Treatment of Non-Response in Longitudinal Network Studies

Mark Huisman*
Department of Psychology
Christian Steglich†
Department of Sociology
University of Groningen

15th November 2007

Abstract

The collection of longitudinal data on complete social networks often faces the problem of actor non-response. The resulting incomplete data pose a challenge to statistical analysis, as there typically is no natural way how to treat the missing cases. This paper examines the problems caused by actors missing as nominators, but still occurring as nominees, in complete, directed networks measured in a panel design. In the framework of stochastic actor-driven models for network change ("SIENA models"), different methods to cope with such incomplete data are investigated. Data on a friendship network among female high school students are used to illustrate the procedures. Similar problems related to early panel exit and late panel entry are not addressed.

Keywords: Missing data; Wave non-response; Imputation; Network evolution.

*Department of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9724 TS, Groningen, email: j.m.e.huisman@rug.nl, tel: +31 50 3636345, fax: +31 50 3636304.

†The second author was funded by the Netherlands Organization for Scientific Research (NWO) under grant 401-01-550.

1 Introduction

Data analysis in social sciences is often hampered by non-response. In the analysis of social networks, non-response results in missing network information. This means that ties from one actor to another are not observed and/or information on actor attributes is not available. According to Burt "*missing data are doubly a curse to survey network analysis*" (1987, p. 63), compared to other types of analyses (see also Borgatti and Molina, 2003). First, the complexity of items in network surveys are more likely to generate missing data (e.g., Marsden, 2003), and second, network analysis is especially sensitive to missing data because of the dependence structure of the data. If a network tie, or worse, an actor is missing, there is limited capacity to describe the network context of the actors whose ties are missing and there is lack of information on the context of neighboring actors (Robins et al., 2004).

The effects of non-response and missing data on the structural properties of networks are investigated in several studies (Burt, 1987; Costenbader and Valente, 2003; Kossinets, 2006). The general conclusion is that missing data have a negative effect on network mapping (Borgatti and Molina, 2003) and estimating structural network properties: the strength of relationships is underestimated (Burt, 1987), centrality measures become unstable as well as degree measures (Costenbader and Valente, 2003, Kossinets, 2006), and clustering coefficients are underestimated (Kossinets, 2006). Still, Costenbader and Valente (2003) find that measures based on indegrees are reasonably robust for small proportions of missing data when the observed incoming ties of non-respondents are used in the analysis. This latter result shows an unique property of social networks: non-participation by respondents does not necessarily mean that they are not included in the study (Borgatti and Molina, 2003). Respondents report ties to non-respondents, that is, the incoming ties of non-respondents are available.

Missing data treatment methods can use the information on non-respondents from the nominations of observed actors. Stork and Richards (1992) propose using these partially described ties between respondents and non-respondents to reconstruct the missing outgoing ties: substitute the missing ties by the value of the tie in the opposite direction. This imputation method is appropriate if ties tend to match across actors, for instance in undirected networks. For directed networks, this (ad hoc) imputation method seems less suitable. Another imputation method is suggested by Burt (1987), who finds that missing relations are strongly associated with weak relations and therefore can be replaced with values indicating such weak relations.

More recent missing data methods are proposed by Robins et al. (2004), Gile and Handcock (2006), Handcock and Gile (2007), and Koskinen (2007). These methods are also based on all available data, including the incoming ties of non-respondents. The proposed methods are model-based treat-

ment methods within the framework of exponential random graph models (ERGMs). Robins et al. (2004) model the ties from respondents to non-respondents separately from the fully described ties, which allows exploring the structural effects for the entire network. The model is especially helpful when the non-respondents systematically differ from the respondents with respect to ties. Gile and Handcock (2006), Handcock and Gile (2007), and Koskinen (2007) use Markov chain Monte Carlo methods to fit ERGMs to incomplete network data. This is a more traditional missing data approach based on the marginal distribution of the observed data (e.g., see Schafer and Graham, 2002), allowing for proper inferences for network properties for both respondents and non-respondents. As the methods repeatedly sample from the conditional distribution of the missing data, they can also be used to impute the data sets.

All these methods are designed for modeling *single*, incomplete observations of a network. Moreover, possible treatments are either simple ad hoc procedures (the imputation methods of Stork and Richards, 1992, and Burt, 1987), or are embedded within ERGMs (Robins et al., 2004; Gile and Handcock, 2006). For the case of longitudinal network data, studies on the effect of non-response or the effect of treatment procedures are lacking. In this paper we examine the effect of non-response and missing data techniques on longitudinal network data. The effect of missing data treatments are investigated within the framework of the actor-driven models for network evolution proposed by Snijders (2001, 2005), using simulations under a known evolution model. The missing data treatments that are used in the simulation study are the analysis of complete cases, two ad hoc imputation methods based on reconstruction (Stork and Richards, 1992) and preferential attachment (Barabasi and Albert, 1999), respectively, and an hybrid imputation procedure based on simulating networks with the actor-driven network evolution models (Snijders, 2005).

The paper is organized as follows. Section 2 addresses the problem of non-response in longitudinal network data, defining the missing data patterns that are considered in this study. In Section 3 the family of actor-driven models for network evolution of Snijders (2001, 2005) is briefly described. Section 4 presents the missing data treatments, of which the performance (i.e., the effects of the treatments on modeling the data with actor-driven models) is investigated in a simulation study. The design of this study is presented in Section 5 and in Section 6 the results in terms of convergence of the estimation procedure and the absolute and relative bias in the parameter estimates. The paper ends with a discussion of the results and some general recommendations.

2 Non-response in longitudinal network studies

In missing data research usually two types of non-response are distinguished: unit non-response, where complete cases are missing, and item non-response, where the unit participated but data on particular items are missing. For social network data, *unit non-response* means that an actor does not participates in the study and therefore all his or her outgoing ties are unavailable for analysis. *Item non-response* means that only particular (outgoing) ties are unavailable in the analyses.

In the case of longitudinal data where respondents are repeatedly contacted at successive time points, non-response patterns can be further distinguished by including partial non-response (De Leeuw et al., 2003), or wave non-response (Lepkowski, 1989). *Wave non-response* is characterized by time dependency and means that only at certain time points data are available. This is often due to panel mortality or attrition, which results in completely missing cases after a certain time point. For social network data, we define wave non-response as complete non-response at one or more time points, which results in completely missing outgoing ties of some actors for these time points.

The present study restricts analyses to unit and wave non-response, and considers the case of longitudinal network data with two observations moments and completely missing actors at one or both time points. This means we assume that at a certain observation moment non-response results in completely missing outgoing ties of some actors. The incoming ties of these non-responding actors are observed. Moreover, occasionally missing ties (item non-response) are not considered. This results in four kinds of missingness patterns, which are presented in Figure 1. For more than two observations of the network the procedures are generally the same. The main difference is in the number of possible missing data patterns.

Figure 1 shows the sociomatrices of a network at two observations moments. As the rows and columns of the matrices represent the actors, the subsets of actors distinguish mutually exclusive subsets of rows and columns. The observed missingness patterns consist of four set of actors: A_1, actors observed at both time points, A_2, actors observed at T_1 and missing at T_2 (wave non-response), A_3, actors missing at T_1 and observed at T_2 (wave non-response), and A_4, actors missing at both time points (unit non-response). At observation moment T_1, the set of observed actors is $A_1 \cup A_2$ and the data set consists of the corresponding rows in the sociomatrix. These are the white areas in first adjacency matrix in Figure 1, consisting of incoming and outgoing ties of observed actors in $A_1 \cup A_2$, and incoming ties of non-respondents in $A_3 \cup A_4$. At observation time T_2 the set of observed actors is $A_1 \cup A_3$.

Although the non-response patterns play an important role in treating the missingness, the most important question is whether the non-response

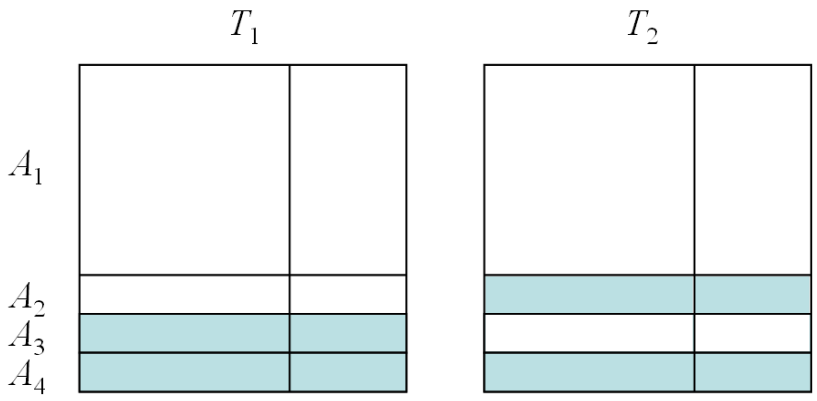


Figure 1: Partial non-response in a network observed at two time points (T_1, T_2), identifying four subsets of actors (A_1 – A_4). Gray areas indicate missing outgoing ties.

causes systematic bias in the analysis due to systematic differences between respondents and non-respondents. This leads to the difficult task of assessing whether the data are missing at random and, as a result, the missingness mechanism can be ignored (Rubin, 1976). Data are called *missing at random* (MAR) if the missingness is unrelated to the (unknown) value of the missing item itself. For network analysis this means that missingness is unrelated to the value of the tie. In this case the missingness may be related to completely observed actor characteristics, but not to network properties.

If the missingness is related to the value of the missing tie itself, the data are *missing not at random* (MNAR). In this case, network properties are biased because they are computed from the network in which tie values are systematically missing. The extent in which the structural properties are affected depends on the property itself. Measures based on indegrees, for instance, are found to be less affected than other measures, because incoming ties are only partially missing for all respondents.

3 Actor-driven models for network evolution

The prominent tool for modeling and analyzing longitudinal, complete network data is the family of stochastic, actor-driven models introduced by Snijders (1996, 2001, 2005). Estimation of these models is implemented in the SIENA software (shorthand for Simulation Investigation for Empirical Network Analysis; Snijders et al., 2007). The present study refers to this model family and software package, and the way parameter estimates are affected by missing data and missing data treatment.

In the actor-driven approach, the dynamics of a social network are modeled as a stochastic process $X(t)$ in the state space $\mathcal{X}$ consisting of all possible network configurations on a given set of n actors. In this paper, the networks are directed graphs, coded as binary-valued, possibly asymmetric adjacency matrices x with $x_{ij} = 1$ indicating presence of a tie and $x_{ij} = 0$ indicating absence; the state space then is $\mathcal{X} = \{0, 1\}^{n(n-1)}$. Furthermore, we treat the case of a two-wave panel, so two networks $x(t_1)$, $x(t_2) \in \mathcal{X}$ are given as observed data. Because the exact trajectory of network changes that occur in-between the two observations is unobserved, it is appropriate to model these data as resulting from a continuous-time process. This is achieved by constructing a continuous-time Markov chain.

The compound change that occurs between the two observations is modeled as the aggregate outcome of a series of unobserved, stochastically spaced, small changes called *micro steps*. The first observation is taken as starting value of the stochastic process and hence is not modeled itself. The micro steps consist of the change of one tie variable x_{ij} between two actors i and j in the network, and is modeled as maximization by actor i (the 'sender' of the tie) of an *objective function*

$$f_i(X(t)) = \sum_k \beta_k b_{ki}(X(t)) \tag{1}$$

plus a random term ϵ_i with a conveniently chosen distribution[1]. Parameters β_k are weighting actor-specific network statistics $b_{ki}(X)$. Commonly, these statistics correspond to local subgraph counts or non-linear transformations thereof. Examples of network effects and the corresponding statistics for the objective function are given in Table 1. These statistics and the estimated parameter values are are used in the simulation study to generate incomplete data sets. The included effects are *outdegree*, for measuring actor i's tendency to randomly establish ties to any other actor, *reciprocity*, measuring tendencies to reciprocate ties, *transitive triplets* and *geodesic distance 2*, both for measuring tendencies toward transitive closure, and the *attribute-related similarity*, measuring patterns of homophile selection on an actor attribute z, in this case alcohol consumption. The model is further discussed in Section 5.1.

Maximization of the objective function takes place over a choice set consisting of micro steps and the option of no change. The distribution of waiting times between these small changes is modeled by a parametric family of exponential distributions called the *rate functions*. For the present purposes, we assume rates to be constant across actors. When increasing the number n of network actors, the cardinality of state space $\mathcal{X}$ rises at a squared exponential rate, which renders explicit calculations of expectations

[1]The choice of convenience here is independent drawing from the extreme value type I (or Gumbel) distribution, which allows to express choice probabilities in multinomial logit shape (McFadden, 1974).

Table 1: The included effects, statistics, and estimated parameter values of the 'true' evolution model used in the simulation study.

Effect	Statistic $b_{ki}(X)$	Parameter value
Outdegree	$\sum_j x_{ij}$	−2.01
Reciprocity	$\sum_j x_{ij}x_{ji}$	2.11
Transitivity	$\sum_{jk} x_{ij}x_{ik}x_{kj}$	0.27
Geodesic distance 2	$\sum_j (1-x_{ij}) \max_k x_{ik}x_{kj}$	−0.79
Alcohol-related similarity	$\sum_j x_{ij}(1-\frac{\lvert z_i-z_j\rvert}{range(z)})$	0.92
Constant rate		6.87

and likelihoods practically impossible. Estimation of the actor-driven models therefore has to rely on simulation-based inference. The SIENA software instantiates simulation-based *method of moments* estimation of the models, which we use for the estimations in this paper (newer versions also allow for simulation-based *maximum likelihood* and *Bayesian* estimation; Snijders et al., 2007; Koskinen and Snijders, 2006).

4 Missing data treatments

There are several ways to deal with missing data. Two general, popular approaches are likelihood-based estimation based on the available data and imputation (Schafer and Graham, 2002). The ERGM-based procedures proposed by Robins et al. (2004), Gile and Handcock (2006), Handcock and Gile (2007) and Koskinen (2007) are examples of the former group of treatments, although the latter three can also be used to produce imputed data sets. The reconstruction method suggested by Stork and Richards (1992), and the replacement of missing data with values representing weak relations suggested by Burt (1987) are also examples of imputation procedures.

In this section four missing data treatments are discussed, two of which are imputation methods. All four treatments are investigated in the simulation study described in Section 5. The techniques are *(i)* complete case analysis, i.e., reduction of the data set to the completely observed cases, *(ii)* imputation by reconstruction, *(iii)* imputation by preferential attachment, and *(iv)* missing data treatment within the framework of actor-driven evolution models. The two imputation procedures are ad hoc procedures that impute each observation of the network independently from other observations and result in completed networks at both time points, that is, all actors in $A_1 \cup ... \cup A_4$ are available for analyses. The fourth procedure is based on the simulation of micro steps in the estimation procedure of the actor-driven models described in Section 3. As its primary concern is model

estimation and uses only initial imputations at T_1, this hybrid imputation procedure does not automatically result in a completed data set.

While these techniques can be used for all types of non-response, this paper is only concerned with missingness due to unit and wave non-response – a situation we believe is close to what empirical network researchers typically face. For this situation Robins et al. (2004) remark that "*imputation is unlikely to be very successful*" (p. 206). This may be particularly true for the imputation methods *(ii)* and *(iii)*, but it has never been investigated in a longitudinal context. In any case, these methods are acceptable benchmarks to investigate the effectiveness of other methods. Our focus of interest, naturally, lies on assessing the quality of the model-based hybrid imputation method available in the SIENA software.

4.1 Complete case analysis

Complete case analysis (CC) is based on the smaller network of completely observed actors, i.e., those who gave valid responses at all measurement points ('listwise' deletion of actors). The analyzed data set consists of the incoming and outgoing ties of these actors, that is, the upper-left white block in the sociomatrices depicted in Figure 1 for actor set A_1. The observed incoming ties for missing actors are ignored in this procedure. The data reduction which this method implies can be considerable. If at k observation points independently the response rate is ρ, the probability for any tie variable x_{ij} to be retained in the complete case data set is ρ^{2k}. This implies that in a two-wave design, already a response rate of 71% delivers network matrices containing but 25% of the original number of cells. It can be expected that this method will deliver results that are highly sensitive to the fraction of missing cases, which should reasonably be taken serious only for very low levels of missingness.

A straightforward strategy to avoid this loss of data is to impute artificial observations for the missing values. The following three methods all employ variants of this theme.

4.2 Imputation by reconstruction

Stork and Richards (1992) suggest *reconstructing* the missing part of the network by using the observed incoming relations of the missing actors. As the procedure does not allow reconstruction of ties between non-respondents, additional imputations are necessary in order to reconstruct the whole network. The following procedure is used:

1. For all ties between non-respondents i and respondents j, impute the observed value of the opposite tie: $x_{ij}^{imp} = x_{ji}$.

2. For all ties between non-respondents, randomly impute a tie proportional to the observed density (i.e., the probability of a tie is equal to the observed density of the network).

The reconstruction procedure (RE) generates imputations for the two observation moments separately, that is, the missing actors $A_2 \cup A_4$ at T_1 are treated independently from the missing actors $A_3 \cup A_4$ at T_2. It is assumed that reported ties match across actors. This is true for undirected networks (as those studied by Stork and Richards), but may not be the case for directed networks. Even in networks with strong reciprocity effects, a large number of ties may not be reciprocated. Hence, we can expect that the degree of reciprocity in the data is exaggerated by this method, which will have an impact on analytical results.

4.3 Imputation by preferential attachment

This procedures uses the concept of preferential attachment, which states that the probability that an actor will link to another actor is dependent on the connectivity of other actors (Barabasi and Albert, 1999). Preferential attachment is incorporated in terms of indegrees by assuming that the probability that a missing actor i will be connected to another (observed or missing) actor j is proportional to the indegree k_j of actor j:

$$\Pi(k_j) = \frac{k_j}{\sum_{j \neq i} k_j}. \tag{2}$$

The following two-step procedure is used to replace the missing ties by randomly drawn zeros or ones:

1. For each missing actor i in either $A_3 \cup A_4$ or $A_2 \cup A_4$, randomly draw an outdegree d_i from the observed outdegree distribution at the observation moment under consideration.

2. For the missing actor i, randomly draw a total of d_i actors j ($j \neq i$), without replacement, from the total set of actors $A_1 \cup ... \cup A_4$ using the the preferential attachment probabilities $\Pi(k_j)$. For the ties between i and j impute the value $x_{ij}^{imp} = 1$, otherwise $x_{ij}^{imp} = 0$.

The preferential attachment procedure (PA) generates imputations for two observation moments separately, such that the observed degree distributions at both moments are retained.

4.4 Missing data treatment within actor-driven models

This fourth procedure is highly contingent on the analytical tool with which the data are analyzed, i.e., a model-based method. It differs from the previous procedures in two ways. First, different subsets of missing actors (at

different time points) are not treated similarly, as they were in all other methods. And second, model estimation under this treatment is based on the set of completely observed actors at both time points, A_1, only, instead of the completed data of all actors.

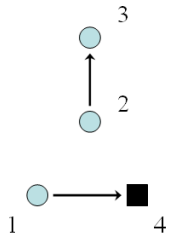
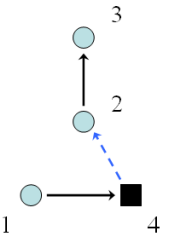
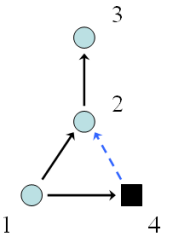
The procedure is based on the simulation of the network evolution process within the estimation procedure of the actor-driven models. This estimation procedure is based on the simulation of continuous-time Markov chains of networks. Starting from the first observation at T_1, a Markov chain of networks (i.e., a series of micro steps) is simulated using the specified model (objective function and rate function) and the current values of the model parameters. At the second observation moment T_2 the difference between the simulated and observed network (expressed in a vector of differences on network statistics) is used to update the parameter estimates. Then, a new Markov chain of networks is simulated with the updated parameter estimates, and the process is repeated until parameter values converge (see Snijders, 2001, 2005, for details).

The missing data treatment starts with initial imputations at T_1: all missing ties are treated as being absent, that is, impute the value $x_{ij}^{imp} = 0$. This seems a reasonable choice, considering that the networks under study typically are sparse (the mode of the tie values is zero), and that missingness is found to coincide with weak ties (Burt[2], 1987). The missing ties at T_2 are not replaced, but imputed by way of simulation. In the simulation phase of the estimation procedure, all actors – observed and missing – are allowed to make changes in their outgoing ties. As all actors have the opportunity to interact between two observations, all ties are free to change, including the imputed ones. This results in a simulated network at the second observation moment T_2, in which all ties have meaningful simulated values. These can be used to impute the missing ties at T_2. However, the parameter update step (and hence model estimation) is based on the observed ties at both time points only, that is, the network statistics used for the updating step are calculated on this reduced data set, just as for the complete case treatment. This way, the impact of missing actors at T_1 and T_2 on the estimates is minimized. Still, the missing actors at T_1 have an indirect effect on the results by acting as constraints and opportunities for tie changes during the simulations, thus affecting what happens in the non-missing part of the data.

In Table 2 an example of an indirect effect of a model-imputed tie is presented. The table illustrates two micro steps in a network region involving four actors: three observed (1, 2, 3) and one missing (4). For the first micro step from τ_1 to τ_2, missing actor 4 is randomly chosen to apply a change to his outgoing ties, with the result that a tie to actor 2 is created. In the second micro step from τ_2 to τ_3, observed actor 1 makes a change by initiating a

[2]Note that Burt refers to missing ties in an ego-centered survey and his finding may not be true for missing actors in a complete network.

Table 2: Example of the indirect effect of a model-imputed tie on network structure.

Time	τ_1	τ_2	τ_3
			

Circles represent observed actors, boxes missing actors.

tie to actor 2, closing the triplet (1, 2, 4). This change may be induced by a preference for transitive triplets, but this particular triplet will not be counted in the network statistics used for the parameter updates because it involves a missing actor. However, the second micro step did result in an increase of the distance 2 statistic (between 1 and 3 via 2), which will be used in parameter updating because all three actors involved are observed.

The data used for estimating the model are the network statistics calculated on the actor set A_1 only, i.e., those observed at both time points. Different from the CC treatment method, though, the other data are not discarded. All other actors are used in the simulation of the network, this way providing structural constraints for how the reduced network among actors in A_1 evolves. For actor set A_3, this inclusion is straightforward, as the values of their outgoing ties at T_1 are known. For actors in $A_2 \cup A_4$, the inclusion requires a prior imputation of tie variables at T_1, here by setting them to zero. This missing data treatment is currently implemented in the SIENA software to handle missing ties (i.e., both completely missing actors and individual ties; see also Snijders, 2005), and will be referred to as the SIENA method (SM) in the remainder of the paper.

5 Simulation study

In order to investigate the sensitivity of parameter estimates of the actor-driven models to the various types of missing data treatments, a simulation study is performed. The general pattern of the study is:

1. generate complete data under a known evolution model,
2. generate missing data by erasing a fraction of actors (i.e., all outgoing ties of the actors),
3. treat the missing data using the procedures outlined in Section 4,

4. re-estimate the evolution model on the data treated for missingness,

5. investigate the effect of the treatments on the estimation procedure and the estimates.

5.1 Generating longitudinal network data

For generating simulated network evolution processes, the first two waves of a sample data set of 50 actors are used. These sample data are provided together with the SIENA software[3]. On these data, an actor-driven evolution model was estimated that is used as the 'known' evolution model to generate the data in the study. This way, our 'true' evolution model is close to what can be encountered in actual research (our simulations are 'empirically informed') – and does not incur overly long computer runs.

The model contains parameters for outdegree, reciprocity, transitive triplets, geodesic distance 2, and similarity on the covariate dimension of alcohol consumption. The parameter values are presented in Table 1. The outdegree parameter (-2.01) indicates that actors generally avoid ties, which is no surprise in a sparse network. They do have a preference for reciprocated ties (2.11), transitive closure (the transitivity parameter equals 0.27, and the distance 2 parameter equals -0.79, together indicating that actors prefer direct to indirect relations), and ties to others with the same score on alcohol consumption (0.92). The rate parameter equals 6.87, indicating the average frequency in-between observations by which network actors can apply changes to their network neighborhood. All parameters are significant at $\alpha = 0.01$ in the original data.

In the simulations of the network evolution process, the first observation of the network is taken as initial state of the process, and the observed data on alcohol consumption at first measurement as a constant actor covariate. Using the the estimated evolution model (based on the 'true' second observation), 500 times an actor-driven evolution process was simulated. This resulted in 500 simulated networks at the end of the simulation period. Note that these simulated evolution processes deliver different trajectories due to the stochastic nature of the model. The simulated end networks were taken as second observations in the simulation study.

5.2 Generating missing data

As we restrict our study to unit and wave non-response, missing data were created by randomly selecting actors and deleting all outgoing ties of these actors. This amounts to specifying rules by which actors are allocated to sets A_1 through A_4, introduced in Section 2. Missing actors were selected

[3]The data are a 'cleansed' subset of the girls' subnetwork in the *Teenage Health and Lifestyle* study, as discussed by Michell and Amos (1997), Pearson and West (2003), Steglich et al. (2006), and Steglich et al. (2007).

independently at both time points, using the same selection mechanism. The fractions missing actors at each time point are $(1 - \rho) = 0.2$, 0.4, and 0.6, where ρ is the response rate at each wave. Independence between time points in this procedure implies that the fraction of missing actors at both waves (set A_4) is, in expectation, equal to the product of the fractions missing at the single time points ($A_3 \cup A_4$ at T_1 and $A_2 \cup A_4$ at T_2). Four different missingness mechanisms were used, which define the probability that an actor is missing in the following way:

- completely random selection of actors,
- probability proportional to $\frac{1}{(alcohol\ score)^2}$
- probability proportional to $\frac{1}{(indegree+1)^2}$,
- probability proportional to $\frac{1}{(outdegree+1)^2}$.

Each of these mechanisms can be viewed as operationalization of assumptions about how missing data may occur in real-world network studies. Random deletion is a simple but coarse model of missingness and may be realistic when there is no reason to assume that actors differ in their propensity to fill in network questionnaires (or otherwise deliver their local part of the data). The data are MCAR, as the missingness is unrelated to network or actor characteristics.

However, often non-response will be related to network or actor characteristics, resulting in data that are MAR or even MNAR. In the second mechanism, deletion of actors is proportional to the covariate alcohol consumption. The mechanism is such that actors with lower scores on the covariate have a larger probability to be missing[4]. As the covariates are completely observed, the data are MAR.

The third and fourth type of missingness are both related to network properties, that is, the indegrees and outdegrees of the actors: actors with low degrees have a larger probability to be missing. This reflects the ideas that popular actors are more inclined to participate in a network study (indegree) and that network data of socially active actors are collected more easily than network data of inactive actors (outdegree), as inactive actors are more difficult to recruit or care less to respond. The square was added in the mechanism to make the distinction even more pronounced.

Both degree-mechanisms result in data that are MNAR, but the missingness patterns and biases may be quite different. As ties are missing completely for non-respondents, outdegrees cannot be computed for missing actors. Indegrees, however, can always be estimated using the partially

[4]This means that the less alcohol respondents consume, the less they are inclined to participate in the network study. It may not be overly realistic in every context, but certainly does not diminish usefulness of the mechanism for illustrative purposes.

observed incoming data of all respondents. This means that the (maybe biased) estimates of the indegrees can be used to treat the missingness due to the fourth mechanism in a MAR-based procedure. This might, at least partly, correct for the non-randomness of the mechanism.

In order to minimize the impact of random noise on the estimation of the evolution model, the missing actors of lower missingness levels are chosen to be subsets of the actors missing at higher levels. In practice, this was accomplished by establishing, for each of the networks analyzed and under each type of missingness, a sequence of 'dropping out' for the actors. The first ten actors in such a sequence then constitute the missing actors at the 20% level, the first twenty actors constitute the missing actors at the 40% level, etc. As indicated in the beginning of this section, these dropping out sequences are established independently for all networks. To further reduce the impact of random noise on the estimation results, the same sets of missing actors were used across the treatments procedures.

5.3 Model convergence

The generation of the networks and the missing data resulted in 500 complete and 3 (missingness levels) $\times 4$ (missingness mechanisms) $\times 500 = 6,000$ incomplete data sets (consisting of two waves). The complete data, generated according to a known ('true') evolution model, are used as a reference category. On all data, the actor-driven evolution model is re-estimated according to the four treatment methods proposed in Section 4, which amounts to a total of $500 + 4 \times 6,000 = 24,500$ estimation runs.

The chosen model specification, however, cannot be fit to each data set. Even without missing data, the possible mismatch between model and data needs to be monitored. Convergence diagnostics indicating the mismatch between model-consistent data and the to-be-analyzed data are implemented in the SIENA software. They are used to monitor possible convergence problems. When estimating actor-driven models from deliberately mutilated data sets, monitoring convergence problems is even more important because we expect a growing mismatch as the fraction of missing actors in the data increases.

Unfortunately, the SIENA convergence diagnostics do not necessarily detect all 'inaccurate' solutions. Most importantly, during the estimation process, parameter values may reach a region of the parameter space where model-derived expected data are not sensitive to changes in specific parameters any more, and where accordingly neither the parameter nor its standard error can be estimated accurately[5]. Therefore, we decided to classify estimation runs as divergent when at least one out of three conditions was

[5]This is similar to logistic regression models, where differences in the very high or the very low region of parameter values have barely any impact on the modeled choice probabilities, because the tails of the logistic link function are very flat.

satisfied:

1. SIENA diagnosed divergence based on the convergence diagnostics.

2. At least one of the parameter estimates was unreasonably high, that is, the absolute value is larger than 10 for the five parameters in the objective function. For the rate parameter, a more liberal threshold of 50 was chosen, as it already has the quite large value of 6.87 in the true model.

3. At least one of the estimated standard errors was unreasonably high, that is, the absolute value is larger than 10.

The (ad hoc) threshold values of 10 and 50 seem liberal in the sense that they are much higher than what is usually reported for SIENA results, which can also be seen from the results obtained for the complete data, yet they seem to suffice for distinguishing completely meaningless results from reasonably interpretable ones. The distribution of the results as shown in the boxplots of Figures 2–4 does not crucially depend on the chosen thresholds.

6 Results

The effect of the missing data treatments on modeling the longitudinal network data was evaluated using three measures of performance: practicability (operationalized as number of converged estimation runs), absolute size of error (operationalized as median parameter bias), and relative size of error (operationalized as the relative position of the true score in the distribution of estimates). The use of robust measures (percentiles) instead of sensitive ones (like averages or standard deviations) reduces the impact of possibly remaining outliers on the results. Figures 2–4 display, for each parameter under each combination of missingness level, missingness type and treatment method, the distribution of estimates in the shape of a boxplot. The width of the boxes indicates the fraction of convergent projects (the less projects, the narrower the box), the dotted lines indicate the true values of the parameters, and the four missing data treatments are labeled CC (complete case), SM (SIENA method), PA (preferential attachment) and RE (reconstruction). In the following, we aggregate the total information contained in these diagrams in terms of the three criteria identified above.

6.1 Convergence

Convergence crucially depends on the amount of information that the data set provides. It can be expected that the number of divergent projects increases with higher fractions of missing data. In Table 3 the percentages of diverging projects are presented, for each missing data mechanism and

Figure 2: Boxplots of the estimated parameters outdegree (left) and rate (right). From bottom to top, first a plot for the reference category (no missings) and next, four blocks with plots for the missingness mechanisms are rendered (MCAR, Indegree, Outdegree, Alcohol). Within each block, four sets of three plots are presented for the techniques CC, SM, PA, and RE, each for the three missingness levels (0.2, 0.4, 0.6). The width of the boxes represent the number of converged projects.

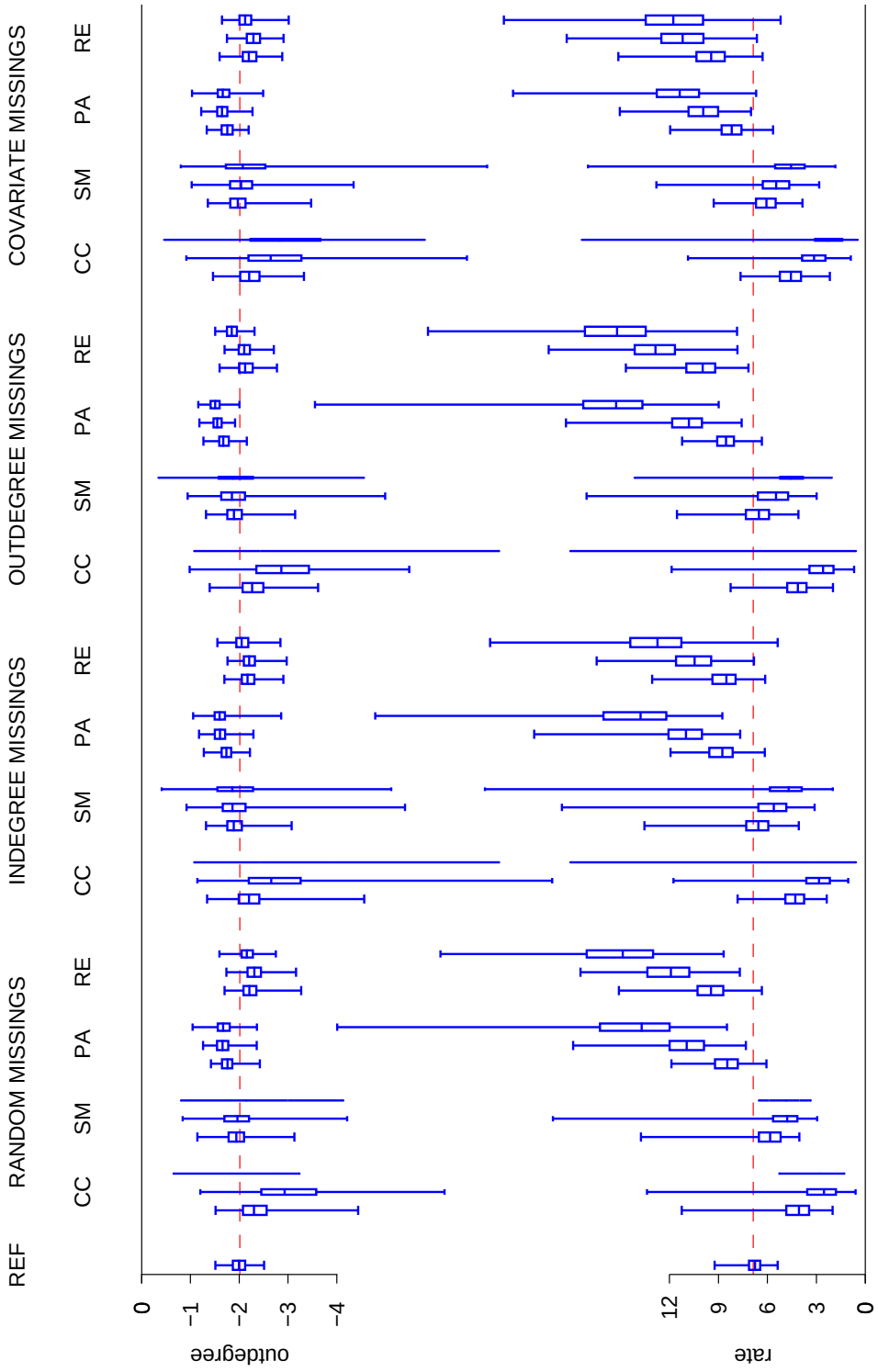


Figure 3: Boxplots of the parameters reciprocity (left) and transitivity (right). From bottom to top, first a plot for the reference category (no missings) and next, four blocks with plots for the missingness mechanisms are rendered (MCAR, Indegree, Outdegree, Alcohol). Within each block, four sets of three plots are presented for the techniques CC, SM, PA, and RE, each for the three missingness levels (0.2, 0.4, 0.6). The width of the boxes represent the number of converged projects.

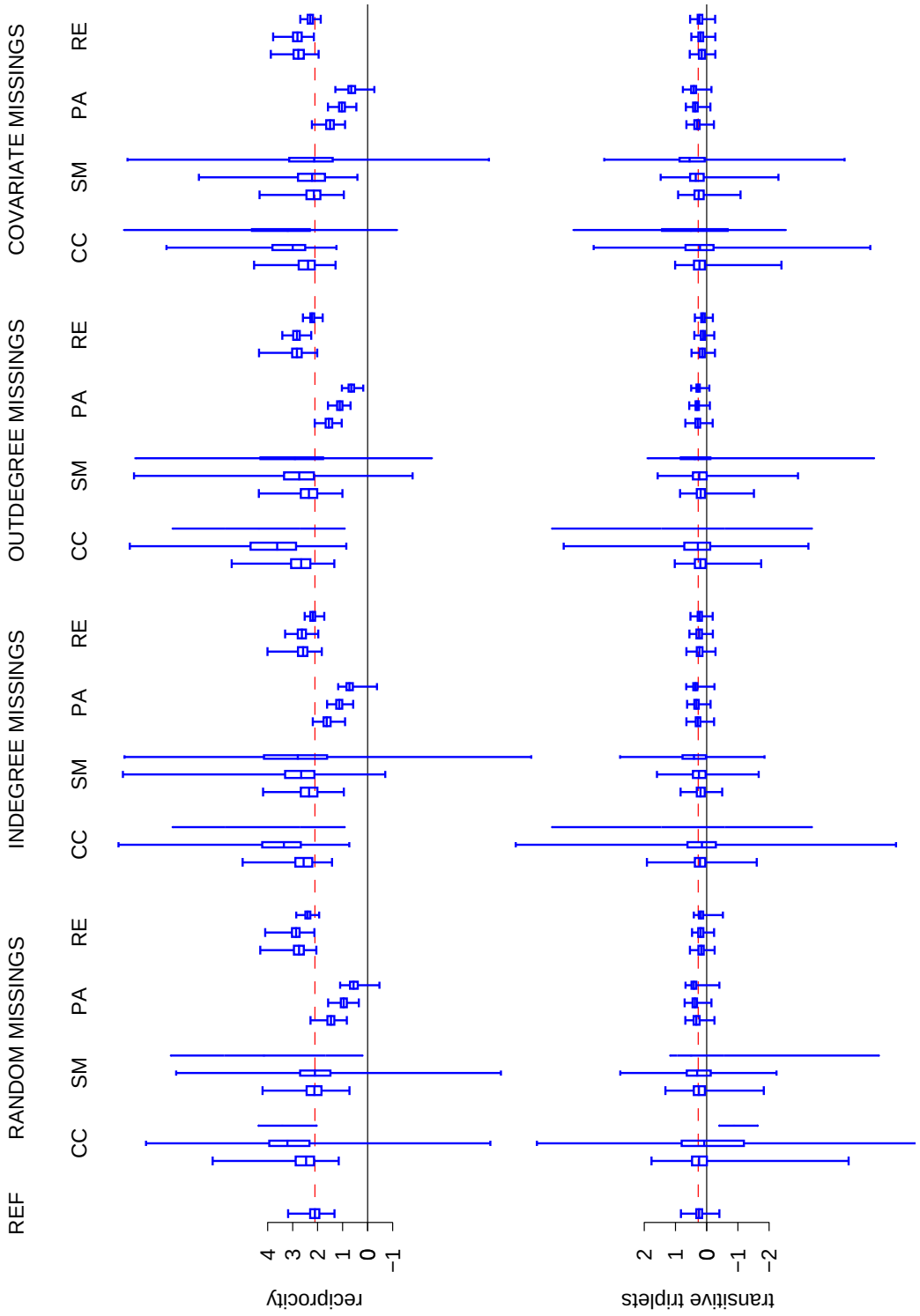


Figure 4: Boxplots for parameters distance-2 (left) and alcohol similarity (right). From bottom to top, first a plot for the reference category (no missings) and next, four blocks with plots for the missingness mechanisms are rendered (MCAR, Indegree, Outdegree, Alcohol). Within each block, four sets of three plots are presented for the techniques CC, SM, PA, and RE, each for the three missingness levels (0.2, 0.4, 0.6). The width of the boxes represent the number of converged projects.

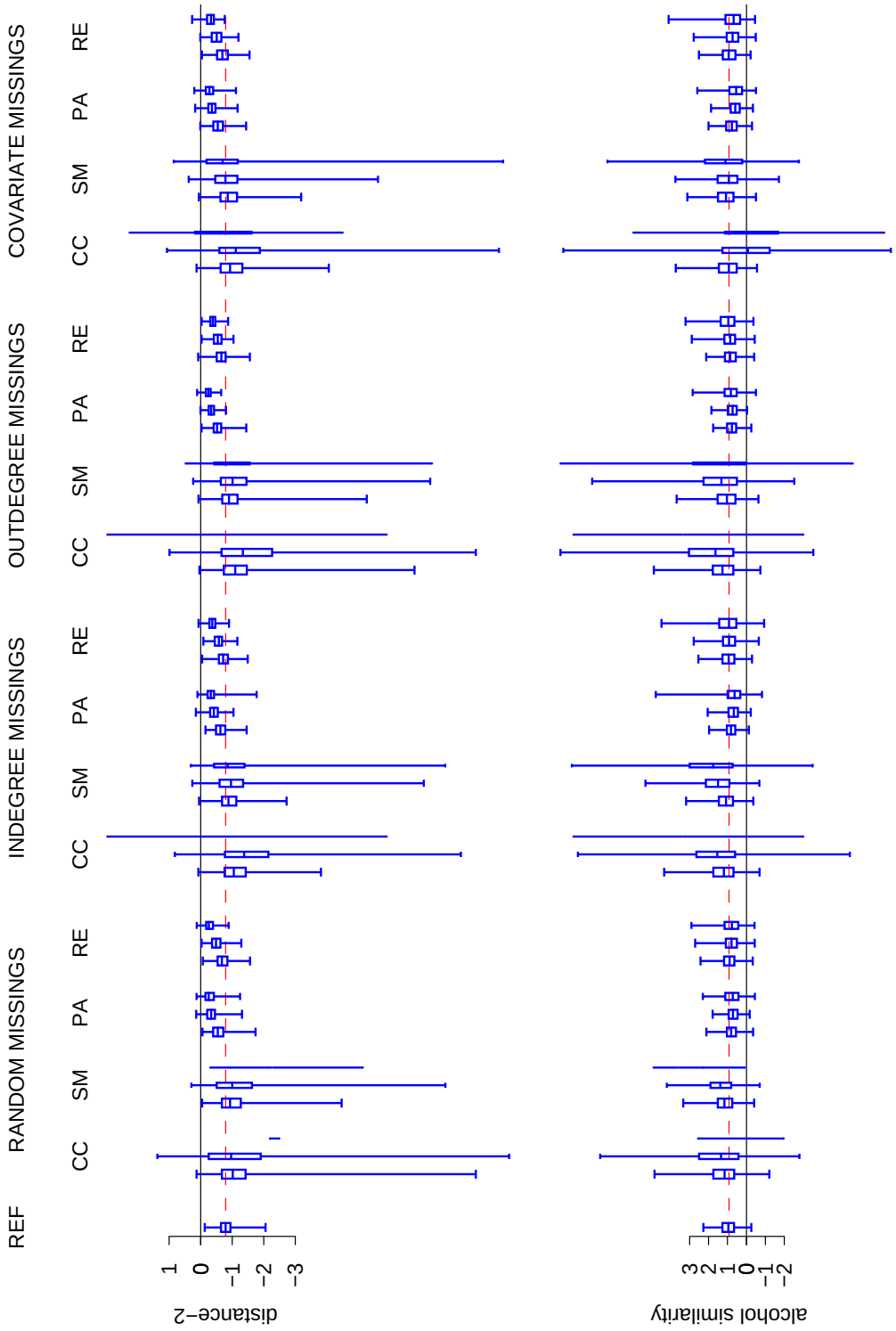


Table 3: Percentage of divergent projects per cell of the simulation design (500 projects per cell). For the reference condition without missing data, the corresponding number is 0.4%.

Missingness mechanism	Fraction missing	Method			
		CC	SM	PA	RE
MCAR	0.2	4.2	4.2	<0.1	<0.1
	0.4	40.8	33.4	4.4	0.4
	0.6	99.4	97.8	15.4	6.4
Alcohol	0.2	1.2	3.6	0.8	0.2
	0.4	35.6	16.0	2.8	0.2
	0.6	74.6	47.0	9.8	2.6
Indegree	0.2	0.6	4.0	<0.1	0.2
	0.4	29.4	18.4	2.0	0.2
	0.6	86.4	58.4	12.4	1.8
Outdegree	0.2	1.6	3.0	0.2	1.8
	0.4	21.4	21.2	1.6	0.4
	0.6	86.4	74.4	13.0	1.0

fraction of missing data. From the table it follows that indeed, the number of divergent projects increases with higher missingness fractions. This holds in the first place for the CC method, and to a slightly lesser degree for the SM method, because in these methods, the missing actors are not (or not fully) used in the estimation procedure. For the two imputation methods PA and RE, convergence is less of an issue, as the resulting data sets are essentially complete, though increasingly distorted.

Compared to completely random missings, the more systematic missingness mechanisms result in less divergent projects, due to the fact that these mechanisms leave more network structure intact. For indegree and outdegree missings, this is so because predominantly the less network-involved actors are eliminated. For alcohol missings, one needs to consider that similar alcohol consumption is a determinant of friendship selection (homophily). Elimination of predominantly non-drinkers therefore is likely to affect only a relatively self-contained non-drinkers' subnetwork, while the also relatively self-contained drinkers' subnetwork remains in the data, carrying considerable part of the original network structure.

6.2 Parameter bias

Bias in the parameter estimates can be caused by two distinct sources. On the one hand, parameters of the actor-driven models have a built-in sensitivity to network size. Coupled with the reduction of the effectively handled number of respondents in the methods CC and SM, bias can be expected. On the other hand, there can be a mismatch of the real type of missingness

Table 4: Main effects of fraction missing actors, missing data mechanism and missing data treatment: median parameter bias.

	Outd.	Recip.	Trans.	Dis. 2	Al.sim.	Rate
True value	−2.010	2.110	0.270	−0.790	0.920	6.870
Reference	0.016	−0.001	−0.026	−0.001	0.040	−0.109
$p = 0.2$	0.001	0.210	−0.037	0.049	0.015	0.588
$p = 0.4$	−0.031	0.431	−0.002	0.240	−0.073	2.473
$p = 0.6$	0.134	−0.015	0.002	0.457	−0.132	5.808
MCAR	−0.024	0.161	−0.006	0.252	−0.043	2.252
Alcohol	−0.026	0.125	0.008	0.239	−0.159	1.420
Indegree	0.040	0.151	−0.006	0.167	0.019	1.620
Outdegree	0.110	0.214	−0.059	0.230	0.002	2.198
CC	−0.368	0.647	−0.044	−0.284	0.233	−3.148
SM	0.084	0.196	−0.016	−0.108	0.251	−1.077
PA	0.358	−1.023	0.064	0.400	−0.202	3.422
RE	−0.141	0.479	−0.091	0.288	−0.064	4.278

and the assumptions about missingness made by applying the treatment method. When comparing treatment methods, both causes should be taken into account, although the extent of each separate cause here is not determined.

Table 4 shows the median biases of each parameter, separately grouped according to the different factors of our design. This way, we identify main effects of the fraction of missing actors, of the missing data mechanisms, and of the treatment methods. A positive bias indicates that the estimated parameter value is larger than the true value. For instance, at the 20% missingness level, the median estimate of the rate parameter (over all converging projects at this missingness level, regardless of treatment method and missing data type) equals $0.59 + 6.87 = 7.46$. The table shows that for the parameters distance 2, alcohol similarity and rate, bias grows with the fraction of missing actors, as was expected. At this level of aggregation, it is difficult to say why the other three parameters deviate from this pattern. By averaging over treatment methods and missingness types, several factors may be at work that cancel out on the aggregate level.

A closer look at Figure 3 reveals that the reconstruction method (RE) may be the culprit: for missingness levels of 20% and 40%, the reciprocity parameter is strongly inflated by this method. This is no surprise because at these missingness levels, the majority of imputed outgoing ties (80% and 60%, respectively) are reconstructed based on known incoming ties, and thus automatically reciprocal. For a missingness level of 60%, however, this is no

longer the case, as now only a minority of imputed ties (40%) can be reconstructed based on known incoming ties, the rest is reconstructed at random (see description of the method in Section 4). This model-induced reciprocity bias of the RE method implies that the other parameters in the model are calibrated against the artificially inflated reciprocity. The boxplots indicate such a calibration effect on the outdegree parameter. Presumably, also the transitivity parameter was affected, explaining that in Table 4, also these two parameters show a nonlinear dependency of bias on missingness level.

Inspection of the main effects of missingness mechanisms shows that the largest bias occurs when the parameter is sensitive to the mechanism. This way, bias of the alcohol similarity parameter is largest when missingness is related to alcohol consumption, and bias of the outdegree parameter is largest when missingness is related to outdegree. In all other cases, bias is largest either for random missings (MCAR, for parameters distance 2 and rate) or for missings based on outdegree (MNAR, for parameters reciprocity and transitivity). These are two extreme cases which have the largest negative effect on network structure and therefore lead to more extreme problems than the MAR cases in-between.

Finally, the most interesting of all main effects is the comparison between treatment methods. Here, the clear winner is the SIENA method (SM), which has smallest median bias for all parameters except the alcohol similarity parameter – for this one, the reconstruction method (RE) performs best. Again, the aggregate result of Table 4 in principle could be misleading. A look at the boxplot diagrams, though, gives relief here. By comparison across treatment methods, the SM boxplots are neatly centered around the true value, for all missingness types and levels. Which brings us to our third and final criterion for treatment method quality.

6.3 Centrality of true score

While in the previous section, the absolute size of the difference between true parameters and median estimates was investigated, we now address the relative size of this bias in the distribution of estimates. Looking again at Table 4, consider the median bias of the outdegree parameter under CC and under PA treatment, which are -0.368 and 0.358, respectively. The two methods thus have about the same absolute bias – however, when looking at Figure 2, it is obvious that the PA method scores much more consistently above the true value than the CC method scores below it. This difference is what we want to capture as third criterion of treatment performance: the probability to estimate the true score – or, more precisely, the probability that the estimate will surpass the true value. This can be done by studying the percentile in the distribution of estimates at which the true score is located.

In Table 5, we render it in terms of percentages relative to the median po-

Table 5: Main effects of fraction missing actors, missing data mechanism and missing data treatment: position of true value relative to the median in the distribution of the estimates; (50 + cell entry)% = percentile of the true value.

	Outd.	Recip.	Trans.	Dis. 2	Al.sim.	Rate
Reference	−4.1	+0.1	+8.1	+0.5	−3.1	+7.1
$p = 0.2$	−0.2	−9.9	+7.7	−5.5	−1.3	−7.1
$p = 0.4$	+2.0	−11.1	+0.6	−9.9	+4.6	−11.2
$p = 0.6$	−13.4	+1.1	−0.4	−39.2	+8.2	−32.4
MCAR	+2.1	−6.1	+1.0	−20.4	+3.1	−16.4
Alcohol	+2.4	−5.8	−1.8	−19.9	+10.3	−11.3
Indegree	−4.0	−8.3	+1.5	−15.1	−1.2	−15.0
Outdegree	−10.4	−10.6	+12.5	−19.1	−0.1	−15.7
CC	+32.0	+16.5	+4.2	+16.6	−8.4	+47.8
SM	−10.9	−11.8	+1.9	+9.2	−12.2	+28.9
PA	−47.2	−21.1	−21.1	−42.9	+9.0	−48.8
RE	+25.2	−44.6	+27.1	−35.8	+5.3	−49.4

sition. To come back to our example: the outdegree parameter's true value is located 32% above the median in the distribution of outdegree estimates under CC treatment. That means that it lies inside two-sided confidence intervals of confidence level $2 \times 32\% = 64\%$ or higher that can be constructed based on this distribution. For PA treatment, the true value is located 47.2% below the median, meaning that only for much higher confidence levels ($\geq 94.4\%$), the true value will be included in confidence intervals.

Concerning the main effects of missingness level and missingness mechanism, results in Table 5 differ little from what Table 4 reported. The main difference between the tables lies, as illustrated, in the comparison of treatment methods. While the imputation methods PA and RE for some parameters deliver comparatively small absolute median bias, this is not reflected in centrality of the true score in the distribution of estimates under the method. In general, these methods impose their own structure on the data – enhanced levels of reciprocity in the RE case, and enhanced popularity of few actors in the PA case – which makes it extremely difficult to recover the true parameters with any reliability. Inspection of Figures 2–4 suggests that there is no need to further refine these results. Overall, also here, SM treatment is evaluated as the best treatment method in the field, independent of missingness level or missingness type. Unexpectedly, also the CC method is somewhat rehabilitated. This treatment is second-best for five of the six parameters – so we may conclude that if it is possible to

obtain estimates by this method (which can be difficult), these at least do not depart from the true value as systematically as corresponding estimates obtained by methods PA and RE would.

7 Discussion

Missing actors have a large effect on analyzing longitudinal network data. The simulations show that ignoring the missing data and restricting the analysis to completely observed cases leads to problems when using actor-driven network evolution models. These problems are twofold. First, the reduced sample size and the loss of information leads to problems in fitting a model to the data (convergence problems). With large fractions of missing data it is hardly possible to find a fitting evolution model. Second, ignoring the missing data generally leads to biased parameter estimates.

Imputation of the missing data may solve the first problem: the data set is completed and no information seems to be missing. However, this nice feature of artificially completed data – or, as Dempster and Rubin (1983) remark, "the pleasurable state of believing that the data are complete" – has a major shortcoming: single imputation underestimates uncertainty levels, because predictions are treated as observed values and the actual sample size is overestimated (e.g., Schafer and Graham, 2002). Also, the second problem of biased parameter estimates still exists, it may even be enhanced by the imputation method when this method artificially injects network features that do not correspond to real network structure. Imputation thus has to be done in a more sophisticated way, should it result in trustworthy estimates.

A model-based approach based on the available data (e.g., ERGM-based procedures, Robins et al., 2004; Gile and Handcock, 2006; Handcock and Gile, 2007; and Koskinen, 2007) does not underestimate uncertainty levels, but uses a smaller network than originally intended. Although the reduction may not be as large as analyzing only complete actors (the method CC in the simulations), it still suffers from convergence problems. Moreover, the methods assume MAR and non-random missing data lead to biased results.

In the simulations, these shortcomings of missing data treatments were found. Imputation by preferential attachment (PA) and imputation by reconstruction (RE; Stork and Richards, 1992) lead to completed data sets at both observation moments, and hardly have any convergence problems. The parameters of the actor-driven models, however, are generally severely biased. In the boxplots of Figures 2–4, the range of estimates under imputation methods PA and RE decreases for increasing missingness fraction – creating the false impression that accuracy of estimates increases with severity of the missing data problem. What happens, though, is that the estimates are more and more determined by the artificially imputed struc-

ture, which can be seen in median bias (Table 4) being very high and the true score lying far out in the tails of the distribution of estimates (Table 5).

Handling the missing actors within the actor-driven model (SM: the SIENA method) is a model-based approach based on available data. The uncertainty levels are not underestimated, but convergence problems do arise. These problems, however, are not as large as for the complete case method, and relatively minor for small to medium fractions missing actors. SM treatment generally resulted in small biases in model parameters, especially for small to medium missingness levels. The distribution of estimates under this method does reflect the increased uncertainty due to missing data, as can be seen by an increased range of estimates for higher missing data fractions. Median bias, however, is lowest among all methods compared, both in absolute (Table 4) and relative terms (Table 5).

In this study we compared a model-based treatment of missing data within actor-driven network evolution models with complete case analysis and two naive imputation methods. In case of wave non-response there are at least two other popular approaches: weighting and imputation by last value carried forward/backward (Lepkowski, 1989). Weighting seems less suitable for network data, as weights are usually computed using selection probabilities and need auxiliary (non-network) information. Imputation by last value carried forward leads to a reduction of the amount of change between the observation moments and the imputed values do not add to the estimation of the model parameters. In this respect it is similar to the method SM, but has the shortcoming of underestimated uncertainty levels (artificially reduced range of estimates). Moreover, extra imputations are needed for actors who are missing at both observation moments.

In the case of actors joining or leaving the network, that is, missing data that emerge due to network composition changes, a model-based approach was proposed by Huisman and Snijders (2003; and implemented in the SIENA software). This specific form of wave non-response (drop-out due to leaving the network, or new entry) remains outside the scope of this investigation, as it is a qualitatively different type of missing data – if one can speak of 'missing' at all. In SIENA, it currently can be handled by modeling the joining and leaving times as exogenous events in the continuous-time Markov chain of micro steps in the actor-driven model.

¿From the simulations in this paper, it can be concluded that the model-based approach within the actor-driven models is the best method to use: parameter biases are not too large for small to medium fractions missing and standard errors are not underestimated. For small networks, though, convergence problems may arise due to the reduction of effectively handled actors. For the SIENA users, therefore the best recommendation at this stage of software development is to employ the software's own missing data treatment method, instead of mutilating the data sets by reduction to complete cases, or falsifying it by imputation of alien structure. Possible improve-

ments that need to be studied relate to the fine-tuning of this method. In this vein, better initial imputations can be used (currently, the mode is imputed) and partially observed actors or even imputed actors can be included in the estimation procedure.

References

Barabasi, A-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Borgatti, S.P., Molina, J.L., 2003. Ethnical and strategic issues in organizational social network analysis. Journal of Applied Behavioral Science 39, 337–349.

Burt, R.S., 1987. A note on missing network data in the general social survey. Social Networks 9, 63–73.

Costenbader, E., Valente, T.W., 2003. The stability of centrality measures when networks are sampled. Social Networks 25, 283–307.

De Leeuw, E.D., Hox, J.J., Huisman, M., 2003. Prevention and treatment of item nonresponse. Journal of Official Statistics 19, 153–176.

Dempster, A.P., Rubin, D.B., 1983. Overview. In: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.), Incomplete data in sample surveys, Vol. II: Theory and bibliographies. Academic Press, New York, pp. 3–10.

Gile, K., Handcock, M.S., 2006. Model-based assessment of the impact of missing data on inference for networks. CSSS Working paper no. 66, University of Washington, Seattle.
(`http://www.csss.washington.edu/Papers/wp66.pdf`)

Handcock, M.S., Gile, K., 2006. Modeling social networks with sampled or missing data. CSSS Working paper no. 75, University of Washington, Seattle.
(`http://www.csss.washington.edu/Papers/wp75.pdf`)

Huisman, M., Snijders, T.A.B., 2003. Statistical analysis of longitudinal network data with changing composition. Sociological Methods & Research 32, 253–287.

Koskinen, J., Snijders, T.A.B., 2006. Bayesian inference for dynamic network data. To be published.

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

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

Marsden, P.V., 2005. Recent developments in network measurement. In: Carrington, P.J., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge, pp. 8–30.

McFadden, D.L., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Michell, L., Amos, A., 1997. Girls, pecking order and smoking. Social Science and Medicine 44, 1861–1869.

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

Robins, G., Pattison, P., Woolcock, J., 2004. Missing data in networks: exponential random graph ($p*$) models for networks with non-respondents. Social Networks 26, 257–283.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychological Methods 7, 147–177.

Schweinberger, M., 2007. Statistical methods for studying the evolution of networks and behavior. Unpublished PhD thesis. University of Groningen, ICS, Groningen.

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: Sobel, M.E., Becker, M.P. (Eds.), Sociological Methodology. Blackwell, London, pp. 361–395.

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

Snijders, T.A.B., Steglich, C.E.G., Schweinberger, M., Huisman, M., 2007. Manual for SIENA version 3. University of Groningen, ICS, Groningen. University of Oxford, Oxford. (`http://stat.gamma.rug.nl/stocnet/`)

Steglich, C.E.G., Snijders, T.A.B., Pearson, M., 2007. Dynamic networks and behavior: Separating selection from influence. (Submitted.)

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

Stork, D., Richards, W.D., 1992. Nonrespondents in communication network studies. Group & Organization Management 17, 193–209.