.76	6.15	7.21
058	0 = never	53.03	56.14	53.18
	1 = 1-3days	10.02	9.38	9.12
	2 = 4-21 days	12.94	11.62	11.58
	3 = 22 or more days	24.01	22.85	26.13
007	0 = never	57.85	58.43	56.85
	1 = 1-3days	7.44	8.43	8.12

Smoking level at t_2				
School	**Smoking level (past 30 days, %)**	**Strategy 1**	**Strategy 2**	**Strategy 3 & 4**
	2 = 4-21 days	11.57	11.80	13.20
	3 = 22 or more days	23.14	21.35	21.83
008	0 = never	68.57	66.93	65.10
	1 = 1-3days	8.57	7.87	8.72
	2 = 4-21 days	11.43	8.66	8.72
	3 = 22 or more days	11.43	16.54	17.45
002	0 = never	95.84	94.45	92.31
	1 = 1-3days	0.00	1.85	2.56
	2 = 4-21 days	2.08	1.85	1.28
	3 = 22 or more days	2.08	1.85	3.85
126	0 = never	92.86	91.30	90.32
	1 = 1-3days	2.38	4.35	4.84
	2 = 4-21 days	2.38	2.17	1.61
	3 = 22 or more days	2.38	2.17	3.23

Smoking level at t_3				
School	**Smoking level (past 30 days, %)**	**Strategy 1**	**Strategy 2**	**Strategy 3 & 4**
077	0 = never	75.21	72.22	71.76
	1 = 1-3days	8.93	9.08	9.37
	2 = 4-21 days	8.23	8.72	8.91
	3 = 22 or more days	7.64	9.98	9.96
058	0 = never	47.60	47.82	45.39
	1 = 1-3days	10.65	12.29	11.68
	2 = 4-21 days	11.69	10.04	10.55
	3 = 22 or more days	30.06	29.85	32.38
007	0 = never	49.59	46.63	47.21
	1 = 1-3days	7.44	8.43	8.12
	2 = 4-21 days	8.26	10.11	9.64
	3 = 22 or more days	34.71	34.83	35.03
008	0 = never	44.29	43.31	46.31
	1 = 1-3days	10.00	10.24	9.40
	2 = 4-21 days	17.14	15.75	14.77
	3 = 22 or more days	28.57	30.71	29.53
002	0 = never	85.41	81.48	83.33
	1 = 1-3days	4.17	5.56	7.69
	2 = 4-21 days	6.25	5.56	3.85
	3 = 22 or more days	4.17	7.41	5.13
126	0 = never	80.95	82.61	80.65
	1 = 1-3days	9.52	8.70	9.68
	2 = 4-21 days	7.14	6.52	4.84

Smoking level at t_3				
School	**Smoking level (past 30 days, %)**	**Strategy 1**	**Strategy 2**	**Strategy 3 & 4**
	3 = 22 or more days	2.38	2.17	4.84

Table A3
Covariate statistics

School		Strategy 1	Strategy 2	Strategy 3 & 4
077	Female (%)	51.00	47.97	47.52
	Grade level (%)			
	10th grade	47.00	38.05	37.24
	11th grade	43.83	32.74	33.43
	12th grade	9.17	29.21	29.34
	Depression, mean (sd)	0.14(0.51)	0.11(0.52)	0.14(0.53)
	Home smoking environment, mean (sd)	0.79(0.76)	1.11(0.77)	1.09(0.77)
	Parental support, mean (sd)	-0.02(0.28)	-0.03(0.29)	-0.05(0.30)
058	Female (%)	47.18	48.48	48.46
	Grade level (%)			
	9th grade	35.28	28.79	28.79
	10th grade	33.40	30.38	28.48
	11th grade	24.01	21.14	21.72
	12th grade	7.31	19.68	21.00
	Depression, mean (sd)	-0.05(0.51)	-0.04(0.51)	0.00(0.53)
	Home smoking environment, mean (sd)	1.24(0.78)	1.38(0.74)	1.42(0.73)
	Parental support, mean (sd)	-0.04(0.29)	-0.03(0.28)	-0.04(0.29)
007	Female (%)	42.98	47.19	48.22
	Grade level (%)			
	7th grade	17.36	15.17	14.72
	8th grade	19.01	16.29	15.74
	9th grade	22.31	22.47	22.84
	10th grade	18.18	14.04	14.21
	11th grade	17.36	15.17	15.23
	12th grade	5.79	16.85	17.26
	Depression, mean (sd)	-0.13(0.47)	-0.11(0.47)	-0.09(0.48)
	Home smoking environment, mean (sd)	1.33(0.80)	1.42(0.74)	1.43(0.74)
	Parental support, mean (sd)	0.07(0.23)	0.07(0.25)	0.06(0.24)
008	Female (%)	45.71	50.39	48.99
	Grade level (%)			
	7th grade	27.14	21.26	21.48
	8th grade	18.57	19.69	20.81
	9th grade	12.86	14.17	14.09

School		Strategy 1	Strategy 2	Strategy 3 & 4
	10th grade	22.86	15.75	16.78
	11th grade	14.29	12.60	11.41
	12th grade	4.29	16.54	15.44
	Depression, mean (sd)	-0.08(0.49)	-0.07(0.51)	-0.06(0.50)
	Home smoking environment, mean (sd)	1.30(0.73)	1.44(0.66)	1.42(0.68)
	Parental support, mean (sd)	0.03(0.28)	0.02(0.27)	0.02(0.27)
002	Female (%)	56.25	53.70	52.56
	Grade level (%)			
	7th grade	18.75	18.52	21.79
	8th grade	22.92	20.37	19.23
	9th grade	20.83	20.37	21.79
	10th grade	12.50	12.96	14.10
	11th grade	18.75	20.37	16.67
	12th grade	6.25	7.41	6.41
	Depression, mean (sd)	-0.14(0.34)	-0.15(0.35)	-0.16(0.38)
	Home smoking environment, mean (sd)	0.59(0.68)	0.69(0.72)	0.87(0.81)
	Parental support, mean (sd)	0.13(0.19)	0.13(0.19)	0.13(0.20)
126	Female (%)	52.38	50.00	53.23
	Grade level (%)			
	7th grade	45.24	45.65	46.77
	8th grade	50.00	50.00	48.39
	9th grade	4.76	4.35	4.84
	Depression, mean (sd)	-0.10(0.48)	-0.12(0.47)	-0.03(0.51)
	Home smoking environment, mean (sd)	1.14(0.73)	1.14(0.71)	1.16(0.75)
	Parental support, mean (sd)	0.12(0.20)	0.13(0.20)	0.08(0.22)

Table A4
Four Models for school 077

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Constant friendship rate (period 1)	8.40^{***} (0.69)	9.95^{***} (0.56)	15.51^{***} (2.57)	22.87^{***} (0.91)
Constant friendship rate (period 2)	4.92^{***} (0.24)	6.15^{***} (0.29)	8.74^{***} (0.86)	9.47^{***} (0.44)
Out-degree (density)	-4.66^{***} (0.09)	-4.33^{***} (0.08)	-4.25^{***} (0.30)	-4.39^{***} (0.06)
Reciprocity	2.94^{***} (0.12)	2.90^{***} (0.11)	2.48^{***} (0.13)	2.89^{***} (0.09)
Transitive triplets	0.95^{***} (0.06)	1.03^{***} (0.05)	0.98^{***} (0.21)	0.83^{***} (0.04)
3-cycles	-0.89^{***} (0.14)	-0.94^{***} (0.13)	-1.07^{***} (0.15)	-0.81^{***} (0.14)
In-degree - popularity	0.10^{*} (0.04)	0.10^{**} (0.03)	0.07^{*} (0.03)	0.06^{***} (0.02)
In-in degree^(1/2) assortativity	-0.14^{*} (0.07)	-0.19^{**} (0.06)	-0.11^{*} (0.05)	$-0.05^{\dagger}$ (0.02)

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Race similarity	1.41*** (0.06)	1.20*** (0.05)	1.09*** (0.04)	1.09*** (0.04)
Grade similarity	0.66*** (0.04)	0.54*** (0.03)	0.51*** (0.03)	0.50*** (0.02)
Smoking alter	0.09 (0.07)	0.15* (0.06)	0.22** (0.07)	0.09** (0.03)
Smoking ego	-0.02 (0.06)	0.01 (0.08)	-0.03 (0.07)	0.07† (0.04)
Smoking similarity	0.26** (0.09)	0.36*** (0.09)	0.35* (0.14)	0.16* (0.08)
Limited nomination ego	-0.48*** (0.12)	-0.46*** (0.10)	-0.48*** (0.08)	-0.52*** (0.07)
Behavior decision: influence processes				
Rate smoking behavior (period 1)	3.47*** (0.47)	3.09*** (0.42)	13.45*** (2.65)	12.77*** (1.23)
Rate smoking behavior (period 2)	5.82*** (1.57)	17.29*** (3.27)	24.71*** (3.62)	22.33*** (2.10)
Smoking behavior linear shape	-2.39*** (0.15)	-2.43*** (0.11)	-2.44*** (0.10)	-2.46*** (0.08)
Smoking behavior quadratic shape	0.71*** (0.04)	0.73*** (0.03)	0.72*** (0.06)	0.74*** (0.03)
Smoking behavior in-degree	-0.02 (0.04)	-0.01 (0.02)	0.00 (0.02)	0.00 (0.01)
Smoking behavior new similarity	0.54* (0.27)	0.51** (0.16)	0.51* (0.24)	0.56** (0.20)
Effect from gender (female=1)	-0.18* (0.08)	-0.12* (0.05)	-0.13*** (0.04)	-0.13** (0.05)
Effect from Black	-0.26* (0.11)	-0.33*** (0.07)	-0.25*** (0.07)	-0.25*** (0.04)
Effect from depression	0.13† (0.07)	0.07 (0.06)	0.09 (0.06)	0.09* (0.04)
Effect from home smoking environment	0.18*** (0.05)	0.18*** (0.03)	0.14*** (0.02)	0.14*** (0.02)
Effect from parental support	-0.32* (0.14)	-0.19* (0.09)	-0.16† (0.09)	-0.17* (0.08)

† Two-sided p<0.1;
* Two-sided p<0.05;
** Two-sided p<0.01;
*** Two-sided p<0.001

Table A5
Four Models for school 058

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Constant friendship rate (period 1)	10.66*** (0.49)	13.59*** (0.56)	16.80*** (0.66)	23.47*** (0.82)
Constant friendship rate (period 2)	9.47*** (0.46)	10.68*** (0.55)	13.79*** (0.46)	15.15*** (0.52)
Out-degree (density)	-2.53*** (0.06)	-2.66*** (0.06)	-2.62*** (0.06)	-2.76*** (0.05)
Reciprocity	2.59*** (0.08)	2.47*** (0.06)	2.35*** (0.11)	2.51*** (0.06)
Transitive triplets	0.70*** (0.03)	0.60*** (0.03)	0.66*** (0.02)	0.55*** (0.02)
3-cycles	-0.47*** (0.07)	-0.51*** (0.06)	-0.63*** (0.06)	-0.46*** (0.04)
In-degree - popularity	0.07*** (0.01)	0.06*** (0.01)	0.06*** (0.02)	0.05*** (0.01)
In-in degree^(1/2) assortativity	-0.15*** (0.02)	-0.09*** (0.02)	-0.09** (0.03)	-0.06** (0.01)
Grade similarity	0.50*** (0.03)	0.49*** (0.02)	0.47*** (0.02)	0.43*** (0.02)
Smoking alter	0.04† (0.02)	0.06** (0.02)	0.08** (0.02)	0.08*** (0.02)
Smoking ego	-0.04* (0.02)	-0.01 (0.02)	0.00 (0.02)	0.00 (0.02)

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Smoking similarity	0.20*** (0.03)	0.24*** (0.03)	0.27*** (0.03)	0.26*** (0.02)
Limited nomination ego	-0.71*** (0.08)	-0.80*** (0.07)	-0.70*** (0.06)	-0.71*** (0.06)
Behavior decision: influence processes				
Rate smoking behavior (period 1)	2.06*** (0.23)	2.45*** (0.30)	9.41*** (1.00)	9.18*** (1.28)
Rate smoking behavior (period 2)	6.03** (1.22)	11.06*** (1.33)	14.55*** (1.42)	14.50*** (1.83)
Smoking behavior linear shape	-2.31*** (0.16)	-2.29*** (0.12)	-2.28*** (0.09)	-2.25*** (0.10)
Smoking behavior quadratic shape	0.72*** (0.04)	0.71*** (0.03)	0.67*** (0.03)	0.67*** (0.02)
Smoking behavior in-degree	0.02 (0.02)	0.01 (0.01)	0.02† (0.01)	0.01 (0.01)
Smoking behavior new similarity	0.60*** (0.16)	0.41*** (0.12)	0.82*** (0.11)	0.80*** (0.10)
Effect from gender (female=1)	0.03 (0.09)	-0.00 (0.06)	-0.01 (0.04)	-0.01 (0.04)
Effect from depression	0.15† (0.08)	0.13* (0.06)	0.13** (0.04)	0.13** (0.04)
Effect from home smoking environment	0.11* (0.05)	0.11** (0.04)	0.12*** (0.03)	0.12*** (0.03)
Effect from parental support	0.23 (0.15)	0.15 (0.11)	-0.04 (0.07)	-0.02 (0.07)

† Two-sided p<0.1;

* Two-sided p<0.05;

** Two-sided p<0.01;

*** Two-sided p<0.001

Table A6
Four Models for school 007

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Constant friendship rate (period 1)	7.46*** (0.59)	10.73*** (0.84)	10.85*** (0.91)	13.53*** (1.13)
Constant friendship rate (period 2)	5.12*** (0.47)	7.68*** (0.80)	7.64*** (0.85)	8.19*** (1.13)
Out-degree (density)	-2.37*** (0.17)	-2.25*** (0.12)	-2.19*** (0.13)	-2.25*** (0.14)
Reciprocity	2.22*** (0.19)	1.92*** (0.15)	1.96*** (0.16)	1.97*** (0.17)
Transitive triplets	0.74*** (0.10)	0.51*** (0.05)	0.55*** (0.06)	0.38*** (0.06)
3-cycles	-0.49** (0.18)	-0.54*** (0.12)	-0.51*** (0.14)	-0.30** (0.11)
In-degree - popularity	0.15*** (0.03)	0.10*** (0.02)	0.11*** (0.02)	0.08*** (0.02)
In-in degree^(1/2) assortativity	-0.24** (0.09)	-0.11* (0.05)	-0.16** (0.06)	-0.10* (0.06)
Grade similarity	0.58*** (0.06)	0.51*** (0.04)	0.51*** (0.04)	0.44*** (0.06)
Smoking alter	0.09† (0.05)	0.04 (0.04)	0.01 (0.04)	0.04 (0.04)
Smoking ego	0.03 (0.06)	0.01 (0.05)	0.03 (0.05)	-0.02 (0.05)
Smoking similarity	0.11* (0.05)	0.11* (0.06)	0.10† (0.06)	0.10† (0.06)
Limited nomination ego	-0.89*** (0.10)	-0.73*** (0.09)	-0.74*** (0.08)	-0.68*** (0.07)
Behavior decision: influence processes				
Rate smoking behavior (period 1)	2.98** (1.02)	6.05** (1.83)	8.41* (3.52)	8.44* (3.65)
Rate smoking behavior (period 2)	7.25*** (1.87)	18.23* (4.38)	17.63** (6.36)	17.61* (7.00)

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Smoking behavior linear shape	-2.90*** (0.35)	-2.61*** (0.19)	-2.52*** (0.33)	-2.53*** (0.27)
Smoking behavior quadratic shape	0.85*** (0.09)	0.77*** (0.06)	0.76*** (0.06)	0.76*** (0.07)
Smoking behavior in-degree	0.05 (0.06)	0.02 (0.03)	0.00 (0.02)	0.00 (0.02)
Smoking behavior new similarity	-0.28 (0.30)	0.09 (0.19)	0.15 (0.23)	0.17 (0.36)
Effect from gender (female=1)	-0.03 (0.17)	-0.02 (0.08)	-0.11 (0.09)	-0.11 (0.08)
Effect from depression	0.52* (0.20)	0.08 (0.11)	0.05 (0.10)	0.06 (0.09)
Effect from home smoking environment	0.25* (0.11)	0.19* (0.08)	0.19* (0.09)	0.19** (0.06)
Effect from parental support	0.16 (0.42)	-0.11 (0.19)	-0.20 (0.18)	-0.20 (0.20)

† Two-sided p<0.1;

* Two-sided p<0.05;

** Two-sided p<0.01;

*** Two-sided p<0.001

Table A7
Four Models for school 008

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Constant friendship rate (period 1)	5.65*** (0.61)	9.18*** (0.84)	10.82*** (1.09)	12.96*** (1.78)
Constant friendship rate (period 2)	4.06*** (0.62)	6.29*** (0.77)	6.34*** (0.70)	6.94*** (0.64)
Out-degree (density)	-1.73*** (0.35)	-1.54*** (0.35)	-1.67*** (0.29)	-1.88*** (0.19)
Reciprocity	1.86*** (0.33)	1.80*** (0.26)	1.82*** (0.26)	1.97*** (0.26)
Transitive triplets	0.93*** (0.18)	0.53*** (0.11)	0.47*** (0.12)	0.32** (0.10)
3-cycles	-0.53* (0.27)	-0.62** (0.19)	-0.53** (0.19)	-0.27* (0.13)
In-degree - popularity	0.23*** (0.07)	0.14*** (0.03)	0.11** (0.04)	0.09** (0.03)
In-in degree^(1/2) assortativity	-0.42† (0.23)	-0.24* (0.10)	-0.16 (0.13)	-0.13 (0.10)
Grade similarity	0.57*** (0.09)	0.64*** (0.06)	0.59*** (0.05)	0.49*** (0.05)
Smoking alter	0.20† (0.12)	0.42† (0.25)	0.45† (0.24)	0.44* (0.19)
Smoking ego	-0.26 (0.18)	-0.52 (0.34)	-0.47 (0.29)	-0.38* (0.18)
Smoking similarity	0.41* (0.18)	0.36 (0.28)	0.45 (0.30)	0.46* (0.23)
Limited nomination ego	-1.57*** (0.45)	-1.49*** (0.30)	-1.34*** (0.28)	-1.21*** (0.25)
Behavior decision: influence processes				
Rate smoking behavior (period 1)	3.22** (1.00)	8.15** (2.51)	11.83** (3.92)	11.32*** (2.35)
Rate smoking behavior (period 2)	10.53* (5.16)	42.50** (13.16)	40.18*** (10.43)	37.01*** (7.68)
Smoking behavior linear shape	-2.08*** (0.33)	-2.36*** (0.24)	-2.50*** (0.23)	-2.44*** (0.27)
Smoking behavior quadratic shape	0.66*** (0.10)	0.71*** (0.08)	0.70*** (0.08)	0.70*** (0.07)
Smoking behavior in-degree	-0.04 (0.06)	-0.01 (0.02)	0.01 (0.02)	0.00 (0.02)
Smoking behavior new similarity	0.19 (0.32)	0.13 (0.21)	0.34 (0.44)	0.32† (0.19)
Effect from gender (female=1)	0.06 (0.23)	-0.03 (0.08)	-0.02 (0.08)	-0.02 (0.09)

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Effect from depression	-0.12 (0.21)	-0.08 (0.08)	-0.08 (0.08)	-0.09 (0.08)
Effect from home smoking environment	0.10 (0.12)	0.10 (0.07)	0.16† (0.08)	0.16* (0.07)
Effect from parental support	-0.36 (0.36)	-0.29 (0.19)	-0.33* (0.14)	-0.31† (0.18)

† Two-sided p<0.1;
* Two-sided p<0.05;
** Two-sided p<0.01;
*** Two-sided p<0.001

Table A8
Four Models for school 002

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Constant friendship rate (period 1)	4.71*** (0.63)	5.78*** (0.62)	7.82*** (0.74)	9.77*** (0.97)
Constant friendship rate (period 2)	3.41*** (0.50)	4.38*** (0.58)	5.48*** (0.58)	6.28*** (0.68)
Out-degree (density)	-1.94*** (0.32)	-2.03*** (0.28)	-1.74*** (0.38)	-1.95*** (0.20)
Reciprocity	1.92*** (0.42)	1.75*** (0.32)	1.87*** (0.32)	1.85*** (0.32)
Transitive triplets	0.93*** (0.22)	0.75*** (0.15)	0.76*** (0.13)	0.60*** (0.10)
3-cycles	-0.95** (0.44)	-1.06*** (0.31)	-0.99*** (0.26)	-0.62** (0.19)
In-degree - popularity	0.31*** (0.09)	0.21** (0.07)	0.21** (0.05)	0.14*** (0.03)
In-in degree^(1/2) assortativity	-0.62* (0.30)	-0.28 (0.20)	-0.39† (0.23)	-0.21* (0.09)
Grade similarity	1.02*** (0.17)	0.95*** (0.13)	0.86*** (0.10)	0.74*** (0.09)
Smoking similarity	2.04* (0.89)	1.05** (0.38)	0.14 (0.15)	0.17* (0.08)
Limited nomination ego	-1.08** (0.38)	-1.04*** (0.21)	-1.07*** (0.25)	-0.89*** (0.16)

† Two-sided p<0.1;
* Two-sided p<0.05;
** Two-sided p<0.01;
*** Two-sided p<0.001

Table A9
Four Models for school 126

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
Constant friendship rate (period 1)	6.94*** (0.82)	7.91*** (0.84)	9.17*** (0.98)	9.69*** (0.97)
Constant friendship rate (period 2)	3.51*** (0.50)	3.76*** (0.50)	4.91*** (0.61)	5.15*** (0.61)
Out-degree (density)	-2.51*** (0.29)	-2.58*** (0.31)	-2.28*** (0.24)	-2.47*** (0.20)
Reciprocity	2.01*** (0.47)	1.85*** (0.44)	1.65*** (0.31)	1.65*** (0.28)
Transitive triplets	0.71*** (0.16)	0.74*** (0.18)	0.68*** (0.14)	0.59*** (0.09)
3-cycles	-0.12 (0.25)	0.00 (0.30)	-0.28 (0.21)	-0.18 (0.16)

Effect name	Model 1	Model 2	Model 3	Model 4
Network decision: selection processes	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**	**beta (s.e.)**
In-degree - popularity	0.21*** (0.06)	0.19** (0.06)	0.14*** (0.04)	0.12*** (0.03)
In-in degree^(1/2) assortativity	-0.44* (0.21)	-0.44† (0.24)	-0.33† (0.18)	-0.22* (0.09)
Grade similarity	0.89*** (0.25)	0.78*** (0.22)	0.93*** (0.20)	0.84*** (0.18)
Smoking similarity	0.06 (0.16)	0.20 (0.17)	0.02 (0.12)	0.06 (0.09)
Limited nomination ego	-0.97*** (0.28)	-1.14*** (0.31)	-0.93*** (0.26)	-0.88*** (0.19)

† Two-sided p<0.1;

* Two-sided p<0.05;

** Two-sided p<0.01;

*** Two-sided p<0.001

References

Agneessens, Filip; Wittek, Rafael. Social capital and employee well-being: disentangling intrapersonal and interpersonal selection and influence mechanisms. Revue francaise de sociologie. 2008

Alexander, Cheryl; Piazza, Marina; Mekos, Debra; Valente, Thomas. Peers, Schools, and Adolescent Cigarette Smoking. Journal of Adolescent Health. 2001; 29:22–30. [PubMed: 11429302]

Anderson, Brigham S.; Butts, Carter T.; Carley, Kathleen M. The Interaction of Size and Density with Graph-level Indices. Social Networks. 1999; 21:239–267.

Baerveldt, Chris; Volker, Beate; Van Rossem, Ronan. Revisiting selection and influence an inquiry into the friendship networks of high school students and their association with delinquency. Canadian Journal of Criminology and Criminal Justice. 2008

Berger, Christian; Dijkstra, Jan Kornelis. Competition, Envy, or Snobbism? How Popularity and Friendships Shape Antipathy Networks of Adolescents. Journal of Research on Adolescence. 2013

Borgatti, Stephen P.; Everett, Martin G. A graph-theoretic perspective on centrality. Social networks. 2006; 28:466–484.

Burk, William J.; Kerr, Margaret; Stattin, Håkan. The co-evolution of early adolescent friendship networks, school involvement and delinquent behaviors. Revue francaise de sociologie. 2008

Burk, William J.; Steglich, Christian EG.; Snijders, Tom AB. Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. International Journal of Behavioral Development. 2007; 31:397–404.

Butts, Carter T. Exact bounds for degree centralization. Social Networks. 2006; 28:283–296.

Butts, Carter T. A relational event framework for social action. Sociological Methodology. 2008; 38:155–200.

Cheadle, Jacob E.; Schwadel, Philip. The ‘friendship dynamics of religion,’ or the ‘religious dynamics of friendship’? A social network analysis of adolescents who attend small schools. Social Science Research. 2012; 41:1198–1212. [PubMed: 23017927]

Cheadle, Jacob E.; Goosby, Bridget J. The Small School Friendship Dynamics of Adolescent Depressive Symptoms. Soc Ment Health. 2012; 2:99–119. [PubMed: 23599906]

Christakis, Nicholas A.; Fowler, James H. The Spread of Obesity in a Large Social Network over 32 Years. The New England Journal of Medicine. 2007; 357:370–379. [PubMed: 17652652]

de Cuyper, Ruben; Weerman, Frank; Ruiter, Stijen. The co-evolution of friendship and relationships delinquent behavior among Dutch youth. People & Society. 2009

de Klepper, Maurits; Sleebos, Ed; van de Bunt, Gerhard; Agneessens, Filip. Similarity in friendship networks: Selection or influence? The effect of constraining contexts and non-visible individual attributes. Social Networks. 2010; 32:82–90.

de la Haye K, Green HD, Kennedy DP, Pollard MS, Tucker JS. Selection and Influence Mechanisms Associated with Marijuana Initiation and Use in Adolescent Friendship Networks. Journal of Research on Adolescence. 2013

Dijkstra, Jan Kornelis; Lindenberg, Siegwart; Veenstra, Rene; Steglich, Christian; Isaacs, Jenny; Card, Noel A.; Hodges, Ernest VE. Influence and selection processes in weapon carrying during adolescence The roles of status, aggression, and vulnerability. Criminology. 2010

Ennett, Susan T.; Bauman, Karl E. Peer group structure and adolescent cigarette smoking: A social network analysis. Journal of Health and Social Behavior. 1993; 34:226–236. [PubMed: 7989667]

Faust, Katherine. Very local structure in social networks. Sociological Methodology. 2007; 37:209–256.

Flashman, Jennifer. Academic Achievement and Its Impact on Friend Dynamics. Sociology of Education. 2012; 85:61–80.

Gile, Krista J.; Handcock, Mark S. CSSS Working Paper 66. Seattle, WA: Center for Statistics and the Social Sciences; 2006. Model-based assessment of the impact of missing data on inference for networks; p. 18

Goodreau, Steven M.; Handcock, Mark S.; Hunter, David R.; Butts, Carter T.; Morris, Martina. A statnet Tutorial. Journal of Statistical Software. 2008; 24

Handcock, Mark S. Missing Data for of Social Networks. Center for Statistics and the Social Sciences, University of Washington; 2002.

Handcock, Mark S.; Hunter, David R.; Butts, Carter T.; Goodreau, Steven M.; Morris, Martina. Statnet: software tools for the representation, visualization, analysis and simulation of network data. Journal of Statistical Software. 2008; 24:1548–7660.

Harris, Kathleen Mullan. The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007-2009 [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill; 2009.

Huisman, Mark; Snijders, Tom AB. Statistical Analysis of Longitudinal Network Data With Changing Composition. Sociological Methods & Research. 2003; 32:253–287.

Huisman, Mark; Steglich, Christian. Treatment of non-response in longitudinal network studies. Social Networks. 2008; 30:297–308.

Hunter, David R.; Handcock, Mark S.; Butts, Carter T.; Goodreau, Steven M.; Morris, Martina. Ergm: a package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software. 2008; 24

Lerner, Jürgen; Indlekofer, Natalie; Nick, Bobo; Brandes, Ulrik. Conditional Independence in Dynamic Networks. Journal of Mathematical Psychology. 2013; 57:S 275–283.

Light, John M.; Dishion, Thomas J. Early adolescent antisocial behavior and peer rejection: a dynamic test of a developmental process. New Dir Child Adolesc Dev. 2007:77–89. [PubMed: 18157876]

Light, John M.; Greenan, Charlotte C.; Rusby, Julie C.; Nies, Kimberley M.; Snijders, Tom AB. Onset to First Alcohol Use in Early Adolescence: A Network Diffusion Model. Journal of Research on Adolescence. 2013

Logis, Handrea A.; Rodkin, Philip C.; Gest, Scott D.; Ahn, Hai-Jeong. Popularity as an Organizing Factor of Preadolescent Friendship Networks: Beyond Prosocial and Aggressive Behavior. Journal of Research on Adolescence. 2013

Mathys, Cecile; Burk, William J.; Cillessen, Antonius HN. Popularity as a Moderator of Peer Selection and Socialization of Adolescent Alcohol, Marijuana, and Tobacco Use. Journal of Research on Adolescence. 2013

Mouw, Ted; Entwisle, Barbara. Residential Segregation and Interracial Friendship in Schools. American Journal of Sociology. 2006; 112:394–441.

Mundt, Marlon P.; Mercken, Liesbeth; Zakletskaia, Larissa. Peer selection and influence effects on adolescent alcohol use: a stochastic actor-based model. BMC Pediatrics. 2012

Ojanen, Tiina; Sijtsema, Jelle J.; Rambaran, Ashwin J. Social Goals and Adolescent Friendships: Social Selection, Deselection, and Influence. Journal of Research on Adolescence. 2013; 23:550–562.

Osgood, D Wayne; Ragan, Daniel T.; Wallace, Lacey; Gest, Scott D.; Fienberg, Mark E.; Moody, James. Peers and the Emergence of Alcohol Use Influence and Selection Processes in Adolescent Friendship Networks. Journal of Research on Adolescence. 2013

Pearson, Michael; Steglich, Christian; Snijders, Tom AB. Homophily and assimilation among sport-active adolescent substance users. Connections. 2006

Radloff, Lenore Sawyer. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. Applied Psychological Measurement. 1977; 1:385–401.

Rambaran, Ashwin J.; Dijkstra, Jan Kornelis; Stark, Tobias H. Status-based influence processes: The role of norm salience in contagion of adolescent risk attitudes. Journal of Research on Adolescence. 2013

Ripley, Ruth M.; Snijders, Tom AB.; Boda, Z.; Vörös, A.; Preciado, Paulina. Manual for Siena version 4.0 (version June 26, 2014). University of Oxford, Department of Statistics, Nueld College; 2014. Available online at

Rook, Karen S.; August, Kristin J.; Sorkin, Dara H. Social network functions and health. In: Contrada, R.; Baum, A., editors. Handbook of stress science: Biology, psychology, and health. New York: Springer; 2011. p. 123-135.

Rubin, Donald B. Inference and Missing Data. Biometrika. 1976; 63:581–592.

Schaefer, David R.; adams, jimi; Haas, Steven A. Social Networks and Smoking: Exploring the Effects of Influence and Smoker Popularity through Simulations. Health Education & Behavior. 2013; 40:24–32. [PubMed: 22491009]

Schaefer, David R.; Haas, Steven A.; Bishop, Nicholas J. A dynamic model of US adolescents' smoking and friendship networks. Am J Public Health. 2012; 102:e12, 8. [PubMed: 22515861]

Schafer, Joseph L. Analysis of Incomplete Multivariate Data. New York: Chapman and Hall; 1997.

Shakya, Holly B.; Christakis, Nicholas A.; Fowler, James H. Parental influence on substance use in adolescent social networks. Arch Pediatr Adolesc Med. 2012; 166:1132–9. [PubMed: 23045157]

Shoham, David A.; Tong, Liping; Lamberson, Peter J.; Auchincloss, Amy H.; Zhang, Jun; Christakis, Nicholas A.; Dugas, Lara; Kaufman, Jay S.; Cooper, Richard S.; Luke, Amy. An actor-based model of social network influence on adolescent body size, screen time, and playing sports. PLoS One. 2012; 7:e39795. [PubMed: 22768124]

Simpkins, Sandra D.; Schaefer, David R.; Price, Chara D.; Vest, Andrea E. Adolescent Friendships, BMI, and Physical Activity: Untangling Selection and Influence through Longitudinal Social Network Analysis. Journal of Research on Adolescence. 2013

Snijders, Tom AB. The Statistical Evaluation of Social Network Dynamics. Sociological Methodology. 2001

Snijders, Tom AB.; Pattison, Philippa E.; Robins, Garry L.; Handcock, Mark S. New Specifications for Exponential Random Graph Models. Sociological Methodology. 2006; 36:99–153.

Snijders, Tom AB. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. Journal of Social Structure. 2002; 3:2–40.

Snijders, Tom AB.; van de Bunt, Gerhard G.; Steglich, Christian EG. Introduction to Stochastic Actor-Based Models for Network Dynamics. 2010

Steglich, Christian; Snijders, Tom AB.; Pearson, Michael. Dynamic networks and behavior: Separating selection from influence. Sociological Methodology. 2010

Valente, Thomas W.; Unger, Jennifer B.; Unger, Jennifer B.; Johnson, C Anderson. Do Popular Students Smoke? The Association between Popularity and Smoking among Middle School Students. Journal of Adolescent Health. 2005; 37:323–329. [PubMed: 16182143]

Van Workum, Nicole; Scholte, Ron HJ.; Cillessen, Antonius HN.; Lodder, Gerine MA.; Giletta, Matteo. Selection, deselection, and socialization processes of happiness in adolescent friendship networks. Journal of Research on Adolescence. 2013

Veenstra, Rene; Steglich, Christian. Actor-based model for network and behavior dynamics. 2011

Wasserman, Stanley; Robins, Garry. An introduction to random graphs, dependence graphs, and p*. In: Carrington, PJ.; Scott, J.; Wasserman, S., editors. Models and methods in social network analysis, Structural analysis in the social sciences. Cambridge, UK: Cambridge University Press; 2005. p. 148-161.

Biographies

John R. Hipp is a Professor in the departments of Criminology, Law and Society, and Sociology, at the University of California Irvine. His research interests focus on how neighborhoods change over time, how that change both affects and is affected by neighborhood crime, and the role networks and institutions play in that change. He approaches these questions using quantitative methods as well as social network analysis. He has published substantive work in such journals as *American Sociological Review, Criminology*, *Social Forces*, *Social Problems*, *Mobilization*, *City & Community*, *Urban Studies* and *Journal of Urban Affairs*. He has published methodological work in such journals as *Sociological Methodology*, *Psychological Methods*, and *Structural Equation Modeling*.

Cheng Wang is a post-doctoral scholar in the program of Public Health at the University of California, Irvine. He is good at using social network analysis and machine learning techniques to explore big data from cellphone networks and other social media such as Twitter, Facebook, and Amazon over multiple time periods. He is now working on the relationship between friendship networks and substance use behaviors of U.S. adolescents with stochastic actor-based models and Exponential Random Graph Models.

Carter T. Butts is a Professor in the departments of Sociology, Statistics, and EECS, and the Institute for Mathematical Behavioral Sciences at the University of California, Irvine. His research involves the application of mathematical and computational techniques to theoretical and methodological problems within the areas of social network analysis, mathematical sociology, quantitative methodology, and human judgment and decision making. His work has appeared in a range of journals, including *Science*, *Sociological Methodology*, the *Journal of Mathematical Sociology*, *Social Networks*, and *Computational and Mathematical Organization Theory*.

Rupa Jose is a doctoral candidate in the department of Psychology and Social Behavior, at the University of California Irvine. She studies the significance of social relationships and institutional structures at determining health, desistence from abuse, and delinquency engagement. Since 2012, she has served as a graduate affiliate of the Networks, Computation, and Social Dynamics (NCSD) team.

Cynthia M. Lakon is an Assistant Professor in the Program in Public Health at the University of California Irvine and Principal Investigator of the grant which funds this study (Cascades of Network Structure & Function: Pathways to Adolescent Substance Use. National Institute on Drug Abuse (NIDA) Health, grant #1 R21 DA031152-01A1). Her research focuses on adolescent social networks and drug use behaviors. She has published substantive work in the *American Journal of Public Health*, *Social Science and Medicine*, *and Health and Place*.

Highlights

- The consequences of missing longitudinal network data are under-studied
- We estimate SAB models using four different missing data strategies
- The estimated parameters often differ considerably across missing data techniques
- The influence and selection effects can differ over missing data techniques

Table 1

Pattern of data missingness, and inclusion in various missing data methods

Missing pattern			In sample?			
Wave 1	**Wave 2**	**Wave 3**	**Method 1**	**Method 2**	**Method 3**	**Method4**
X	X	X	Yes	Yes	Yes	Yes
X	O	X	No	Yes	Yes	Yes
X	X	O	No	Yes	Yes	Yes
X	O	O	No	Yes	Yes	Yes
O	O	X	No	No	W1=0	W1=impute
O	X	O	No	No	W1=0	W1=impute
O	X	X	No	No	W1=0	W1=impute

Note: In waves, X indicates present and O indicates missing observation. “Yes” or “No” indicates whether the person would be present in the sample in the given technique. W1=0 indicates that network ties are set to 0 in wave 1; W1=impute indicates that network ties are imputed to a plausible value based on an ERG model for wave 1.

Table 2

Comparing coefficients across strategies for handling missing data (averaged estimates from six different networks)

	Strategy 1 vs. strategy 4		Strategy 2 vs. strategy 4		Strategy 3 vs. strategy 4	
	Bias	**Error**	**Bias**	**Error**	**Bias**	**Error**
Constant friendship rate (period 1)	-49.9%	49.9%	-34.6%	34.6%	-20.4%	20.4%
Constant friendship rate (period 2)	-40.4%	40.4%	-22.9%	22.9%	-8.2%	8.2%
Out-degree (density)	-0.8%	4.9%	-2.5%	5.3%	-6.8%	6.8%
Reciprocity	6.3%	8.2%	-1.1%	5.1%	-4.7%	5.0%
Transitive triplets	68.2%	68.2%	31.4%	31.4%	28.9%	28.9%
3-cycles	31.8%	42.6%	34.3%	67.7%	57.8%	57.8%
In-degree - popularity	91.2%	91.2%	47.2%	47.2%	30.1%	30.1%
In-in degreeˆ(1/2) assortativity	169.2%	169.2%	95.0%	95.0%	68.7%	68.7%
Race similarity	28.7%	28.7%	10.0%	10.0%	-0.3%	0.3%
Grade similarity	23.9%	23.9%	15.0%	17.5%	12.5%	12.5%
Smoking alter	-0.2%	35.9%	6.8%	17.7%	20.6%	39.2%
Smoking ego	(b)	(b)	(b)	(b)	(b)	(b)
Smoking similarity (a)	9.2%	22.7%	70.5%	82.1%	6.8%	35.8%
Limited nomination ego	14.3%	16.6%	13.0%	16.9%	6.0%	9.0%
Behavior decision: influence processes						
Rate smoking behavior (period 1)	-71.7%	71.7%	-51.4%	51.4%	3.0%	3.2%
Rate smoking behavior (period 2)	-65.7%	65.7%	-7.0%	16.2%	4.9%	4.9%
Smoking behavior linear shape	-0.1%	8.5%	0.1%	2.2%	0.8%	1.2%
Smoking behavior quadratic shape	2.1%	6.9%	1.3%	2.5%	-0.6%	1.0%
Smoking behavior in-degree	(b)	(b)	(b)	(b)	(b)	(b)
Smoking behavior similarity	-84.5%	84.5%	-41.6%	41.6%	-3.3%	7.1%
Effect from gender (female=1) (c)	38.9%	38.9%	-5.1%	5.1%	-0.5%	0.5%
Effect from Black	5.2%	5.2%	31.4%	31.4%	1.8%	1.8%
Effect from depression	228.0%	228.0%	1.5%	17.1%	-2.0%	2.0%
Effect from home smoking environment	4.5%	26.1%	-5.1%	20.8%	0.3%	1.3%
Effect from parental support (d)	57.1%	57.1%	4.0%	9.8%	2.4%	6.1%

Note: (a) excluding school 002 from calculation given extreme value. (b) excluding values given parameters were estimated close to zero. (c) only using school 077, since other schools had very small estimates. (d) only including schools 077 and 008 since other schools had very small estimates.

Table 3

Comparing determinations of statistical significance across strategies for handling missing data (estimates from six different networks)

	Strategy 1 vs. strategy 4		Strategy 2 vs. strategy 4		Strategy 3 vs. strategy 4	
	sig 1, not 4	sig 4, not 1	sig 2, not 4	sig 4, not 2	sig 3, not 4	sig 4, not 3
Constant friendship rate (period 1)	-	-	-	-	-	-
Constant friendship rate (period 2)	-	-	-	-	-	-
Out-degree (density)	-	-	-	-	-	-
Reciprocity	-	-	-	-	-	-
Transitive triplets	-	-	-	-	-	-
3-cycles	-	-	-	-	-	-
In-degree - popularity	-	-	-	-	-	-
In-in degreeˆ(1/2) assortativity	1	-	1	2	-	2
Race similarity	-	-	-	-	-	-
Grade similarity	-	-	-	-	-	-
Smoking alter	-	3	-	1	-	1
Smoking ego	1	1	-	1	-	1
Smoking similarity (a)	1	-	1	1	-	2
Limited nomination ego	-	-	-	-	-	-
Behavior decision: influence processes	-	-	-	-	-	-
Rate smoking behavior (period 1)	-	-	-	-	-	-
Rate smoking behavior (period 2)	-	-	-	-	-	-
Smoking behavior linear shape	-	-	-	-	-	-
Smoking behavior quadratic shape	-	-	-	-	-	-
Smoking behavior in-degree	-	-	-	-	-	-
Smoking behavior similarity	-	-	-	-	-	-
Effect from gender (female=1) (c)	-	-	-	-	-	-
Effect from Black	-	-	-	-	-	-
Effect from depression	1	2	-	1	-	1
Effect from home smoking environment	-	1	-	1	-	1
Effect from parental support (d)	-	-	-	-	1	1

Note: comparing levels of significance across imputation techniques. Using $p < .05$ as the significance criterion for all decisions. Cells indicate the number of times a strategy produced statistically significant results whereas the compared strategy did not. For example, the first column shows the number of models in which strategy 1 found a statistically significant effect for a particular parameter whereas strategy 4 did not. The second column shows the number of models in which strategy 4 found a statistically significant effect for a parameter whereas strategy 1 did not.