doi.org/10.26398/IJAS.0030-002

MULTIPLE IMPUTATION FOR LONGITUDINAL NETWORK DATA

Robert W. Krause, Mark Huisman [1]
Department of Sociology/ICS, University of Groningen, Groningen, The Netherlands

Tom A.B. Snijders
Department of Sociology/ICS, University of Groningen, Groningen, The Netherlands
Nuffield College, University of Oxford, Oxford, UK

Abstract *Missing data on network ties are a fundamental problem for network analysis. The biases induced by missing edge data are widely acknowledged. In this paper, we present a new method with two variants to handle missing data due to actor non-response in the framework of stochastic actor-oriented models (SAOMs). The proposed method imputes missing tie variables in the first wave either by using a Bayesian exponential random graph model (BERGMs) or a stationary SAOM and imputes missing tie variables in later waves utilizing a SAOM. The proposed method is compared to the standard SAOM missing data treatment as well as recently proposed methods. The multiple imputation procedure provided more reliable point estimates than the default treatment. The results have relevant implications for the analysis of network dynamics under missing data.*

Keywords*: Missing data, Multiple imputation, Longitudinal network data, Stochastic actor-oriented models, Bayesian exponential random graph models.*

1. INTRODUCTION

Missing data have always been a problem for empirical social scientists. It reduces power and can induce biases into the data analysis. While missing data constitute problems for all social science research, the field of longitudinal network research is affected on multiple fronts. On the one hand, longitudinal research is likely to produce more missing data, because the same people are followed over time, making dropout more likely. On the other hand, network questionnaires are complex and often ask sensitive questions from the respondents, thus increasing the poten tial for missing data. Additionally, the strong dependence between actors makes missing data. Missing tie variables do not only

[1] Corresponding author: Robert W. Krause, email: r.w.krause@rug.nl

mean less information about the sending actors, but also less information about all receiving actors in the network, given that missing actors could have nominated any given number of the observed actors. Research into missing data and missing data treatments in networks is an ongoing field of research (e.g., de la Haye et al., 2017; Huisman and Krause, 2017; Smith et al., 2017). One model family used to analyze network dynamics in a longitudinal setting are stochastic actor oriented models (SAOMs). In this paper, we present a multiple imputation procedure for missing data in longitudinal network research in the framework of SAOMs. It extends previous work on multiple imputation for longitudinal networks that focused on the imputation of missing data in the first observation point (first wave) of the study (Hipp et al., 2015). The proposed procedure is an imputation method applicable to missing data at all waves.

The paper is organized as follows. In Section 2, we describe the two network model families relevant for this paper, the stochastic actor-oriented model and the exponential random graph model (ERGM). In Section 3, we detail the non-response problem and its specifics for missing data in networks. Section 4 presents the proposed multiple imputation procedure for longitudinal network data, and a (simulated) example application is presented in Section 5, including comparisons to completely observed networks and benchmark procedures. We end the paper with a discussion of the findings and according recommendations.

2. STATISTICAL MODELS FOR NETWORK ANALYSES

Social network analysis is the study of relational data between social actors using statistical models. A core issue in analyzing these relations is their embeddedness in the larger network structure. Any analysis model must take these dependencies into account. Two commonly used model families for analyzing networks are stochastic actor-oriented models and exponential random graphs.

2.1. STOCHASTIC ACTOR-ORIENTED MODELS

Researchers studying the co-evolution of social relations (e.g., friendships) and behaviors (or attitudes) encounter the problem that usually the social relations and the behaviors are only observed at discrete points in time. It is unrealistic to assume that the changes made in a friendship network observed M times all happened at once between observations. It is more likely that the changes between the network states from $m-1$ to m are the result of a dynamic process consisting of a sequence of small changes.

A common model to analyze these network dynamics is the stochastic actor-oriented model introduced by Snijders (1996, 2001, 2005). The SAOM assumes that each actor has control of its outgoing ties and is aware of the ties between other actors. The SAOM models the change between the networks as a series of mini steps, each constituting the creation or deletion of a tie, or no change. At each step, a certain actor i, stochastically chosen with frequencies determined by a rate function, evaluates her choice set based on the current state of the network. Usually the rate function is constant for all actors, meaning actors are chosen at random, however, the rate function can be estimated (or set) to incorporate endogenous or exogenous effects.

Let x denote the $n \times n$ adjacency matrix where n is the number of actors, with $x_{ij} = 1$ when there is a directed tie from actor i to actor j and $x_{ij} = 0$ when there is no tie[2]. Self nominations are not allowed ($x_{ii} = 0$). The chosen actor i can either create a tie to an unconnected actor, drop a tie to a connected actor, or do nothing. The probability of each possible actor decision is determined by an objective function, in which actor-specific network statistics and covariates s_{ki} are weighted with parameters of the network evolution θ_k, given the current state of the network x:

$$f_i(\theta, x) = \sum_k \theta_k s_{ki}(x). \tag{1}$$

The network statistics $s_{ki}(x)$ can be subgraph counts (or non-linear transformations thereof) in the network neighborhood of the focal actor i (e.g., reciprocity, outdegree, indegree) or functions of the attributes of the actors sending or receiving the ties, and are always calculated from the network at the current mini step. This allows the model to capture the dynamic process. Two problems arise that make it impossible to directly calculate the likelihoods or expected values of parameters. First, the true sequence of these mini steps is unobserved[3]. Second, the possible states of the network are far too numerous; A binary network of only 30 actors already has $2^{30^2-30} = 7.9 \times 10^{261}$ possible states. Therefore SAOMs are estimated using a simulation approach (hence the name SIENA – Simulation Investigation of Empirical Network Analysis – for the software to estimate SAOMs; The contributed package to the statistical system R is RSiena, Ripley et al., 2017).

[2] Throughout the paper we focus on directed networks. All models discussed also apply to undirected networks.

[3] If this sequence of changes is observed, it is recommended to use relational event models (REM; Butts, 2008) or their actor oriented counter part, dynamic network actor models (dyNAM; Stadtfeld and Block, 2017; Stadtfelt et al., 2017).

Model estimation is typically done by the method of moments, that is, determining parameters such that for a selected set of statistics the expected values are equal to the observed values. The algorithm is split into three phases. Phase 1 determines the sensitivity of the parameters to the given statistics and provides initial estimates for the parameters. Phase 2 estimates the parameters iteratively by simulating the mini steps of the entire process many times, each time calculating the statistics used for the method of moments, and updating the parameters according to Robbins-Monro steps (see Snijders, 2001). Phase 3 takes the resulting parameter estimates to simulate multiple runs of network evolutions (normally at least 1000) to estimate the covariance matrix of the model parameters. This process involves thousands of repeated simulations of the whole dynamic network process, and each of these simulations consists of several hundred or more mini steps. The simulated networks in Phase 3 are used to test the convergence of the model, calculate standard errors for the parameters and the final networks can be used for goodness-of-fit (GoF) testing.

Four different simulation-based estimation methods are implemented in the SIENA software: Method of moments (MoM), generalized method of moments (GMoM), maximum likelihood (ML), and Bayesian estimation (Amati et al., 2015; Koskinen and Snijders, 2007; Ripley et al., 2017; Snijders et al., 2010a). Especially the MoM and the ML estimation algorithm have appealing features that will be utilized in the proposed imputation method. One important difference between MoM and ML estimation lies in how they simulate network evolution trajectories, both in phase 2 and phase 3 of their respective estimation processes. Network trajectories under MoM are simulated conditional on the observed network at wave $m-1$ and on the estimated parameters. In contrast, ML simulations are conditional on the observed network at wave $m-1$, the observed network at wave m and the estimated parameters. Thus, MoM simulations provide a distribution of networks at the end of the simulation, while ML simulations always end in the observed network at wave m. (For a more detailed introduction to SAOMs see Snijders, 2017a; For an introduction to applying SAOMs see Snijders et al., 2010b, and Steglich et al., 2010.)

2.2. STATIONARY STOCHASTIC ACTOR-ORIENTED MODELS

Although SAOMs are mostly used for investigating dynamic change processes over time, they can also be applied to cross-sectional network data (Snijders and Steglich, 2015). While longitudinal SAOMs model the changes in network structure, stationary SAOMs assume that the network structure, although changing, is in a stochastically stable state. This means that it is assumed that the observed network is in a short-term dynamic equilibrium and thus the statistics $s_{ki}(x)$ will have a stationary distribution. The stationary models can be estimated by using the observed network as both starting and end network for the stationary distribution (reflecting that the network statistics remain constant) and fixing the rate parameter to a large value (say 50). The rate parameter cannot be estimated in the stationary SAOM, as it reflects the rate of change and the stationary SAOM assumes no change. However, fixing a large value for the rate function allows the model to simulate network trajectories and estimate the parameters in the objective function such that the observed network statistics remain stable.

2.3. EXPONENTIAL RANDOM GRAPH MODELS

The most common family used to analyze cross-sectional network data is the family of exponential random graph models (ERGMs; Lusher et al., 2013). ERGMs model the observed network as a function of its statistics (mainly counts of subgraphs, e.g., the number of reciprocated ties or the number of transitive triplets). Basic to the ERGM is a linear predictor quite similar to the objective function of the SAOM:

$$\sum_k \theta_k s_k(x), \tag{2}$$

with the key difference that in the SAOM the objective function is actor specific, as can be seen in the actor index i in (1). SAOMs are, as the name states, actor-oriented models, while ERGMs are tie-oriented models. While SAOM parameters focus on the decision of social actors given their network neighborhood, ERGM parameters focus on the presence (or absence) of a tie, given all other ties in the network. For a more detailed comparison between SAOMs and ERGMs see Block et al. (2016).

3. MISSING DATA

3.1. MISSING DATA MECHANISMS

Missing data mechanisms describe the underlying processes for the data to be missing, using the distribution of missingness. Following the framework defined by Rubin (1976), there are three types of missing data mechanisms. Data are missing completely at random (MCAR) if each individual tie variable (or actor) is missing independent of observed and missing data. Data are missing at random (MAR) if the probability to be missing is independent of the missing tie (actor) itself, but is related to other observed variables (e.g., males are less likely to fill out the network part of the survey). These two cases are often summarized as ignorable missing data in the survey research setting, because given proper missing data techniques are applied, they will yield no bias on a resulting analysis. Lastly, data are missing not at random (MNAR) if the missingness is dependent on the (unknown) missing value itself.

3.2. MISSING DATA TYPES

It is not only important to inspect missing data mechanisms, but also the patterns of missing data showing the spread over the data set. Usually, two types of patterns are distinguished: item (or tie) non-response and unit (or actor) non-response (Huisman and Steglich, 2008). Item non-response occurs when a participant is only observed on some items, but not on all. In network research this means that only some ties (outgoing or incoming) are not observed for an actor. Unit non-response occurs when a complete case is missing. In the setting of network research this means that all outgoing ties of the participant are missing. Incoming ties however will still be observed. In some cases unit non-response of an actor will not only lead to missing outgoing ties, but will remove the actor completely from the study, leading to missing incoming ties as well (Borgatti and Molina, 2003).

A special case of non-response in longitudinal research is wave non-response (Huisman and Steglich, 2008). In this case, data are only available for some actors for some waves of the data collection, but not for all. This study will only focus on wave non-response, which is illustrated in Section 5 with networks collected over three time points, including (completely) observed covariates. The findings, however, can be applied to the case of item non-response, as item non-response is less severe and retains more information per actor than wave or unit non-response. For ease of the illustration, in this paper, all data are considered missing completely at random (MCAR).

3.3. MISSING DATA IN LONGITUDINAL NETWORK DATA

For estimating SAOMs, it is important to distinguish between missing data in the first wave and missing data in following waves, because the first wave is the starting point for the simulation and is treated as given by the model. Therefore, it is necessary to impute the missing data in the first wave to provide a starting point for the simulations.

Handling missingness in consecutive waves differs depending on the estimation procedure used in the SIENA software (Ripley et al., 2017). For the method of moments (MoM) procedure, the model-based hybrid imputation procedure described by Huisman and Steglich (2008) is used to handle missing tie variables. It is hybrid because it uses imputation for the simulations but then restricts the use of the imputed values for the estimating equations. For the first wave, it uses the simple method of imputing no-ties (zeros) for missing ties. Social networks are usually sparse and without taking any other information into account a no-tie is the most likely guess for each missing cell. Missing ties in consecutive waves are imputed by last value carried forward (Lepkowski, 1989). In the calculation of the target statistics used for parameter estimation, missing tie variables are excluded. Therefore, the imputations have no direct effect on parameter estimation, although they do have effect on the simulations. Earlier work has shown that for small amounts of missing actors (up to 20%), this method provides only small biases in the parameter estimates under MCAR, MAR and MNAR, and is superior to other simple imputation methods (Huisman and Steglich, 2008).

If maximum likelihood (ML) estimation is chosen, missing data at the end of a period are treated in a model-based way. The procedure is given in Snijders (2017b). As described before, the chain of mini steps between two waves simulated in the ML procedure is conditional on the observed data at both time points, $m-1$ and m. If data for time $m-1$ are complete, this conditioning determines the probability distribution of any missings at time m. If data for time $m-1$ are incomplete, then the extra information inserted is the prior distribution for the missing tie variables, and this assumes independent binary variables with the observed density (among observed variables) as the tie probability. Given all observed variables at times $m-1$ and m and this prior, the chains are simulated and this implies the stochastic model-based imputation of the missing tie variables at both waves. The simulated chains are used for parameter estimation. If there are no missing data at wave $m-1$, the imputed values for missing ties at wave m are draws from their conditional distribution given all observed data. If the missing data are MAR and the estimation model is realistic, this does not introduce any additional biases in parameter estimation.

It should be noted that for $M \geq 3$ waves, in the ML estimation procedure implemented in RSiena, all $M-1$ periods in between waves are treated separately. For example, when analyzing $M = 3$ waves, missing ties in wave 2 are treated in a model-based way only for the first period (wave 1 to wave 2), but are imputed with the observed density of the network for the second period (wave 2 to wave 3).

In the case of wave non-response this is a limitation, and was only chosen to keep the algorithm tractable and for purposes of parallelization. Moreover, in the ML procedure, missing data are not imputed in the traditional sense. Neither are imputed values returned, nor are imputed values directly used for parameter estimation in consecutive periods.

An alternative approach for handling missing data in SAOMs was proposed by Hipp et al. (2015). They propose an imputation procedure using ERGMs to impute the first observation of the network. First, an ERGM is estimated on the network, after which the estimated parameters are used to simulate the missing ties, while keeping all observed ties fixed. This provides realistic starting points that can be used in both the simulation phase and estimation phase of the SAOM estimation procedure. Although the procedure was evaluated without reference to a complete data set (and generating model) and only assessed by comparing different missing data handling methods, it is expected that the method outperforms the default procedures discussed above, provided that the imputations are performed with a well-fitting model. This is because the procedure utilizes far more information for imputation than the standard procedures, imputing the missing ties in wave 1 conditional on the observed network and covariates at wave 1. The authors give suggestions how to use the procedure for multiple imputations.

Another strategy for dealing with missing data in network studies with multiple periods ($M > 2$) was proposed by de la Haye et al. (2017), called *inclusive sampling*. The strategy involves forming subgroups of the data for each period. Each subgroup only includes actors that are fully observed at the start and end of the respective period. Although this procedure disregards some available information, it was specifically designed to increase the likelihood of the SAOM to converge.

4. MULTIPLE IMPUTATION

In this paper, we present a multiple imputation procedure for longitudinal network data. It allows the user to analyze all available data and not only completely observed dyads, which results in increased power for the analysis. Multiple imputation has the advantage over single imputation that it takes into account the increased variability of parameter estimates due to imputation (see Huisman and Krause, 2017, for an overview of imputation methods for network data). The proposed method uses model-based imputation for the first wave, similar to Hipp et al. (2015), but takes some further steps.

First, it allows imputation of missing ties both in the first and later waves. The imputations for later waves are obtained using the ML simulation method for SAOMs. This makes it possible to impute the missing tie variables for a given wave by draws from their conditional distribution, given the observed data for the preceding and the current wave. Second, two options for imputing missing data in the first wave are proposed, which both use data from the first and second wave. The first option is an adjustment of the procedure of Hipp et al. (2015), using Bayesian ERGMs to impute the first wave, rather than ERGMs (Caimo and Friel, 2011, 2013; Koskinen et al., 2010; 2013). The second option is to use a stationary SAOM to impute the first wave.

4.1. MULTIPLE IMPUTATION: GENERAL THEORY

Multiple stochastic imputation consists of performing the following steps (e.g., see van Buuren, 2012):

(1) Specify an imputation model and obtain starting values for the parameters of the model (often estimated from the observed data). With this model, specify the probability distribution of the missing data, given the observed data, and fill in starting imputations by random draws from this distribution.

(2) Obtain a conditional distribution of the parameters of the imputation model, given the observed and imputed data and estimate (draw) new values for the parameters (needed to generate proper multiple imputations, either by using Bayesian methods and specifying posterior distributions of the parameters, or by using bootstrap methods and re-estimating parameters from the re-sampled data). With these new parameters, impute the missing values by drawing values from the conditional distribution of the missing data, given the observed data and the new parameters.

(3) Repeat step (2) until convergence, and retain D imputed data sets from this procedure, differing only in the imputed values.

(4) Analyze each imputed data set separately with standard (complete-case) techniques and combine the results of the analyses following the procedures outlined by Rubin (1987).

Rubin's rules for combining results include combining parameter estimates and covariances. Let $\hat{\gamma}_d$ denote the dth estimate of the parameter γ and $W_d = \mathrm{cov}(\hat{\gamma}_d \,|\, x_d)$ the (within-imputation) covariance matrix of the parameters of data set x_d. The combined estimate for the parameters is the average of the estimates of the D analyses:

$$\bar{\gamma}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\gamma}_d. \tag{3}$$

Obtaining the proper standard errors is a bit less straightforward. The combined estimate for the standard error needs to take into account the variance within and between imputations. It requires the average within-imputation covariance matrix $\bar{W}_D$ and the between-imputation covariance matrix B_D. The average within-imputation covariance matrix is given by

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d \tag{4}$$

and the between covariance matrix by

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\gamma}_d - \bar{\gamma}_D)(\hat{\gamma}_d - \bar{\gamma}_D)'. \tag{5}$$

The total variability for $\bar{\gamma}_D$ is estimated by

$$T_D = \hat{\text{cov}}(\bar{\gamma}_D) = \bar{W}_D + \left(1 + \frac{1}{D}\right) B_D. \tag{6}$$

The standard errors for the parameters are given by the square roots of the diagonal elements of T_D.

4.2. MULTIPLE IMPUTATION: LONGITUDINAL NETWORK DATA

When applying these general steps to the longitudinal network setting, we have to adjust the steps (1) to (3) to the SAOM. For steps (1) and (2), we distinguish between the first wave and later waves of the longitudinal network data, as the SAOM does not model the network in wave 1. To outline the general procedure, we will first discuss imputation of later waves, $m = 2, \ldots, M$ and then return to the imputation of the first wave.

MULTIPLE IMPUTATION: MISSING DATA IN LATER WAVES

For consecutive waves $m = 2, \ldots, M$, missing ties are imputed wave by wave using the SAOM. Given the data for wave $m-1$, we use the MoM algorithm of the SAOM with default treatment of the missing data to estimate the imputation model (step (1)). In this step, the MoM procedure is preferred over ML estimation because it is faster and, more importantly, gives the opportunity to assess the goodness of fit of the imputation model by using the networks simulated in phase 3 of the SIENA algorithm. Imputation should be performed with a well-fitting model that includes all parameters that will be included in the analysis model. The model is estimated and convergence is assessed for period $m-1$ to m, and the fit of the model is inspected. If deficiencies are found, new effects (parameters) can be added and the model is re-estimated by MoM. This process of specifying, estimating, and inspecting imputation models is repeated until a reasonable model fit is obtained. An alternative is to specify, estimate, and inspect the imputation model by considering all waves together.

Once a fitting imputation model is obtained for period $m-1$ to m, we continue to step (2) of the imputation process and utilize ML simulation to impute the missing ties at wave m, conditional on the complete data for wave $m-1$, the observed data in wave m, and the imputation model estimated in step (1) (if there were any missings in wave $m-1$ they were imputed in earlier steps of the procedure). Repeating this procedure wave by wave results in one complete data set. The sequence of steps is executed D times to provide D imputed data sets. These completed data sets are analyzed separately in step (4) of the process giving D estimates that are combined according to the rules outlined above.

MULTIPLE IMPUTATION: MISSING DATA IN THE FIRST WAVE

Standard SAOM models are models of change that take the first wave as a given starting point and model the change to consecutive waves. This allows us to use the regular SAOM framework to impute missing tie variables in later waves, but, as the first wave is not modeled, prevents us from imputing the missing data in the first wave. Therefore we need to draw our first wave imputations from a different distribution. This goes beyond the SAOM model definition, and requires additionally a specification of the distribution for the first wave.

One way would be to follow a completely model-based procedure, assuming a prior distribution from which the first wave network was drawn together with the SAOM assumptions for the transitions to later waves. This was shown by Koskinen et al. (2015) for longitudinal network data using the longitudinal ERGMs in a Bayesian approach with prior distributions for the parameters. This is also used in the ML procedure implemented in the SIENA software (Snijders, 2017b). The implementation is not completely multivariate, however, because each pair of consecutive waves is handled separately from the other waves, using very simplistic prior distributions. Furthermore, the SIENA software currently does not allow to export the imputations for the first wave. Therefore we follow a different approach, less compelling than a model-based approach would be but easier to perform.

We propose two options for first wave imputations, 1) Bayesian ERGMs and 2) stationary SAOMs. These distributions are chosen because both are models for cross-sectional networks, able to provide multiply imputed data sets conditional on the observed data for the first wave, and able to take into account the next wave as a covariate. Thus, the choice is made out of convenience and flexibility rather than being principled.

ERGMs were already proposed as a possible distribution for this purpose by Hipp et al. (2015). Moreover, ERGMs (especially Bayesian ERGMs) can be estimated reliably under missing data (Koskinen et al., 2010, 2013). For Bayesian ERGMs the imputation of missing tie variables is integrated with the parameter estimation. In our first option for imputing the first wave, we estimate a Bayesian ERGM under missing data as described by Koskinen et al. (2010, 2013) and retain D imputed data sets from the converged model. This is different from the non-Bayesian method proposed by Hipp et al. (2015), where all D imputations are created with the same set of estimated parameters. The non-Bayesian procedure underestimates the between-imputation variance B_W, giving a downward bias to the standard errors.

For this we employ the Bergm package in R (Caimo and Friel, 2014), which we adapted to incorporate missing data treatment as described by Koskinen et al. (2010, 2013). The choice of the imputation model is not trivial, and generally the imputation model should always contain all the parameters that will also be used in the analysis model. However, there is no perfect one-to-one comparability between SAOM and ERGM parameters (Block et al., 2016), which can be seen in equations (1) and (2). Parameters in ERGMs are multiplied with the overall network statistics $s_k(x)$, while parameters in the SAOM relate only to the network neighborhood of the focal actor (the i in $s_{ki}(x)$). However, both model families are generally able to model similar structures[4]. Given the strong longitudinal dependence, it will be essential that the network in wave 2 is used as a dyadic covariate for wave 1. If, however, all actors that are missing in wave 1 are also missing in wave 2, including wave 2 as a dyadic covariate will add little to the imputations.

[4] To identify corresponding parameters refer to the package manuals (the ergm package, Handcock et al., 2017; The RSiena package, Ripley et al., 2017).

In our second option, we impute missing data in the first wave by using a stationary SAOM. Imputation with the stationary SAOM for wave 1 is similar to imputation with the SAOM employed for later waves as described above. Here, we first estimate a stationary SAOM from wave 1 to wave 1 with the rate parameter fixed to a large value (50), and then use ML simulation to impute the missing tie variables conditional on the observed ties in wave 1 and our imputation model (which should include wave 2 as a dyadic covariate).

However, two minor complications with the current implementation of the ML algorithm in RSiena arise. First, the ML algorithm does not provide easy access to the network that is internally imputed for the beginning of the simulation, thus imputation is not possible given the simulated trajectories alone. Changes to this are not trivial. To overcome this minor problem we create a copy of wave 1 in which we impute the missing data with a simple ad hoc procedure (imputing ties randomly with the probability of the observed density of incoming ties in the available data). The imputed copies of wave 1 and the observed wave 1 with the missing data are then used as respective start and end points for the ML simulation. Second, the ML algorithm requires that at least one tie variable changes between the networks. This is not the case here, as wave 1 is used both as start and end point to estimate a stationary SAOM. To fix this minor issue we change one randomly selected observed tie (selected independently across the D imputed data sets) in the copy of wave 1 to a no-tie. This will have minimal impact on the ML simulation, because the simulated network at the end of the trajectory will be equal to the observed network at the end of the period, thus the change in the copy of wave 1 does not lead to changes in the imputed data. The tie change is only necessary when using ML (or Bayesian) estimation or simulation of stationary SAOMs, but not for MoM estimation or simulation.

MULTIPLE IMPUTATION: SUMMARY

To summarize the procedure, the specification of steps (2) and (3) of the general multiple imputation procedure of Section 4.1 is as follows:

(2.1) If the network at wave 1 has any missing ties, estimate a Bayesian ERGM or a stationary SAOM to impute the missing ties. The model specification includes the observed network at wave 2 as a dyadic covariate.

(2.2) For $d = 1, \ldots, D$:

   (a) If the network at wave 1 has missing ties, impute by a random simulation draw from the model estimated in (2.1).

(b) For $m = 2, \ldots, M$:

i. Estimate a SAOM using MoM for the period $m-1$ to m, conditional on the completed network at $m-1$.

ii. Impute the missing ties in wave m using the fitted imputation model in the ML simulation procedure, conditional on the complete data for wave $m-1$, the observed ties in wave m, and the fitted imputation model.

(3) Repeating step (2.2) D times leads to D completed data sets.

The advantage of multiple imputation is that it can give unbiased parameter estimates with correct standard errors (and confidence intervals), even when the number of imputations D is low (Rubin, 1987; van Buuren, 2012). An ongoing question of research, however, is the required number of imputations D to obtain good inference properties (e.g., power or p values). Following the general guidelines for multiple imputation (for non-network data; e.g., see van Buuren, 2012), it is recommended to set D equal to the percentage of missing cases, but at least to 20. Theoretically it is always better to set D as high as computation and data storage do allow.

4.3. ESTIMATING THE IMPUTATION MODEL FOR MULTIPLE WAVES

In the multiple imputation procedure described above, missing ties are imputed wave by wave, where for each period $m-1$ to m a new imputation model is estimated. If the network dynamics are not homogeneous across periods this is the appropriate procedure and separate models need to be estimated for each period from $m-1$ to m. Given that the models can be reliably estimated, estimating a new imputation model for each period ensures that differences in the dynamics between waves are preserved by the imputation model (e.g., friendship dynamics in a school classroom might be different right after the transition from middle to high school, compared to dynamics in 3rd or 4th year of high school).

If the network dynamics differ between periods, new parameters need to be added in later periods to obtain proper model fit. It is advised that the respective parameters are added for all imputation models, including the imputation models for previous waves. We recommend doing so for two reasons. First, the model fit of a previous wave, although satisfying, could still be improved by incorporating these new parameters. Second, the general recommendation is to include at least all parameters in the imputation models that will be used in the analysis model (e.g., van Buuren, 2012; Huisman and Krause, 2017). Comparison of the network dynamics in different waves or combining the results of multiple waves is easiest when the same parameters are used in all analysis models. Therefore these parameters should also be included in the imputation models.

It is, however, possible to estimate the imputation model using all waves and then applying it period by period. Using one model is advised if (1) the network dynamics are homogeneous across periods (which can be tested within the SAOM framework) or (2) the networks are small (e.g., school classes). Small networks are more likely to yield unstable results, especially in the case of missing data, because they provide less information to reliably estimate parameters[5]. This means that for small networks an imputation model not incorporating the information of multiple periods might not be estimable.

4.4. MULTIPLE GROUPS

Stochastic multiple imputation reflects the uncertainty due to missing data and due to imputation (i.e., prediction) of the missing data by combining within and between-imputation variance in Equation (6). A multiple imputation procedure is called *proper* if it also takes into account the uncertainty included in the estimation of the parameters of the imputation model when estimating the between-imputation variance B_D in Equation (5) (Rubin, 1987; van Buuren, 2012). Improper procedures do not fully capture the increased uncertainty, which can deflate B_D. This means that for proper multiple imputations, a new draw from the distribution of the parameters of the imputation model is needed for every imputation. This can be accomplished by Bayesian estimation[6]. The proposed method does not provide proper imputations, because imputations are not drawn from the full posterior distribution. However, the Bayesian ERGMs used in he first wave provide more reliable estimations of B_D than would be achieved by imputations drawn from ERGMs.

Currently, Bayesian estimation for SAOMs is only implemented in the SIENA software for the analysis of multiple groups. This means that analyzing multiple groups or networks has an important advantage for multiple imputation, as it provides more reliable standard error estimates if the Bayesian analysis is used. The procedure for multiple groups is in general similar to the procedure outlined in

---

5 The actual size of the network is of secondary importance. The network change in relation to the parameters is the deciding factor. Small networks tend to provide overall less network change for the parameters.

6 In general, bootstrap procedures are an alternative option to obtain a (sampling) distribution of the parameters, however, for network data bootstrapping is not a feasible procedure because of the strong, inherent dependencies between observations.

Section 4.2, with the exception that in step (2.2) a Bayesian SAOM is estimated, from which the parameters are drawn to generate imputations.

In the single group situation, the drawback of the procedure in Section 4.2 is that the parameters of the imputation model are not drawn from their posterior distribution and the method does not yield proper multiple imputations. For non-network data it has been shown that not taking into account the extra uncertainty due to estimating the parameters of the imputation model does yield fairly similar results to those obtained under proper imputation, given that the sample size is large and that the proportion of missing data is small (Allison, 2001). Although the impact of proper imputations for network analysis has yet to be determined, it is advised to obtain imputations as proper as possible.

4.5. MULTIPLE IMPUTATION VS. LIKELIHOOD-BASED TREATMENT

The model-based missing data treatment implemented in the ML estimation in RSiena and the proposed multiple imputation procedure should provide asymptotically similar results in the situation of one period with only missing data in the second wave. However, in other scenarios (e.g., missing data in multiple waves), multiple imputation should lead to more reliable results, because the imputed values from the proposed multiple imputation procedure are based on more information than the internal imputations in the ML procedure. Further, multiple imputation is generally more flexible than likelihood-based missing data treatment. It allows to incorporate information not included in the analysis model (e.g., additional actor covariates) and can be adapted easily to test the sensitivity of the model to variations of the missing data mechanism. The purpose of this paper is to introduce a multiple imputation procedure for SAOMs and apply it to a realistic example, therefore a detailed comparison to the ML missing data treatment is out of the scope of this paper.

5. ILLUSTRATIVE EXAMPLE

5.1. NETWORK DATA

The outlined procedure is demonstrated on an adolescent friendship network of 50 girls observed at three waves. The data set is used in previous SAOM (simulation) studies (e.g., Huisman and Steglich, 2008) and was originally part of the Teenage Health and Lifestyle study (Michell and Amos, 1997; Pearson and West, 2003; Steglich et al., 2006). At every wave, the girls' alcohol consumption was also surveyed, using an ordinal five point scale. To illustrate the method and provide a first comparison to existing methods we will apply the following missing data treatments: 1) the default treatment implemented in RSiena (MoM), 2) single imputation with first wave ERGM imputation (1st-ERGM; Hipp et al., 2015), 3) inclusive sampling (de la Haye et al., 2017), 4) multiple imputation with first wave BERGM imputation (MI-BERGM), and 5) multiple imputation with first wave SAOM imputation (MI-SAOM).

In this example, we generated missing data in each wave separately by randomly selecting 10 actors and removing all outgoing ties for these actors (wave non-response giving 20% MCAR data). For ease of the example, no missing data on alcohol consumption were created. In period 1, 64% of the tie variables and 40% of the dyads were observed at both time points. In period 2, 62% of the tie variables and 37% of the dyads were completely observed. This constitutes a very high proportion of missing data.

5.2. MISSING DATA TREATMENT

After generating the missing data, the five missing data treatments were used to handle the missing actors, and a SAOM was estimated on the treated data. The estimated parameters are compared with the estimates obtained from the same SAOM fitted to the complete data.

SAOM MODEL

We first estimated a SAOM to impute waves 2 and 3, using the default MoM procedure on the incomplete data. The model was estimated on the incomplete data and not on the complete data, because in empirical research the complete data will not be available. The SAOM included the following structural effects: Density, degree related effects (square-root of indegree popularity, square-root of outdegree activity), reciprocity, triadic closure (geometrically weighted edgewise shared partners, GWESP[7]), and the interaction of reciprocity and GWESP. Further, the model contained effects regarding selection on alcohol consumption: Ego alcohol consumption, alter alcohol consumption, and similarity on alcohol consumption. Additionally, we included alcohol consumption as a dependent variable, including a linear and quadratic effect of previous alcohol consumption on future alcohol consumption, as well as an effect for friends' influence on alcohol consumption (average similarity to friends alcohol consumption). The model is generally similar to other models estimated on the network (e.g., Huisman and Steglich, 2008). Model fit was evaluated on outdegree, indegree, and geodesic

[7] This parameter and all other geometrically weighted parameters had a decay parameter of log(2).

distance distributions, and on the triad census. The model showed good fit on the incomplete data, and also on the complete data the fit was adequate.

MULTIPLE IMPUTATION

Following the presented procedure, two methods were used to impute the missing data in the first wave: using Bayesian ERGMs and using stationary SAOMs. In both methods, the first wave was multiply imputed $D = 50$ times.

In the Bayesian ERGM procedure, the imputation model included parameters to model similar structures as the SAOM: parameters for edges, reciprocity, triadic closure (geometrically weighted edgewise shared partners, GWESP), two-paths (geometrically weighted dyadwise shared partners, GWDSP[8]), in- and out-degree distribution (geometrically weighted indegree and outdegree) and a term specifically modeling the reciprocated transitive triad. Further, a homophily parameter for alcohol consumption (based on absolute difference), as well as in- and outdegree related effects of alcohol consumption were included. Additionally, the observed network at wave 2 was included as a dyadic covariate. Missing data at wave 2 were substituted with zeros, as the current implementation of dyadic covariates in the ergm package does not allow missing data on dyadic covariates (the ergm package is used for estimating BERGMs). After multiply imputing the first wave, the procedure proposed in Section 4.2 was applied to impute waves 2 and 3 to obtain $D = 50$ imputed data sets using the SAOM with the parameters identified earlier. On each of the 50 imputed data sets, the same SAOM was then estimated using the MoM estimator, and the results were combined using Rubin's rules.

Multiple imputation was also employed using a stationary SAOM for the first wave, followed by the outlined procedure for waves 2 and 3. The imputation model for the first wave included the same parameters as the imputation model for waves 2 and 3 (the regular, non-stationary SAOM) and, additionally, the observed network at wave 2 as dyadic covariate. Missing data at wave 2 were substituted with zeros, as before. Again, $D = 50$ imputed data sets were obtained and analyzed using the SAOM, and the results were combined.

ERGM IMPUTATION

Following the procedure proposed by Hipp et al. (2015), the first wave was imputed multiple times using the ergm package (Handcock et al., 2017). We were

[8] A parameter for two-paths was included to aid the proper estimation of triadic closure.

unable to obtain a converged model with the same parameters as used in the BERGM. Therefore, the parameters for the in- and outdegree distributions (geometrically weighted indegree and outdegree) and the reciprocated transitive triad were removed from the imputation model. Further, GWDSP was replaced by the not geometrically weighted regular two-path parameter. After the first wave was imputed $D = 50$ times, the regular SAOM model (MoM estimation) was estimated, not treating the missing data in wave 2 or 3 (i.e., using the default missing data treatment in RSiena). The results were combined according to Rubin's rules.

INCLUSIVE SAMPLING

The inclusive sampling method was applied by excluding, period by period, all actors who had any missing values within that period. Then, RSiena's multigroup analysis was performed on the separate periods (MoM estimation), treating the two periods as separate groups.

5.3. RESULTS

The estimated model parameters are graphically presented in Figure 1 (including error bars representing 1 standard error of uncertainty); The estimated values are given in Table 1. The default treatment, inclusive sampling and 1st-ERGM imputation show lower estimates for the rate functions and biases for some of the structural effects. Selection effects, however, are well estimated (although inclusive sampling shows a much lower estimate for selection of others with similar alcohol consumption) and so are all parameters related to the evolution of alcohol consumption (here again, inclusive sampling shows a much lower influence effect). The estimated standard errors for these three methods are often substantially larger than the complete data estimates, and sometimes so large that they would influence model inferences (e.g., the standard errors for outdegree activity are so large that the parameter would no longer be considered significant).

The proposed MI procedures are in stark contrast to these three methods. Both of them perform very well and all parameter and standard error estimates are very close to the complete data estimates. Only the rate functions for the first period are underestimated. The 50 rate parameters of the imputed data sets show considerable variation, as can be seen in the large proportion of the between-imputation variance B_D on the total variance T_D, presented in Table 2. The other structural parameters also show reasonably large proportions of between-imputation variance, ranging from 13% to 26% for the MI-BERGM procedure and 13% to 28% for the MI-SAOM procedure.

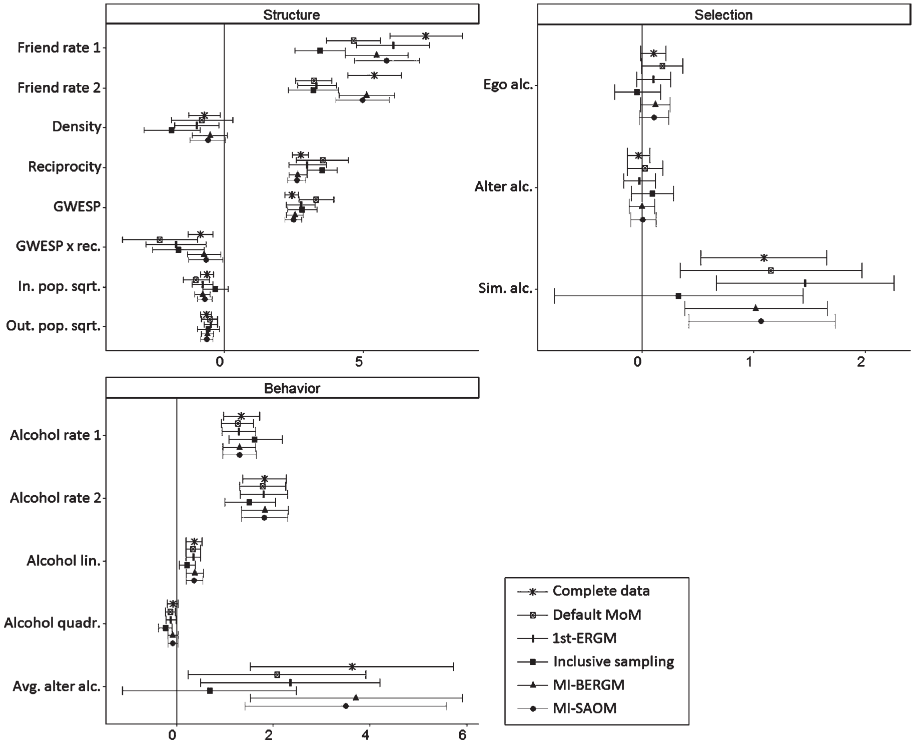

Fig. 1: Estimated SAOM parameters for the complete data and five treated incomplete data sets. The error bars represent ±1 standard error.

In the presented example data, it did not make a meaningful difference if the first wave was imputed by a Bayesian ERGM or a stationary SAOM. The differences in parameter estimates and standard errors are negligible and do not indicate any advantage for one of the two options.

It is important to emphasize that this limited illustration is not an exhaustive comparison of the methods and performance depends on a multitude of factors, such as missing data mechanism, type of non-response, and non-response rate. Especially inclusive sampling was not designed as a primary missing data treatment, but as last resort when the SAOM does not converge due to a large proportion of missing data.

Tab. 1: Estimated SAOM parameters (and standard errors) for the complete data and five treated incomplete data sets.

	Complete data	Default MoM	1st-ERGM	Inclusive sampling	MI-BERGM	MI-SAOM
Friendship rate 1	7.25 (1.30)	4.65 (0.97)	6.08 (1.31)	3.45 (0.89)	5.48 (1.13)	5.85 (1.17)
Friendship rate 2	5.41 (0.96)	3.23 (0.64)	3.34 (0.70)	3.21 (0.90)	5.13 (1.00)	4.98 (0.97)
Density	−0.70 (0.56)	−0.79 (1.10)	−0.98 (0.80)	−1.87 (1.02)	−0.51 (0.64)	−0.58 (0.64)
Reciprocity	2.75 (0.29)	3.54 (0.94)	3.00 (0.67)	3.53 (0.54)	2.66 (0.33)	2.61 (0.33)
GWESP	2.43 (0.24)	3.32 (0.62)	2.76 (0.52)	2.81 (0.53)	2.54 (0.29)	2.50 (0.30)
GWESP × Reciprocity	−0.85 (0.44)	−2.30 (1.34)	−1.72 (1.09)	−1.63 (0.93)	−0.71 (0.60)	−0.65 (0.62)
Indegree popularity square root	−0.60 (0.23)	−1.00 (0.47)	−0.77 (0.37)	−0.31 (0.46)	−0.77 (0.27)	−0.69 (0.27)
Outdegree activity square root	−0.64 (0.20)	−0.52 (0.30)	−0.47 (0.24)	−0.56 (0.39)	−0.59 (0.21)	−0.61 (0.21)
Ego alcohol consumption	0.10 (0.11)	0.18 (0.19)	0.11 (0.15)	−0.04 (0.21)	0.12 (0.13)	0.11 (0.13)
Alter alcohol consumption	−0.03 (0.10)	0.03 (0.16)	−0.02 (0.14)	0.09 (0.19)	< 0.01 (0.11)	0.01 (0.11)
Alcohol consumption similarity	1.09 (0.56)	1.16 (0.82)	1.72 (1.09)	0.33 (1.12)	1.02 (0.64)	1.08 (0.66)
Alcohol consumption rate 1	1.34 (0.37)	1.27 (0.33)	1.29 (0.35)	1.63 (0.55)	1.30 (0.34)	1.31 (0.34)
Alcohol consumption rate 2	1.82 (0.45)	1.78 (0.48)	1.80 (0.49)	1.52 (0.52)	1.83 (0.49)	1.81 (0.48)
Alcohol linear effect	0.36 (0.17)	0.35 (0.15)	0.36 (0.15)	0.23 (0.17)	0.38 (0.17)	0.37 (0.17)
Alcohol quadratic effect	−0.07 (0.11)	−0.12 (0.10)	−0.11 (0.10)	−0.22 (0.14)	−0.07 (0.11)	−0.07 (0.11)
Average simil. alter alcohol	3.63 (2.10)	2.08 (1.84)	2.36 (1.86)	0.69 (1.80)	3.72 (2.19)	3.50 (2.09)

Tab. 2: Ratio of between imputation variance on the total variance (B_D/T_D).

	1st-ERGM	MI-BERGM	MI-SAOM
Friendship rate 1	0.11	0.26	0.25
Friendship rate 2	< 0.01	0.20	0.22
Density	0.01	0.12	0.13
Reciprocity	0.01	0.26	0.23
GWESP	0.01	0.16	0.21
GWESP × Reciprocity	0.01	0.23	0.28
Indegree popularity square root	0.01	0.19	0.20
Outdegree activity square root	0.01	0.13	0.13
Ego alcohol consumption	0.02	0.14	0.15
Alter alcohol consumption	0.01	0.16	0.14
Alcohol consumption similarity	0.03	0.13	0.17
Alcohol consumption rate 1	< 0.01	< 0.01	< 0.01
Alcohol consumption rate 2	< 0.01	< 0.01	< 0.01
Alcohol linear effect	< 0.01	0.01	0.01
Alcohol quadratic effect	0.01	0.03	0.03
Average simil. alter alcohol	0.03	0.10	0.09

6. DISCUSSION

In this study, we introduced a multiple imputation method for SAOMs and demonstrated it on an empirical data set with simulated missing data. The results suggest that multiple imputation performs as well and often better than the default procedure within the SIENA software and the other missing data procedures. Especially standard errors seem to be estimated more reliably. The large standard errors found for the simpler methods were not surprising, given that the estimation of the parameters was based on considerably less data.

This study gives an introduction and small demonstration of the proposed procedure and not a thorough investigation. Future studies are required to determine the actual reduction in bias and verify the impact of the various choices a researcher can make in the imputation model. The proposed procedure needs to be tested on a larger sample of different networks. This will allow reliable estimation of the reduction in biases of parameters and standard errors.

Additionally, the current study only explored the procedure under MCAR. Future research has to investigate the performance under other missing data mechanisms. Although some participants are likely to be missing completely at random, there also might be structural reasons for participants to not participate in the study or withhold information. It is important to investigate how vulnerable multiple imputation is to biases when the data are missing not at random or missing depending on a covariate.

Further research is required to determine the influence of the imputation model and the procedure with which the parameters used for imputation are obtained. Theoretically, proper multiple imputation, using Bayesian estimation of SAOMs to generate distributions of the model parameters of the imputation models, should lead to unbiased results under MCAR or MAR.

In addition, the impact of the first wave imputation needs to be evaluated. While BERGMs provide overall better draws from the respective parameter distribution, stationary SAOMs fit conceptually better to the SAOM used for further modeling of the data. They are also easier to adapt to incorporate imputation of co-evolving behavioral variables or multiplex network structures. The proposed procedure should also be able to impute missing behavioral data. The simulated network evolution trajectories do not only simulate tie changes, but simulate changes on all dependent variables, including behaviors. Maximum likelihood simulations can therefore also be used to multiply impute missing behavior variables.

In spite of these unanswered questions and the limited illustration, the proposed procedure seems theoretically superior to current alternatives (the default MoM estimation implemented in the SIENA software and the procedures proposed by de la Haye et al., 2017, and Hipp et al., 2015), which is also supported by our small example. It utilizes more of the available information and conserves the relationships between all variables.

ACKNOWLEDGMENTS

We thank the two anonymous reviewers for their helpful comments that significantly improved this study.

REFERENCES

Allison, P.D. (2001). *Missing Data.* (In Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136). Sage, Thousand Oaks, CA.

Amati, V., Schönenberger, F. and Snijders, T.A.B. (2015). Estimation of stochastic actor-oriented models for the evolution of networks by generalized method of moments. In *Journal de la Société Française de Statistique.* 156: 140-165.

Block, P., Stadtfeld, C. and Snijders, T.A.B. (2016). Forms of dependence: comparing SAOMs and ERGMs from basic principles. In *Sociological Methods & Research.* 3: 1-38.

Borgatti, S.P. and Molina, J.L. (2003). Ethical and strategic issues in organizational social network analysis. In *Journal of Applied Behavioral Science.* 39: 337-349.

Butts, C.T. (2008). A relational event framework for social action. In *Sociological Methodology.* 38: 155-200.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. In *Social Networks.* 33: 41-55.

Caimo, A. and Friel, N. (2013). Bayesian model selection for exponential random graph models. In *Social Networks.* 35: 11-24.

Caimo, A. and Friel, N. (2014). Bergm: Bayesian exponential random graphs in R. In *Journal of Statistical Software.* 61: 1-25. R package version 4.0.0

de la Haye, K., Embree, J., Punkay, M., Espelage, D.L., Tucker, J.S. and Green, H.D. (2017). Analytic strategies for longitudinal networks with missing data. In *Social Networks.* 50: 17-25.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., Krivitsky, P. and Morris, M. (2017). Package ergm: Fit, simulate and diagnose exponential-family models for networks. In *The Statnet Project.* R package version 3.8.0.

Hipp, J.R., Wang, C., Butts, C.T., Jose, R. and Lakon, C.M. (2015). Research note: The consequences of different methods for handling missing network data in stochastic actor based models. In *Social Networks.* 41: 56-71.

Huisman, M. and Steglich, C.E.G. (2008). Treatment of non-response in longitudinal network studies. In *Social Networks.* 30: 297-308.

Huisman, M. and Krause, R.W. (2017). Imputation of missing network data. In R. Alhajj, and J. Rokne, editors, *Encyclopedia of Social Network Analysis and Mining.* Springer, New York: 1-10.

Koskinen, J.H., Caimo, A. and Lomi, A. (2015). Simultaneous modeling of initial conditions and time heterogeneity in dynamic networks: An application to Foreign Direct Investments. In Network Science. 3: 58-77.

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

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

Koskinen, J.H. and Snijders, T.A.B. (2007). Bayesian inference for dynamic social network data. In *Journal of Statistical Planning and Inference.* 137: 3930-3938.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, editors, *Panel Surveys.* Wiley, New York: 348-374.

Lusher, D., Koskinen, J.H. and Robins, G. (2013). *Exponential Random Graph Models for Social Networks.* Cambridge University Press, New York.

Michell, L. and Amos, A. (1997). Girls, pecking order and smoking. In *Social Science and Medicine.* 44: 1861-1869.

Pearson, M. and West, P. (2003). Drifting smoke rings: Social network analysis and Markov processes in a longitudinal study of friendship groups and risk-taking. In *Connections.* 25: 59-76.

Ripley, R., Boitmanis, K., Snijders, T.A.B. and Schoenenberger, F. (2017). In *Package RSiena: SIENA - Simulation Investigation for Empirical Network Analysis*. R package version 1.2-3.

Rubin, D.B. (1976). Inference and missing data. In *Biometrika.* 63: 581-592.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in *Surveys.* Wiley, New York.

Smith, J.A., Moody, J., and Morgan, J.H. (2017). Network sampling coverage II: The effect of non-random missing data on network measurement. In *Social Networks.* 48: 78-99.

Snijders, T.A.B. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology.* 21: 149-172.

Snijders, T.A.B. (2001). The statistical evaluation of social network dynamics. In *Sociological Methodology.* 31:, 361-395.

Snijders, T.A.B. (2005). Models for longitudinal network data. In P.J. Carrington, J. Scott, and S. Wasserman, editors, *Models and Methods in Social Network Analysis.* Cambridge University Press, New York: 215-247.

Snijders, T.A.B. (2017a). Stochastic actor-oriented models for network dynamics. In *Annual Review of Statistics and Its Application.* 4: 343-363.

Snijders, T.A.B. (2017b). *Siena Algorithms.* http://www.stats.ox.ac.uk/~snijders/siena/Siena_algorithms.pdf.

Snijders, T.A.B., Koskinen, J.H. and Schweinberger, M. (2010a). Maximum likelihood estimation for social network dynamics. In *Annals of Applied Statistics.* 4: 567-588.

Snijders, T.A.B. and Steglich, C.E.G. (2015). Representing micro-macro linkages by actor-based dynamic network models. In *Sociological Methods & Research.* 44: 222-271.

Snijders, T.A.B., van de Bunt, G.G. and Steglich, C.E.G. (2010b). Introduction to stochastic actor-based models for network dynamics. In *Social Networks.* 32: 44-60.

Stadtfeld, C. and Block, P. (2017). Interactions, actors, and time: Dynamic network Actor models for relational events. In Sociological Science. 4: 318-352.

Stadtfeld, C., Hollway, J. and Block, P. (2017). Dynamic network actor models: Investigating coordination ties through time. In *Sociological Methodology.* 47: 1-40.

Steglich, C.E.G, Snijders, T.A.B. and Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. In *Sociological Methodology*. 40: 329-393.

Steglich, C.E.G., Snijders, T.A.B. and West, P. (2006). Applying SIENA: An illustrative analysis of the co-evolution of adolescents' friendship networks, taste in music, and alcohol consumption. In *Methodology*. 2: 48-56.

van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Chapman & Hall/CRC Press, Boca Raton.