

Robert W Krause

Multiple Imputation for Missing Network Data



Multiple Imputation for Missing Network Data

Robert Wilhelm Krause

© Robert W. Krause

ISBN (print):

978-94-034-1984-8

ISBN (digital):

978-94-034-1983-1

Print: Ridderprint | www.ridderprint.nl



university of
 groningen

Multiple Imputation for Missing Network Data

PhD thesis

to obtain the degree of PhD at the
 University of Groningen
 on the authority of the
 Rector Magnificus Prof. C. Wijmenga
 and in accordance with
 the decision by the College of Deans.

This thesis will be defended in public on

Thursday 19 December 2019 at 11:00 hours

by

Robert Wilhelm Krause

born on 2 December 1989
 in Witten, Germany

Supervisor

Prof. dr. T.A.B. Snijders

Co-supervisors

Dr. J.M.E. Huisman

Dr. C.E.G. Steglich

Assessment committee

Prof. dr. C.J. Albers

Prof. dr. S. van Buuren

Prof. dr. R. Veenstra

Contents

1	Introduction	1
1.1	Missing Data	1
1.2	Missing Network Data	3
1.3	Network Models	5
1.3.1	ERGMs and BERGMs	5
1.3.2	SAOMs	6
1.3.3	Multivariate Network Models	8
1.4	Multiple Imputation for Network Data	8
1.4.1	Important Existing Missing Data Treatments for Networks	9
1.5	Overview	12
2	Missing Data in Cross-Sectional Networks	15
2.1	Introduction	15
2.2	Network Analysis	16
2.2.1	ERGMs and BERGMs	17
2.3	Missing Data	18
2.3.1	Missing Data Mechanisms	18
2.3.2	Missing Data Types	19
2.3.3	Effects of Missing Data	19
2.3.4	Missing Data Treatments	20
2.4	Tested Treatments	21
2.4.1	Deletion Methods	21
2.4.2	Single Imputation	22
2.4.3	Multiple Imputation	22
2.5	Simulation Study	25
2.5.1	Network Simulation	26
2.5.2	Missing Data Creation	26
2.6	Results	27

2.6.1	Descriptive Network Statistics	27
2.6.2	Link Reconstruction	30
2.6.3	Model Parameters and Inference	33
2.7	Discussion	39
3	Missing Data in Multiplex Networks	43
3.1	Introduction	43
3.2	Bayesian ERmGMs	44
3.2.1	Bayesian inference for ERGMs	44
3.2.2	Multiplexity	45
3.2.3	Posterior Parameter Estimation for BERmGMs	45
3.2.4	Cross-Network Effects	46
3.3	Missing Data Imputation	47
3.4	Illustration - Florentine Families	49
3.5	Discussion	50
4	Missing Data in Longitudinal Networks	53
4.1	Introduction	53
4.2	Statistical Models for Network Analysis	54
4.2.1	Stochastic Actor-Oriented Models	54
4.2.2	Stationary Stochastic Actor-Oriented Models	56
4.2.3	Exponential Random Graph Models	57
4.3	Missing Data	57
4.3.1	Missing Data Mechanisms	57
4.3.2	Missing Data Types	58
4.3.3	Missing Data in Longitudinal Network Data	58
4.4	Multiple Imputation	60
4.4.1	Multiple Imputation: General Theory	61
4.4.2	Multiple Imputation: Longitudinal Network Data	62
4.4.3	Estimating Imputation Models for Multiple Waves	66
4.4.4	Multiple Groups	67
4.4.5	Multiple imputation vs. Likelihood-Based Treatment	68
4.5	Illustrative Example	68
4.5.1	Network Data	68
4.5.2	Missing Data Treatments	69
4.5.3	Results	71
4.6	Discussion	74
5	Missing Network and Attribute Data	77
5.1	Introduction	77
5.2	Coevolution SAOMs	78

5.3	Missing Data	80
5.3.1	Missing Data Mechanisms	80
5.3.2	Missing Data Types	81
5.3.3	Missing Data in SAOMs	81
5.4	Multiple Imputation with SAOMs	82
5.4.1	Imputing Behavior	82
5.4.2	MICE	83
5.4.3	Stationary SAOM Imputation	85
5.4.4	Later Waves	86
5.4.5	Multiple Groups	86
5.5	Illustrative Example	87
5.5.1	Data Description	87
5.5.2	Imputation Model	88
5.5.3	Results	89
5.6	Discussion	91
6	Extensions for Missing Network Data	93
6.1	Introduction	93
6.2	Missing Data	94
6.2.1	Missing Data Mechanisms	94
6.2.2	Missing Data Types	94
6.3	Stochastic Actor-Oriented Models and Missing Data	95
6.3.1	Stochastic Actor-Oriented Models	95
6.3.2	Stationary Stochastic Actor-Oriented Models	96
6.3.3	Missing Data in SAOMs	96
6.3.4	Multiple Imputation with SAOMs	97
6.4	Extensions	97
6.4.1	Multigroup Network Models	97
6.4.2	Multiplex Networks	98
6.4.3	Bayesian Estimation	100
6.5	Extended Multiple Imputation	101
6.5.1	First Wave	102
6.5.2	Later Waves	103
6.5.3	Obtaining Results	103
6.5.4	Network and Behavior Co-evolution	104
6.6	Illustrative Example – Friendship and Helping	105
6.6.1	Stationary SAOM imputation	106
6.6.2	Longitudinal SAOM	108
6.6.3	Results	108
6.6.4	Time Heterogeneity	115

6.7	Discussion	118
7	Conclusion and Discussion	121
7.1	Summary of the Research	121
7.2	Practical Usage of Multiple Imputation	125
7.2.1	BERGM	125
7.2.2	SAOM	125
7.3	Future Research	125
7.3.1	BERmGMs Implementation	125
7.3.2	Exponential Random Network Models	126
7.3.3	Evaluation of SAOM imputation	126
7.3.4	Sensitivity Analysis	127
7.4	Implementations	128
	Samenvatting	129
	References	135
	Acknowledgments	143
	About the author	145
	ICS dissertation series	147

Introduction

1.1 Missing Data

A problem when conducting research in the social sciences is that the object of study, usually people or organizations formed by people, is not always willing or capable to fully cooperate with the researcher, leading to no or incomplete information about the participant (or organization). Incomplete information, or missing data, are often seen as nuisance by researchers, and often treated as such, that is, missing data are mostly ignored. Participants that dropped out of the study are excluded from the analysis and participants for whom no data at all is available are, if at all, only mentioned as the overall response rate to, say, a questionnaire. This treatment, however, at best only lowers the power of the statistical analysis and at worst introduces biases into the results.

Several options for treating missing data are available when treating missing data researchers need to consider the unknown missing data mechanism. Missing data mechanisms describe the probability distribution of the missingness. Following the framework defined by Rubin (1976), there are three types of missing data mechanisms. Data are missing completely at random (MCAR) if the probability of a value to be missing is independent of the observed data and the value of the missing data. Data are missing at random (MAR) if the probability to be missing is independent of the missing value itself, but is related to other observed variables (e.g., older participants are less likely to fill out parts of the survey). These two cases are often summarized as ignorable missing data in the survey research setting, because, given that proper missing data techniques are applied, they will yield no bias in a resulting analysis. Lastly, data are

Parts of this chapter are based on Krause et al. (2018b,a).

missing not at random (MNAR) if the missingness is dependent on unknown missing values (e.g., high income participants are less likely to provide information about their income). Data missing not at random will lead to biased results and are therefore called non-ignorable.

Researchers have several options for handling missing data. These options can broadly be separated into three categories¹: deletion, likelihood-based estimation, and imputation (for a general overview of missing data handling see Schafer and Graham, 2002). Deletion methods reduce the data to a fully observed subsample. In the case of listwise deletion, the same fully observed subset is used for all statistical calculations (i.e., every participant with any missing data is removed from the data set). In pairwise deletion different fully observed subsets are used for each statistical analysis. Deletion methods are commonly used and the default for most statistical programs, because they are straightforward in their application and explanation.

To avoid loss in statistical power, researcher sometimes recruit more participants, until the desired sample size is obtained. In some cases this is easily feasible, only requiring some minor investments in recruiting new participants. In other cases, though possible, recruiting more participants can become very expensive (e.g., medical trials or neuroscientific studies), or very difficult (e.g., studies of rare diseases or disorders, indigenous secluded people, or high profile organizations). For other studies it can, however, be impossible to recruit new people. In, for instance, a study following a cohort of people over multiple years (e.g., Dijkstra et al., 2015) one cannot simply add new people and inquire retrospectively about experiences and contacts they had years, or even decades, ago, at least not with any reliability comparable to the data collected in the original sample.

If the missing data are MAR or MNAR, deletion methods will likely introduce bias into the analysis. Likelihood-based methods, however are capable of obtaining approximately unbiased estimates in larger samples under MAR (Schafer and Graham, 2002). The marginal distribution of the observed data provides the correct likelihood of the unknown model parameters θ , if the model is a realistic model of the complete data and the data are missing (completely) at random (Schafer and Graham, 2002).

Imputation methods replace the missing values with plausible guesses (Rubin, 1987; Schafer and Graham, 2002). The methods differ in the amount of information they take into account and how this information is used for the replacement

¹A fourth category, re-weighting, is not applicable to network research because of the strong dependencies inherent to network data. This thesis focuses primarily on missing data in the context of network data, and thus, re-weighting of cases will not be discussed.

of the missing values. Stochastic imputation methods use draws from probability distributions to replace missing values. These methods can be used for multiple imputation, where missing values are imputed multiple times based on a conditional probability model. The obtained imputed data sets are analyzed separately leading to a distribution of model parameters. These are then combined to obtain parameter estimates and standard errors. For the calculation of the standard errors both within and between imputation variance are combined. This allows to take the uncertainty about the missing data imputation into account for the estimation of standard errors.

Both single and multiple imputation allow model estimation using all observed information and the calculation of descriptive statistics. While both can provide (on average) unbiased parameter estimation under MCAR and MAR, only multiple imputation is able to provide unbiased standard error estimates given that a correct model is used for the imputation. For non-network data, likelihood-based estimation and multiple imputation are considered the state of the art (Schafer and Graham, 2002).

1.2 Missing Network Data

While there has been extensive research on missing data treatments for panel data (for an overview see Schafer and Graham, 2002), missing data treatments for network data have been far less studied (for an overview on missing data treatments for network data see Huisman and Krause, 2017). A network here constitutes a set of nodes (or actors) and their connections, usually expressed as the random $n \times n$ adjacency matrix x with $x_{ij} = 1$ when there is a tie from node i to node j and $x_{ij} = 0$ when there is no tie². Edges connecting nodes to themselves are usually not allowed ($x_{ii} = 0$). The networks can be directed or undirected (in the latter case $x_{ij} = x_{ji}$). These networks can constitute friendships in a classroom, collaborations between work colleagues, money transfers between banks, or treaties between countries. For an introduction into network analysis see Wasserman and Faust (1994) or Robins (2015).

The effects of missing data on descriptive network statistics depend on the amount of missing data, on the structure of the network, on the descriptive statistic in question, and how the missing data are treated. Note that there is no effect of missing data without the effect of a missing data treatment. Researchers always have to make a decision about missing data. The default

²Many authors in the ERGM literature use y to denote the network, while x is standard in the SAOM literature. For consistency we use x throughout this thesis.

treatments for networks are listwise or pairwise deletion, or imputation of unconditional means, meaning imputation of no-ties (zeros), as most social structures are sparse (density $< .5$) and no-tie being the most likely value. Note that listwise deletion means the complete removal of one or more nodes from the network, including all their outgoing and incoming ties. For these treatments some combinations of statistic and overall network structure are more robust to missingness than others. Larger and more centralized networks are usually more robust against missing data (Smith and Moody, 2013). Measures based on indegree are found to be overall more reliable (Costenbader and Valente, 2003; Smith and Moody, 2013; Smith et al., 2017). A notable difference between network and non-network data can be seen under the MCAR mechanism. While sample estimates of means, variances and model parameters are usually unbiased for non-network data under MCAR with listwise deletion, the same does not apply to network data. There can be considerable biases, even if data are missing completely at random, in descriptive statistics or estimated model parameters of statistical models (Huisman and Steglich, 2008; Smith and Moody, 2013; Huisman, 2009).

Likelihood-based estimation methods are available for various families of network models. For the exponential random graph family see Robins et al. (2004), Gile and Handcock (2006), Handcock and Gile (2007, 2010), Koskinen et al. (2010, 2013). For the family of stochastic actor oriented models see Snijders et al. (2010a) and Snijders (2017a). These methods are by definition model-based, and thus cannot aid the estimation of other network models (e.g., block-models) or the calculation of descriptive statistics.

Both, single and multiple, imputation procedures are available for networks. The properties of single imputations have been extensively studied (e.g., Stork and Richards, 1992; Huisman, 2009; Žnidaršič et al., 2012), and found to provide overall only small improvements to deletion methods, if any, and in some cases they introduce severe biases. Multiple imputation methods for networks are relatively new and only available for the exponential random graph model family of network models (Koskinen et al., 2010; Wang et al., 2016). They have, as of yet, not been systematically studied, and multiple imputation procedures for longitudinal network data (models) have not yet been developed.

The problem of missing network data becomes a double-edged sword when likelihood-based methods or multiple imputation are used to treat it. A peculiar feature that distinguishes missing data in network studies where the network nodes are individual persons who provide information about their outgoing relations from missing data in non-network studies is best highlighted in the case of unit non-response, where no information is provided by some participants. On

the one hand, missing data, that is, missing outgoing network nominations, do not only constitute missing data for the non-responding participants, but they also constitute missing data for the incoming ties of some (in case of partial non-response) or all (in case of complete non-response) other members of the network. The true indegree becomes unknown, after all, the non-responding actors could have send ties to the observed actors. This makes missing data in the network setting seemingly more severe, and induces biases in some measures even under MCAR.

On the other hand, missing data can be, potentially, better salvaged in networks. For undirected networks it is sufficient if only one side provides information about the relation (if $x_{ij} = 1$ then $x_{ji} = 1$). In such cases, missing data only occur for the relation between two missing actors, or, if for legal, ethical, or methodological reasons, information can only be used if both sides provide an observation about the relationship. Directed networks, however, do not have this straightforward solution for missing data. Still, if some members of the network do not provide any information about their contacts (no outgoing ties are observed), there is information about these missing participants, because others in the network could provide information about their relation to the missing actors (incoming ties are observed). Unlike for regular panel data, the participants in a network are not randomly sampled and independent of each other. Their inter-dependencies constitute the subject of the analysis and can be leveraged to better handle missing data. Thus, also for directed networks, complete non-response by some members of the network does not mean that no information is available about them, which would be the case in non-network data.

In this thesis, we will systematically analyze the most prominent existing missing data treatments for networks, extend multiple imputation for missing network data to multiplex network structures, longitudinal network data, and actor attributes. To do so we will rely on two generative network models, Exponential Random Graph Models (ERGMs; Frank and Strauss, 1986; Wasserman and Pattison, 1996; Robins et al., 2007; Lusher et al., 2013) and Stochastic Actor-oriented Models (SAOMs; Snijders, 1996, 2001, 2005, 2017b).

1.3 Network Models

1.3.1 ERGMs and BERGMs

ERGMs (Exponential Random Graph Models) are probability models for cross-sectional network data (for longitudinal versions of ERGMs see Hanneke et al.,

2010; Koskinen et al., 2015) where the probabilities depend on the frequency of occurrence of substructures in the network such as subgraph counts, or other statistics. Network structures are highly dependent upon each other, therefore testing hypotheses about structural properties of a network (e.g., girls are more likely to form cliques than boys) require to also model other network properties (e.g., the general tendency to form friendships, the gender specific tendencies to send and receive ties). A sophisticated approach is needed because the dependencies between nodes and ties need to be taken into account. Let \mathbf{X} denote the set of all possible networks on n nodes and let x be a realization of the random network X . ERGMs represent the probability distribution density of X as

$$P(X = x|\theta) = \frac{\exp[\theta^T s(x)]}{z(\theta)}, \quad (1.1)$$

with θ being a vector of model parameters, $s(x)$ a vector of corresponding sufficient statistics (e.g., number of edges or number of reciprocated ties) and $z(\theta)$ the normalizing constant. The normalizing constant is very difficult to calculate or even intractable in moderate to large graphs. For an introduction into ERGMs see Lusher et al. (2013).

Bayesian estimation of ERGMs (BERGMs) was introduced by Caimo and Friel (Caimo and Friel, 2011). The BERGM samples from the following probability distribution:

$$p(\theta', x', \theta|x) \propto p(x|\theta) \pi(\theta) \epsilon(\theta'|\theta) p(x'|\theta'), \quad (1.2)$$

in which θ' are proposed parameters and x' are networks simulated with these proposed parameters, $p(x'|\theta')$ is the likelihood on which the simulated data x' are defined and belongs to the same exponential family of densities as $p(x|\theta)$, $\epsilon(\theta'|\theta)$ is any arbitrary proposal distribution for the parameter θ' , and π is the prior probability density function of θ . This method employs auxiliary variables θ' and x' , which turns out to be helpful for dealing with the intractable normalizing constant in the estimation process. The proposal distribution is set to be a normal centered at θ . The marginal distribution of θ in the Metropolis-Hastings algorithm is the posterior distribution from which inference is drawn, which can be obtained after integrating out x' and θ' . ERGMs and BERGMs are discussed in more detail in Chapters 2 and 3.

1.3.2 SAOMs

SAOMs (Stochastic Actor-oriented Models) are stochastic network models developed for modeling the (unobserved) change processes between two (or more)

observed time points in a network and potentially co-evolving behavior variables or co-evolving networks. A key assumption of the SAOM is that the change between the observed network at time points m and $m + 1$ can be decomposed into multiple small steps. Let $x(m)$ be the observation of network x at wave m . Not all tie variables change at once between the observations, but the tie variables change in small steps (so-called mini steps) one after the other. Most often this chain of changes is not observed, making it impossible to easily estimate a model for the observed change (for SAOMs for data with fully observed chains of mini steps see Stadtfeld et al., 2017). SAOM estimation solves this problem via simulation. During the estimation hundreds, or thousands, of potential network evolution processes are simulated, each consisting of a series of small changes. Hence the name SIENA for the software to estimate SAOMs – Simulation Investigation of Empirical Network Analysis; **RSiena** is a contributed package (Ripley et al., 2017) to the statistical system R (R Core Team, 2019).

These evolution processes are modeled by two functions. The rate function determines which actor makes a decision, and when, according to an exponential model for waiting times; and the objective function models which decision is made by the chosen actor according to a multinomial logit discrete choice model. The rate function assigns waiting times to all actors. Then the shortest waiting time is chosen and the actor has the chance to either drop one of its existing outgoing ties, create a new tie to a yet unconnected actor, or do nothing and let the network remain as it is, resulting in n possible choices. The probability for each of the possible actor decisions is determined by an objective function, in which actor-specific network statistics (including effects of covariates) $s_{ki}(x)$ are weighted with parameters of the network evolution θ_k , given the current state of the network x ,

$$f_i(\theta, x) = \sum_k \theta_k s_{ki}(x). \quad (1.3)$$

The network statistics $s_{ki}(x)$ can be, for instance, 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 change 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. Second, the possible states of the network, and thus the possible transitions between two network observations, are far too numerous

– a binary network of only 30 actors already has $2^{30^2-30} = 7.9 \times 10^{261}$ possible states. However, the estimation via simulation allows to avoid these problems.

Although SAOMs are primarily used for longitudinal data, a cross-sectional variant also exists. Here, it is assumed that the observed network is the outcome of a continuous, stationary process in (at least short-term) equilibrium. The model assumes that the observed network statistics $s(x)$ are stochastically stable (e.g., the number of ties, the number of reciprocated ties, or the number of triangles). In the estimation procedure, however, actors are allowed to change their relations and the objective function is estimated such that the network statistics $s(x)$ remain overall stable. Stationary SAOMs can be estimated using the observed network as both starting and end network for the stationary distribution (reflecting that the network statistics remain constant). This means that a rate function cannot be estimated, because no change is observed in the network. Estimation requires that the rate function is fixed to an arbitrarily large value. For a more detailed introduction into stationary SAOMs see Snijders and Steglich (2015).

SAOMs and their extensions are discussed in more detail in Chapters 4, 5, and 6.

1.3.3 Multivariate Network Models

Network structures are often not studied in isolation, but together with other dependent variables. These can either be other network relations (e.g., friendship and gossip; Ellwardt et al., 2012) on the same set of actors, so-called multiplex networks, or node-level variables (attributes). Multiplex network structures in relation to ERGMs will be discussed in Chapter 3, multiplex SAOMs in Chapter 6. The co-evolution of networks and attributes (in this context usually called behaviors) will be discussed in Chapters 4 and 5.

1.4 Multiple Imputation for Network Data

In this thesis, we address the development, implementation, and evaluation of multiple imputation algorithms for network data. Multiple imputation methods for non-network data are not applicable for handling missing network data, because they rely on the independence between observations. Thus, imputation methods built on generative network models should be used to properly maintain the structure of the network. The main ingredients for these models have already been provided by Handcock and Gile (2007), Koskinen et al. (2010), Snijders et al. (2010a), Hipp et al. (2015), Wang et al. (2016), and Snijders (2017a).

1.4.1 Important Existing Missing Data Treatments for Networks

ERGM Family

Handcock and Gile (2007) showed how to estimate ERGMs on missing network data using a model-based missing data treatment. Their procedure is implemented in the `ergm` package (Handcock et al., 2007) for R (R Core Team, 2019). The implemented algorithm allows for unbiased estimation of ERGMs under missing data, if the data are missing at random and the chosen model is well fitting. Wang et al. (2016) proposed to utilize the model-based estimation for imputation in two steps. First an ERGM is estimated on the network with missing data. Second, the estimated model is used to simulate the missing network data, conditional on the observed data. Their proposed imputation algorithm has the caveat that all imputations are simulated using the same parameters. The imputations thus do not take the uncertainty around the imputation parameter into account, which is required for multiple imputation parameters to be considered *proper* in the sense of Rubin (1987).

Koskinen et al. (2010) provide an algorithm capable of obtaining proper multiple imputed network data using Bayesian ERGMs. The proposed algorithm imputes the missing network during the estimation. In short, Bayesian estimation of ERGMs is an iterative process following three steps at each iteration. First, parameters for the network structure are proposed. Second, the proposed parameters are used to simulate networks and calculate a set of sufficient network statistics. Third, the statistics calculated on the simulated data are compared to the observed statistics, and the parameter is accepted with a probability dependent on the difference between the observed and simulated statistics, with parameters leading to simulations closer to the observed data having a higher probability of being accepted. Koskinen et al. (2010) added an additional step to this estimation procedure, in which each time, after a proposed parameter is accepted, the parameter is used to obtain an imputation of the network, similar to the imputation procedure proposed by Wang et al. (2016). The imputed network is then used as the new comparison in the estimation procedure, that is, in the next iteration the network statistics calculated on the simulated networks, are compared to those of the imputed network. If a new parameter is accepted, a new imputation is formed and passed on to the next iteration.

This procedure is primarily used as a model-based estimation procedure, as it allows unbiased estimation of Bayesian ERGMs under missing data. However, it is possible to retain the imputed networks and use them for further analysis (e.g., blockmodels). The imputed networks further are *proper* in the sense of Rubin (1987), because each imputation is drawn using a different vector of

parameters from the estimated posterior distribution of the parameters. In the case of Bayesian ERGM estimation, model-based estimation and multiple imputation are thus the same procedure. The algorithm will be explained in further detail in Chapter 2.

SAOM Family

The default treatment of missing data using SAOMs depends on the chosen estimation algorithm. The two most important estimation options are method of moments and maximum likelihood estimation. The default method for SAOM estimation is method of moments (MoM; Bowman and Shenton, 1985; Snijders, 2001), that is, parameters for the network evolution from time point $m - 1$ to m are estimated such that for a vector of target statistics corresponding to the model parameters the expected value, approximated by simulation with these parameters, is equal to the observed values of the target statistics at time point m .

Another way of estimating parameters and simulating networks is by maximum likelihood (ML; Snijders et al., 2010a). ML estimation maximizes the likelihood for the estimated set of parameters to link two consecutive observation waves, $x(m - 1)$ to $x(m)$. This means that ML simulation always ends in the observed network $x(m)$ and the exact target statistics, whereas networks simulated by the method of moments procedure lead to a distribution of networks, which is on average similar to the observed network on the target statistics.

For estimating SAOMs under missing data, 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 by the model as given. Therefore it is necessary to impute data in the first wave to provide a starting point for simulations.

Handling missingness in consecutive waves differs depending on the estimation procedure used in the *RSiena* software (Ripley et al., 2017). For the 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 tie variables. 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 tie variables in consecutive waves are imputed by last value carried forward (Lepkowski, 1987). 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 some MNAR situations, and it is superior to other simple imputation methods (Huisman and Steglich, 2008).

If 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 (2017a). Using ML, the chain of mini steps between two waves 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, missing data are imputed also in a model-based way, where the prior distribution for the unobserved tie indicators at time $m - 1$ is defined as independent binary variables with the observed density as the tie probability. Given all observed variables at times $m - 1$ and m and this imputation of the first wave, the chains are simulated, which leads to 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 tie variables 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 bias in the parameter estimation.

It should be noted that in the ML estimation in **RSiena** for $M \geq 3$ waves, all $M - 1$ periods from time $m - 1$ to m are treated separately. For example, when analyzing $M = 3$ waves, missing tie variables 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. 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.

Two alternatives to the default treatment have been proposed. The first option, proposed by Hipp et al. (2015), is an extension to this default procedure, in which the first wave missing data is imputed using ERGM imputation as introduced by Wang et al. (2016), and consecutive waves are treated by the above described default (MoM) procedure. Missing data in later waves remains untreated, and thus this procedure provides only limited help if many waves are collected. Only the estimation of the first period ($m = 1$ to $m = 2$) profits from the imputation.

The second proposed treatment tackles the convergence problems that can occur under missing data. The default missing data treatment can, in cases with high missing data, fail to converge (within reasonable time). Therefore, de la Haye et al. (2017) propose a pairwise deletion procedure in which for each analyzed period $m-1$ to m only fully observed actors are used. This procedure, however, might lead to biased results. It has been shown by studies investigating the effects of missing data on network structures that deletion methods distort the network structure, even when data are missing completely at random (e.g., Huisman and Steglich, 2008; Smith and Moody, 2013; Huisman, 2009). This might lead to biased target statistics in the estimation procedure. However, in some cases, such a procedure might be helpful to obtain any converged model. In this thesis, we will implement, extend, and test the existing procedures for ERGMs and SAOMs to obtain proper multiple imputation of missing network data.

1.5 Overview

The following chapters can be broadly separated into two groups. Chapters 2 and 3 focus on missing data in cross-sectional network studies and discuss missing data in the context of the ERGM family. Chapters 4, 5, and 6 introduce a new multiple imputation procedure for longitudinal network data within the SAOM family. Chapter 7 presents the summary and conclusions.

Chapter 2 compares several missing data treatment methods for missing network data on a diverse set of simulated networks under several missing data mechanisms. We focus the comparison on three different sets of outcomes: descriptive statistics, link reconstruction, and model parameters. The chapter focuses primarily on multiple imputation using Bayesian Exponential Random Graph Models. This chapter is based on, and an extension to Krause et al. (2018b).

Chapter 3 presents an estimation algorithm for Bayesian Exponential Random multiplex Graphs Models (BERmGMs) under missing network data. The BERmGM is an extension of the ERGM family for multiplex network data, that is, networks where multiple types of relations (e.g., friendship and advice seeking) are observed on the same set of nodes. The new model is implemented in R (R Core Team, 2019)³, an open source software environment for statistical computing. The model is tested on a small network. This chapter is based on Krause and Caimo (2019).

³The model has not yet been implemented in an R package.

Chapter 4 introduces 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 (BERGM) or a stationary SAOM and imputes missing tie variables in later waves utilizing a longitudinal SAOM. The proposed method is compared to the standard SAOM missing data treatment as well as recently proposed methods. The chapter is based on Krause et al. (2018a).

Chapter 5 extends the multiple imputation procedure for SAOMs introduced in Chapter 4 to the case of network and behavior co-evolution. This extension provides joint multiple imputation of both behavior and network, maintaining the relationship between the variables. The method is demonstrated on the example of the coevolution of a friendship network with alcohol drinking and tobacco smoking (Pearson and West, 2003).

Chapter 6 gives an additional extension to the multiple imputation procedure for SAOMs introduced in Chapter 4, that is, an extension for multiplex networks. It further details how to analyze multiple groups, and provides an imputation algorithm based on Bayesian estimation of SAOMs. The extended algorithm is applied to an empirical study, analyzing the coevolution of friendship and helping in 41 classrooms (van Rijsewijk et al., 2019).

Missing Data in Cross-Sectional Networks

An Extensive Comparison of Missing Data Treatment Methods

2.1 Introduction

Previous work has established the detrimental effects of missing network data for studies of network structures (Costenbader and Valente, 2003; Kossinets, 2006; Huisman, 2009). The problem is often more severe than in non-network research, because the refusal of one member of the network to participate will automatically lead to missing data for all members of the network due to the strong dependencies within the network structure. When participants provide information about their outgoing links, they also provide information about the incoming links of other members of the network. If it is not possible to obtain the missing information in some other way (e.g., approach the missing participant), then network researchers either have to find a way to handle the missing data, or start collecting an entirely new network. The problem is simpler for non-network studies, as one can generally reach a complete data set of the desired sample size by simply recruiting new participants.

The effects of missing data on network structure and analysis and the investigation into treatment procedures constitute an ongoing field of research (de la Haye et al., 2017; Smith et al., 2017; Krause et al., 2018a; Huang et al., 2019; Krause and Caimo, 2019)¹. Missing data treatments for networks range from simple deletion procedures and ad hoc imputations of the missing tie variables,

This chapter is based on Krause et al. (2018b) with major extensions.

¹Krause and Caimo (2019) and Krause et al. (2018a) constitute Chapters 3 and 4 of this dissertation.

to complex multiple imputation models and model based procedures for estimation of model parameters (for an overview of missing data imputation methods in networks see Huisman and Krause, 2017). In this study, we compare various techniques in their ability to capture key network level characteristics, how well they are able to reconstruct ties correctly, and how they perform in regard to model parameters and inference. The methods (deletion, single imputation and multiple imputation using ERGMs and Bayesian ERGMs) are compared with respect to their performance on a diverse set of simulated networks. A short version of this paper focusing only on descriptive statistics was published in the proceedings of the ASONAM conference 2018 (Krause et al., 2018b). This extended version includes more missing data treatment methods and compared the techniques in their ability to reconstruct links correctly and estimate model parameters reliably. Previous work on missing data in networks either focused on the comparison of simple treatments under various conditions (e.g., Smith et al., 2017; Huang et al., 2019), or on the introduction of advanced treatments (e.g., Koskinen et al., 2010; Wang et al., 2016). To our knowledge this study is the first to compare both simple and advanced treatment methods under a variety of conditions.

The paper is organized as follows. In Section 2.2, we briefly introduce the exponential random graph model family, which is fundamental for our advanced imputation method. In Section 2.3, we describe the non-response problem and its specifics for missing data in networks. Section 2.4 introduces the tested missing data treatment methods. We continue with a description of the simulation study in Section 2.5. In Section 2.6 we present the results on descriptive network statistics, link reconstruction, and model parameters and inference. We close the paper with a discussion of the findings and corresponding recommendations.

2.2 Network Analysis

The most common model family used to analyze the structure of cross-sectional social networks in the social sciences is the exponential random graph model (ERGM; Frank and Strauss, 1986; Wasserman and Pattison, 1996; Robins et al., 2007; Lusher et al., 2013). We start with introducing this model family for three reasons. First, we will test the performance of the treatment methods on their ability to retain similar ERGM estimates for models estimated on the complete data. Second, sophisticated missing data treatments rely on generative models of the data. In the case of network data this requires a network generative model, like ERGM. Lastly, we used ERGMs to simulate the networks used to test the performance of different treatments.

2.2.1 ERGMs and BERGMs

ERGMs are probability models for networks where the probabilities depend on the frequency of occurrence of substructures in the network such as subgraph counts, or other statistics. Network structures are highly dependent upon each other, therefore testing hypotheses about structural properties of a network (e.g., girls are more likely to form cliques than boys) require to also model other network properties (e.g., the general tendency to form friendships, the gender specific tendencies to send and receive ties). A sophisticated approach is needed because the dependencies between nodes and ties need to be taken into account. Networks can be expressed by the random $n \times n$ adjacency matrix x with $x_{ij} = 1$ when there is a tie from node i to node j and $x_{ij} = 0$ when there is no tie. Edges connecting nodes to themselves are usually not allowed ($x_{ii} = 0$). The networks can be directed or undirected (in the latter case $x_{ij} = x_{ji}$). Let \mathbf{X} denote the set of all possible networks on n nodes and let x be a realization of the random network X . ERGMs represent the probability distribution density of X as

$$P(X = x|\theta) = \frac{\exp[\theta^T s(x)]}{z(\theta)}, \quad (2.1)$$

with θ being a vector of model parameters, $s(x)$ a vector of corresponding sufficient statistics (e.g., number of edges or number of reciprocated ties) and $z(\theta)$ the normalizing constant. The normalizing constant is very difficult to calculate or even intractable in moderate to large graphs. Therefore, ERGMs are usually estimated via simulation. These simulation consist of iterations of swaps of single ties ($x_{ij} = 1$ to $x_{ij} = 0$ or vice versa), conditional on the rest of the network. Tie swaps can be made according to Gibbs or Metropolis-Hastings sampling (Lusher et al., 2013). For an introduction into ERGMs see Lusher et al. (2013).

Bayesian estimation of ERGMs (BERGMs) was introduced by Caimo and Friel (Caimo and Friel, 2011). The posterior conditional probability is given by

$$P(\theta|x) = \frac{\exp[\theta^T s(x)] \pi(\theta)}{z(\theta) q(x)}, \quad (2.2)$$

where $\pi(\theta)$ is the prior density of the parameters and $q(x)$ is the marginal probability function of the observed graph. For an introduction into BERGMs see Caimo and Friel (2011), we will elaborate the estimation algorithm of BERGMs later in more detail, as it is integral to one of the treatment methods. We also include Bayesian estimation in this study, as it has several advantages in the treatment of missing data, which we discuss below.

2.3 Missing Data

Let I be the indicator matrix of whether a tie variable is observed or missing, with $I_{ij} = 1$ if x_{ij} is observed and $I_{ij} = 0$ if x_{ij} is missing. Further we use the convention that u represents the observed part of the data ($I_{ij} = 1$) and v represents the unobserved part of the data ($I_{ij} = 0$). Thus the network x can be reassembled from u and v . With the given network we can define an observation model for I , $f(I | x, \zeta)$, which is a probability model for what is observed and what is not, depending on the network x and some statistical parameter ζ .

2.3.1 Missing Data Mechanisms

For an appropriate treatment of missing data in statistical modeling, Rubin (1976) made it clear that it is of fundamental importance to consider the probability distribution of the missingness. He defined three types of mechanisms for this probability distribution, which can be translated to the network data context (Huisman and Steglich, 2008). First, data are missing completely at random (MCAR) if the probability of it to be missing is independent of any observed variable and also independent of the missing value itself, $f(I | u, v, \zeta) = f(I | \zeta)$. A special case of MCAR can arise when survey methods set a limit to the outdegree of a node (e.g., by asking to name three friends in your class). Any respondent giving the maximum allowed answer has, strictly speaking, missing data on all other outgoing ties, because the respondent might have nominated them if they had been allowed to. This is usually disregarded by researchers, and the remaining ties are set to no-ties.

Second, data are called missing at random (MAR) if the probability of being missing is independent of the missing value but is dependent on other observed variables (e.g., men are less likely to fill out the network questionnaire, assuming gender is a completely observed attribute), $f(I | u, v, \zeta) = f(I | u, \zeta)$. For non-network data, treatment methods have been developed which yield unbiased estimates under these two mechanisms (for an overview see Schafer and Graham, 2002).

The third mechanism is data missing not at random (MNAR). Data are MNAR if the probability of being missing is related to the missing value itself (e.g., isolates are less likely to participate in a network study), $f(I | u, v, \zeta)$. Missing data related to specific tie variables can follow complex patterns. For instance, i 's probability to drop out of the study can be related to nodal attributes of specific alters j with attribute k_j (e.g., being linked to someone who is not

participating might increase the probability for drop out). Missing data mechanisms may also be related to structural embeddedness (e.g., being in a triad makes missing participants less likely to participate). In both examples the probability of a tie variable being missing depends both on the tie variable but also on other (tie) variables.

This study will incorporate examples of all three missing data mechanisms.

2.3.2 Missing Data Types

While missing data mechanisms describe the probability distribution of the missing data, missing data types describe how the missingness is spread over the network. In cross-sectional network research two types of missing data can be distinguished: actor non-response and tie non-response (Huisman and Steglich, 2008). Actor non-response occurs if all outgoing tie variables of an actor are missing, $\sum_{j=1}^n I_{ij} = n-1$. In tie non-response only some, but not all tie variables of an actor are missing, $0 < \sum_{j=1}^n I_{ij} < n$. The terminology of ‘non-response’ implies that data is collected via self-reports of network actors and stems from classical survey research. With self-reports actor non-response is the most likely type of missing data distribution. However, other data collection methods, for instance link tracing or snowball sampling, might lead more often to item non-response. This study will focus only on actor non-response. The findings should also generalize to tie non-response, as this retains more information per actor and is thus less severe than actor non-response.

2.3.3 Effects of Missing Data

The effects of missing data on descriptive network statistics depend on the amount of missing data, on the network structure, on the descriptive statistic in question, and how the missing data is treated. Note that there is no effect of missing data without the effect of a missing data treatment. Researchers always have to make a decision about missing data. The default treatments for networks are listwise or pairwise deletion, or imputation of unconditional means, meaning imputation of no-ties, as most social structures are sparse (density $< .5$) and no-tie being the most likely value. For these treatments some combinations of statistic and overall network structure are more robust to missingness than others. Larger and more centralized networks are usually more robust against missing data (Smith and Moody, 2013). Measures based on indegree are found to be overall more reliable (Costenbader and Valente, 2003; Smith and Moody, 2013; Smith et al., 2017). A notable difference between network

and non-network data can be seen under the MCAR mechanism. While sample estimates of means, variances and model parameters are usually unbiased for non-network data under MCAR with listwise deletion, the same does not apply to network data. There can be considerable biases, even if data is missing completely at random, with parameters of statistical models and descriptive statistics, e.g., density, being biased (Huisman and Steglich, 2008; Smith and Moody, 2013; Huisman, 2009).

2.3.4 Missing Data Treatments

Researchers have several options for handling missing data in networks. These options can broadly be separated into three categories²: deletion, likelihood-based estimation, and imputation (for a general overview of missing data handling see Schafer and Graham, 2002). Deletion methods reduce the network to a fully observed subsample (listwise deletion of actors; Huisman and Steglich, 2008) or ignore the missing data for some, but not all statistical calculations (pairwise deletion). Deletion methods are commonly used and the default for most statistical programs, because they are straightforward in their application and explanation. However, they do not perform well in most situations, as they discard too much information (Huisman and Steglich, 2008; Huisman, 2009; Žnidaršič et al., 2012). In non-network data, cases are usually presumed to be independent (or conditionally independent when conditioning on some social context, e.g., school classroom or company), thus removing participants with missing values will not affect the overall outcome of the model under MCAR. However, removing actors from a network will also remove information about the remaining actors, because incoming ties of the removed actors are outgoing ties of observed actors. These remaining actors will be left with a lower out-degree compared to what was actually observed. Further removal of nodes can affect more complex structures like stars or transitive triads. Despite these limitations listwise (pairwise) deletion can be an adequate missing data treatment if only a small amount of nodes is affected.

Likelihood-based methods estimate the model parameters from the marginal distribution of the observed data. Under M(C)AR this will lead to approximately unbiased estimates in larger samples, given that the model used is correct (Schafer and Graham, 2002). Likelihood-based estimation methods are available for various families of network models; for the exponential random graph family see Robins et al. (2004); Gile and Handcock (2006); Handcock

²A fourth category, re-weighting, is not applicable in network research because of the strong dependencies inherent to network data.

and Gile (2007, 2010); Koskinen et al. (2010, 2013); for the family of stochastic actor oriented models see Snijders et al. (2010a). However, these methods are by definition model-based, and thus cannot aid the estimation of other models (e.g., blockmodels).

Imputation methods replace the missing values with plausible guesses (Rubin, 1987; Schafer and Graham, 2002). For an overview of imputation methods for network data see Huisman and Krause (2017). The methods differ in the amount of information they take into account for the replacement of the missing values. Stochastic imputation methods use draws from probability distributions to replace missing values. These methods can be used for multiple imputation, where missing values are imputed multiple times based on a conditional probability model. This leads to a set of imputed data sets, which are analyzed separately leading to a distribution of model parameters. These are then combined to obtain parameter estimates and standard errors. For the calculation of the standard errors both within and between imputation variance is taken into account. This allows to take the uncertainty about the missing data imputation into account for the estimation of standard errors.

Both single and multiple imputation allow model estimation using all observed information and the calculation of descriptive statistics. While both provide unbiased parameter estimation under MCAR, only multiple imputation is able to provide unbiased standard error estimates, and that both under MCAR and MAR, given the a correct model. For non-network data, likelihood-based estimation and multiple imputation are considered the state of the art (Schafer and Graham, 2002).

2.4 Tested Treatments

In this study, we evaluated the performance of five imputation methods and one deletion method.

2.4.1 Deletion Methods

Although the effectiveness of deletion methods has already been explored in multiple studies (Huisman and Steglich, 2008; Huisman, 2009; Žnidaršič et al., 2012), we incorporate listwise deletion (available cases) in this study, because it is commonly used in network research. It is therefore important to contrast its performance with other methods.

2.4.2 Single Imputation

We compare the performance of both single and multiple imputation methods. The two single imputation methods are null-tie imputation (Žnidaršič et al., 2012) and reconstruction (Stork and Richards, 1992). In null-tie imputation, all missing links are replaced with zeros. This is comparable to imputing unconditional modes in non-network data, as social networks tend to be sparse with a density below 50%, thus not observing a tie between two actors is the most likely case, ignoring everything else.

In reconstruction, missing outgoing tie variables are imputed with the respective incoming tie variables ($x_{ij} = x_{ji}$). An additional step is required for missing links between non-respondents. In this study these ties are imputed stochastically with the probability of a tie equal to the nodal indegree density, that is, the probability for a tie from missing actor i to any actor j is given by $p(x_{ij} = 1) = \sum_{j=1}^{n_u} x_{ji} (\sum_{j=1}^{n_u} I_{ji})^{-1}$, where n_u is the number of observed actors (Žnidaršič et al., 2012).

2.4.3 Multiple Imputation

This study investigates the performance of two multiple imputation methods: Multiple imputation using ERGMs and multiple imputation using Bayesian ERGMs. Imputation by ERGM simulation, as introduced by Wang et al. (2016) works as follows: (1) estimate an ERGM on the observed data using likelihood-based estimation under missing data (Robins et al., 2004; Gile and Handcock, 2006; Handcock and Gile, 2007, 2010); (2) simulate the missing values conditional on the observed ties and the estimated model. By repeating the second step, using the same parameters of the imputation model, multiple imputations can be obtained. However, this procedure is not considered proper multiple imputation as defined by Rubin (1987). In *proper* multiple imputation, the uncertainty about the parameters of the imputation model is reflected by drawing each imputed value using a different parameter vector, where the parameter vectors are draws from their posterior distribution given the observed data. This allows to take the uncertainty about the imputation properly into account. By repeatedly imputing with the same parameter vector it is likely that standard errors will be underestimated.

Imputation using BERGMs was performed using the procedure outlined by Koskinen et al. (2010) and is implemented in the `Bergm` package³ in R (R Core Team, 2019) using an approximate exchange algorithm (Caimo and Friel, 2011,

³The procedure is only implemented since package version 4.2

2014). In this procedure, the missing network data are imputed using draws from the posterior distribution during model estimation. This procedure was developed for estimation of BERGMs under missing data, however, it is possible to retain the augmented networks, thus achieving proper multiple imputation. The BERGM samples from the following probability distribution:

$$p(\theta', x', \theta | x) \propto p(x | \theta) \pi(\theta) \epsilon(\theta' | \theta) p(x' | \theta'), \quad (2.3)$$

in which θ' are proposed parameters and x' are networks simulated with these proposed parameters, $p(x' | \theta')$ is the likelihood on which the simulated data x' are defined and belongs to the same exponential family of densities as $p(x | \theta)$, $\epsilon(\theta' | \theta)$ is any arbitrary proposal distribution for the parameter θ' , and π is the prior probability density function of θ . The proposal distribution is set to be a normal centered at θ . The marginal distribution of θ is the posterior distribution from which inference is drawn.

In the case of missing data, x is not fully observed and the algorithm needs to be extended. The extended algorithm presented below is limited to the setting where v (the set of unobserved tie variables) is known and fixed, and all covariates are known and fixed. Extensions for missing data in multiplex network models exist (Krause and Caimo, 2019). The algorithm will work properly, that is, generate draws from the predictive posterior distribution of the missing data given the prior distribution and the observed data, if missingness is at random (MAR or MCAR) and the parameter for the missingness model ζ is unrelated to the parameter for the network model θ . Thus we do not model the missing data mechanism ζ here. We augment the observed data u by draws v^* from the full conditional posterior $[v | u, \theta]$ of the unobserved data, creating the augmented network $x^* = (u, v^*)$. The algorithm alternates between draws from $[\theta | u, v]$ and $[v | u, \theta]$. The BERGM under missing data thus samples from this adjusted probability distribution:

$$p(\theta', x', \theta, x^* | x) \propto p(x^* | x, \theta) p(x | \theta) \pi(\theta) \epsilon(\theta' | \theta) p(x' | \theta'). \quad (2.4)$$

The marginal distribution of θ is the posterior of interest, which can be obtained after integrating out x' , θ' , and x^* .

This is implemented in the `Bergm` package in R in the following way: At each MCMC iteration, the exchange algorithm has four main steps. First, a new value of θ' is generated. Second, with this θ' a new value of x' is generated. by drawing from $p(\cdot | \theta')$ with an MCMC algorithm (Hunter et al., 2008). Third,

an exchange probability is calculated with,

$$\min \left(1, \frac{p(x'|\theta) p(\theta') \epsilon(\theta|\theta') p(x^*|\theta')}{p(x^*|\theta) p(\theta) \epsilon(\theta'|\theta) p(x'|\theta')} \times \frac{z(\theta) z(\theta')}{z(\theta') z(\theta)} \right), \quad (2.5)$$

and with this probability θ is replaced by θ' . Fourth, if the replacement has taken place, $x^* = (u, v^*)$ is updated by generating v^* from the conditional distribution $p(\cdot | \theta, u)$.

Note that the intractable normalizing constants in (5) cancel each other out. It was shown by Everitt (2012) that this exchange algorithm samples asymptotically from the desired posterior distribution of (θ, v) given u . Further, the algorithm starts with an initial simple imputation of the missing data, by estimating the sufficient statistics $s(x)$ only from the observed data $s(u)$, which in later steps are replaced by the sufficient statistics estimated on the augmented data $s(x^*)$. The algorithm is implemented in the following way for K iterations in Algorithm 1.

Algorithm 1 Approximate exchange algorithm for BERGMs under missing data

Set $s(u)$ as starting values for $s(x^*)$

Initialize θ

for $k = 1, \dots, K$ **do**

 Generate θ' from $\epsilon(\cdot|\theta)$

 Simulate x' from $p(\cdot|\theta')$

 With the log of the probability:

$$\min \left(0, [\theta - \theta']^T [s(x') - s(x^*)] + \log \left[\frac{p(\theta')}{p(\theta)} \right] \right) \quad (2.6)$$

 Replace θ with θ' , and

 impute the missing tie variables v by simulating tie swaps v^*

 from $p(\cdot | \theta', u)$, and form a new realization of $x^* = (u, v^*)$

end for

We employed two imputation models, a simple dyadic independence model with parameters for density, reciprocity, and homophily, and a more complex model with the previous three parameters and parameters for triadic closure (GWESP – geometrically weighted edgewise shared partners), and for two-paths (GWDSF – geometrically weighted dyadwise shared partners). In general, multiple imputation should be performed with a model that is at least as complex as the data generating process and contains all parameters that are to be tested in a later step. This ensures that the relationship between the variables is preserved in the imputation (Huisman and Krause, 2017). The larger imputation model

is equal to the data generating model, the smaller is a less complex, misspecified, model. This allows us to investigate the impact of the complexity of the imputation model on the quality of the obtained imputations.

Due to identification and estimation problems, the larger, more complex imputation model could only be used with the BERGMs. First, we were unable to identify one model using ERGMs that converged in a reasonable time on all networks even without missing data. Specifically, identifying one value for the decay parameter of the geometrically weighted parameters was problematic. Using the same imputation model on all networks is important to make the results comparable and reduce variance in the results. Second, some network structures were hardly observed under large percentages of missing data (e.g., in some networks with 50% missing nodes and missingness mechanisms based on high outdegree, there was only one reciprocated tie). This made it impossible to reliably estimate the complex ERGMs. We were able to solve this problem for BERGMs by using weakly informative priors: $N(0, \sigma = 2)$ (Gelman et al., 2008). Setting the prior standard deviation to 2 ensured that even with little information the estimated parameters remained in a plausible and meaningful range (~ -5 to 5). For reasons of comparability we applied the priors to all estimations, although models on smaller proportions of missing data obtained reliable and smooth posterior distributions using less informative priors ($\sigma = 10$). These problems do not affect the estimation of the simple dyadic independence model.

2.5 Simulation Study

To be able to compare the performance of missing data treatment techniques for different networks, missing data mechanisms, and missing data rates, we simulated network data. Although results obtained from simulated data are harder to extrapolate to real, empirical data, they have several advantages over real world networks in the study of missing data.

Simulating the data generating process gives us full control over the network boundaries, relevant covariates, missing data distribution, and we know the true parameters of the data generating model. This gives us experimental control over the network compositions in this study, allowing us to investigate the performance of the treatment methods under experimentally varying, but controlled conditions. Further, it allows us to use the data generating model for imputation and estimation of parameters, and also enables us to investigate the performance of misspecified imputation models. Lastly, using simulated networks ensures that there is no missing data in the complete observed network.

Empirical network studies are likely to encounter missing data. Although it is vital to study empirical patterns of missing data in networks, they are a hindrance in evaluating missing data handling techniques and may even bias results of studies such as this. Knowing the true complete data allows the researcher to evaluate how well the treatment method performs and gives complete control over the missing data type and mechanism. In short, simulating the networks ensures that we can test the missing data techniques under optimal conditions.

2.5.1 Network Simulation

Directed networks were simulated using the `ergm` package in R (Hunter et al., 2008; R Core Team, 2019). The simulation model included parameters for reciprocity, homophily, GWESP (geometrically weighted edgewise shared partners; Snijders et al., 2006; Hunter, 2007) and GWDSP (geometrically weighted dyadwise shared partners) while keeping the number of ties fixed. The networks differ in size (30 vs. 80 nodes), density (average degree 3 vs. 6), reciprocity (30% vs. 50% reciprocated ties) and homophily on a binary nodal covariate with half the group having the value 0 and the other half having the value 1 (50% vs. 70% homophilous ties). All networks have 30% closed two-paths (transitive ties). This leads to 16 different configurations. For each configuration, ten complete networks were simulated, leading to 160 networks in total. Only simulated networks were selected that did not differ by more than 2.5% at the most on any of the mentioned descriptive statistics. These configurations were selected such that the resulting simulated networks are similar in their structure to social networks that are often observed in small groups (e.g., helping relations in schools).

2.5.2 Missing Data Creation

Missing data were created using six different mechanisms and five different missing data rates in steps of 10% (10-50%). All missing data were generated as actor non-response (i.e., missing all outgoing tie variables of an actor) and, for simplicity, the binary covariate was always observed. The six missing data mechanisms are MCAR, MAR related to the covariate, and MNAR related to high and low in- and outdegree. For missingness related to the covariate, the probability of an actor being missing was set to 0.8 in the first group, and to 0.2 in the second group. It was necessary to allow some members of the second group to be missing to prevent the first group from being completely missing, especially for the higher missing data rates, in which case estimation of homophily parameters would be impossible. A similar process was used for

degree-related missing. Nodes were first ordered by the target mechanism (e.g., low indegree), and then split into three groups: The first group consists of the 50% strongest scoring actors on the target mechanism (e.g., 50% with lowest indegree), the second group was formed by the next 20% of actors, and the third group was formed by the remaining and lowest scoring 30%. Actors in group one had an 80% probability to be missing, actors in group two a 50% probability, and group three was completely observed. This process was chosen to ensure that the missingness was strongly related to the desired mechanism, but also guarantee that the mechanism was not too deterministic. This prevents the observed networks from becoming too dense or too sparse. All missing data was cumulative, nodes missing at 10% were also missing at 20% and higher rates.

2.6 Results

The generation of the networks resulted in $16 \times 10 = 160$ complete data sets and 160×5 (rates) $\times 6$ (mechanisms) = 4800 incomplete data sets. All missing data were treated using the six methods described in the Section IV: Available cases, null-tie imputation, reconstruction, MI simple model (ERGM), MI simple model (BERGM), and MI complex model (BERGM). The performance of the missing data treatment methods was evaluated on (1) their ability to capture descriptive network statistics, (2) how well they are able to impute missing ties/no-ties (link reconstruction), and (3) how well they capture model parameters and lead to similar model inference.

2.6.1 Descriptive Network Statistics

The performance of the imputation models was inspected for the following descriptive statistics: Average degree, reciprocity (proportion of reciprocated ties), transitivity (proportion of closed two-paths), homophily (proportion of within-group ties on all ties). Further, because the network is directed, we evaluate the degree distribution on both indegree and outdegree variance. To measure how the connectivity of the network is preserved by the treatment methods, the average inverse geodesic distance (shortest path) in both the directed and undirected version was chosen. We chose the inverse geodesic, because although none of the complete networks have isolated nodes or subgraphs, with larger amounts of missing data these structures will inevitably appear, thus making the shortest path between subgraphs undefined (usually seen as infinite). By taking the inverse these distances will be set to 0. The directed version only

follows paths in the direction of the ties, while the undirected version first symmetrizes the network by reciprocating all incoming ties, and then calculates the geodesics.

We only present the results on an aggregate level combining the results for the 160 networks for each combination of missing data mechanism and missing data rate. The pattern of results did not meaningfully differ for the 16 network configurations. A detailed analysis of the structural properties of the network configurations revealed that in most situations, the effect on bias was negligible ($< 5\%$), thus an aggregation was deemed justifiable⁴. The results are presented as average relative bias compared to the statistic calculated on the complete network to obtain comparable results across different network structures⁵: $bias = (s(x^*) - s(x))/s(x)$, with $s(\cdot)$ the statistics and x^* the data of the treated network. Taking the relative bias is applicable for these statistics because all of them have non-negative scales with 0 as a meaningful endpoint (i.e., they are all ratio scales).

All results are presented graphically using grid plots. An example is presented in Figure 2.1. Positive values of the relative bias (shown in green) represent an overestimation of the statistics by the treatment and negative values (shown in red) represent an underestimation. For each combination of descriptive statistic and treatment (imputation) method, we created six plots, one for each missing data mechanism, showing the average relative bias for each missing data rate (10-50%). The level of bias is expressed by the saturation of the color, with higher saturation indicating stronger bias. The resulting 48 plots (8 statistics \times 6 treatments) were combined in Figure 2.2.

Unsurprisingly, biases increased with larger missing data rates. Three main conclusions can be drawn. First, multiple imputation using the complex BERGM performs on average better or equally well compared to any of the other treatment methods for all descriptive statistics. The biases remained very small even with very high missing data rates, especially under MCAR, covariate related missing, and high and low indegree missing. Low outdegree missing also led to small positive biases, unlike the other missing data mechanisms, where statistics were underestimated. Second, the results indicate that low amounts of missing data (10-20%) can be handled reasonably well by all methods (on average the

⁴The exception is bias in the reciprocity statistic. Here, the level of reciprocity in the complete network is a relevant predictor of the overall performance – complete networks with 30% reciprocated ties had overall a 12% larger bias. However, this was mainly driven by the performance of the reconstruction imputation. Excluding imputation by reconstruction reduced the main effect of reciprocity on bias in reciprocity to 5%.

⁵We also investigated the average relative absolute bias. Overall, the pattern of results did not change meaningfully.

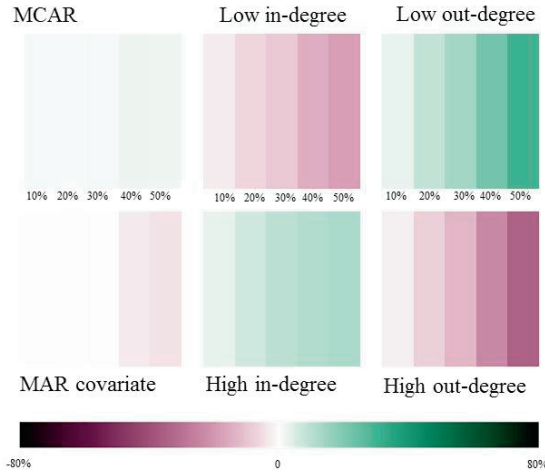


Figure 2.1: **Example of a grid plot**

The six groups consisting of five bars represent the six missing data mechanisms. Each bar stands for 10% missing data, ranging from 10% to 50% missing data. A color scale is given for the interpretation of the colors. The saturation was scaled to 80% as this was the largest average bias observed. Stronger saturation represents larger bias.

absolute bias is below 20% over all networks, mechanisms, and imputations for all descriptives. Third, homophily was estimated without any relevant bias with all methods under all mechanisms. Note that this does not mean that missing data generally has no effect on measurements of homophily. Specific missing data mechanisms targeting hetero- or homophilous actors, or ties, can still lead to biased estimates.

Overall, the simple treatments (available cases, null-tie, reconstruction) did not perform well with higher missing data rates. Imputation with a simple ERGM or BERGM generally behaved similarly⁶. Both methods using the smaller imputation model illustrate the importance of proper model specification for the imputation model. They performed very well for average degree and reciprocity,

⁶The average relative bias under ERGM imputation is slightly smaller compared to BERGM imputation for average degree and reciprocity. This difference disappears when looking at the absolute bias.

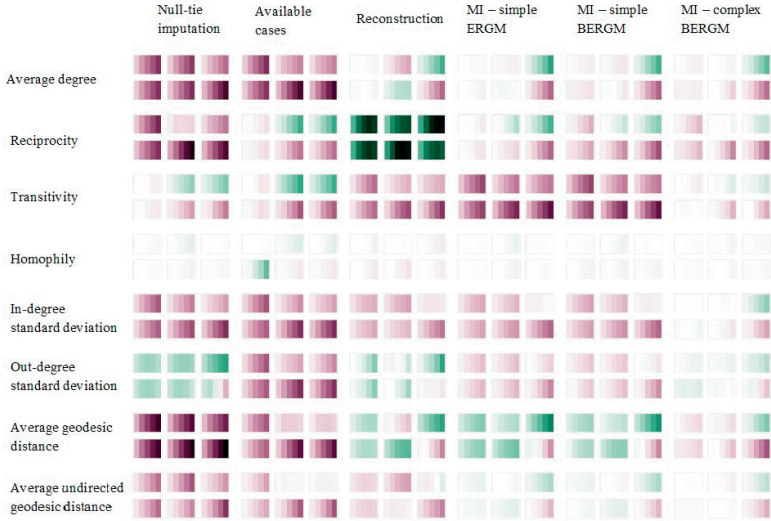


Figure 2.2: Average relative bias for each descriptive statistic by treatment (imputation) method

The 48 grid plots show the average relative bias for each of the descriptive statistics (rows) by missing data treatment (columns). For detailed interpretation of the smaller grid plots, consult Figure 2.1. The level of bias is expressed by the saturation of the color, with higher saturation indicating stronger bias. Positive values of the relative bias (shown in green) represent an overestimation of the statistics by the treatment and negative values (shown in red) represent an underestimation.

but the complex BERGM with the GWESP and GWDSP parameters performed far better on transitivity, degree variance and directed geodesic distance.

2.6.2 Link Reconstruction

Another criterion for the evaluation of missing data handling techniques is link reconstruction (Wang et al., 2016). How well are the treatments able to correctly impute missing ties and missing no-ties? Available cases and null-tie imputation will not be evaluated on this criterion as the one does not provide any imputation at all, and the other always imputes $x_{ij} = 0$. To estimate the link reconstruction capabilities of the multiple imputation methods we combined the multiple imputations obtained for each missing tie variable. We calculated the proportion of imputed scores (either 0 or 1) for a given missing tie variable (e.g., with 10 out of 50 imputations of x_{ij} imputing a tie $x_{ij} = 1$, we get a proportion score of $x_{ij} = .20$). We then chose a cut-off to decide whether this

probability suggest an imputation of a tie ($x_{ij} = 1$) or an imputation of a no-tie ($x_{ij} = 0$). This cut-off was set to the density of the observed network with missing data using case-wise deletion (usually around .10). The observed density seems a reasonable cut-off, because it reflects imputation by the unconditional mean. The performance of the imputation methods is presented in Figures 2.3 and 2.4. Figure 2.3 shows the proportion of correctly imputed zeroes from the number of possible correctly imputed zeros (i.e., correctly imputing a no-tie). Figure 2.4 shows the same of the imputation of ties, that is, the proportion of correctly imputed ones.

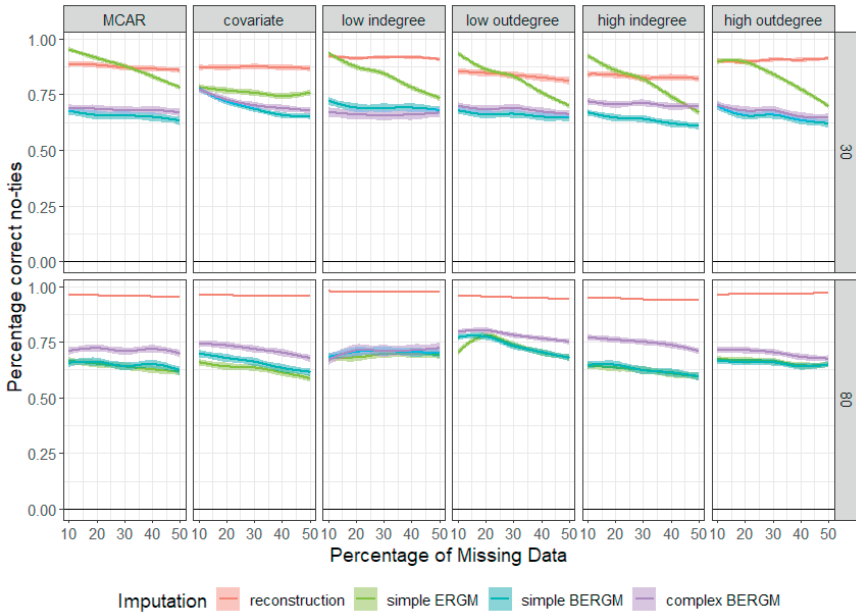


Figure 2.3: **Average percentage of correctly imputed no-ties**

The two rows show the results for 30 nodes (upper) and 80 nodes (lower) networks. The columns show the results for each of the six missing data mechanisms. The percentage of correctly imputed no-ties can be seen on the y -axis. The x -axis shows the missing data rate. The different imputation methods are shown in the different colors.

Overall, no-ties were imputed correctly in the majority of cases (Figure 2.3), with on average over all networks and missing data rates and mechanisms at least 65% of no-ties correctly imputed. The reconstruction method performed best, with on average 87% of no-ties imputed correctly for small networks and 96% correct for larger networks. Multiple imputation with simple ERGMs performed better than the two Bayesian methods for smaller networks with an

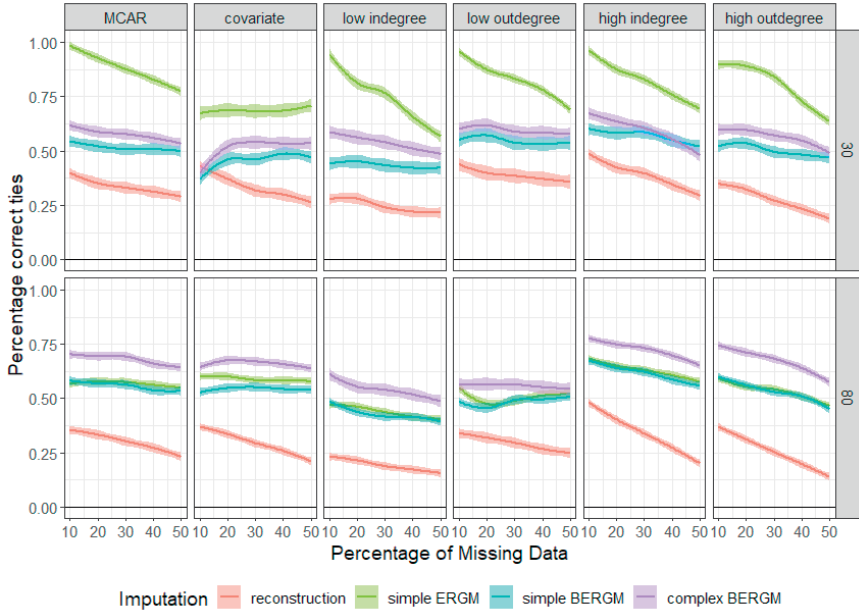


Figure 2.4: **Average percentage of correctly imputed ties**

The two rows show the results for 30 nodes (upper) and 80 nodes (lower) networks. The columns show the results for each of the six missing data mechanisms. The percentage of correctly imputed ties can be seen on the y -axis. The x -axis shows the missing data rate. The different imputation methods are shown in the different colors.

average of 82%, in some situations even outperforming reconstruction. Simple and complex BERGMs performed similarly to each other around 67-82%. Complex BERGMs performed better than the simple ERGMs and BERGMs for larger networks, but are outperformed by reconstruction. This high performance on no-ties is not surprising, given the sparsity of the networks and the vast majority of missing tie variables being no-ties, combined with the fact that all methods are far more likely to impute no-ties than ties, and is in line with previous work on ERGM imputation (Wang et al., 2016). The missing data rate had hardly any effect on the performance of the models

For proportion of correctly imputed ties (Figure 2.4), the performance of multiple imputation is far better. While reconstruction on average only imputes around 30% of the ties correctly, multiple imputation procedures yielded over 50% correctly imputed ties in most cases. Again, for small networks, simple ERGMs performed better than the Bayesian methods, with an average of 79%. Complex BERGMs outperformed the other models slightly for larger networks,

with on average 64% correctly imputed ties. Higher missing data rates lead to a worse performance of the reconstruction method. These results are again consistent with previous work on tie imputation with ERGMs (Wang et al., 2016).

We conclude from these results that the multiple imputation methods lead overall to more reliable imputations, especially simple ERGMs for smaller networks and complex BERGMs for larger networks. However, they are more likely to yield false positive results than reconstruction (as can be seen in lower performance on imputation of no-ties in Figure 2.3). Reconstruction performed poorly in imputation of ties. However, the performance of reconstruction is highly dependent on the general rate of reciprocity in the network, and reconstruction might yield reliable ad hoc imputations with low amounts of missing data in highly reciprocated networks.

2.6.3 Model Parameters and Inference

Finally, we investigated how well model parameters can be recaptured, and how well the techniques are able to come to the same inferential conclusions as would have been drawn from the complete data. In this section, we only compare BERGMs estimated on available cases with the complex Bayesian ERGM with missing data augmentation. The simple ERGM and BERGM are not compared on this dimension, because the resulting parameters would necessarily be different, as the model is not the same. Using the simple models only for imputation and then estimating the complex model on the imputed data goes against the general recommendation for multiple imputation that the imputation model should be at least as complex as the analysis model. Thus, only the complex model is competing on this dimension. It has already been shown that ERGMs can be estimated approximately unbiasedly under M(C)AR (Handcock and Gile, 2010). Earlier work on BERGMs under missing data has suggested that the algorithm presented here is also approximately unbiasedly under M(C)AR (Koskinen et al., 2010). However, we also wanted to investigate the performance of the algorithm when missing data is not MCAR. Available cases is taken as the comparison, because it is often used to handle missing data, although it is not recommended for ERGMs, or BERGMs. The two methods were compared on the means and standard deviations of the estimated posteriors for each parameter, model wide, and on Bayesian p -values. We calculated the relative deviation ($\frac{\text{treated}-\text{complete}}{SD_{\text{complete}}}$) for the comparison of mean posterior parameter estimates and standard deviations. Figures 2.5 and 2.6 show the deviations for the posterior means and the standard deviations of the posterior parameter

distributions, respectively. The deviations for parameters were standardized by dividing by standard deviations of the complete data estimates. The deviations in standard deviations were also standardized by dividing by the standard deviations of the complete data estimate. The results were quite homogenous for the different network configurations with the only important characteristic being the size of the networks, thus the figures are averaged over all networks with the same size. The results for the different network sizes are represented by the shape of the lines (solid for 30 nodes, dotted for 80 nodes). The color of the lines represent the treatment method (red for available cases, blue for BERGM imputation).

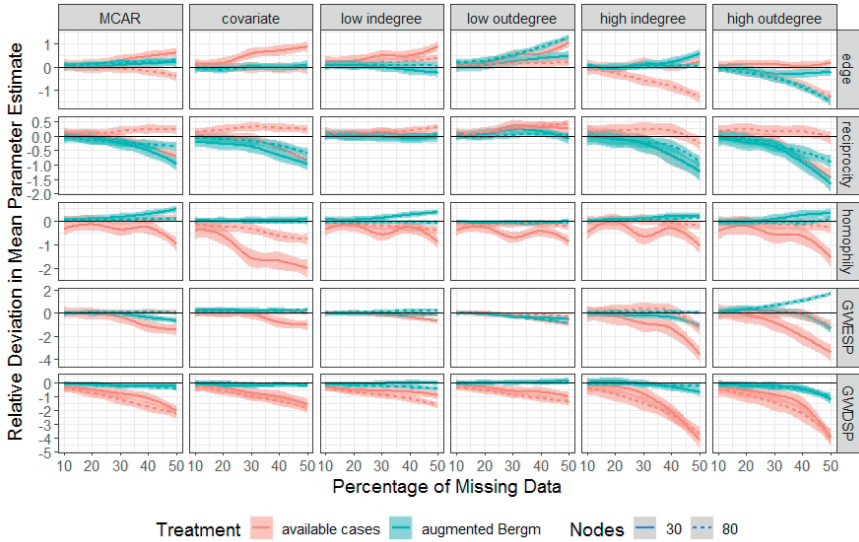


Figure 2.5: **Relative deviation in mean parameter estimates**

Results are presented separately for each combination of parameter (rows) and missing data mechanism (columns). The y -axis in each smaller figure shows the relative deviation, the x -axis the missing data rate. The line color reflects the missing data treatment, the line type the size (solid: 30 nodes; dotted: 80 nodes).

The edge parameter was generally estimated quite well, with deviations between 1.5SD above and below the complete data estimate. Augmented BERGMs generally performed better, in most cases only slightly. They only showed larger deviations than available cases in larger networks for low and high outdegree missings. Under low outdegree missings the parameter was overestimated and under high outdegree the parameter was underestimated. These results are in line with what could be expected under outdegree related missingness. Miss-



Figure 2.6: **Relative deviation in posterior standard deviations**

Results are presented separately for each combination of parameter (rows) and missing data mechanism (columns). The y -axis in each smaller figure shows the relative deviation, the x -axis the missing data rate. The line color reflects the missing data treatment, the line type the size (solid: 30 nodes; dotted: 80 nodes).

ing nodes will have their ties imputed following the patterns observed in the data, thus, under low outdegree missing, too many ties will be imputed, and under high outdegree missing too few will be imputed. Available cases performed similarly poorly under high outdegree missing.

Both treatments performed similarly for the reciprocity parameter, albeit that augmented BERGMs were generally more likely to underestimate the parameter, especially under high in- and outdegree related missing, while available cases, especially for larger networks, performed generally well. In contrast, augmented BERGMs performed very well for the homophily parameter, while available cases, especially for smaller networks, was more likely to have deviating estimates. The GWESP parameter was estimated with negligible deviations by augmented BERGMs, except for high outdegree related missings. Available cases again performed worse with smaller networks and showed severe deviations under high in- and outdegree missings (up to 4SD below the the complete data estimate). Finally, the GWDSP parameter was estimated without any relevant deviation by the augmented BERGM, with the exception of higher missing data rates under high outdegree missing. Available cases, however, showed larger de-

viations with increased missing data rate for all missing data mechanisms. The deviations were, again, particularly large for high in- and outdegree missings (up to 4.5SD below the the complete data estimate).

The deviations in the estimation of the posterior standard deviations follow a similar pattern for all parameters and missing data mechanisms. Augmented BERGMs generally show smaller deviations and available cases show larger deviations for smaller networks. The deviation in the edge parameter is small, with maximally around 50% larger standard errors at 50% missing data, compared to the up to four times larger standard errors that can be found for the GWESP parameter under available cases.

Summarizing the results when looking at single parameters we can conclude that augmented BERGMs generally performed better, with often no deviations, or deviation similar to those under available cases. Both methods showed hardly any deviation with 10% missing data and deviations at 20% missing data were mostly small.

However, comparing single parameters of a (B)ERGM in isolation can be misleading. The parameters are often highly correlated and deviations in the estimation may only be seen in the numeric values of some parameters, while distorting the sufficient statistics belonging to other parameters as well. We also want to compare the methods by taking all parameters in the model into account. For this, models were also compared using Mahalanobis distances:

$$M = \sqrt{(P - \mu_D)^T S_D^{-1} (P - \mu_D)}. \quad (2.7)$$

The Mahalanobis distance is a measure of distance between a point P and the centroid μ_D of a multi-dimensional cloud D . It gives the distance between P and μ_D , the center of mass, adjusted for the width of data cloud in the direction of P , using the covariance matrix of D , S_D . For the comparison in our study, we used the posteriors of Bayesian ERGMs estimated on the complete data as the data cloud D . The points P are the centroids of the posteriors of BERGMs (the center of mass of the multivariate cloud formed by the posteriors) estimated either on the missing data with augmentation, or estimated on the available cases. Thus, we compared the distance of the mean estimates of both treatments to posterior distributions estimated on the complete data. The results for Mahalanobis distances are presented in Figure 2.7.

There was no meaningful difference between the treatments for missing data rates of 10%, and while augmented BERGMs performed systematically better, missing data rates of 20% did not yield a large difference. The differences become striking with increased missing data rates, however, and are very large

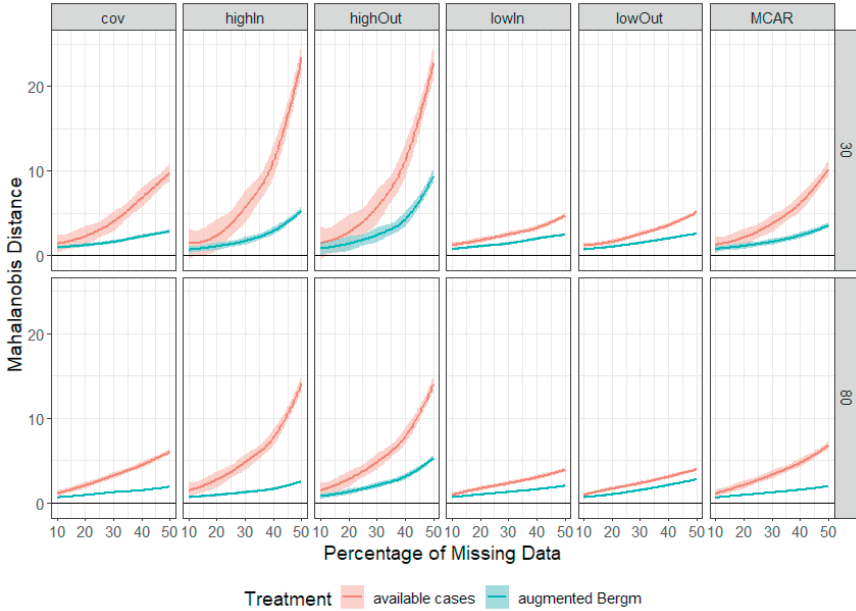


Figure 2.7: Mahalanobis distances between posterior estimates on complete and treated data

Results are presented separately for the two network size (upper row: 30 nodes; lower row: 80 nodes) and missing data mechanism (columns). The y -axis in each smaller figure shows the Mahalanobis distance to the complete data estimate, the x -axis the missing data rate. The line color reflects the missing data treatment.

under high in- and outdegree missings for larger percentages of missing data. Larger networks were estimated better, which is reflects on overall smaller Mahalanobis distances. Overall, the augmented BERGMs perform far better than using available cases.

Although comparing the estimated model parameters simultaneously by investigating Mahalanobis distances is better than comparing individual parameters, comparing the two missing data treatments based on the estimated parameter values is still difficult. Parameters of B/ERGMs are dependent on the density and the size of the network, and both change drastically when using available cases. Thus, some differences in the parameters can be expected, even without deviations in parameter estimation due to missing data. Therefore we also compare the performance on a third measure.

Researchers in the social sciences are often less interested in the exact size of a parameter, but often primarily care about the inference drawn from a model,

the direction of estimated parameters, and their corresponding uncertainties. To compare the difference in inference drawn between the complete data models and the treatments, we are evaluating Bayesian p -values. Bayesian p -values give the mass of the posterior probability distribution that is on one side of 0. A Bayesian p -value of, for instance, .03 can thus be interpreted as: The probability that the parameter is negative, given the priors, the model, and the data, is 3%. However, difference in Bayesian p -values are not very interesting in themselves, because moving from 0 to .10 constitutes a more substantial difference than, say, moving from .40 to .50. Thus we dichotomized Bayesian p -value in categories of classical significance, below .05 vs above .05, and a more lenient significance level of below .10 and above .10. Note that in the complete data nearly all parameters were significant⁷, with exception to the homophily parameter in networks with no homophily tendencies. The results are presented in Figures 2.8 and 2.9, summed over all networks. We depict the count of false positive and false negative results, meaning results where the complete data estimate had a Bayesian p -value of $p > .05$, but the treatment lead to a Bayesian p -value of $p < .05$ for false positives and the reverse for false negatives. Figure 2.8 shows the results for Bayesian p -values of $p < .05$ and Figure 2.9 shows the results for Bayesian p -values of $p < .10$. The shape of the lines indicate false negatives (solid) and false positives (dotted) counts, the color the treatment method (red for available cases, blue for BERGM imputation).

The counts for false results for the edge parameter were generally low under $p < .05$, with only at most only 20 out 160 models giving false negative results, and the one possible false positive results occurred in some cases. Neither false positives nor false negatives are found under the more lenient criterion of $p < .10$. This pattern is even more striking for the reciprocity parameter. With $p < .05$ we find up to 60 false negative results. However, all of these are significant at the $p < .10$ level. Here only one result is sometimes found as false positive. The same pattern appears for the GWDSP parameter. There were no false positives or negatives found for the GWESP parameter under either criterion. The homophily parameter showed several false negatives under $p < .05$, which all vanished under $p < .10$. However, there were several false positives which persisted even under the more lenient p -value.

Both treatments either performed similarly or augmented BERGMs showed a lower number of false results. This was the case for the homophily parameter for both p -values, as well as the GWESP parameter under $p < .05$. In our

⁷The edge parameter in one of the networks has a Bayesian p -value of .052, the reciprocity parameter in three networks has p -values of .071, .076, and .117, the GWESP parameter is not significant in two cases (.053 and .077), and the GWDSP parameter in two networks has Bayesian p -value of .0502 and .086.

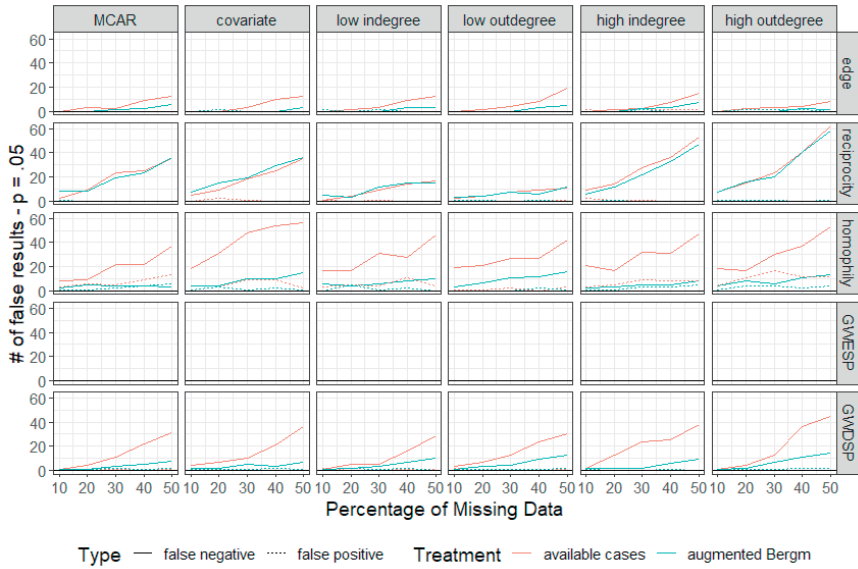


Figure 2.8: Number of false results with $p = .05$

Results are presented separately for each combination of parameter (rows) and missing data mechanism (columns). The y -axis in each smaller figure shows the count of false results, the x -axis the missing data rate. The line color reflects the missing data treatment, the line type the type of error (solid: false negatives; dotted: false positives).

simulation most effects were considerably strong and the models were correctly specified. It is possible that effects will not be as accurately identified, especially using available cases, if the effect is small, shows larger variance, or the model is misspecified. However, it seems that available cases does not necessarily lead to poor inferences.

2.7 Discussion

In this study, we evaluated several missing data treatments for missing network data in a cross-sectional setting. We show that under some circumstances, using the right treatment method and focusing on specific outcome variables, even very high rates (50%) of missing data can be handled adequately. The results indicate that multiple imputation with a sufficiently complex Bayesian ERGM outperforms commonly used techniques on several criteria. It showed less bias than alternative methods in estimating descriptive network statistics, smaller deviation on model parameters and variances, and yielded better model infer-

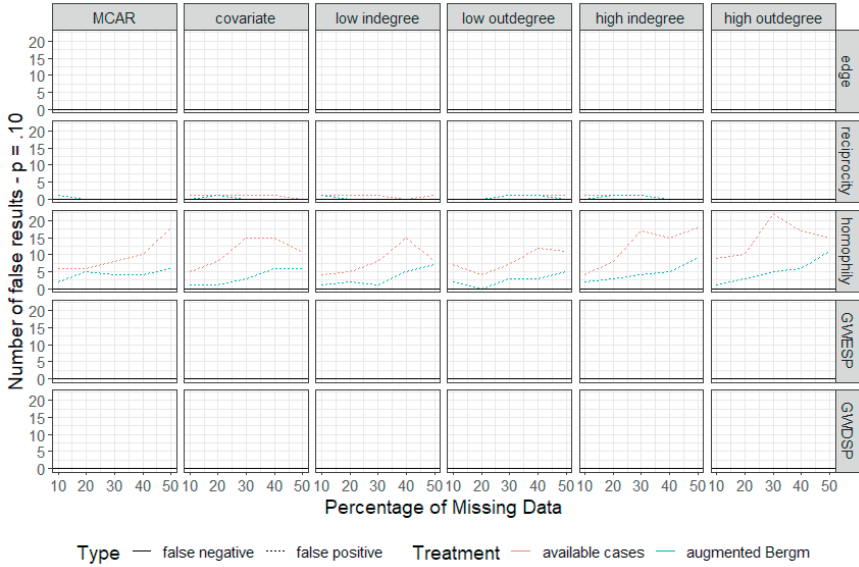


Figure 2.9: Number of false results with $p = .10$

Results are presented separately for each combination of parameter (rows) and missing data mechanism (columns). The y -axis in each smaller figure shows the count of false results, the x -axis the missing data rate. The line color reflects the missing data treatment, the line type the type of error (solid: false negatives; dotted: false positives).

ences. Especially available case analysis and imputation of zeros performed poorly compared to multiple imputation. Multiple imputation by complex BERGM was suboptimal only in imputation of specific ties and no-ties. For this purpose, the reconstruction method and imputation with simple ERGMs performed better for no-ties (i.e., correctly imputing zeros), and simple ERGMs performed better for ties in smaller networks (i.e., correctly imputing ones), multiple imputation by complex BERGMs was, however, more accurate in imputing existing ties in larger networks.

This study focused on a limited number of networks from a restricted set of possible configurations. Moreover, we tested these methods under ideal situations where the data generating model is known. This is usually not the case in empirical research. The results indicated that an insufficiently specified model will not be able to lead to the same reduction in bias as a more complete model of the data. However, an insufficient but still suitably complex model is likely to lead to more accurate results than simpler methods. Further, the simulated data showed only a limited amount of complexity, using only five parameters

for the simulation. Empirical data is often far richer. Our results indicate that multiple imputation performs better in more structured networks.

Despite these limitations, multiple imputation using BERGMs seems superior to current alternatives for missing data treatment in providing less biased descriptive statistics and more accurate model estimates. Future research needs to develop guidelines on the selection of the imputation model and on assessing the sensitivity of the results to model specifications. The results suggest that the imputation with a too simple model will still lead to less bias than the most commonly used procedures (null-tie imputation and available cases) for the majority of analyzed descriptives.

Missing Data in Multiplex Networks

Imputation for Bayesian Exponential Random Multiplex Graph Models

3.1 Introduction

In recent years, it is becoming more and more apparent that the understanding of social structure often requires to take more than just one type of social relation into account, so called multiplex networks. Notable examples include the important interrelations between friendships and advice seeking behavior (Snijders et al., 2013), the importance of antipathy-ties in the maintenance of friendship group structures (Stadtfeldt et al., 2018), and the relationships between joined drug use, sexual relations, and co-visitation of social venues (Fujimoto et al., 2015). These papers, and many others, show that to understand one type of relation it is often necessary to also take other types of relations between actors into account.

Also, for reconstructing missing information, multiplex networks have the potential to be better predictors (e.g., an indirect connection in one network can predict a tie in another layer). And for that aim, generative models for the analysis of multiplex networks offer a great advantage for the treatment of missing data. While missing data is a problem for all (social) sciences, network models suffer particularly under missing data, because of the strong dependencies within the data. Non-response by one participant does not only mean we know less about this participant, but we also know less about the social network of all other participants, after all, the missing participant could have nominated any of the other participants thus potentially changing the network structure drastically. Previous work has primarily focused on single layer networks. Statistical tools have been developed to handle missing social network data, both

This chapter is an adaptation of Krause and Caimo (2019)

to obtain more reliable model estimates (Handcock and Gile, 2007; Snijders et al., 2010a; Koskinen et al., 2010; Krause et al., 2018a), as well as reliable descriptive statistics (Krause et al., 2018b)¹. Extending this work to multiplex models is likely to lead to less biased model estimations (less biased statistics), because multiplexity allows to use information of observed layers for the missing data augmentation of incompletely observed layers. Even if information about the same relations is missing from all layers, multiplex models can access more data to obtain more accurate imputations of missing network data.

In this paper we propose an extension of previous work on missing data augmentation to the context of exponential random multiplex models (ERmGMs). We advance the literature twofold, first, by proposing an estimation procedure for Bayesian ERmGMs, and second by proving proper multiple imputation of missing multiplex network data. In Section 3.2 of this paper we will introduce the exponential random graph model family and the proposed multiplex extension algorithm. Section 3.3 details the missing data problem for networks and how the algorithm is extended to properly estimate BERmGMs under missing data, and an example application is presented in Section 3.4. We end the paper with a discussion of our findings and according recommendations.

3.2 Bayesian ERmGMs

Before we describe the proposed algorithm for BERmGMs, we will first introduce the ERG-family and network multiplexity.

3.2.1 Bayesian inference for ERGMs

The exponential random graph model family (ERGMs; Lusher et al., 2013) is most commonly used to analyze cross-sectional network data. ERGMs model an observed network, or graph, as a function of sufficient network statistics (primarily counts of subgraph configurations, e.g., number of ties, number of reciprocated ties or number of transitive triplets). A network graph is expressed as a random $n \times n$ adjacency matrix x with $x_{ij} = 1$ when there is tie from node i to node j and $x_{ij} = 0$ when there is no tie. Usually, edges connecting a node with itself are not allowed ($x_{ii} = 0$). Networks can be directed or undirected ($x_{ij} = x_{ji}$). Let \mathcal{X} denote the set of all possible networks on n nodes and let x be a realization of $x \in \mathcal{X}$. Then, in Bayesian ERGMs (BERGMs) the posterior

¹Chapters 2 of this dissertation.

probability of the parameters conditional on the data is given by

$$p(\theta|x) = \frac{\exp[\theta^T s(x)] p(\theta)}{z(\theta) p(x)}, \quad (3.1)$$

with θ being a vector of model parameters, $s(x)$ a vector of corresponding sufficient network statistics, $z(\theta)$ the normalizing constant, $p(\theta)$ the prior distribution of the parameters and $p(x)$ is the marginal probability of the data. See Lusher et al. (2013) for an introduction to ERGMs. (Bayesian) ERGMs are usually estimated via simulation. These simulation consist of iterations of swaps of single ties ($x_{ij} = 1$ to $x_{ij} = 0$ or vice versa), conditional on the rest of the network. Tie swaps can be made according to Gibbs or Metropolis-Hastings sampling (Lusher et al., 2013).

3.2.2 Multiplexity

Multiplex networks are structures with multiple different types of relations on the same set of nodes. Multiplex networks can thus be expressed as a random $n \times n \times l$ adjacency array x with $x_{ij}^l = 1$ when there is tie from node i to node j on network layer l and $x_{ij}^l = 0$ when there is no such tie on layer l . Each layer l of the multiplex network can be either directed or undirected. Multiplex ERGMs were first introduced by Pattison and Wasserman (1999) and later extended by Wang (2012). Multiplexity increases the complexity of network models by an additional factor, while a single layer directed network has $2^{n(n-1)}$ possible configurations (e.g., a network of 20 nodes has $\sim 2.5 \times 10^{114}$ possible configurations), this number increases exponentially to the number of layers, $2^{(n(n-1)) \times l}$ (e.g., a multiplex network of 20 nodes with 2 layers has $\sim 6.1 \times 10^{228}$ possible configurations).

3.2.3 Posterior Parameter Estimation for BERmGMs

The Markov-Chain Monte-Carlo (MCMC) estimation algorithm of the posterior $p(\theta|x)$ is an extension of the approximate exchange algorithm introduced by Caimo and Friel (2011) and currently implemented in the `Bergm` package (Caimo and Friel, 2014) in R (R Core Team, 2019). The algorithm samples from the following distribution:

$$p(\theta', x', \theta|x) \propto p(x|\theta) p(\theta) \epsilon(\theta'|\theta) p(x'|\theta'), \quad (3.2)$$

with $p(x'|\theta')$ being the likelihood on which the simulated data x' are defined and belongs to the same exponential family of densities as $p(x|\theta)$, $\epsilon(\theta'|\theta)$ is any

arbitrary proposal distribution for the parameter θ' . This proposal distribution is set to be a normal centered at θ .

At each MCMC iteration, the exchange algorithm consists of three main steps: First, proposing a Gibbs update of θ' , followed by a Gibbs update of x' , a draw from $p(\cdot|\theta')$ with an MCMC algorithm (Hunter et al., 2008). Third an exchange from the current state θ to the proposed new parameter θ' is taken. This deterministic proposal is accepted with the following probability:

$$\min \left(1, \frac{q_\theta(x') p(\theta') \epsilon(\theta|\theta') q_{\theta'}(x)}{q_\theta(x) p(\theta) \epsilon(\theta'|\theta) q_{\theta'}(x')} \times \frac{z(\theta) z(\theta')}{z(\theta') z(\theta)} \right) \quad (3.3)$$

where q_θ and $q_{\theta'}$ indicate the unnormalized likelihoods for parameters θ and θ' , respectively. The intractable normalizing constants cancel each other out in this equation, thus avoiding the problem of calculating them. The exchange algorithm samples asymptotically from the desired posterior distribution (Everitt, 2012).

The key change to the regular **Bergm** algorithm is in the network simulation loop, which is here sampling from a multiplex network space. Instead of directly simulating a new multiplex network x' with the proposed parameter θ' , the simulation is performed iteratively (in total H times) for each of the L layers of ties by proposing one, or a few, tie swaps on each layer, conditional on the proposed parameter vector θ' and on all tie swaps simulated on all layers of the network in this and previous iterations, that is conditional on ${}^h x'^{l=1}, \dots, {}^h x'^{l-1}, {}^{h-1} x'^l, {}^{h-1} x'^{l+1}, \dots, {}^{h-1} x'^L$. This is repeated H times and a sample is drawn from $p(\cdot|\theta')$. For small networks $H = 1000$ is sufficient, while larger networks require higher settings of H (e.g., $H = 5000$).

The implementation of the algorithm is presented below. Here, K is an arbitrarily large number of iterations of the algorithm, $K = 2000$ is often sufficient.

Adaptive procedures such as the adaptive direction sampling (Caimo and Friel, 2011; Thiemichen et al., 2016) or the delayed rejection sampling (Caimo and Mira, 2015) can be adopted for this algorithm.

3.2.4 Cross-Network Effects

Currently three fundamental dyadic cross-network effects are implemented for the algorithm. These effects are: 1) co-occurrence, 2) entrainment, and 3) cross-network reciprocity. Co-occurrence expresses the tendency of edges on one layer to occur with edges on another layer in an undirected graph and entrainment

Algorithm 2 Approximate exchange algorithm for BERmGMs

Initialize θ
for $k = 1, \dots, K$ **do**
 Generate θ' from $\epsilon(\cdot|\theta)$
 for $h = 1, \dots, H$ **do**
 for $l = 1, \dots, L$ **do**
 Simulate one (or a few) tie swaps in x^{l-1} from
 $p(\cdot|\theta', x^{l-1}, \dots, x^{l-2}, x^{l-1}, x^{l-1}, \dots, x^{l-1})$
 end for
 end for
 Update $\theta \rightarrow \theta'$ with the log of the probability:

$$\min \left(0, [\theta - \theta']^T [s(x') - s(x)] + \log \left[\frac{p(\theta')}{p(\theta)} \right] \right) \quad (3.4)$$

end for

is its directed counterpart. The corresponding sufficient statistic can thus be calculated similarly for both:

$$s_{\text{co|ent}}(x) = \sum_{i < j} x_{ij}^{l=1} x_{ij}^{l=2}. \quad (3.5)$$

cross-network reciprocity models the co-occurrence of outgoing ties of one type with incoming ties of another type on the same dyad.

$$s_{\text{cross-recip}}(x) = \sum_{i < j} x_{ij}^{l=1} x_{ji}^{l=2}. \quad (3.6)$$

3.3 Missing Data Imputation

The proposed missing data augmentation procedure is an extension of the work by Koskinen et al. (2010). In short, every time a new θ' is accepted in the algorithm outlined above, the missing network data are imputed conditional on the observed data and θ' . The imputation follows a similar simulation procedure as the parameter estimation. However, only tie-swaps for missing tie variables are proposed. The obtained imputed multiplex network x^* is then used as the starting point for the next iteration and treated as new baseline. Thus equation (3.4) optimizes $[s(x') - s(x^*)]$, and not $[s(x') - s(x)]$. We present the algorithm below. Here, A is an arbitrarily large number of auxiliary iterations. In the algorithm presented below, x^{*-l} are all imputed network layers of x except x^l (e.g., in case of $L = 2$ and $l = 2$, $x^{*-l} = x^{*-2} = x^{*1}$). Let ux denote the observed part of the network x .

Algorithm 3 Approximate exchange algorithm for BERmGMs under missing data

Use naive imputation obtain starting values for $s(x^*)$

Initialize θ

for $k = 1, \dots, K$ **do**

 Generate θ' from $\epsilon(\cdot|\theta)$

for $h = 1, \dots, H$ **do**

for $l = 1, \dots, L$ **do**

 Simulate one (or a few) tie in swaps x'^l from

$p(\cdot|\theta', {}^h x'^{l=1}, \dots, {}^h x'^{l-1}, {}^{h-1} x'^l, {}^{h-1} x'^{l+1}, \dots, {}^{h-1} x'^L)$

end for

end for

 With the log of the probability:

$$\min \left(0, [\theta - \theta']^T [s(x') - s(x^*)] + \log \left[\frac{p(\theta')}{p(\theta)} \right] \right) \quad (3.7)$$

 Replace θ with θ'

for $a = 1, \dots, A$ **do**

for $l = 1, \dots, L$ **do**

 Simulate one (or a few) tie swaps of missing tie variables in x^l

 from $p(\cdot|\theta', {}^a x^{*l=1}, \dots, {}^a x^{*l-1}, {}^{a-1} x^{*l}, {}^{a-1} x^{*l+1}, \dots, {}^{a-1} x^{*L})$, and form a new realization of x^{*l}

end for

end for

end for

The imputed networks x^* can be retained after the estimation of the posterior $p(\theta|x)$ and used for additional analyzes, because they constitute proper multiple imputations of x , assuming a well fitting model.

This algorithm has been shown to provide reliable estimates of $p(\theta|x)$ (Koskinen et al., 2010), and low biases in descriptive statistics (Krause et al., 2018b) in the single-layer network setting. However, if x is a multiplex network, it is important to impute missing data with a multiplex network model to guarantee that the observed relationships between the layers are maintained in the imputation process. Even if one is only interested in one layer of x it we recommended to use a multiplex imputation model to use all available information for the imputation. Using a multiplex model can yield better imputations as it allows to use information of one layer to aid in the imputation of another layer. Using the other layers only as covariates has the downside that, if the other layers are analyzed later, imputations made for these layers are independent of the imputations of the layer currently under study, which, in the worst case, can lead to contradictory results. The multiplex imputation algorithm, however

comes with a clear advantage. If multiple layers are imputed simultaneously, the imputed information one layer will contribute to more reliable imputations on the other layer. This is especially the case if one layer is more structured, that is, more explainable by the used imputation model, than the other layer. Therefore, the proposed algorithm provides an important advancement in the treatment of missing network data.

The algorithm is expected to perform better in cases where missing tie variables are only missing on some, but not all layers, because it will use the available information of the observed layers for the imputation of the missing data. However, it should provide reliable imputations under complete non-response of an actor. The proposed algorithm is an extension of the existing algorithm by Koskinen et al. (2010), and all recommendations regarding missing data augmentation concerning model selection for regular (B)ERGMs also apply to BERmGMs. Multiple imputation in general requires that the imputation model is well fitting and at least as complex as the final estimation model (Krause et al., 2018b).

3.4 Illustration - Florentine Families

As a simple illustration we present Padgett's network of the Florentine banking families, a classical example of network analysis (Padgett and Ansell, 1993). The network consists of 16 nodes (the banking families), their business relations, and their marital connections (Figure 3.1). We present only a simple model for the multiplex graph for illustrative purposes. The within-layer effects are similar for both business and marriage relations. The model consists of a set of parameters for edges (modeling the density), geometrically weighted degree (GWDEGREE - modeling the degree distribution) and geometrically weighted edgewise shared partners (GWESP - modeling triadic closure; Hunter and Handcock, 2006; Snijders et al., 2006). Additionally, the model includes a parameter for the co-occurrence of ties between the layers.

Missing data in most data collection designs will be complete non-response by an actor, leading to missing data on all network layers. Thus, missing data were created by randomly selecting three ($\sim 20\%$) of the families and setting their outgoing and incoming ties and no-ties for both layers to missing. For this illustration only one set of missing data was created.

The probability density plots for the posterior distributions for the complete data model as well as for the missing data model are presented jointly in Figure 3.2. The missing data augmentation algorithm performs well in approxi-

inating the posterior of the full data model. The probability densities largely overlap, and both models lead to substantively the same conclusions about the direction and size of the effects. This performance of the missing data augmentation was expected given its theoretical similarity to existing work on missing data augmentation with BERGMs (Koskinen et al., 2010; Krause et al., 2018b).

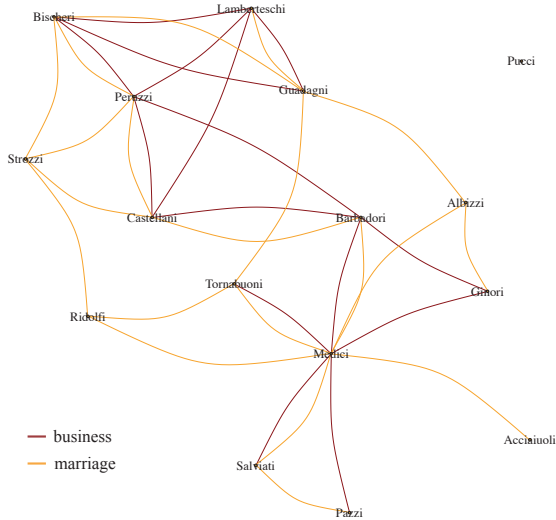


Figure 3.1: **Business and marriage relations of the 16 Florentine families.**

The nodes represent the banking families. Business relations between the families are shown in red, marriage relations in yellow.

3.5 Discussion

In this paper, we present a Bayesian computational algorithm for the estimation of multiplex exponential random graphs under missing data. The code implementing the methodology is currently available on GitHub and in future will be part of the `Bergm` package in R. It is thus far the only implementation of multiplex random graphs in R. The algorithm theoretically extends to networks of any size, any given number of layers, with the theoretical restrictions that apply to ERG-family models. However, the algorithm is currently completely implemented in R, which means that estimating larger networks will very computationally intensive.

Currently, `Bergm`, and by extension the proposed algorithm, are heavily reliant on the `ergm` package. Unfortunately, `ergm` does not facilitate estimation

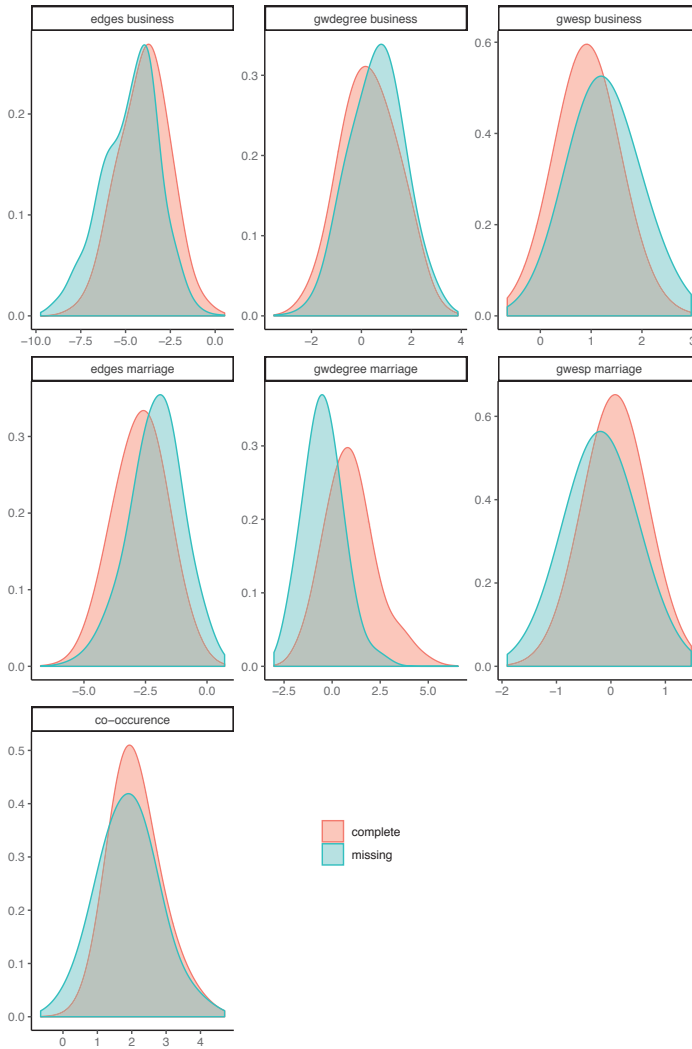


Figure 3.2: **Posterior Density Distribution of BERmGM estimates for complete and missing data.**

The results for the BERmGMs estimated on the missing data are shown in blue, results estimated on the complete data in red.

of ERmGMs, which limits the availability of cross-network effects. The proposed algorithm can be easily adapted to estimate Bayesian exponential random (multiplex-)network models (Fellows and Handcock, 2012a), an extension of the ERG-family models where also nodal attributes are random and dependent on the connectivity structure of the network. The estimation of this joint network and attribute distribution can be implemented similarly to the estimation of the multiplex structure.

Missing Data in Longitudinal Networks

Multiple Imputation for Longitudinal Network Data

4.1 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 handicapped 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 potential for missing data. Additionally, the strong dependence between actors makes network analysis more sensitive to missing data. Missing tie variables do not only 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

This chapter has been published as Krause et al. (2018a) with minor alterations.

et al., 2015). The proposed procedure is an imputation method applicable to missing data at all waves¹.

The paper is organized as follows. In Section 4.2, we describe the two network model families relevant for this paper, the stochastic actor-oriented model (SAOM) and the exponential random graph model (ERGM). In Section 4.3, we detail the non-response problem and its specifics for missing data in networks. Section 4.4 details the proposed multiple imputation procedure for longitudinal network analysis. Section 4.5 presents a (simulated) example application and comparison to the original (complete) networks and a benchmark procedures. We end the paper with a discussion of the findings and according recommendations.

4.2 Statistical Models for Network Analysis

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 (SAOMs) and exponential random graphs (ERGMs).

4.2.1 Stochastic Actor-Oriented Models

Researchers studying the co-evolution of social relations (e.g., friendships) and behaviors (or attitudes) over time 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 SAOM 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,

¹For their helpful comments that significantly improved this study, we thank the two anonymous reviewers.

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. The chosen actor i can either create a tie to an unconnected actor, drop a tie to a connected actor, or do nothing². Let x denote the $n \times n$ adjacency matrix where n is the number of actors, with $x_{ij} = 1$ when there is a tie from actor i to actor j and $x_{ij} = 0$ when there is no tie. Self nominations are not allowed ($x_{ii} = 0$). Then the probability for each of the possible actor decisions is determined by an objective function, in which actor-specific network statistics (including effects of covariates) s_{ki} are weighted with parameters of the network evolution θ_k , given the current state of the network x

$$f_i(\theta, x) = \sum_k \theta_k s_{ki}(x). \quad (4.1)$$

The network statistics s_{ki} can be, for instance, subgraph counts (or non-linear transformations) 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 unobserved³. 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 for the software to estimate SAOMs – Simulation Investigation of Empirical Network Analysis; RSiena is a contributed package (Ripley et al., 2017) to the statistical system R (R Core Team, 2017).

Model estimation is typically done by the Method of Moments, that is, determining parameters so 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 rough estimates for the parameters. Phase 2 estimates the parameters iteratively by simulating the mini steps entire process many times, each time

²Throughout this paper we focus on directed networks. All models discussed in this paper also apply to undirected networks.

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

calculating the statistics used for the Method of Moments, and updating the parameters according to Robbins-Monro steps. 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 RSiena software: Method of Moments (MoM), Generalized Method of Moments (GMoM), Maximum Likelihood (ML) and Bayesian estimation (Koskinen and Snijders, 2007; Amati et al., 2015; 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. 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 simulation provide a distribution of networks at the end of the simulation, while ML simulations always ends in the observed network at wave m . For a more detailed introduction to SAOMs see Snijders (2017b); for an introduction to applying SAOMs see Snijders et al. (2010b) or Steglich et al. (2010).

4.2.2 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 $s(x)$ 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.

4.2.3 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 sub-graphs, for instance, 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

$$\sum_k \theta_k s_k(x), \quad (4.2)$$

with the key difference that in the SAOM the objective function is actor specific, as can be seen in the actor index i in (4.1). 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).

4.3 Missing Data

4.3.1 Missing Data Mechanisms

Missing data mechanisms describe the underlying processes for the data to be missing, using the distribution of missingness. Using the framework defined by Rubin (1976) and Little and Rubin (1987), 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 variable (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 missing data.

4.3.2 Missing Data Types

It is not only important to distinguish missing data mechanisms, but also how the missing data are spread over the data set. Usually, two types of missing data 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 also to missing incoming ties (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 with networks collected over three time points, including (completely) observed covariates. The findings can, however, 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 missing completely at random (MCAR).

4.3.3 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 by the model as given. Therefore it is necessary to impute data in the first wave to provide a starting point for simulations.

Handling missingness in consecutive waves differs depending on the estimation procedure used in the *RSiena* 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 tie variables. 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 tie variables in consecutive waves are imputed by last value carried forward (Lepkowski, 1987). 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 (2017a). 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 tie variables 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 bias in the parameter estimation.

It should be noted that in the ML estimation in *RSiena* for $M \geq 3$ waves, all $M - 1$ periods from time $m - 1$ to m are treated separately. For example, when analyzing $M = 3$ waves, missing tie variables 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. 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), using ERGMs. 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 tie variables, while keeping all observed ties fixed. This provides realistic starting points that can be used not only in the simulation phase of the SAOM estimation, but also in the estimation phase. 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 tie variables in wave 1 conditional on the observed network and covariates at wave 1. The authors show how the procedure can be used for multiple imputations, but do not actually apply multiple imputation.

Recently, a new 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.

4.4 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, like Hipp et al. (2015), but takes some further steps.

First, it allows imputation of missing tie variables 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 to the procedure of Hipp et al. (2015), by using Bayesian ERGMs to impute the first wave, rather than

ERGMs (Koskinen et al., 2010; Caimo and Friel, 2011, 2013; Koskinen et al., 2013). The second option is to use a stationary SAOM to impute the first wave.

4.4.1 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 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 d th estimate of the parameter γ and $W_d = \text{cov}(\hat{\gamma}_d | x_d)$ 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:

$$\bar{\gamma}_D = \frac{1}{D} \sum_{d=1}^D \hat{\gamma}_d. \quad (4.3)$$

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

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^D W_d \quad (4.4)$$

and the between covariance matrix by

$$B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\gamma}_d - \bar{\gamma}_D)(\hat{\gamma}_d - \bar{\gamma}_D)'. \quad (4.5)$$

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

$$T_D = \text{cov}(\bar{\gamma}_D) = \bar{W}_D + \left(1 + \frac{1}{D}\right) B_D. \quad (4.6)$$

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

4.4.2 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, $m = 2, \dots, M$ and then return to the imputation of the first wave.

Multiple imputation: Missing data in later waves

For consecutive waves $m = 2, \dots, M$, missing tie variables 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 in step (1) to estimate the imputation model. 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 regular RSiena 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 estimate the model, and determine a good model specification, by considering all waves together in one analysis.

Once a fitting imputation model is obtained for the period $m - 1$ to m , we continue to step (2) and utilize ML simulation to impute the missing tie variables at wave m , conditional on the complete data for wave $m - 1$ (if there were any missings in wave $m - 1$, they were imputed in earlier steps of the procedure) and the observed data in wave m , and the imputation model estimated in step (1). 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) using the regular MoM procedure, or estimator of choice, 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 done for longitudinal network data according to the Longitudinal ERGM by Koskinen et al. (2015), in a Bayesian approach with prior distributions for the parameters. It also used in the ML procedure implemented in *RSiena* (Snijders, 2017a), but incompletely because each pair of consecutive waves is handled separately from the other waves, using very simplistic prior distributions. Further, the *RSiena* software currently does not allow to export the stochastic 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, ERMGs (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 (Caimo and Friel, 2014) in R, 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., 2019), which can be seen in equations (4.1) and (4.2). 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 structures⁴. Given the strong longitudinal dependence, it will be essential that the network in wave 2 is used as a dyadic covariate. 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.

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 (e.g., 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 the imputation of the model is not possible given the simulated trajectories alone. Changes to this are not trivial. To overcome this minor problem we create

⁴To identify corresponding parameters refer to the package manuals, for ERGMs: Handcock et al. (2007); for `RSiena`: Ripley et al. (2017).

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 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.

Multiple imputation: Summary

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

- (2.1) If the network at wave 1 has any missing tie variables, estimate a Bayesian ERGM or a stationary SAOM to impute the missing tie variables. The model specification includes the observed network at wave 2 as a dyadic covariate.
- (2.2) For $d = 1, \dots, D$:
- a) If the network at wave 1 has missing tie variables, impute by a random simulation draw from the model estimated in (2.1).
 - b) For $m = 2, \dots, M$:
 - 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 tie variables in wave m using the fitted imputation model in the ML simulation procedure, conditional on 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 set.

The advantage of multiple imputation is that it can give unbiased estimates with correct standard errors (and confidence intervals), even when the number of imputations D is low (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 by 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.

4.4.3 Estimating Imputation Models for Multiple Waves

In the multiple imputation procedure described above, missing tie variables 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 homogenous across periods (which can be tested within the SAOM framework) or (2) the networks are small (e.g., school classes). Small networks (especially with missing data) are more likely to yield unstable results, because they provide less information to reliably estimate parameters⁵. This means that

⁵The 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.

for small networks an imputation model not incorporating the information of multiple periods might not be estimable.

4.4.4 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 (4.5). 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 (4.4) (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 estimation⁶. The proposed method does not provide proper imputations, because imputations are not drawn from the full posterior distribution. However, the Bayesian ERGMs used in the first wave provide more reliable estimations of B_D than would be achieved by imputations drawn from ERGMs.

Currently in *RSiena*, Bayesian estimation for SAOMs is only implemented 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 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 by drawing them from the conditional distribution of the missing data.

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.

⁶In 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.

4.4.5 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.

4.5 Illustrative Example

4.5.1 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 (MCAR) selecting 10 (20%) of the actors and removing all outgoing ties for these actors (wave non-response). 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.

4.5.2 Missing Data Treatments

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.

Multiple imputation of wave 2 and 3

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 edge-wise shared partners, GWESP⁷, Snijders et al., 2006), 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 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 of wave 1

Following the presented procedure, two methods were used to impute the missing data in the first wave was: 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. It included parameters for

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

edges, reciprocity, triadic closure (geometrically weighted edgewise shared partners, GWESP), two-paths (geometrically weighted dyadwise shared partners, GWDSP⁸), 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 (used for estimating the BERGMs) does not allow missing data on dyadic covariates. After multiply imputing the first wave, the proposed procedure of Section 4.2 was applied to impute waves 2 and 3 using the SAOM with the parameters identified earlier to obtain $D = 50$ imputed data sets. 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., 2007). We were unable to obtain a converged model with the same parameters used in the BERGM. Therefore, the parameters for the in- and outdegree distributions (geometrically weighted indegree and outdegree) and the reciprocated transitive triad were excluded. 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 multi-

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

group analysis was performed on the separate periods (MoM estimation), treating the two periods as separate groups.

4.5.3 Results

The estimated models are presented in Figure 4.1, model estimates can be found 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 other 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, for each imputed data set, 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 between imputation variance of the 1st-wave imputation with ERGM are very small. This is the case for two reasons. First, missing data are only treated in the first wave, and thus missing data in waves 2 and 3 stay unaltered and are handled by *RSiena*'s default treatment. Therefore the target statistics are the same for all D estimations, leading to very similar estimated parameters⁹. Second, the differences between the imputed tie variables are smaller compared to the imputation with BERGMs, because only a single parameter vector was used for the imputation, while BERGM imputation draws D parameter vectors from the estimated posterior distribution. The small variance could lead to underestimation of the uncertainty about the estimated parameters.

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.

⁹The estimated parameters are not identical, because the starting points are different due to imputation of the first wave, and because the estimation is a stochastic process.

Table 4.1: Estimated model parameters (and standard errors) for the complete and treated incomplete data.

	Complete data	Default MoM	MI-BERGM	MI-SAOM	1st-ERGM	Inclusive Sampling
Friend rate 1	7.25 (1.30)	4.65 (0.97)	5.48 (1.13)	5.85 (1.17)	6.08 (1.31)	3.45 (0.89)
Friend rate 2	5.41 (0.96)	3.23 (0.64)	5.13 (1.00)	4.98 (0.97)	3.34 (0.70)	3.21 (0.90)
Density	-0.70 (0.56)	-0.79 (1.10)	-0.51 (0.64)	-0.58 (0.64)	-0.98 (0.80)	-1.87 (1.02)
Reciprocity	2.75 (0.29)	3.54 (0.94)	2.66 (0.33)	2.61 (0.33)	3.00 (0.67)	3.53 (0.54)
GWESP	2.43 (0.24)	3.32 (0.62)	2.54 (0.29)	2.50 (0.30)	2.76 (0.52)	2.81 (0.53)
GWESP \times Rec.	-0.85 (0.44)	-2.30 (1.34)	-0.71 (0.60)	-0.65 (0.62)	-1.72 (1.09)	-1.63 (0.93)
Indeg. pop. sqrt.	-0.60 (0.23)	-1.00 (0.47)	-0.77 (0.27)	-0.69 (0.27)	-0.77 (0.37)	-0.31 (0.46)
Outdeg. act. sqrt.	-0.64 (0.20)	-0.52 (0.30)	-0.59 (0.21)	-0.61 (0.21)	-0.47 (0.24)	-0.56 (0.39)
Ego alcohol	0.10 (0.11)	0.18 (0.19)	0.12 (0.13)	0.11 (0.13)	0.11 (0.15)	-0.04 (0.21)
Alter alcohol	-0.03 (0.10)	0.03 (0.16)	< .01 (0.11)	0.01 (0.11)	-0.02 (0.14)	0.09 (0.19)
Alc. similar.	1.09 (0.56)	1.16 (0.82)	1.02 (0.64)	1.08 (0.66)	1.72 (1.09)	0.33 (1.12)
Alcohol rate 1	1.34 (0.37)	1.27 (0.33)	1.30 (0.34)	1.31 (0.34)	1.29 (0.35)	1.63 (0.55)
Alcohol rate 2	1.82 (0.45)	1.78 (0.48)	1.83 (0.49)	1.81 (0.48)	1.80 (0.49)	1.52 (0.52)
Alc. linear	0.36 (0.17)	0.35 (0.15)	0.38 (0.17)	0.37 (0.17)	0.36 (0.15)	0.23 (0.17)
Alc. quadratic	-0.07 (0.11)	-0.12 (0.10)	-0.07 (0.11)	-0.07 (0.11)	-0.11 (0.10)	-0.22 (0.14)
Avg. alter Alc.	3.63 (2.10)	2.08 (1.84)	3.72 (2.19)	3.50 (2.09)	2.36 (1.86)	0.69 (1.80)

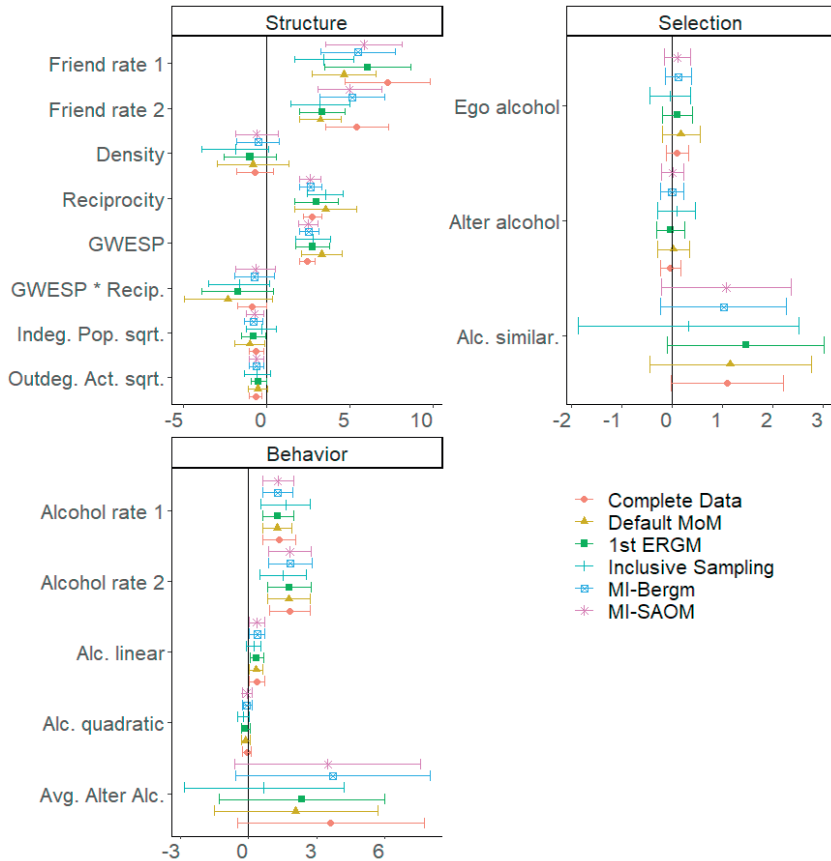


Figure 4.1: **Estimated parameters for the complete data and the five treatment procedures.**

The estimated parameters are presented in three blocks, structural network parameters (upper left), network alcohol selection parameters (upper right), and alcohol evolution parameters (lower left). The parameter estimate for the complete data is given with a dot (confidence interval in red), the estimate for the default missing data treatment with a triangle (confidence interval in yellow), the estimate for first wave ERGM imputation with a square (1st-ERGM; confidence interval in green), inclusive sampling with a vertical bar (confidence interval in light blue), multiple imputation with first wave BERGM imputation with crossed square (MI-Bergm; confidence interval dark blue), and the estimate for multiple imputation with with first wave stationary SAOM imputation is given with a star (MI-SAOM; confidence interval in pink).

It is important to emphasize that this limited illustration is not an exhaustive comparison of the methods and the performance depends on a multitude of factors, such as missing data mechanism, type and frequency. 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.

Table 4.2: Ratio of between imputation variance on the total variance (B_D/T_D).

	MI-BERGM	MI-SAOM	1st-ERGM
Friend rate 1	0.26	0.25	0.11
Friend rate 2	0.20	0.22	< 0.01
Density	0.12	0.13	0.01
Reciprocity	0.26	0.23	0.01
GWESP	0.16	0.21	0.01
GWESP \times Rec.	0.23	0.28	0.01
Indeg. pop. sqrt.	0.19	0.20	0.01
Outdeg. act. sqrt.	0.13	0.13	0.01
Ego alcohol	0.14	0.15	0.02
Alter alcohol	0.16	0.14	0.01
Alc. similar.	0.13	0.17	0.03
Alcohol rate 1	< 0.01	< 0.01	< 0.01
Alcohol rate 2	< 0.01	< 0.01	< 0.01
Alc. linear	0.01	0.01	< 0.01
Alc. quadratic	0.03	0.03	0.01
Avg. alter Alc.	0.10	0.09	0.03

4.6 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 *RSiena* software and other proposed alternatives. Especially standard errors seem to be estimated more reliably. The large standard errors 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 a distribution of 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 *RSiena* and the procedures proposed by Hipp et al. (2015); de la Haye et al. (2017), which is also supported by our small example. It utilizes more of the available information and conserves the relationships between all variables.

Missing Network and Attribute Data

Multiple Imputation for Longitudinal Coevolution Models

5.1 Introduction

Recently a new algorithm for multiple imputation of missing tie variables in longitudinal network research using stochastic actor oriented models (SAOMs; Snijders, 2001, 2017b) has been introduced (Krause et al., 2018a). Most studies using SAOMs, however, do not only model the evolution of a network through time, but focus on the coevolution of a network and an actor behavior, attitude, or attribute - from now on referred to as behavior (for an introduction into coevolution modeling see Steglich et al., 2010). In this study we extend the multiple imputation algorithm to also handle missing data in behavior variables. We extend work by Zandberg and Huisman (2019) on missing data handling for missing behavior variables in network and behavior coevolution. Zandberg and Huisman (2019) evaluated several treatment methods for missing behavior observations, among those notably multiple imputation by chained equations (MICE; van Buuren, 2012). Their work, however, showed that this proposed procedure alone does not outperform the currently established default treatment of missing behavior data, which we will detail below. We will thus combine their proposed treatment with the algorithm of Krause et al. (2018a), that is, we will utilize the proposed MICE imputation for missing data in the first observation wave and continue with imputation by SAOM in later waves. This study is organized as follows. Section 5.2 introduces coevolution SAOMs, Section 5.3 details the missing data problem and the established default treatment. Section 5.4 details the proposed multiple imputation procedure for network and behavior

This chapter is co-authored by Anna Iashina, Mark Huisman, Christian Steglich, and Tom Snijders.

missings. In Section 5.5 we apply the proposed procedure to an illustrative example. We end the paper with a discussion of the findings.

5.2 Coevolution SAOMs

SAOMs are stochastic network models developed for modeling the (unobserved) change processes between two (or more) observed time points in a network and potentially co-evolving behavior variables. A key assumption of the SAOM is that the change between the observed network at time points m and $m + 1$ can be decomposed into multiple small steps. Not all tie variables change at once between the observations, but the tie variables change in small steps (so called mini steps) one after the other. And similarly, not all actors change their behavior simultaneously, but actors in the network change their behavior in mini steps, increasing or decreasing their behavior step by step. SAOMs assume a mixture of these two continuous time processes, with one process governing network change and one process governing behavior change. Most often this chain of changes is not observed, for SAOMs for data with fully observed chains of mini steps see Stadtfeld et al. (2017).

In a coevolution model we do not only model the change in a network variable x between two (or more) observed time points, but also in a coevolving behavior variable z . As before x denotes the $n \times n$ adjacency matrix where n is the number of actors with $x_{ij} = 1$ if there is a tie from actor i to actor j and $x_{ij} = 0$ when there is no tie (self nominations are not allowed, $x_{ii} = 0$). We further define $x(m)$ as the m th observation of x .

In this paper, we only focus on the ordinal variant of the SAOM, because for the continuous behavior SAOM (Niezink et al., 2019) currently lacks an implementation of maximum likelihood simulation, a feature crucial for the proposed imputation procedure, as explained below; see also Krause et al. (2018a). The behavior z can be expressed as an ordinal vector of size n with a value for each of the n participants, $z(m)$ being the m th observation of z .

The evolution of these variables (x and z) is modeled as two coupled processes, one expressing the network change given the behavior, the other expressing the behavior change given the network. Each of these processes consists of a rate function, that determines which actor and when makes a decision according to an exponential model for waiting times, and an objective function that models which decision is made by the chosen actor according to a multinomial (or conditional) logit discrete choice model. A model with two dependent variables thus has two rate functions and two objective functions. The rate functions for

behavior and network compete, that is, both functions assign waiting times for the respective change opportunities (network and behavior) to all actors. Then the shortest waiting time is chosen and the actor has the chance to either change the network or behavior, depending on whether this shortest waiting time was assigned by the network rate or the behavior rate function. After choosing an actor using the network rate function, the objective function for the network x determines which network decision is made by the actor. The options are to create a tie to a yet unconnected actor, to drop a tie to a connected actor, or to do nothing, resulting in n possible choices. The objective function for the behavior variable z models the actors decision to either increase or decrease the behavior by one step, or to remain at the current level of the behavior, resulting in three possible choices. Actors with z values at the maximum or minimum of the scale have only two choices, to decrease/increase or remain. However, a model variant exist that allows for three choices (Ripley et al., 2019). Here, choices beyond the extreme range remain part of the choice set. If a value beyond the minimum/maximum is chosen by the objective function, the actor will remain at the extreme value. We will use this so-called absorbing model in the example below.

At each mini step the probability for each possible decision of an actor is determined by the objective function, which is a function of actor-specific network and behavior statistics (including effects of covariates) s_{ki} weighted by parameters of the evolution process θ_k given the current state of the network and behavior variables:

$$f_i^x(\theta, x, z) = \sum_k \theta_k^x s_{ki}^x(x, z), \quad (5.1)$$

$$f_i^z(\theta, x, z) = \sum_k \theta_k^z s_{ki}^z(z, x). \quad (5.2)$$

The network statistics s_{ki}^x for the network evolution can be subgraph counts (or non-linear transformations thereof) in the neighborhood of focal actor i (e.g., outdegree, or the number of reciprocated ties) or functions of attribute values of i or the (potential) receiving actor j . The network statistics s_{ki}^z for the behavior evolution likewise can be subgraph counts and attribute values, and often include effects of attribute levels of connected alters (e.g., the average level of z of all alters i is connected to). These statistics are always calculated based on the current state of the coevolution process when the mini-step takes place.

Because the true sequence of mini steps is unobserved and the potential network states are far too numerous in all but trivially small networks, it is not possible to give an analytic expression for estimating θ . SAOMs solve this issue by using simulation. The default estimation is done via method of moments (MoM; Robbins and Monro, 1951; Snijders, 2001), that is, parameters are estimated so that for a set of target statistics corresponding to the model parameters the expected value, approximated by simulation with these parameters, is equal to the observed values of the target statistics. Another way of estimating parameters and simulating networks is by maximum likelihood (ML; Snijders et al., 2010a). The proposed missing data imputation procedure uses ML simulation. ML simulation differs from MoM simulation in a crucial way. Network and behavior trajectories simulated with MoM are only conditional on the starting network at $m - 1$ and the estimated parameters θ . This means that the simulations start from the observed data at wave $m - 1$ but do not end necessarily at the observed data at m . However, the target statistics of the simulated network at the end of the process will, on average, be similar to those observed at wave m (e.g., the network will have the same amount of ties, reciprocated dyads, triangles, and homophilous ties as the observed data).

In contrast, simulation with ML is conditional on the observed network at time points $m - 1$ and m . ML requires that all simulated trajectories meet the observed data exactly, and not only in expectation on some specified target statistics. This means that networks simulated with ML for the period $m - 1$ to m always are identical to the network observed at m , and what is simulated is the sequence of mini steps connecting them.

Although SAOMs are primarily used for longitudinal data, a cross-sectional variant also exists. Here it is assumed that the observed network and behavior are the outcome of a continuous, stationary process in (at least short-term) equilibrium. For a short introduction into stationary SAOMs see Krause et al. (2018a), for a more detailed introduction see Snijders and Steglich (2015).

5.3 Missing Data

5.3.1 Missing Data Mechanisms

Missing data mechanisms describe the probability 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 the probability of a tie (or individual behavior score) variable to be missing is independent of the observed data and the value of the missing tie variable (or

individual behavior score). Data are missing at random (MAR) if the probability to be missing is independent of the missing tie variable (or individual behavior score) itself, but is related to other observed variables (e.g., older participants 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 that proper missing data techniques are applied, they will yield no bias in a resulting analysis. Lastly, data are missing not at random (MNAR) if the missingness is dependent on missing data, leading to biased results unless the missing data mechanism is modeled correctly. MNAR data are therefore called non-ignorable.

5.3.2 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. 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.

5.3.3 Missing Data in SAOMs

The default method used in estimation of SAOMs as implemented in *RSiena* (Ripley et al., 2019) is explained in detail in Huisman and Steglich (2008); Krause et al. (2018a). Here, we will only focus on missing data in the behavior variable. The missing data treatment for behavior missings, unlike the one for network missings, is independent from the estimation method. The procedure follows three steps. First, missing data is imputed with last observed value carried forward. Second, if there is no previous observation, then missing data is imputed with next observation carried backward. Third, if there is still missing data, that is, if there is complete actor non-response, missing data is imputed with the mode at each wave. In case of multiple modes, the lowest is imputed. Similar to the default imputations for network missings, the imputed behavior

scores do not contribute to the calculation of target or change statistics, and have only indirect influence on the estimation process via the simulations.

5.4 Multiple Imputation with SAOMs

We adapt the multiple imputation algorithm proposed by Krause et al. (2018a) to incorporate missing data in behavior variables. As a reminder, missing data on the network variable are imputed wave by wave, using imputations of previous waves as starting point for imputations on consecutive waves. The imputation for each wave was split into two parts: First, estimate an imputation model from the previous wave to the wave with the missing data that is to be imputed using MoM estimation. Second, use the estimated model for ML simulation and retain the simulated missing tie variable at the target wave. For the first wave we proposed either using imputation by Bayesian ERGMs (Caimo and Friel, 2011; Krause et al., 2018b), or by a stationary SAOM (Krause et al., 2018a).

5.4.1 Imputing Behavior

Imputation of the behavior variable adds an additional layer to the above outlined multiple imputation procedures. SAOMs are models of change, thus the starting point of the estimation can have a strong influence on the estimated change process. In networks with binary tie-variables there are only two possible starting points per tie variable ($x_{ij} = 0$ or $x_{ij} = 1$). Individual behavior scores, however, have generally more possible starting points (often the range of behavior variables is between three and seven). Thus the impact of the starting values can be much higher for the behavior variable, compared to the network variable. While a wrongly imputed tie variable is only one step away from the correct value, a wrongly imputed behavior variable can easily be four, five, or six steps away. Therefore, missing data in the behavior variable in the first wave are imputed in a two step process. First, missing data are imputed using a multiple imputation by chained equations (MICE) algorithm (van Buuren, 2012), using the procedure proposed by Zandberg and Huisman (2019). Then, missing data in the behavior and network in the first wave are jointly imputed with a stationary SAOM, using the imputed behavior values from the first step as starting points. Missing data in later waves are imputed jointly with missing network data using the proposed SAOM procedure. That is, first, a coevolution SAOM is estimated for period $m - 1$ to m . Then missing data in wave m is imputed with ML simulation for both the behavior and the network. This joint

imputation process maintains the dependence between behavior variables and network in both directions.

5.4.2 MICE

The MICE procedure exploits the idea that multiple imputation may be done as a sequence of numerous steps. With MICE we are able to account for missing data mechanisms, preserve the relationships between the variables, and estimate the uncertainty about those relationships efficiently. The algorithm follows a variable-by-variable logic to impute each variable conditionally on the observed (and imputed) other variables. First, a model, usually a normal regression model, for the first behavior variable containing missing data is estimated using all available data. This includes observed covariates, observed behavior data at the second wave $z(2)$ and higher waves, as well as measures derived from the network (e.g., indegree). Next, this model is used to impute the missing data. Then, a model for the next variable containing missing data is estimated. From the second step onward, the data used for the estimations is composed of the observed data (the same in all steps) and the provisionally imputed data (potentially changing from step to step). This process is continued for every behavior variable with missing data, as well as any of the included predictor variables with missing data. The whole process is repeated T times until it is converged. The T th imputation of the variables is retained and the process is repeated until D imputations are obtained. Let ${}^d z(1)$ be the d th retained MICE imputation of $z(1)$.

The MICE algorithm is implemented in the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011) in R (R Core Team, 2019). The package offers several options for imputation models depending on the nature of the variable to impute, for instance, linear regression for continuous variables or logistic regression for binary or categorical variables. It also offers Predictive Mean Matching (*pmm*; Little, 1988), which we, in line with Zandberg and Huisman (2019), propose to use. In *pmm* the imputation is based on a linear regression model. However, the imputed value is not the one predicted by the regression, but the predicted value is used to identify the 3 closest observed data points with similar prediction. Then, randomly, one of these is taken as the imputation. This method is proven to be a good imputation method in general and is suitable for both numeric and categorical data. More importantly, SAOMs are generally used with ordinal behavior variables, thus only possible observed scores should be imputed, even if the imputation model predicts scores outside the observed range. Below we present the algorithm in detail. Here,

θ_k is the parameter vector specifying the distribution of the variable Y_k given $p(\theta_k | Y_1, \dots, Y_{k-1}, Y_{k+1}, \dots, Y_K)$, that is, the parameter of the imputation model. In mice, θ_k is drawn from a Bayesian posterior distribution. This Bayesian estimation uses non-informative priors for the estimation of the posterior, see van Buuren (2012) for more details on the mice procedure. The MICE algorithm is presented in Algorithm 4.

Algorithm 4 MICE algorithm

```

i. Create  $Y$  that contains  $z$  and additional predictor variables  $V$ , with a total
of  $K$  variables
ii. Initialize the procedure by imputing missing data in  $Y_k$  with random draws
from the observed values in  $Y_k$ 
for  $d = 1, \dots, D$  do
  for  $t = 1, \dots, T$  do
    for  $k = 1, \dots, K$  do
      draw  $\theta_k^{(t)}$  from  $p(\theta_k^{(t)} | Y_1^{(t)}, \dots, Y_{k-1}^{(t)}, Y_{k+1}^{(t-1)}, \dots, Y_K^{(t-1)})$ 
      impute  $Y_k^{(t)} \sim pm(Y_k | Y_1^{(t)}, \dots, Y_{k-1}^{(t)}, Y_{k+1}^{(t-1)}, \dots, Y_K^{(t-1)}, \theta_k^{(t)})$ 
    end for
  end for
  Retain  ${}^d z(1)$  from  $Y^{(T)}$ 
end for

```

Generally T can be set to a rather low number of iterations, say 20 (van Buuren, 2012), but convergence of the algorithm should be assessed before the imputations can be used. Further, V should include all relevant predictors. However, including too many predictors is not recommended. One should consider the number of observations per variable, and the number of complete cases providing information for the estimation of the parameters. Predictors may contain missings themselves, which will be also imputed during the procedure.

In the concrete case of using MICE to impute the first wave behavior z for a SAOM, the most important variable to include in Y are z values at the second wave, and potentially later waves. We further suggest to include several network measures in Y . These measures should reflect, as closely as possible, all relevant z network processes related to z . Sometimes including the exact network measure is not possible, for instance, the average z value of alters cannot be calculated if all outgoing ties of an actor are missing. We suggest to use incoming ties instead of outgoing ties to identify alters, for instance, the average z behavior of alters sending ties to the missing actor, $v_i = \sum_j z_j x_{ji} / \sum_j x_{ji}$ (and the average z if the denominator is 0). If the analysis model assumes an influence effect of the sum of the behavior values of all alters' i is connected to, this can

best be captured by calculating the sum of all z values of all alters sending ties to the missing actor ($v_i = \sum_{j=1} z_j x_{ji}$)¹.

If the coevolution of more than one behavior variable (z^1, \dots, z^B) is modeled, missing data in these variables should be imputed jointly, that is, Y should contain all the variables z^1, \dots, z^B .

5.4.3 Stationary SAOM Imputation

After D imputations of the first wave behavior have been obtained with MICE, missing network and behavior data of the first wave are imputed jointly using stationary SAOMs. Ideally, a stationary SAOM is estimated for each of the D imputations. However, for better convergence it is also possible to conduct a combined SAOM estimation using all D imputations together with the multi-group option of *RSiena* (Ripley et al., 2019). In the multi-group option the estimation algorithm assumes homogeneity of the D groups and jointly estimates one set parameters, using the data from all groups. This procedure uses the observed network data D times in the estimation process, each time paired with a different behavior imputation retained from the MICE procedure. This multiplication of data will lead to a much more stable estimation of the stationary SAOM. The resulting standard errors are highly underestimating the true uncertainty, however, the proposed algorithm only relies on reliable estimation of the parameters.

This has the downside that only one set of parameters is estimated for the imputation, instead of the usual D sets of parameters. One problem regarding multiple imputation is the accurate estimation of standard errors of parameters. This estimation needs to take the uncertainty about the missing data into account. Rubin (1987) solved this problem by taking the variance between the D results estimated on the D imputed data sets into account. However, if all imputations are drawn with the same set of parameters, then it is likely that this between imputation variance is underestimated, leading to an underestimated standard errors in the final results. A solution that both utilizes the data from all D imputations but still takes the uncertainty surrounding the missing data properly into account would be a Bayesian estimation procedure for SAOMs (Koskinen and Snijders, 2007). Bayesian imputation, its advantages, and difficulties for network models are discussed in Krause et al. (2019b)². Let $d_x(m)$

¹Similarity scores cannot be calculated for actors with missing z values. If the analysis model contains similarity effects the recommendation is to use approximations like the ones described above (e.g., if the hypothesis is about the average similarity effect the closest approximation would be the average value of incoming alters).

²Chapters 6 of this dissertation.

be the d th stationary SAOM imputation of x at wave m and ${}^d z(m)$ the d th stationary SAOM imputation of z at wave m . Further, we use the convention that u represents the observed part of the data. Let ux be the observed network data and uz be the observed behavior data. In this paper we present the algorithm for the non-Bayesian, combined analysis in Algorithm 5.

Algorithm 5 Algorithm for imputing missing behavior and network data in wave $m = 1$

I Impute missing data in $z(1)$ with MICE and obtain D imputations $\tilde{z}(1)$
II Estimate θ for a stationary SAOM for $(x(1), \tilde{z}(1))$
 with MoM using the second wave network data $x(2)$ and the second wave behavior $z(2)$ as dyadic covariate with multi-group option for D data sets combined
for $d = 1, \dots, D$ **do**
 Draw a joint imputation $({}^d x(1), {}^d z(1))$ with ML simulation
 from $p(x(1), z(1) \mid \theta, ux(1), ux(2), {}^d \tilde{z}(1), uz(2))$
end for

5.4.4 Later Waves

The following waves are imputed wave by wave. Imputations are performed in parallel, each using the previous imputed wave as a starting point for the imputation in the current wave. Network and behavior are imputed jointly. A new parameter vector ${}^d \theta(m)$ is estimated for each wave m and imputation d . The algorithm is presented below in Algorithm 6 and is an extension of the algorithm proposed in Krause et al. (2018a).

Algorithm 6 Algorithm for imputing missing behavior and network data in waves $m \geq 2$

for $d = 1, \dots, D$ **do**
 for $m = 2, \dots, M$ **do**
 (I) Estimate ${}^d \theta(m)$ for with a longitudinal SAOM from wave $m - 1$ to m with MoM using ${}^d x(m - 1)$ and ${}^d z(m - 1)$
 (II) Draw a joint imputation $({}^d x(m), {}^d z(m))$ with ML simulation from $p(x(m), z(m) \mid {}^d \theta(m), {}^d x(m - 1), {}^d z(m - 1), ux(m), uz(m))$
 end for
end for

5.4.5 Multiple Groups

Often researchers collect data in more than one group, and, usually in some, or all, groups missing data occur. The imputation procedure described above

follows the same principle when more than one group is analyzed. If a multi-level or multigroup estimation procedure is planned for the final analysis, the same procedure should be followed for the imputation. This means that the imputation parameters for the different groups are estimated jointly. It is advisable to include data sets without missing data (if present), in this process, to obtain more reliable estimates of the imputation parameters. If each group is analyzed separately and results are to be combined in a meta-analysis later, then imputation can also be done separately following the described procedure for each group. However, it might be necessary to combine the data for the estimation of parameters, and it might even be preferable to obtain more reliable estimates. Multiple imputation for multiple groups with Bayesian estimation of SAOMs is discussed in Krause et al. (2019b).

5.5 Illustrative Example

5.5.1 Data Description

For illustration purposes we apply the described procedure to a friendship network of 103 school students that was originally a part of the Teenage Health and Life Style study data set (Michell and Amos, 1997; Pearson and West, 2003; Steglich et al., 2006), observed at two time points ($M = 2$). Two dependent behavior variables were included in the analysis: smoking and drinking. Smoking was measured on a three-point scale and drinking was measured on a five-point scale. Gender was used as nodal covariate. The original data set contained 160 students, however, 57 of these had missing data on either the network or one of the behavior variables in at least one measurement point. We decided to remove these students, because it is not possible to investigate the effect of missing data and the efficiency of treatment methods without knowing the true values.

In this demonstration, we set 20% of the actors to missing completely at random (MCAR) in each wave both for the behavior and network parts (actor non-response, i.e., for 20% of the actors all outgoing tie variables and values for both behavior variables were removed in each wave). This random selection of actors led to five actors (of the 21 missing in each wave) to be missing both in the first and second wave. For simplicity, no missing data were created on the variable gender. Generally, covariates may contain missing data and could also be imputed within an extension of procedure described in this paper. This extension would treat the missing covariates as dependent behavior variables and impute with the described procedure. This could, however, lead to very

complex models with multiple dependent behavior variables. Further, some covariates (e.g., gender, ethnicity) cannot meaningfully be seen as dependent variables in a longitudinal SAOM. They could, however, be meaningfully imputed using stationary SAOMs. Alternatively, covariates could also be imputed using MICE; such an imputation must incorporate the relationship between the covariate and the network and behavior variables. We set $D = 100$ imputations.

5.5.2 Imputation Model

The procedure required to formulate three imputation models: one for the MICE imputation, one for the stationary SAOM, and one for the longitudinal SAOM.

MICE Imputation

We used the `mice()` function of the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011) in R (R Core Team, 2019). The following variables were included in the MICE imputation: Both behavior variables of first and second wave, gender, indegree and squared indegree of each node, as well as the average behavior values of indegree alters in waves one and two. The number of iterations before imputations were retained was set to $T = 50$. This large value for T was chosen to better to evaluate the convergence of the MICE procedure, however, much smaller values (e.g., $T = 10$) would suffice.

Stationary SAOM Imputation

The model for the stationary SAOM was nearly identical to the analysis model, which we describe below. There were two differences in the model configuration: First, triadic closure was modeled with Geometrically Weighted Edgewise Shared Partners (GWESP Snijders et al., 2006). Transitive triplet closure can lead to degenerate networks in long run stationary network models. This does not necessarily constitute a problem with rate functions fixed to small values. However, to be on the safe side we used GWESP instead of regular closure for the stationary model. In this illustration we fixed the network rate function for to 5 and the behavior rate functions both to 3. As of yet, large simulation studies exploring the importance of the value of the stationary rate function for SAOM imputation are missing. We assume that small values, such as 5 or 10, should lead to reasonable imputation models. An advantage of smaller rate function is the computation time; smaller rate functions mean fewer decisions per actor, leading to faster estimation and imputation. Second, the stationary model

uses the network and the behavior observed at the second wave ($x(2), z(2)$) as covariates, outlined the imputation algorithm for wave one above.

Longitudinal SAOM Imputation

The longitudinal SAOM imputation model was equal to the analysis model. The model contained several structural effects: Density, reciprocity, transitive closure, reciprocated transitive closure, reciprocal degree activity, effects modeling the tendency of nodes to receive more ties when receiving ties (indegree popularity squared), and modeling the tendency for actors to send more ties when sending many ties (outdegree activity squared). Further, effects modeling the sender, receiver, and homophily effects for gender and the behavior variables for selection were included.

On the behavior side the model included effects modeling the general linear and quadratic trends of the behavior. Further, the main effects of gender and the respective other behavior were included, as well as the average behavior of outgoing alters to model social influence.

5.5.3 Results

Of the $D = 100$ imputed data sets two were unable to reach convergence. A further 19 data sets, although satisfying the usual convergence criterion of a maximal t -convergence ratio of $t_{covmax} < .25$ did not yield meaningful results, with parameters in the objective function estimated as 20 or larger. Estimates of this size are not meaningful for SAOM models (or any logistic model), because they lead to tie probabilities of $p(x_{ij} = 1|x, y, \theta) = 1$. These estimates are the consequence of very unlikely imputations due to failed imputation models. We recommend to exclude such models from the final analysis. If such an exclusion would lead to only very few imputations remaining this might imply that the imputation model was not optimally chosen and should be improved. It is also possible to simply increase the number of imputations such that eventually sufficient useable imputations remain. In this example, after the exclusion $D = 79$ imputations remained, a number sufficiently large.

The results based on the $D = 79$ imputations, together with the estimates from the complete data and those from the RSiena default missing data treatment are presented in Figure 5.1 and Figure 5.2. Both the default treatment and the multiple imputation overall provide estimates close to those obtained with the complete data. The results indicate the multiple imputation leads to estimates closer to the complete data for parameters and standard errors compared to the

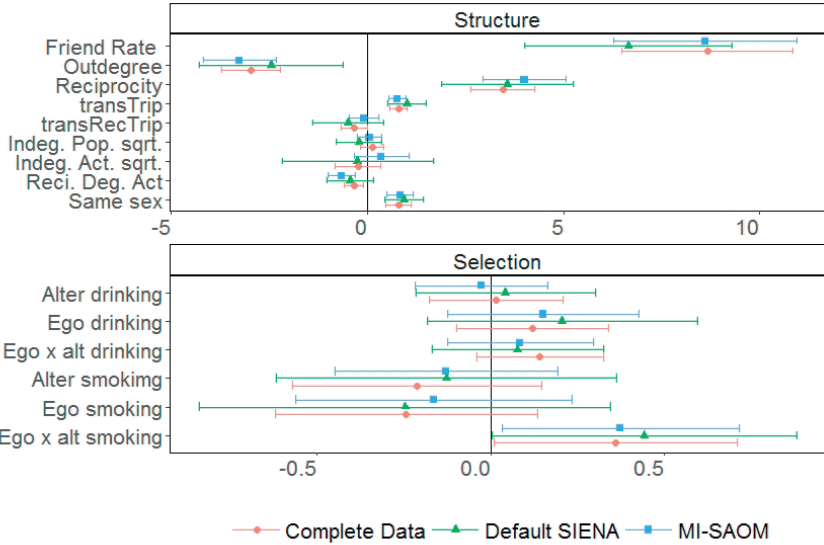


Figure 5.1: Results for the network evolution model

This figure shows the estimated parameters and 95% confidence intervals (1.96 times standard error) for the network evolution. Estimates are separated between general structural effects (upper) and behavior specific selection effects (lower). The complete data estimates are presented with circles (confidence interval in red), default missing data treatment with triangles (confidence intervals in green), and the multiple imputation procedure with squares (confidence interval in blue).

default treatment. However, multiple imputation does not provide standard errors that are closer to the complete data estimate for the behavior variables compared to the default treatment, with the exception of the rate functions.

Multiple imputation is likely to perform worse for behavioral variables compared to network variables for several reasons. First, the network generally contains more data, more observed change, than the behavioral variables, thus more observed information is available to support the imputation. Second, network models are generally more complex than behavior models. This allows that more of the available information is used. In this example two parameters guiding behavior change are reliant on other imputed data. The effect from the other behavior relies on a good imputation of the other behavior, the effect of the average behavior score of friends is reliant on reliable network imputations. The network on the other hand, although the network model had several behavior related selection effects, had strong predictors unrelated to the behavior scores (e.g., reciprocity and triadic closure). Thus, variance in the behavior

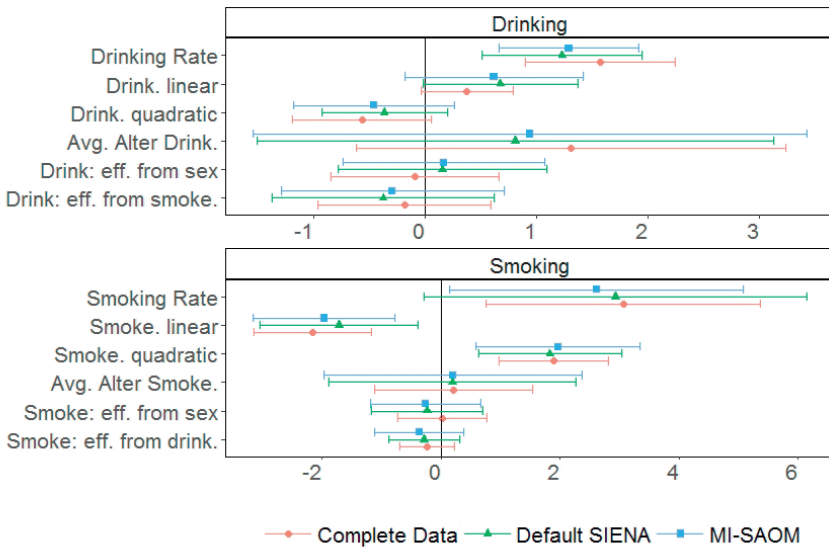


Figure 5.2: Results for the behavior evolution models

This figure shows the estimated parameters and 95% confidence intervals (1.96 times standard error) for the behavior evolutions. Estimates are separated between the drinking behavior (upper) and smoking behavior (lower). The complete data estimates are presented with circles (confidence interval in red), default missing data treatment with triangles (confidence intervals in green), and the multiple imputation procedure with squares (confidence interval in blue).

imputations had less impact on the network model, compared to variance in the network and behavior imputations on the behavior model. In general, multiple imputation will lead to larger standard errors compared to those obtained with the complete data, after all, multiple imputation has less observed data available. The results presented here are not an exhaustive investigation of the performance of the MI procedure, but only a small demonstration.

5.6 Discussion

In this study we presented an extension of the multiple imputation procedure for missing data in longitudinal network studies introduced by Krause et al. (2018a). The extended algorithm is capable of handling missing nodal attributes jointly with missing tie variables. The algorithm still needs to be tested under different missing data rates and mechanisms, as well as on differently sized and structured networks. Further, the imputation model for the behavior variable

can potentially be improved. Here, the standard errors obtained via multiple imputation do not appear to be closer to the complete data estimates than those obtained by the default procedure. These findings are, however, only based on a single case. More research is necessary to properly estimate the effect of the imputation procedure. It is further important to note that the example contained no strong predictor of the behavior changes. Neither the direct effects of gender or the other behavior, nor the influences effects were strong. We expect that the imputation procedure will lead to more reliable imputations, that is, imputations closer to the observed data, in cases where the link between behavior and network is stronger, and thus more information is utilized in the imputation model. In general, it might be necessary to formulate more complex models for the dependent behavior variables to allow the use of more information during the imputation. The imputation model can exceed the final analysis model in complexity, but should always contain all parameters of the final analysis model. Finally, the results using multiple imputation must not always be closer to the complete data estimate, even under ideal conditions. The uncertainty created by the missing data is taken into account in the estimation of the final standard errors (see e.g. Rubin, 1987; Krause et al., 2018a), thus the resulting standard errors are necessarily larger, reflecting the added uncertainty, while standard errors obtained by the *RSiena* default procedure are too small, given the available information.

Despite these limitations we are confident that the proposed algorithm is an improvement to the default procedure and the only network focused multiple imputation algorithm.

Extensions for Missing Network Data

Multigroup, Multiplex, and Bayesian Procedures for Longitudinal Network Data

6.1 Introduction

In this study we extend a recently introduced algorithm for multiple imputation of missing tie variables in networks (Krause et al., 2018a, 2019a) in three important ways. First, the proposed algorithm will be applied in a multigroup setting, where the same model is investigated with more than one group. Second, multiple imputation for the multiplex case (more than one dependent network variable) will be introduced. Third, the imputation and estimation are fully Bayesian. In addition to these algorithmic advances the applied example will highlight several issues that can be encountered when applying multiple imputation to empirical network data and how these can be overcome. The three additions to the algorithm each come with benefits to the imputation procedure, but also add their own complications. We assume the reader is familiar with Stochastic Actor Oriented Models (SAOMs; Snijders, 2001, 2017b), multiple imputation (Schafer and Graham, 2002; van Buuren, 2012) and, specifically multiple imputations for SAOMs (Krause et al., 2018a, 2019a).

This study is structured as follows. Section 6.2 gives a brief introduction to the topic of missing data. A short introduction into SAOMs and missing data treatment with SAOMs is given in Section 6.3. In Section 6.4, we discuss the advantages and problems of multigroup, multiplex, and Bayesian analysis. Section 6.5 introduces the extended imputation procedure. An illustration of

This chapter is co-authored by Mark Huisman, Christian Steglich, Loes van Rijsewijk, and Tom Snijders.

the procedure is provided in Section 6.6. The study finishes with a discussion and according recommendations.

6.2 Missing Data

6.2.1 Missing Data Mechanisms

Missing data mechanisms describe the probability 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), in the case of network data, if the probability of a tie variable to be missing is independent of the observed data and the unknown value of the missing tie variable. Data are missing at random (MAR) if the probability of a tie to be missing is independent of the missing tie variable itself, but is related to other observed variables (e.g., older participants 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 that proper missing data techniques are applied, they will yield no bias in a resulting analysis. Lastly, data are missing not at random (MNAR) if the missingness is dependent on the missing data, leading to biased results, unless the analysis models the (unknown) missing data mechanism correctly. MNAR data are therefore called non-ignorable. The example presented in Section 6.6 is an empirical case, thus the true missing data mechanisms is unknown. However, we will assume M(C)AR for the demonstration of our algorithm.

6.2.2 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. 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. In the example below both complete actor non-response, as well as wave non-response are present.

6.3 Stochastic Actor-Oriented Models and Missing Data

6.3.1 Stochastic Actor-Oriented Models

SAOMs are a model family for the analysis of change processes in longitudinal network data (Snijders, 2001, 2017b). A core assumption of SAOMs is that the observed change between two network observations can be separated into a series of smallest possible steps (changes of single tie variables). The change process is modeled by two functions, the rate and the objective function. The rate function determines when which member of the network is allowed to change one of its outgoing tie variables. The objective function determines which decision is taken by the chosen actor (dropping an existing tie, creating a new tie, or no change). Let x denote the $n \times n$ adjacency matrix where n is the number of actors with $x_{ij} = 1$ if there is a tie from actor i to actor j and $x_{ij} = 0$ when there is no tie (self nominations are not allowed, $x_{ii} = 0$). We further define $x(m)$ as the m th observation of x . The objective function is given by:

$$f_i^x(\theta, x) = \sum_k \theta_k s_{ki}(x), \quad (6.1)$$

The network statistics $s_{ki}(x)$ can be subgraph counts (or non-linear transformations thereof) in the neighborhood of focal actor i (e.g., reciprocity, or out-degree) or functions of attribute values of i or the (potential) receiving actor j . These statistics are always calculated based on the current state of the evolution process.

The two most important estimation options for SAOMs are estimation by method of moments (MoM; Snijders, 2001) and maximum likelihood (ML) estimation (Snijders et al., 2010a). Estimation by MoM yields parameters for which the resulting expected value of a set of chosen target statistics $s(x)$ obtained via simulation is equal to the observed target statistics. ML estimation maximizes the likelihood for the estimated set of parameters to link two consecutive observation waves ($x(m-1)$ to $x(m)$). The important difference is that simulation by MoM for period $m-1$ to m yields a distribution of networks at m , for which network statistics are on average similar to the statistics of the observed network $x(m)$. In contrast, simulation with ML always ends in the observed network at wave $x(m)$. It yields a distribution of sequences of tie changes that could have taken $x(m-1)$ to $x(m)$.

6.3.2 Stationary Stochastic Actor-Oriented Models

Stationary SAOMs assume that the observed network x is the outcome of a stable, continuous, and stationary process (Snijders and Steglich, 2015). The model assumes that the observed network statistics $s(x)$ are stable (e.g., the number of ties, the number of reciprocated ties, or the number of triangles). In the estimation procedure, however, actors are allowed to change their relations and the objective function is estimated so that the network statistics $s(x)$ remain overall stable. Stationary SAOMs can be estimated using the observed network as both starting and end network for the stationary distribution (reflecting that the network statistics remain constant). This means that a rate function cannot be estimated, because no change is observed in the network. This requires that the rate function is fixed to an arbitrarily large value. It has been shown that relatively small values (5 to 10) are sufficient to obtain reliable estimates of the objective function for the purpose of missing data imputation (Krause et al., 2018a, 2019a).

6.3.3 Missing Data in SAOMs

The default procedure for missing data treatment implemented in the R package (R Core Team, 2019) *RSiena* (Ripley et al., 2019) depends on the chosen estimation method. With estimation by MoM, missing tie variables are imputation by last value carried forward, or, if no previous observation exists, imputed by 0 (i.e., the unconditional mean imputation in a sparse network). The model parameters are estimated such that the imputed values do not contribute to the calculation of the target statistics. In this way the imputed values only influence parameter estimation indirectly via the network simulation (see Huisman and Steglich, 2008; Krause et al., 2018a). For ML estimation in *RSiena*, missing data at wave m are imputed in a model-based way by simulating missing values conditional on the observed data at both time points $m - 1$ and m and the estimated parameters. Here missing data and imputation do not introduce additional bias, if the estimation model is realistic and missing data are missing at random (MAR, Snijders et al., 2010a). Missing data at wave $m - 1$ are imputed internally with random draws assuming a binary independent distribution with the observed network density as the prior distribution for tie variables.

In the case of more than two waves ($M \geq 3$), the variables imputed in wave m during the ML estimation of the period $m - 1$ to m are not utilized in the estimation of the period m to $m + 1$. For period m to $m + 1$ missing data in m will be imputed by draws from the observed behavior distribution at m . This limitation was implemented to keep the estimation more tractable and allow

parallelization of the estimation. Further, simulated values are not retained or used for parameter estimation.

6.3.4 Multiple Imputation with SAOMs

A multiple imputation procedure for SAOMs was proposed by Krause et al. (2018a, 2019a). This procedure imputes missing data wave by wave: First a SAOM is estimated for period $m - 1$ to m using MoM or ML estimation, and second, missing data at wave m are imputed with the estimated SAOM using ML simulation, conditional on the data at $x(m - 1)$ in the previous wave, data at $x(m)$ in the current wave, and the estimated parameters. This procedure is repeated for each wave, using previously imputed data as starting points for the next imputation. Missing data in the first wave ($m = 1$) have to be imputed differently, because the first wave is not modeled in a regular longitudinal SAOM. For the first wave, a stationary SAOM is estimated using the second wave observation $x(2)$ as a dyadic covariate. For a more detailed introduction into multiple imputation with SAOMs see Krause et al. (2018a, 2019a).

6.4 Extensions

6.4.1 Multigroup Network Models

The analysis of multiple groups has several advantages for network models. First, statistical models are estimated more reliably when more data is used. This is, however, not a simple task for network studies, because actors cannot be simply added to an empirical network. The only two options for network researchers to increase statistical power are either to collect more longitudinal observations of the same network, or to analyze multiple similar networks. Second, it is much harder to generalize the results estimated from only a single network. While single network studies can be very informative, they are network case studies and thus not easily generalizable. Results estimated on multiple networks are far more likely to represent the population from which they are drawn.

These advantages of multigroup research also have an effect on multiple imputation. Using multiple groups will lead to more accurate estimation of the imputation parameters, and thus, will lead to more reliable imputations. This advantage is not only due to a larger sample size. But the missing data rates are likely to vary between groups, with some having few or no missing data.

These groups allow for a much more reliable estimation of the (imputation) parameters.

Despite these benefits, the joint analysis of multiple groups comes with a major complication. While one can assume that the general tendencies of the network dynamics are the same across groups, the parameter estimates between groups might differ significantly. This can be due to differences in composition (e.g., ethnic homophily is less important in less ethnically diverse groups), differences in the general tendency to create ties (e.g., groups might differ in density), or substantial differences in network size. SAOM parameters are sensitive to the number of actors in the network. While this sensitivity is small and not relevant for small size differences (e.g., size differences between class rooms), it will lead to different parameters for substantial size differences (e.g., class rooms compared to school wide networks). Another reason can of course be that the groups are behaving systematically different from each other.

Whatever the reason, meaningfully different parameter estimates for each group will make it hard for the model to converge to one parameter estimate for all groups, and such a converged model is likely to not fit any of the groups well. There are three options for solving this issue. First, one can analyze each group separately and combine the results in a meta analysis to draw inferences. This solution, however, does not help with the imputation of missing data. If missing data occurs they would need to be treated separately for each group, applying the algorithm introduced in the Krause et al. (2018a, 2019a) for each group. Second, one can include interactions on the group level for parameters that lead to heterogeneity between the groups. While this option allows to use information from all groups to support the imputation, it might lead to very large models with many group-level interactions. The third, option is a multilevel model. Here some parameters are allowed to randomly vary between the groups allowing for heterogeneity, while other parameters are estimated as fixed effects across all groups. The estimation of the randomly varying parameters in each group also uses information of other groups to support the estimation. Such a model is implemented in the Bayesian estimation procedure in the `RSienaTest` package in R (Ripley et al., 2019; R Core Team, 2019) and will be used in the illustration below.

6.4.2 Multiplex Networks

Multiplex network models are important to better understand the relationships between actors (Krause and Caimo, 2019). Social (as well as professional) life is not bound to only one form of relation, but spans many dimensions. These

different types of relationships are often not independent of each other. To understand one form of relation, it is therefore often helpful to model other relations between the same actors as well. Moreover, multiplex models can also lead to better model fit (e.g., modeling antipathy ties leads to more reliable generation of group structures Stadtfeld et al., 2018). The downside of modeling more than one dependent network variable is that multiplex network structures add more complexity to an already challenging modeling task, as estimating appropriate network models for even one layer can be difficult. Multiplex models are further not as commonly studied as single layer network models, which means they are less well understood.

Multiplex structures can provide crucial benefits for multiple imputation. First, models that fit the data better are likely to lead to more reliable imputation. Second, if the missing data are not spread similarly over all network layers (e.g., friendship ties are observed for an actor, but advice seeking ties of this actor are missing), using the multiplex information will result in more accurate imputations. This situation, however, does not occur often with empirical data, where actor non-response is most common.

Multiplex SAOMS

For multiplex SAOMs we define x^l as the $n \times n$ adjacency matrix where n is the number of actors with $x_{ij}^l = 1$ if there is a tie from actor i to actor j on layer l and $x_{ij}^l = 0$ when there is no such tie (self nominations are not allowed, $x_{ii}^l = 0$). Assuming $L = 2$, the evolution of the two network layers (x^1 and x^2) is modeled as two coupled network evolution processes. One process models the change in x^1 given x^2 , and the other models the change in x^2 given x^1 . Similar to single layer SAOMs, each of these processes is split into a rate and an objective function, resulting in two rate and two objective functions (in the case of $L = 2$). The resulting two objective functions, which determine the changes in the network, are modeled as a weighted sum of actor-specific network statistics s_{ki} on both layers and covariates weighted by parameters of the evolution process θ_k , given the state of the network layers at the current mini step:

$$f_i^{x^1}(\theta, x) = \sum_k \theta_k^1 s_{ki}^1(x), \quad (6.2)$$

$$f_i^{x^2}(\theta, x) = \sum_k \theta_k^2 s_{ki}^2(x). \quad (6.3)$$

Here, $s_{ki}^l(x)$ depend on the current state of the network $x = (x^1, x^2)$, and are not limited to the network layer on which the decision is made. They may include multiplex statistics, such as entrainment $(x_{ij}^1 \times x_{ij}^2)$ or cross network reciprocity $(x_{ij}^1 \times x_{ji}^2)$. We demonstrate multiple imputation of multiplex networks on the case of friendship and helping coevolution in Section 6.6.

6.4.3 Bayesian Estimation

Bayesian theory assumes that there is no one single value estimate for the parameter θ , but that the parameter θ is a random variable following some distribution. Its distribution represents our knowledge and our uncertainty about this parameter. This distribution is estimated by updating a prior distribution of θ , reflecting what is already known about θ , with observed data, resulting in the posterior distribution of θ . For a detailed introduction into Bayesian analysis see, for instance, Gelman et al. (2013). There are different options for choosing prior distributions. Generally, a prior should represent the prior (subjective) belief and/or prior (objective) available information about the parameter distribution. Because of the subjectivity, often so-called uninformative priors are chosen which have only small effect on the result posterior distribution. It is, however, helpful, especially in complicated models, to use weakly informative priors (Gelman et al., 2008). Weakly informative priors do not strongly influence the posterior estimation, but they limit the parameter range. For example, given the scale at which effects are usually defined, a parameter of the size of $\theta_k = 10$ or larger is usually not meaningful in the objective function of a SAOM, and is usually an indicator for a perfect predictor of a tie. The estimation of a model with such a parameter is unlikely to converge because the parameter could equally be $\theta_k = 10$, $\theta_k = 50$, or $\theta_k = 100$. A weakly informative prior is able to bind the estimation with a meaningful parameter range. Moreover, prior knowledge about a parameter can actively contribute to stabilize complex (network) models.

Imputations obtained from parameters drawn randomly from their estimated posterior distribution are generally more reliable than multiple imputations obtained from parameters with fixed estimated values (Schafer and Graham, 2002; van Buuren, 2012). Drawing from the posterior increases the variance between the imputation parameters, and thus between the imputed values. Without this between-imputation variance, the uncertainty about parameter values is underestimated. In non-network data this variance can also be obtained via bootstrapping, however, bootstrapping with a network is not meaningful, be-

cause it distorts the dependency between the nodes and their ties, the modeling of which is the purpose of the analysis.

The recently introduced multiple imputation algorithm for SAOMs (Krause et al., 2018a, 2019a) did not draw imputation parameters from a posterior distribution. Although this non-Bayesian multiple imputation algorithm did not seem likely to seriously underestimate standard errors, the imputations were not proper in the sense of Rubin (1976). The algorithm proposed in this study will draw imputation parameters from their estimated posterior distribution.

Bayesian SAOMs

Bayesian estimation of SAOMs was introduced in Koskinen and Snijders (2007) and implemented in *RSienaTest* (Ripley et al., 2019). Bayesian SAOMs have the additional benefit that they allow for multilevel modeling. In these multilevel models, fixed parameters (η) are assumed to have one posterior distribution, while randomly varying parameters have both a group-specific posterior distribution, as well as a hyper posterior distribution (μ), from which the group specific distributions are drawn. In general, Bayesian estimation is always performed within a likelihood framework. For SAOMs this means that parameters for the network evolution period from wave $m - 1$ to m are estimated on trajectories that connect the observed wave $m - 1$ with m .

In the situation with only missing data in wave m and missing data at wave $m - 1$, Bayesian estimation of SAOMS under missing data is handled statistically optimally (Bright et al., 2019). This is the case because missing data are imputed internally in a model-based way, with prior distributions given by draws from the full conditional posterior given the observed data at waves m and $m - 1$, and given the current draw of the parameters, which leads to proper imputations (Rubin, 1976). To allow the simulations to start, missing data in wave $m - 1$ are imputed internally with independent draws from random binary variables with the density of the observed data $x(m - 1)$.

6.5 Extended Multiple Imputation

The proposed extended algorithm follows a procedure similar to the one proposed by Krause et al. (2018a, 2019a), that is, the first wave is imputed with a stationary SAOM, then later waves are imputed wave by wave with longitudinal SAOMs.

6.5.1 First Wave

Similar to the previously proposed algorithm (Krause et al., 2018a, 2019a), missing data in the first wave are imputed with a stationary multigroup Bayesian SAOM. A single multilevel Bayesian SAOM is estimated jointly for all groups. The imputation, including fixed parameters and random effects structure, should reflect the longitudinal analysis model as closely as possible. In the case of multiplex data, the model is also multivariate, estimating one joint model for the multiplex structure. Let ${}_g x(m)$ denote the network of group g at wave m . Further, we use the convention that u represents the observed part of the data. Let ux be the observed network data, with ${}_g ux(m)$ denoting the observed part of the network of group g at wave m . The stationary SAOM is estimated for the first wave data $x(1)$, using the second wave data $x(2)$ as dyadic covariate. First, the posterior parameter distribution $p(\theta \mid ux(1), ux(2))$ is estimated. Second, for each group a group specific parameter vector ${}_g^d \theta$ is drawn from $p(\theta \mid ux(1), ux(2))$, d indexing the imputations and g the groups. The parameter is group specific, because it reflects the random effects structure. Third, the drawn parameter is used in ML simulation to obtain an imputed network ${}_g^d x(1)$ from $p({}_g x(1) \mid {}_g^d \theta, {}_g ux(1), {}_g ux(2))$. The last two steps, drawing parameter and imputing, are repeated until D imputations are obtained. Algorithm 7 below illustrates this procedure.

Algorithm 7 Algorithm for Bayesian imputation of missing multiplex network data with multiple groups in wave $m = 1$

Estimate the posterior distribution $p(\theta \mid ux(1), ux(2))$ with a stationary Bayesian SAOM for $x(1)$ using the observed network at the second wave $ux(2)$ as dyadic covariate.

for $g = 1, \dots, G$ **do**

for $d = 1, \dots, D$ **do**

 (I) Draw a group specific parameter vector ${}_g^d \theta$ from $p(\theta \mid ux(1), ux(2))$

 (II) Draw a joined imputation for the missing tie variables ${}_g^d x(1)$ with

 ML simulation from $p({}_g x(1) \mid {}_g^d \theta, {}_g ux(1), {}_g ux(2))$

end for

end for

Alternatively, the first wave could also be imputed using Bayesian exponential multiplex graph models as introduced by Krause and Caimo (2019). It is, however, recommended to stay within the SAOM model family to stay as closely as possible to the longitudinal model.

6.5.2 Later Waves

Later waves are imputed wave by wave, each imputation using a separate, already imputed, previous wave ${}^d x(m-1)$ as starting point. For each of the D imputations and for each of the waves m a separate posterior parameter distribution $p({}^d \theta(m) \mid {}^d x(m-1), ux(m))$ is estimated. After estimating $p({}^d \theta(m) \mid {}^d x(m-1), ux(m))$, one parameter vector ${}^d_g \theta(m)$ is randomly drawn for each group and used in ML simulation to obtain the imputed network ${}^d_g x(m)$. This process is repeated for each consecutive wave, leading to D imputed data sets, as detailed in Algorithm 8.

Algorithm 8 Algorithm for Bayesian imputation of missing multiplex network data with multiple groups in waves $m \geq 2$

```

for  $d = 1, \dots, D$  do
  for  $m = 2, \dots, M$  do
    Estimate the posterior distribution  $p(\theta(m) \mid {}^d x(m-1), ux(m))$  with a
    longitudinal Bayesian SAOM
    for  $g = 1, \dots, G$  do
      (I) Draw a group specific parameter vector  ${}^d_g \theta(m)$  from the estimated
      posterior  $p(\theta(m) \mid {}^d x(m-1), ux(m))$ 
      (II) Draw a joined imputation for the missing tie variables  ${}^d_g x(m)$  with
      ML simulation from  $p({}_g x(m) \mid {}^d_g \theta(m), {}^d x(m-1), ux(m))$ 
    end for
  end for
end for

```

6.5.3 Obtaining Results

The previously proposed algorithm by Krause et al. (2018a, 2019a) required that a separate analysis model is estimated on each of the D imputed data sets. The resulting D parameter estimates are combined using Rubin's rules (Rubin, 1987) to obtain the final parameter estimates and standard errors. This is, however, not required in the above outlined procedure. Because each of the longitudinal estimated posterior distributions $p({}^d \theta(m) \mid {}^d x(m-1), x(m))$ is a proper draw from the posterior, these estimates can be directly combined, without having to estimate SAOMs on the completed data sets. Further, the combination of the results is simpler than Rubin's rules, because all that is required is to simply join the different posterior distributions together, leading to one large posterior. In the case of $M > 2$ it is advised to first inspect if the posteriors for the different waves are homogenous, meaning that the network dynamics remain stable over time. If that is the case, the posteriors over all

waves can be combined. If not, the posteriors should be only combined within each wave and results should be presented separately for each wave.

6.5.4 Network and Behavior Co-evolution

In the case of network and behavior co-evolution, the procedure proposed by Krause et al. (2019a) needs to be slightly adjusted. Let z be a vector of size n with a value for each of the n participants representing the dependent behavior, ${}_gz(m)$ being the observation of group g in wave m . First, as in Krause et al. (2019a), the imputation begins with multiple imputation by chained equations (MICE van Buuren, 2012) to impute missing data in the behavior variable and obtain D imputed behavior sets $\tilde{z}(1)$, with ${}^d_g\tilde{z}(1)$ being the d th MICE imputed behavior vector (see Krause et al., 2019a, for more details on MICE imputation of behavior variables). Second, follow the algorithm for Bayesian imputation of missing multiplex network data with multiple groups in wave $m = 1$ with a minor change. After estimating the posterior, use a different draw from ${}^d_g\tilde{z}(1)$ as starting points for each of the imputations. Let uz be the observed part of the behavior variable. The posterior distribution for the stationary SAOM is not estimated using the imputed data from MICE, $\tilde{z}(1)$, but using the observed data $p(\theta \mid ux(1), ux(2), uz(1), uz(2))$. Estimating the posterior for $\tilde{z}(1)$ will be very tie consuming with multiple groups. This adapted algorithm is presented in Algorithm 9.

Algorithm 9 Algorithm for Bayesian imputation of missing network and behavior data with multiple groups in wave $m = 1$

```

Impute missing data in  $z(1)$  with MICE obtaining  $\tilde{z}(1)$ 
Estimate the posterior distribution  $p(\theta \mid ux(1), ux(2), uz(1), uz(2))$  with a
stationary Bayesian SAOM for  $(x(1), z(1))$  using the second wave observations
 $(ux(2), uz(2))$  as covariates.
for  $g = 1, \dots, G$  do
  for  $d = 1, \dots, D$  do
    (I) Draw a group specific parameter vector  ${}^d_g\theta$ 
    from  $p(\theta \mid ux(1), ux(2), uz(1), uz(2))$ 
    (II) Draw a joined imputation  $({}^d_gx(1), {}^d_gz(1))$  with ML simulation
    from  $p({}_gx(1), {}_gz(1) \mid {}^d_g\theta, {}_gux(1), {}_gux(2), {}^d_gz(1), {}_guz(2))$ 
  end for
end for

```

All later waves are imputed wave by wave following the algorithm for imputation of waves $m \geq 2$ outlined above.

6.6 Illustrative Example – Friendship and Helping

We demonstrate the outlined algorithm using data from the Dutch SNARE study (Dijkstra et al., 2015) and following the analysis of van Rijsewijk et al. (2019). They investigated the coevolution of friendship and helping relations (networks) in adolescents throughout the course of a year. For a detailed introduction in the data and theory see van Rijsewijk et al. (2019). The data consist of 41 classrooms with 953 students (mean classroom size of 23.2 students, mean age of 12.7, 49.5% girls, and 84.5% identified as Dutch). Data were collected three times (October, December, and April). Missing data were distributed as follows: 11 students were missing at all measurement points, 34 students were missing at wave one (5%), 60 at wave two (7.5%), and 56 at wave three (7%). The missing data was spread evenly across all groups, with some groups not having any missing data in any wave. All missing data are complete actor non-response, meaning no outgoing network information was observed. Additionally, some students nominated (nearly) everyone in their class as either helper or friend at one assessment, but nominated hardly anyone in other waves. These students most likely interpreted the question differently from their peers. All their outgoing nominations ($x_{ij}^l = 1$) were treated as missing data, however, their outgoing no-ties ($x_{ij}^l = 0$) were retained as observed no-ties. In total, this occurred for 1, 13 and 8 students at the respective waves.

The analysis model of van Rijsewijk et al. (2019) contained standard structural effects for both networks: outdegree (modeling the general tendency to create ties), reciprocity (modeling the tendency to reciprocate incoming ties), transitivity (modeling friends/helpers of friends/helpers becoming friends/helpers), outdegree activity (the tendency of actors with high outdegree to send more ties), indegree popularity (the tendency of actors who receive a lot of incoming nominations, to receive even more), and, lastly, a same gender effect (the tendency to send ties to those with the same gender).

Further, the model contained several cross-network effects. The direct effects of friendship on helping and helping on friendship were included ($x_{ij}^{friend} = 1$ predicting $x_{ij}^{help} = 1$ and vice versa), as well the effects of reciprocated ties in one layer on the tie formation in the other layer ($x_{ij}^{friend} = 1$ and $x_{ji}^{friend} = 1$ predicting $x_{ij}^{help} = 1$ and vice versa). van Rijsewijk et al. (2019) further split several parameters in their effects on creation and maintenance. Usually, parameters in the objective function are so-called evaluation parameters, which contribute both to the creation and the maintenance of ties. For example, a positive evaluation parameter for reciprocity in the case of $x_{ij} = 0$ and $x_{ji} = 1$ increases the probability that i will create a new tie to j , in the case of $x_{ij} = 1$ and $x_{ji} = 1$,

the positive parameter makes it more likely that i will maintain the existing tie to j . It is, however, possible that certain parameters contribute differently to the creation and maintenance of ties (e.g., gender homophily might be helpful in initiating a friendship, but less important in maintaining ties). The following parameters were included as both creation and maintenance: reciprocity in friendship, the direct effect of helping on friendship, and the effect of mutual helping ties on friendship. Lastly, all but the cross-network effects were allowed to randomly vary between the networks. For in depth theoretical reasoning see van Rijsewijk et al. (2019).

The only notable difference to the model run by van Rijsewijk et al. (2019) was the choice of prior between group variances. While van Rijsewijk et al. (2019) chose a rather flat prior variance of $\sigma = 1$ for the between group variance, a more narrow prior variance of $\sigma = .01$ was chosen for this demonstration, because recent experiences with Bayesian SAOMs suggest that narrower priors lead to more stable estimation¹. To compare the results of the proposed algorithm with the results of van Rijsewijk et al. (2019), their model was re-estimated with the narrower priors. Further, van Rijsewijk et al. (2019) analyzed all three observation waves jointly, assuming homogeneity for each parameter (except rate parameters) across the observation periods. As detailed above, the proposed missing data imputation algorithm yields estimated posteriors for each period separately. If these posteriors are homogeneous over the waves, they can be combined. We will come back to this in the results section.

6.6.1 Stationary SAOM imputation

We initially set out to follow the described algorithm, that is, estimate a stationary form of the above described model² on the first wave networks, using the second wave observations as covariate. We were, however, unable to obtain a converged estimate for this complex stationary model. Even very simplistic multiplex models did not converge. Thus we fell back to a less attractive alternative for the first wave imputations, which should still yield more reliable imputations than those used in the implemented internal default treatment³. The simpler imputation procedure splits the multiplex model into two separate models, one for friendship, and one for helping. Both of these models were very simplistic, only containing the following parameters: outdegree, reciprocity, same gender,

¹Kappa was set to $\kappa = .01$.

²Note that creation and maintenance cannot be separated in a stationary model.

³Using the Bayesian exponential multiplex graph models (BERmGMs Krause and Caimo, 2019) was not a reasonable alternative, because BERmGMs, in their current implementation, take far too long to estimate and do not facilitate multigroup or even multilevel estimation.

and the direct effect of the other layer. The model for helping further included the second wave observation of helping.

For friendship, a model with both the second wave observation of friendship $x^{friend}(2)$ and the direct effect of helping on friendship $x^{help}(1)$ did not converge, indicating strong multicollinearity of these two variables. Although second wave observations have proven to be important predictors (Krause et al., 2018a, 2019a), we decided to include the direct of effect from helping on friendship in the stationary model to better maintain the relationship between the layers.

After estimating these two stationary Bayesian SAOMs it is possible to impute the missing data in both layers in an iterative alternating procedure. The alternating procedure described below leads to strong dependence between consecutive imputations. To obtain imputations that are less dependent on each other, at least $T = D \times 2$ imputations are drawn with the following procedure. Start with $t = 1$ and, (I) draw a parameter vector ${}^t\theta(1)^{friend}$ from the estimated posterior distribution of friendship and (II) use ML simulation to obtain an imputed first wave friendship network ${}^t x^{friend}(1)$ from $p(\cdot \mid {}^t\theta^{friend}(1), ux^{friend}(1), ux^{help}(1))$. (III) draw a parameter vector ${}^t\theta^{help}(1)$ from the posterior distribution of helping, and (IV) use ML simulation to obtain an imputed first wave helping network ${}^t x^{help}(1)$ from $p(\cdot \mid {}^t\theta^{help}(1), ux^{help}(1), {}^t x^{friend}(1), ux^{help}(2))$, using the previously imputed friendship network ${}^t x^{friend}(1)$ as a predictor. Repeat this procedure for the next value $t + 1$, starting with imputing the friendship network using the previously (step t) imputed helping network as covariate, followed by imputing helping with the newly imputed friendship network as covariate, and continue until there are at least $T = D \times 2$ imputed networks. In this procedure, the friendship network of one imputation step (${}^t x^{friend}$) is directly linked to the helping network of the previous iteration (${}^{t-1} x^{help}$). Thus only every second imputed network ${}^t x(1)$ is retained as an imputed network for later analysis, ${}^d x(1)$. Note that the posterior parameter distributions $p(\theta^{friend}(1) \mid ux^{friend}(1), ux^{help}(1))$ and $p(\theta^{help}(1) \mid ux^{help}(1), ux^{friend}(1), ux^{help}(2))$ in steps (I) and (III) are estimated on the observed data ($ux^{friend}(1), ux^{help}(1)$) using the default missing data treatment. But imputations are drawn conditional on the observed network that is to be imputed and on the previously imputed covariate network layer, that is, for friendship imputations ($ux^{friend}(1)$ and ${}^{t-1} x^{help}(1)$) and for helping imputations ($ux^{friend}(1)$ and ${}^{t-1} x^{help}(1)$), as well as the observed second wave helping network $ux^{help}(2)$. We present the algorithm for this adapted procedure in Algorithm 10.

Algorithm 10 Algorithm for adapted Bayesian imputation of missing multiplex network data with multiple groups in wave $m = 1$

Estimate the posterior distributions $p(\theta^{friend}(1) \mid ux^{friend}(1), ux^{help}(1))$ and $p(\theta^{help}(1) \mid ux^{help}(1), ux^{friend}(1), ux^{help}(2))$

for $t = 1, \dots, T$ **do**

(I) Draw a parameter vector ${}^t\theta(1)^{friend}$ from its estimated posterior distribution $p(\theta^{friend}(1) \mid ux^{friend}(1), ux^{help}(1))$

(II) Obtain the imputed network ${}^tx^{friend}(1)$ with ML simulation drawing from $p(\cdot \mid {}^t\theta^{friend}(1), ux^{friend}(1), {}^{t-1}x^{help}(1))$

(III) Draw a parameter vector ${}^t\theta(1)^{help}$ from its estimated posterior distribution $p(\theta^{help}(1) \mid ux^{help}(1), ux^{friend}(1), ux^{help}(2))$

(IV) Obtain the imputed network ${}^tx^{help}(1)$ with ML simulation drawing from $p(\cdot \mid {}^t\theta^{help}(1), ux^{friend}(1), {}^{t-1}x^{help}(1), ux^{help}(2))$

if t is an even number **then**

 Retain ${}^tx(1)$ as imputation ${}^dx(1)$ with $d = t/2$

end if

end for

6.6.2 Longitudinal SAOM

After D imputations of the first wave were obtained, we proceeded with the imputation algorithm for later waves ($m \geq 2$) described in Section 5. The imputation procedure revealed heterogeneity in the parameters between the two estimation periods. The heterogeneity implies that we cannot simply join the posteriors estimated during the imputation procedure into a single estimate. To be able to compare the results under imputations with the default procedure, a model with default missing data treatment was estimated separately for each wave. It is important to note that the heterogeneity does not have a major impact on the conclusions drawn in van Rijsewijk et al. (2019).

6.6.3 Results

The estimated posterior density distributions for each parameter are presented in Figures 6.1 to 6.6. We only show the distributions for the fixed parameters (Figures 6.3 and 6.6) and the estimated hyper distributions for the random parameters (Figures 6.1, 6.2, 6.3, and 6.4), not the group-specific estimates. The plots of the posterior parameter density distributions in each figure are presented in two rows: the upper row gives the plots for period 1, the lower row for period 2. Each plot consists of two densities, the green distribution shows the posterior parameter density of the results using the default missing data treatment, the red for the results under multiple imputation. The results for

the friendship network dynamics are given in Figures 6.1 and 6.2, the cross-network effects from helping on friendship are Figure 6.3. The results for the helping network dynamics are given in Figures 6.4 and 6.5, the cross-network effects from friendship on helping in Figure 6.6. We did not combine the posterior distributions for the first and second period, because they showed strong heterogeneity for some parameters (notably the same gender selection effects on both networks, see Figures 6.2 and 6.5). This heterogeneity does not generally invalidate the results from van Rijsewijk et al. (2019), where parameters were estimated over all three time points assuming homogeneity, as we will discuss later. The posterior mean parameter estimates, their standard deviations, and Bayesian p -values of the friendship dynamics are presented in Table 6.1, the estimates for the helping dynamics in Table 6.2. Bayesian p -values give the mass of the posterior probability distribution that is on one side of 0. A Bayesian p -value of, for instance, .01 can thus be interpreted as: The probability that the parameter is negative, given the priors, the model, and the data, is 1%.

Overall, the imputation method leads to similar results as those obtained by the implemented default, which can be seen from the largely overlapping posterior distributions. The friendship dynamics only showed some deviation in the cross-network effects. In the first period, the effect of existing mutual helping relations on the formation of friendships ($x_{ij}^{help} = 1$ and $x_{ji}^{help} = 1$ help create $x_{ij}^{friend}(m - 1) = 0$ to $x_{ij}^{friend}(m) = 1$) is estimated as $mean = -1.84$ ($sd = 0.68$) by the default procedure and as $mean = -0.86$ ($sd = 0.62$) with multiple imputation. This difference could be the result of the simplistic imputation model for the first wave. The counts for friendship tie creation with mutual helping are quite low in the observed data (20 in total), thus changes in the imputed networks can lead to differences in the estimation of the parameter. It is the only difference with relevance for the hypothesis in van Rijsewijk et al. (2019). The results in van Rijsewijk et al. (2019) show an overall negative effect, net of the direct effect of helping on friendship. Using multiple imputation, this net negative effect on the creation of friendship ties seems to be primarily driven by the second period, while the first the period shows that mutual helping has no strong negative or positive effect on the creation of friendship ties, net of the direct effect of an existing helping tie (for an interpretation of this effect see van Rijsewijk et al., 2019).

In the second period, two other cross network effects of helping on friendship show differences between the posterior distributions estimated by the default method and those obtained via multiple imputation. These are the effect of

Table 6.1: Posterior means, standard deviations, and Bayesian p -values for friendship dynamics

	Period 1				Period 2			
	Default Procedure		Multiple Imputation		Default Procedure		Multiple Imputation	
	mean (sd)	p	mean (sd)	p	mean (sd)	p	mean (sd)	p
Friendship rate	8.79 (0.51)	> .99	8.81 (0.51)	> .99	8.94 (0.53)	> .99	9.37 (0.55)	> .99
Outdegree	-2.42 (0.14)	< .01	-2.37 (0.13)	< .01	-1.88 (0.14)	< .01	-1.79 (0.13)	< .01
Reciprocity maintenance	0.86 (0.12)	> .99	0.83 (0.13)	> .99	1.01 (0.09)	> .99	0.99 (0.13)	> .99
Reciprocity initiation	0.12 (0.12)	.84	0.13 (0.13)	.85	0.18 (0.10)	.97	0.17 (0.12)	.92
Transitive triplets	0.26 (0.02)	> .99	0.25 (0.02)	> .99	0.28 (0.02)	> .99	0.26 (0.02)	> .99
Indegree popularity	0.002 (0.02)	.53	0.001 (0.02)	.53	-0.03 (0.02)	.04	-0.03 (0.02)	.03
Outdegree activity	-0.004 (0.01)	.39	-0.01 (0.01)	.32	-0.03 (0.01)	.01	-0.03 (0.01)	.02
Same gender	0.82 (0.08)	> .99	0.80 (0.08)	> .99	0.59 (0.07)	> .99	0.55 (0.07)	> .99
Help on friendship maintenance	1.13 (0.20)	> .99	1.13 (0.18)	> .99	1.00 (0.20)	> .99	0.79 (0.20)	> .99
Help on friendship initiation	1.09 (0.28)	> .99	0.84 (0.28)	> .99	1.50 (0.30)	> .99	1.11 (0.36)	> .99
Mutual help on friendship maintenance	1.29 (0.33)	> .99	1.24 (0.30)	> .99	0.64 (0.29)	.98	0.84 (0.31)	> .99
Mutual help on friendship initiation	-1.84 (0.68)	< .01	-0.86 (0.62)	.06	-2.01 (0.65)	< .01	-2.04 (0.99)	< .01

Table 6.2: Posterior means, standard deviations, and Bayesian p -values for helping dynamics

	Period 1			Period 2		
	Default		Multiple	Default		Multiple
	Procedure	Imputation	Imputation	Procedure	Imputation	Imputation
	mean (sd)	p	mean (sd)	p	mean (sd)	p
Help rate	7.46 (0.48)	> .99	7.49 (0.47)	> .99	8.64 (0.56)	> .99
Outdegree	-3.42 (0.13)	< .01	-3.52 (0.16)	< .01	-3.19 (0.13)	< .01
Reciprocity	0.33 (0.08)	> .99	0.35 (0.08)	> .99	0.26 (0.08)	> .99
Transitive triplets	0.28 (0.03)	> .99	0.25 (0.04)	> .99	0.29 (0.04)	> .99
Indegree popularity	-0.02 (0.02)	.20	-0.01 (0.03)	.31	-0.004 (0.02)	.43
Outdegree activity	0.07 (0.02)	> .99	0.08 (0.02)	> .99	0.06 (0.02)	> .99
Same gender	0.58 (0.09)	> .99	0.59 (0.10)	> .99	0.27 (0.08)	> .99
Friendship on help	1.26 (0.12)	> .99	1.35 (0.13)	> .99	1.17 (0.15)	> .99
Mutual friendship on help	0.84 (0.11)	> .99	0.78 (0.10)	> .99	0.96 (0.10)	> .99

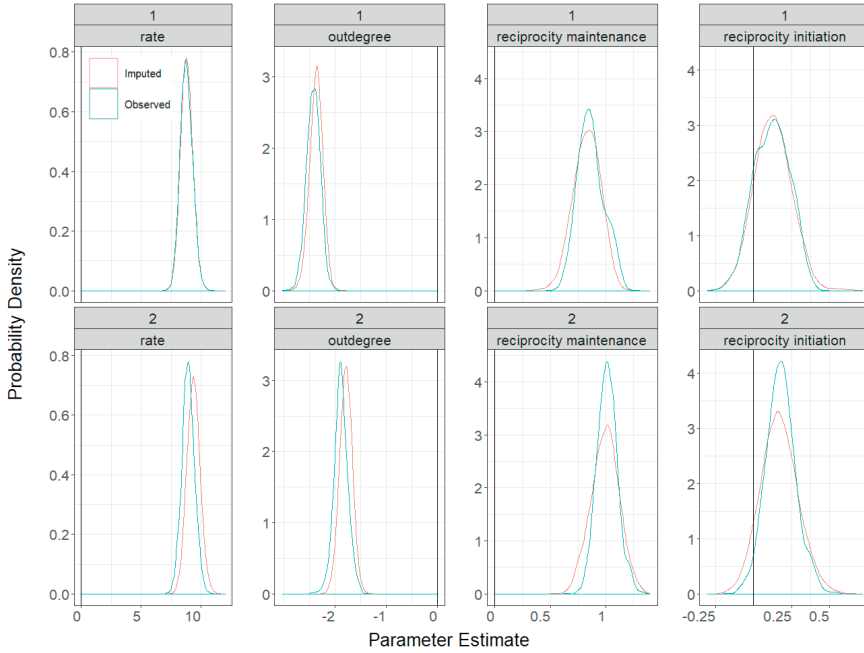


Figure 6.1: **Posterior Parameter Density Distributions for Friendship Dynamics (I).**

The plot shows the results for the friendship network dynamics part (I). The estimated posterior parameter density distributions are presented in two rows: the upper row shows the results estimated for first observation period, the lower row the results estimated for the second observation period. Results obtained on the observed data are shown in green, results obtained with missing data imputation are shown in red. Two parameters for reciprocity were estimated, one modeling the effect of reciprocity on maintaining existing ties (main), the other modeling the effect of reciprocity on creating new ties (create).

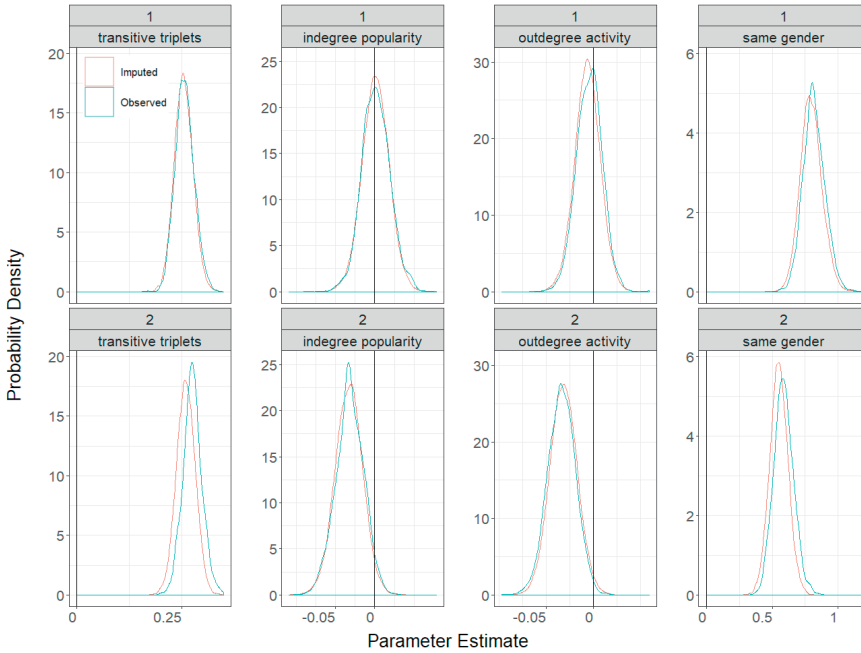


Figure 6.2: **Posterior Parameter Density Distributions for Friendship Dynamics (II).**

The plot shows the results for the friendship network dynamics part (II). The estimated posterior parameter density distributions are presented in two rows: the upper row shows the results estimated for first observation period, the lower row the results estimated for the second observation period. Results obtained on the observed data are shown in green, results obtained with missing data imputation are shown in red.

a helping tie on the maintenance of friendship ($x_{ij}^{help} = 1$ helps maintaining $x_{ij}^{friend}(m-1) = 1$ to $x_{ij}^{friend}(m) = 1$) and the effect of a helping tie on the creation of friendship ties ($x_{ij}^{help} = 1$ helps create $x_{ij}^{friend}(m-1) = 0$ to $x_{ij}^{friend}(m) = 1$). The differences are, however, small and do not effect the inference, leading to similar Bayesian p -values ($p = 1.00$). The creation and maintenance effect of helping on friendship are estimated smaller under the imputed data; maintenance-default: $mean = 1.00$ ($sd = 0.20$) versus maintenance-imputation: $mean = 0.79$ ($sd = 0.20$); and creation-default: $mean = 1.50$ ($sd = 0.30$) versus creation-imputation: $mean = 1.11$ ($sd = 0.36$).

The results for the helping network dynamics estimated with the default procedure are very similar to those estimated under multiple imputation. The only noteworthy difference is in the mean posterior estimates for the same gender effect in the second period. The multiple imputation suggests an on average

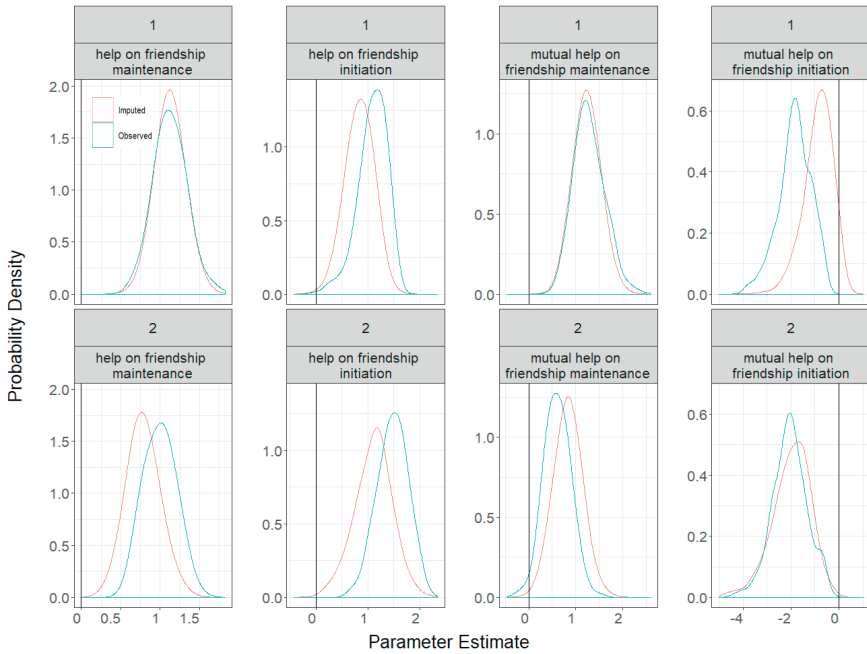


Figure 6.3: **Posterior Parameter Density Distributions for Cross-Network Friendship Dynamics.**

The plot shows the results for the cross-network effects of helping on the friendship network dynamics. The estimated posterior parameter density distributions are presented in two rows: the upper row shows the results estimated for first observation period, the lower row the results estimated for the second observation period. Results obtained on the observed data are shown in green, results obtained with missing data imputation are shown in red. The two parameters for the effect of mutual ties on the helping network were estimated, one modeling the effect of mutual helping ties on the creation of new friendship ties (create), the other modeling the effect of mutual helping ties on the maintenance of existing friendship ties (main).

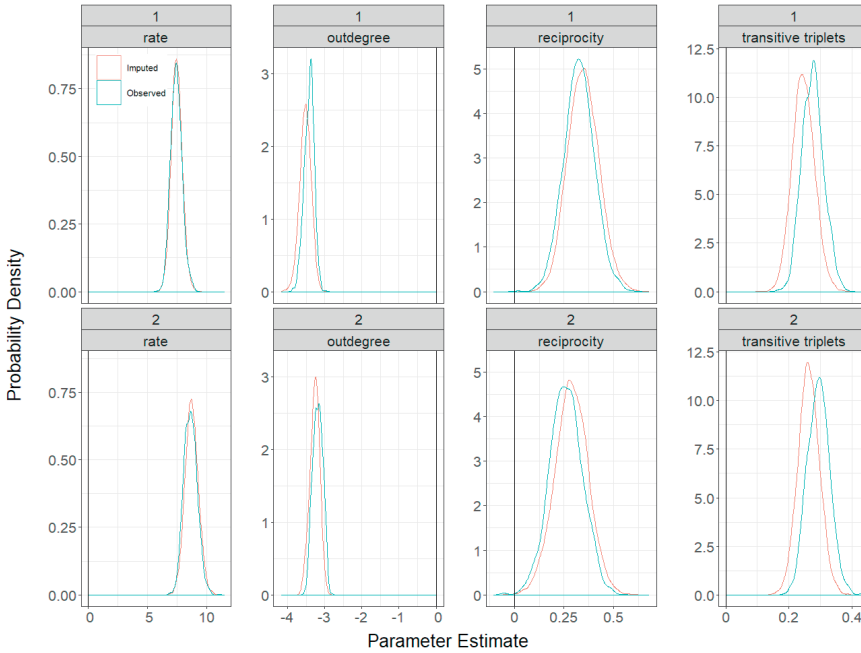


Figure 6.4: **Posterior Parameter Density Distributions for Helping Dynamics (I).**

The plot shows the results for the helping network dynamics part (I). The estimated posterior parameter density distributions are presented in two rows: the upper row shows the results estimated for first observation period, the lower row the results estimated for the second observation period. Results obtained on the observed data are shown in green, results obtained with missing data imputation are shown in red.

smaller effect of same gender, $mean = 0.18$ ($sd = 0.08$), than the default procedure, $mean = 0.27$ ($sd = 0.08$).

6.6.4 Time Heterogeneity

All three time points (both periods) were analyzed simultaneously in van Rijsewijk et al. (2019), assuming homogeneity of the model parameters. Homogeneity assumptions like these are common in the analysis of longitudinal network dynamics (e.g., Krause et al., 2018a), and often justified. The wave-by-wave imputation algorithm, however, revealed that the periods in this study show time heterogeneity. The most notable differences can be seen in the same gender selection effects for both friendship and helping. Gender homophily becomes less important in both network layers (friendship: Period 1 $mean = 0.80$,

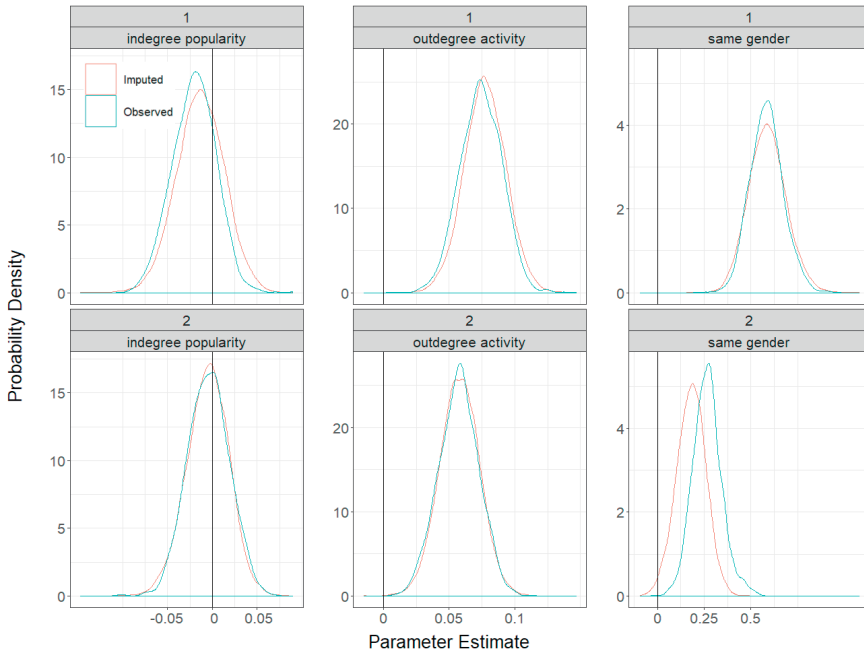


Figure 6.5: **Posterior Parameter Density Distributions for Helping Dynamics (II).**

The plot shows the results for the helping network dynamics part (II). The estimated posterior parameter density distributions are presented in two rows: the upper row shows the results estimated for first observation period, the lower row the results estimated for the second observation period. Results obtained on the observed data are shown in green, results obtained under missing data are shown in red.

Period 2 $mean = 0.55$; helping: Period 1 $mean = 0.59$, Period 2 $mean = 0.18$). The reduction in the importance of gender homophily cannot be seen in van Rijsewijk et al. (2019), because only one parameter is estimated, $mean = 0.74$ ($sd = 0.14$)⁴.

Time heterogeneity also occurs for the degree related effects on the friendship network, indegree popularity (ties to alters who receive a lot of nominations are more likely to be maintained and created) and outdegree activity (actors who send a lot of ties are more likely to maintain and create ties). Both of these effects have no relevant impact on the network dynamics in the first period with indegree popularity $mean = 0.002$ ($sd = 0.02$, Bayes- $p = .53$) and outdegree activity $mean = -0.004$ ($sd = 0.01$, Bayes- $p = .39$). But both of these have

⁴Note that some of the difference in the estimate stems from differences in the priors in van Rijsewijk et al. (2019)

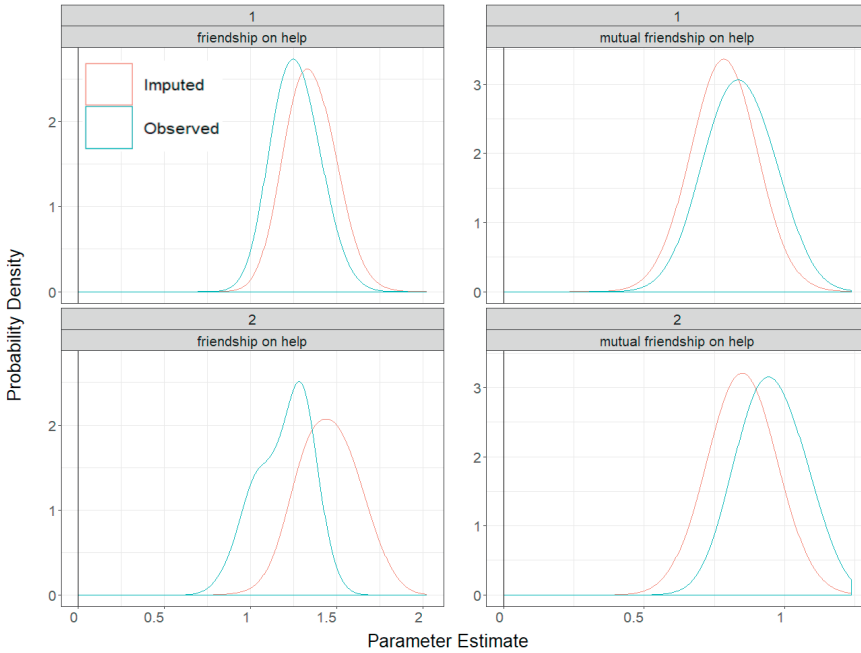


Figure 6.6: **Posterior Parameter Density Distributions for Cross-Network Helping Dynamics.**

The plot shows the results for the cross-network effects of friendship on the helping network dynamics. The estimated posterior parameter density distributions are presented in two rows: the upper row shows the results estimated for first observation period, the lower row the results estimated for the second observation period. Results obtained on the observed data are shown in green, results obtained with missing data imputation are shown in red.

a relevant impact on the friendship dynamics in the second period, indegree popularity $mean = -0.03$ ($sd = 0.02$, Bayes- $p = .04$) and outdegree activity $mean = -0.03$ ($sd = 0.01$, Bayes- $p = .01$). This implies that the number of friendship nominations (both outgoing and incoming) becomes more evenly distributed over time. The parameters van Rijsewijk et al. (2019) both had $mean = -0.01$ ($sd = 0.11$).

As discussed above, the effect of mutual helping relations on the creations on the formation of new friendship ties shows time heterogeneity when data is multiply imputed. Overall, however, none of these difference has a meaningful impact on the conclusions drawn in van Rijsewijk et al. (2019).

6.7 Discussion

This study extends a recently introduced multiple imputation algorithm for missing network data (Krause et al., 2018a, 2019a) in three ways. The algorithm is extended to handle multiple groups of networks (same relation) within a multilevel model with fixed and random effects, multiplex networks (different relations), and introduces a Bayesian procedure, allowing for proper multiple imputation in the sense of Rubin (1987). The proposed extension is demonstrated on an empirical case. This study further illustrates how to handle convergence issues of the stationary SAOM imputation model.

The benefit of the algorithm over Bayesian SAOM estimation with default missing data treatment lies in the imputation of the starting networks. While Bayesian estimation of SAOMs relies by default on a simplistic imputation model of missing data at the beginning of a period, the algorithm proposed in this study utilizes far more information for the imputation of the starting wave of each period, leading to more reliable imputations.

Unlike the non-Bayesian algorithm introduced by Krause et al. (2018a), the extension to Bayesian imputation allows to obtain proper imputations of the missing data. Each imputation is made with a random draw from the estimated posterior distributions of the parameters, allowing for more variance between the imputations, than if the imputations had been drawn using the fixed parameter estimate. The algorithm proposed by Krause et al. (2018a), relying on MoM (or ML) estimation, only uses the same parameter for first wave imputations with stationary SAOMs, and estimates a separate imputation model for each imputation at each of the later waves. However, these estimated imputation models are very close to each other, because they are obtained either via method of moments or maximum likelihood estimation, that is, they are likely to converge to very similar estimates. In contrast, the new proposed method samples the imputation parameter from the estimated posterior distribution, leading to far more variance of the imputation parameters, and thus allowing proper estimation of the between imputation variance. 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). The importance of sampling imputation parameters from their posterior, compared to using only one, or very similar estimates, is as of yet unclear for multiple imputation with SAOMs. Future research is required to investigate the benefit of Bayesian imputation with regard to standard error estimation.

A clear benefit of Bayesian estimation with SAOMs is that, at least for now, multilevel estimation of SAOMs is only implemented with Bayesian estimation. Here, group-specific variation in the parameters can be maintained in the estimation via group-level random effects, while still utilizing the data from all groups for the estimation of the parameters. For imputation with multiple groups, Bayesian estimation should thus be preferred, unless strong homogeneity across the groups can be assumed.

The example indicates that under low amounts of missing data, imputation leads to results similar to those obtained using the default missing data treatment under Bayesian estimation. It has been shown (Koskinen and Snijders, 2007; Bright et al., 2019) that Bayesian SAOM estimation under missing data is reliable given a realistic model and data missing (completely) at random. Obtaining similar estimates can thus be interpreted as good performance of the multiple imputation algorithm. However, this example is not an extensive investigation of the proposed imputation algorithm (or default procedure), but only a single case. Future research is required to better evaluate the performance of the imputation procedure. Such an extensive investigation can lead to helpful guidelines in which cases imputation will lead to improved estimation, and in which cases the extra work is not required. Even in cases where imputation does not directly yield better parameter estimation, it still can aid in the estimation of other models (e.g., blockmodels), or can support the estimation of descriptive statistics.

The similarities in the estimate, however, do not make a multiple imputation algorithm superfluous. The imputed data can be further passed on to other analysis (e.g., blockmodels), used to estimate descriptive statistics, and used in other studies, only focusing on, for instance, a subsample of the data. Further, if multiple models are to be compared on the same data (e.g. to test competing theories), then it will be useful to first multiple impute the data, and then compare the performance of the model on the imputed data sets. Otherwise, each model will internally impute the data with model specific parameters, which means that the two models are compared on two partly different data sets. In such a case it is important that the imputation model contains all parameters that are to be compared in the final model comparison, that is, all competing parameters. Otherwise the imputation might distort the results. If is not possible to estimate such model (e.g., because of complete or strong multicollinearity of the set of competing parameters), then multiple sets of imputed data should be obtained, each obtained by one of the competing models, and the models should be compared in their performance over all the imputed data sets, to best control for the effects of the imputation model.

It can be further seen from the example that researchers need to be careful when estimating SAOM models over multiple waves. While assuming homogeneity and estimating a joint model will stabilize the estimation where the assumption is justified, it can lead to distorted estimates where the assumption is violated. However, the example also illustrates that such violations must not always be harmful for the purpose of the research. The conclusions of van Rijsewijk et al. (2019) stay, mainly, unaltered. This, of course, cannot be known upfront and thus the homogeneity assumption should be thoroughly tested.

Conclusion and Discussion

In this final chapter we will summarize the developments regarding the treatment of missing network data made in this thesis. Further, we will give directions for future research.

7.1 Summary of the Research

The analysis of social networks focuses on how social life is structured, which mechanisms drive change in our social connections, and how social life influences our actions and cognition. These networks are usually formed by people (the nodes) and their connections (the ties), albeit many scholars focus on networks between larger social structures, like organizations or countries. Irrespective of the type of node and the type of relation, missing network data, that is, missing information about connections between nodes, can severely bias the observed network structures, making it more difficult to estimate models, and are likely to lead to biased conclusions.

Most software packages for the statistical analysis of network structures have built-in model-based missing data treatments. While some of these methods lead to unbiased estimates given the right model being used and the data being missing at random (Handcock and Gile, 2007; Koskinen et al., 2010; Snijders, 2017a; Bright et al., 2019), others are only able to handle small amounts of missing data reliably (Huisman and Steglich, 2008). However, model-based procedures only help in the estimation of the model, but they do not help with the calculation of descriptive statistics or facilitate the estimation of other models that lack model-based missing data treatment.

In this dissertation, we focused on two families of network models; Exponential Random Graph Models (ERGMs: Frank and Strauss, 1986; Wasserman and

Pattison, 1996; Robins et al., 2007; Lusher et al., 2013) and Stochastic Actor-oriented Models (SAOMs: Snijders, 1996, 2001, 2005, 2017b). ERGMs are probability models for networks where the probabilities depend on the frequency of occurrence of substructures in the network such as subgraph counts, or other statistics. They are primarily used to analyze cross-sectional network structures (for longitudinal versions of ERGMs see Hanneke et al., 2010; Koskinen et al., 2015). SAOMs on the other hand are primarily used to analyze longitudinal networks. They model the change from one network observation to another as a series of (unobserved) mini steps of decisions of the network actors. Cross-sectional SAOMs are, as of yet, rarely used (Snijders and Steglich, 2015). This thesis starts with an in-depth investigation of missing data treatments mainly using (Bayesian) ERGMs, followed by an introduction of a multiple imputation algorithm for SAOMs.

In Chapter 2, we compared several missing data handling methods for cross-sectional network data. The methods were compared based on their ability to recapture descriptive network statistics, reconstruct missing links, and on estimate model parameters. The competing methods were listwise deletion, imputation of no-ties, reconstruction (single imputation by reciprocating all incoming ties; Stork and Richards, 1992), multiple imputation using ERGMs, and multiple imputation using Bayesian ERGMs (BERGMs; Koskinen et al., 2010; Caimo and Friel, 2011). The methods were tested on a set of simulated networks, varying in size, density, clustering, and homophily on a binary covariate. The results confirmed previous findings, that larger, more structured networks are more robust to missing data than smaller, less structured networks; and statistics based on incoming ties are more robust to missing data (e.g., Smith et al., 2017). The results showed that multiple imputation, especially multiple imputation with complex BERGMs performed best in recapturing descriptive network statistics. However, simple ERGMs were more accurate for the imputation of ties in small networks, while complex BERGMs were more accurate in larger networks. Lastly, the results for estimating model parameters clearly indicate that model estimation with missing data imputation outperforms listwise deletion. However, when more lenient criteria are used to draw inferences (i.e., significance levels of .10 instead of .05), listwise deletion and imputation both lead to conclusions similar to those drawn from the complete data.

Chapter 3 extends missing data imputation with BERGMs to the multiplex case, where several different types of relations are observed between the nodes and analyzed simultaneously. In the applied example, marriage relations and business relations between Florentine banking families were analyzed. To that end, Bayesian Exponential multiplex-Graph Models were developed and imple-

mented in R (R Core Team, 2019). The new model is capable of jointly imputing missing data on two network layers, while modeling both the network structures within each layer, as well as the dependencies between the layers. Multiplex imputation is especially useful, if only some, but not all of the network layers are unobserved for some actors. In that case, the information of the observed layers can contribute to obtaining more reliable imputations. If other layers are only used as covariates in an imputation model, the resulting imputations might not properly maintain their dependencies in a multiplex setting. Further, in the case of complete actor non-response, that is, when no (outgoing) information is available on any layer for some actors, using other layers as covariates can only contribute indirectly to the imputation model (e.g., by taking the observed indegree on the other layers as covariate). However, in a multiplex setting all missing layers are imputed jointly, and imputations on one layer can help provide more reliable imputations on another.

While Chapters 2 and 3 focused on cross-sectional networks, Chapter 4 introduces multiple imputation for longitudinal network studies using the SAOM family. The proposed procedure imputes missing data wave by wave: First a SAOM is estimated for the observation period (from one observation wave to the next), and second, missing data at the end of the period are imputed with the estimated SAOM using maximum likelihood simulation, conditional on the observed and imputed data at the previous wave, on the observed data in the current wave, and the estimated parameters. This procedure is repeated for each wave, using the previously imputed data as starting points for the next imputation. Missing data in the first wave have to be imputed with a different procedure, as the first wave is not modeled in a regular longitudinal SAOM. For the first wave, two options are proposed. Either missing data is imputed with BERGMs, as discussed in Chapter 1, or a stationary SAOM is used to first estimate a model for the first wave, and then impute the data with maximum likelihood simulation. The benefits of imputing the first wave with a stationary SAOM are twofold. First, a stationary SAOM does not require to combine two different network model families in the imputation procedure. Second, stationary SAOMs can also be used with multiple dependent variables, be it dependent actor attributes or a multiplex network structure. After all waves are imputed multiple times, the analysis model is estimated on each of the imputed data sets and the results are combined following the procedure outlined by Rubin (1987). The algorithm is demonstrated on an example of an adolescent friendship network with simulated missing data. The results obtained with multiple imputations are very close to those obtained from the complete data and outperformed both the default treatment (Huisman and Steglich, 2008), as well as

recently proposed alternatives (Hipp et al., 2015; de la Haye et al., 2017).

The multiple imputation algorithm for missing data in longitudinal networks introduced in Chapter 4 is extended to the case of network and behavior coevolution in Chapter 5. SAOMs are often used to model the interdependencies between a network and the actors' behavior (e.g., how does alcohol consumption change the friendship network, or how do friends influence each other in their consumption of alcohol?; Burk et al., 2012). The previously introduced algorithm focused only on missing data on the network side. The algorithm is updated to also incorporate joint imputation of the network and behavior. The procedure has one additional prior step, in which missing data of the behavior variable in the first observation is imputed using multiple imputation by chained equations (MICE; van Buuren, 2012). These imputed values are then used as starting points for the stationary SAOM imputation. All SAOM imputation models are coevolution models, that is, both network and behavior are jointly modeled. The algorithm is demonstrated on an example of the coevolution of friendship, smoking, and drinking with simulated missing data. The extended multiple imputation procedure performed well, leading to reliable standard error and parameter estimates.

In the final chapter of this thesis, the multiple imputation algorithm for missing data in longitudinal networks introduced in Chapter 4 is further extended in three ways, (1) to accommodate the analysis of multiple networks (i.e., multiple groups), (2) to accommodate joint imputation for multiplex network structures (i.e., multiple types of network relations between the same set of nodes), and (3) to provide a fully Bayesian imputation algorithm. The analysis of multiple groups allows to use the data of more than one evolving network to estimate the complex network evolution process. This means that data from multiple groups that are assumed to be homogenous (e.g., multiple classrooms) are used to estimate one model for the network evolution, increasing the power of the analysis. Further, the extended algorithm estimates one imputation model using the data of all groups, which provides better estimates of the parameters of the imputation model, especially when some of the groups are fully observed. The second extension of the algorithm incorporates multiplex network structures. Here, both network layers are jointly imputed, estimating one coevolving SAOM for both network layers. The third improvement of the algorithm is its extension to Bayesian estimation. Bayesian estimation of the imputation model helps to obtain more reliable imputation and assessment of uncertainty. Further, Bayesian SAOMs allow for a multilevel analysis in a multigroup setting, estimating both group-specific random effects and effects fixed across all networks. All three extensions are demonstrated on an applied case (van Rijsewijk

et al., 2019), and it is further illustrated how to handle convergence issues with the stationary SAOM imputation model. The results indicate that the extended multigroup, multiplex, and Bayesian procedure performs well.

7.2 Practical Usage of Multiple Imputation

7.2.1 BERGM

Estimating Bayesian ERGMs on networks with missing data requires the use of the missing data augmentation algorithm introduced by Koskinen et al. (2010). Otherwise, the estimation will be biased. This need not be of concern for applied researchers, because we have implemented the algorithm in the `Bergm` package (Caimo and Friel, 2014) in R (R Core Team, 2019) in the `bergmM()` function. Obtaining multiple imputations with BERGMs is thus as simple as obtaining BERGM estimates.

7.2.2 SAOM

Multiple stochastic imputation with SAOMs requires more work and time investment than the default methods implemented in `RSiena`. However, this investment will be worth the effort in cases with moderate missing data (20% or more), because it will lead to more reliable estimates than the default method (using MoM or ML estimation). Further, network data collection is very resource intensive and usually multiple studies are run with the same data. Thus investing time once in obtaining multiple imputed data sets is worthwhile. The multiple imputation algorithm is further not more difficult than running SAOMs.

7.3 Future Research

7.3.1 BERmGMs Implementation

The BERmGM algorithm is currently completely implemented in R leading to very long computation times. Estimating trivial models on very small graphs (say 16 nodes) can take more than a day. Thus, BERmGMs are as of yet not practically applicable. An implementation in C++ or similarly efficient programming languages would highly speed up the process. Further, only very few cross-network effects are actually implemented in the model. More work is required to test more complex cross-network effects, like, for instance, cross

network reputation. Additionally, the implementation is currently restricted to two layers. The algorithm itself, however, is easily adaptable to more network layers. For now, empirical researchers might be better suited using the **XP-net** (Wang et al., 2009) software to analyze cross-sectional multiplex structures with missing data, which, however does not (easily) allow to obtain multiple imputations of the data. Alternatively, it is possible to use stationary multiplex SAOMs, however, this would imply leaving the ERGM family, which might not be desirable (see Block et al., 2019, for a comparison of ERGMs and SAOMs).

7.3.2 Exponential Random Network Models

Fellows and Handcock (2012b) introduced the family of Exponential Random Network Models (ERNMs). ERNMs are quite similar to ERGMs, but they extend the ERGM framework to incorporate actor attributes as random variables. They thus allow to model both a network and an actor attribute as dependent variables. The algorithm introduced for BERmGMs could be adapted to model Bayesian Exponential Random Network Models (BERNMs). BERNMs would allow joint missing data imputation, similar to that of BERmGMs. The main hurdle for future researchers will be the (time) efficient implementation of the algorithm. Similarly to cross-sectional multiplex models, stationary SAOMs are an alternative to ERNMs, with, however, different assumptions.

7.3.3 Evaluation of SAOM imputation

Unlike BERGM imputation, SAOM imputation has not been thoroughly tested in this dissertation. An extensive simulation study is required to better understand the performance of the developed multiple imputation algorithm. The number of imputations required to obtain reliable results should be systematically investigated. The algorithm can become very time intensive, thus reducing the number of imputations required for a reliable estimate is important to better facilitate the use of the algorithm.

Further, it is yet unclear how the algorithm performs under missingness at random, or how large biases under non-random missingness will be. Additionally, it is important to identify at which missing data percentage the algorithm will no longer yield reliable results. A study similar to Chapter 2 for SAOMs is needed. Future research should also investigate the effects of using a misspecified imputation model, and how such misspecification can, potentially, be detected.

7.3.4 Sensitivity Analysis

The imputation algorithms tested in this study assume that data is either missing at random or completely at random, that is, they assume that the probability to be missing is independent of the missing values themselves. It is impossible to verify this assumption in empirical data, as the missing value is unknown. However, it is possible to test the sensitivity of the estimated model to specific missing data mechanisms. To do so a slight alteration needs to be made to the imputation algorithm. First, the imputation model is estimated on the observed data. Second, a binary node-level covariate is added to the data, indicating whether an actor has missing data or not. Third, a parameter θ expressing the hypothesized missing data mechanism is added to the estimated model and fixed to a value expressing the direction and strength of the missing data mechanism. Fourth, this updated model is used to impute the missing data. For instance, if it is hypothesized that nodes with a low outdegree (e.g., students who do not see anyone as their friend in the class) are less likely to participate, then one can add an ego effect of the new covariate for outgoing ties, and fix it to a negative value of, say, $\theta = -1$. The value for θ should be varied (e.g., under the assumption that missing actors have lower outdegree $\theta = (-.5, -1, -1.5, -2)$ could each be tried). The estimation model, however, remains unchanged, only the imputation model is adjusted.

For multiple imputation, or model-based treatment, with BERGMs, this procedure needs to be implemented within the algorithm estimating the model (see Algorithm 1 in Chapter 2). Such altered BERGMs will provide both adjusted imputations as well as a posterior distribution of the parameters conditional on the hypothesized missing data mechanism.

For multiple imputation with SAOMs the estimation of the imputation model remains unchanged, and θ is added to the estimated model before the imputation. After D imputations have been obtained, each imputed data set is analyzed separately and the results are combined following Rubin's rules (Rubin, 1987). The process is repeated for each of the different tested values of θ .

In either case, the researchers obtain multiple estimates of the final model. One without sensitivity testing, and several for each hypothesized missing data mechanism. If the inference drawn from each of these results is similar, the researchers can be confident that the results are robust to the specific missing data mechanism. However, if the results differ, the researchers should report all of the estimates, and detail how the results differ depending on the hypothesized missing data mechanism.

Both, an investigation of this proposed procedure, and a tutorial for researchers would be a valuable contribution to social network research.

7.4 Implementations

The BERGM estimation under missing data introduced in Chapter 2 has been implemented in the `bergmM()` function in the `Bergm` package (Caimo and Friel, 2011, 2014) in R. The `bergmM()` function provides estimates for posterior parameter distributions under missing data and can also provide proper multiple imputations. A tutorial for the multiple imputation algorithm for SAOMs is available on the `RSiena` website. Tutorial scripts for SAOM imputation with behavior variables, multiplex data, multiple groups, and Bayesian estimation will be added in the future.

Samenvatting

De analyse van sociale netwerken is gericht op hoe het sociale leven is gestructureerd, welke mechanismen verandering in onze sociale verbindingen sturen, en hoe het sociale leven onze acties en kennis beïnvloedt. Deze netwerken worden meestal gevormd door mensen (de actoren of *nodes*) en hun relaties (de banden of *ties*), hoewel veel wetenschappers zich ook richten op netwerken tussen grotere sociale structuren, zoals organisaties of landen. Ongeacht welk type actor of relatie wordt onderzocht, kunnen ontbrekende gegevens, dat wil zeggen, ontbrekende informatie over relaties tussen actoren, de waargenomen netwerkstructuren ernstig beïnvloeden, waardoor het moeilijker is om modellen te schatten en juiste conclusies te trekken.

De meeste softwarepakketten voor de analyse van netwerkstructuren hebben standaardmethoden om datasets met ontbrekende gegevens te analyseren. Deze standaardmethoden zijn vaak gebaseerd op een modelmatige aanpak (zgn. *model-based* methoden). Hoewel sommige van deze methoden leiden tot zuivere schattingen, gegeven dat het juiste analysemodel wordt gebruikt en de ontbrekende gegevens niet selectief zijn (*Missing at Random*; Handcock and Gile, 2007; Koskinen et al., 2010; Snijders, 2017a; Bright et al., 2019), is een deel van de (meest simpele) methoden alleen in staat kleine hoeveelheden ontbrekende gegevens op een betrouwbare manier te verwerken (Huisman and Steglich, 2008). De *model-based* methoden helpen echter alleen bij het schatten van de parameters van het beoogde (eind)model, maar ze ondersteunen noch het berekenen van beschrijvende statistieken, noch het schatten van andere modellen waarvoor geen modelmatige aanpak van ontbrekende gegevens is ontwikkeld. Iets wat met wel mogelijk is met imputatiemethoden, waarbij ontbrekende gegevens worden geschat en ingevuld op de lege plekken in de dataset.

Dit proefschrift richt zich op twee families van netwerkmodellen: *Exponential Random Graph Models* (ERGMs; Frank and Strauss, 1986; Wasserman and Pattison, 1996; Robins et al., 2007; Lusher et al., 2013) en *Stochastic Actor-oriented*

Models (SAOMs; Snijders, 1996, 2001, 2005, 2017b). ERGM's zijn waarschijnlijkheidsmodellen voor netwerken waarbij de kans op wel of geen relatie tussen twee actoren afhankelijk is van de frequentie van substructuren in het netwerk, zoals tellingen van subgrafen of andere statistieken. Ze worden voornamelijk gebruikt om cross-sectionele netwerkstructuren te analyseren (voor longitudinale versies van ERGM's zie Hanneke et al., 2010; Koskinen et al., 2015). SAOM's daarentegen worden voornamelijk gebruikt om longitudinale netwerken te analyseren. Ze modelleren de verandering van de ene netwerkobservatie naar de andere als een reeks (niet-waargenomen) ministappen van beslissingen van de netwerkactoren. Er zijn ook cross-sectionele (stationaire) SAOM's ontwikkeld, maar deze worden tot nu toe zelden gebruikt (Snijders and Steglich, 2015). Dit proefschrift begint met een onderzoek naar de behandeling van ontbrekende gegevens in cross-sectionele netwerken met behulp van (Bayesiaanse) ERGM's, gevolgd door de introductie van een algoritme voor multi-pele imputatie voor longitudinale netwerkdata met behulp van SAOM's.

In hoofdstuk 2 worden verschillende methoden voor het omgaan met ontbrekende gegevens vergeleken voor cross-sectionele netwerken. De methoden worden vergeleken op drie criteria: het (terug)schatten van beschrijvende netwerkstatistieken, het reconstrueren van ontbrekende relaties (*ties*) en het schatten van modelparameters. De onderzochte methoden zijn *listwise deletion*, het behandelen van ontbrekende gegevens als 'geen relatie' (imputeren van de waarde 0), *reconstruction* (reconstructie, dat wil zeggen, imputeren met behulp van binnenkomende relaties; Stork and Richards, 1992), multi-pele imputatie met ERGM's en multi-pele imputatie met Bayesiaanse ERGM's (BERGM's; Koskinen et al., 2010; Caimo and Friel, 2011). De methoden worden onderzocht met behulp van gesimuleerde netwerken, variërend in grootte, dichtheid, clustering en similariteit op een binaire actor-variabele. De resultaten bevestigen eerdere bevindingen: grotere, meer gestructureerde netwerken zijn beter bestand tegen ontbrekende gegevens dan kleinere, minder gestructureerde netwerken en ook zijn statistieken op basis van inkomende relaties beter bestand tegen ontbrekende gegevens (e.g., Smith et al., 2017). De resultaten tonen verder aan dat multi-pele imputatie, in het bijzonder multi-pele imputatie met complexe BERGM's, het best presteert bij het schatten van beschrijvende netwerkstatistieken. Eenvoudige ERGM's zijn daarentegen nauwkeuriger voor het terug-schatten van relaties in kleine netwerken, terwijl complexe BERGM's nauwkeuriger zijn in grotere netwerken. Ten slotte geven de resultaten voor het schatten van modelparameters duidelijk aan dat imputatiemethoden beter presteren dan het verwijderen van de ontbrekende actoren (*listwise deletion*). Wanneer echter minder strenge criteria worden gebruikt om conclusies te trekken over de

modelparameters (d.w.z., significantieniveaus van .10 in plaats van .05), leiden *listwise deletion* en imputeren beide tot conclusies die vergelijkbaar zijn met de conclusies die worden getrokken uit analyses op de volledige gegevens (zonder missing data).

In hoofdstuk 3 wordt de imputatiemethode voor ontbrekende netwerkdata gebaseerd op BERGM's uitgebreid voor zogenaamde multipelexe netwerken. multipelexe netwerken zijn netwerken waarin verschillende typen relaties tussen dezelfde groep actoren worden geobserveerd, en waarin deze verschillende relaties (de netwerklagen) gelijktijdig worden geanalyseerd. In het hoofdstuk wordt een toegepast voorbeeld gepresenteerd waarin huwelijks- en zakelijke relaties tussen Florentijnse bankfamilies worden geanalyseerd. Hiervoor zijn Bayesiaanse *Exponential multiplex-Graph Models* ontwikkeld en geïmplementeerd in de statistische software R (R Core Team, 2019). Het nieuwe model is in staat om ontbrekende gegevens in twee netwerklagen gezamenlijk te imputeren, terwijl zowel de netwerkstructuren binnen elke laag als de afhankelijkheden tussen de lagen worden gemodelleerd. multiplex-imputatie is vooral nuttig als niet elke netwerklaag ontbrekende gegevens bevat. In dat geval kan de informatie van de waargenomen lagen worden gebruikt om de ontbrekende gegevens in incomplete netwerklagen betrouwbaarder te imputeren. Als de netwerklagen niet gezamenlijk worden gemodelleerd, maar bijvoorbeeld de ene laag alleen als covariaat in een imputatiemodel voor een andere laag wordt gebruikt, reflecteren de geïmputeerde waarden de afhankelijkheden in de data mogelijk niet goed en worden netwerkstructuren niet goed geschat. In het geval van volledige actor-non-respons, dat wil zeggen wanneer er voor bepaalde actoren in geen enkele netwerklaag gegevens zijn, levert het gebruik van informatie uit andere lagen als covariaat (bijvoorbeeld de geobserveerde *indegree*) alleen een indirecte bijdrage aan het imputatiemodel. In een multiplex model worden echter de ontbrekende gegevens in alle lagen gezamenlijk geïmputeerd en helpen imputaties in de ene laag om betrouwbaardere imputaties in een andere laag te krijgen.

Waar in hoofdstukken 2 en 3 cross-sectionele netwerken worden besproken, introduceert hoofdstuk 4 een multipele-imputatieprocedure voor longitudinale netwerken, gebaseerd op *Stochastic Actor-oriented Models* (SAOM's). In de voorgestelde procedure worden de ontbrekende gegevens per observatiemoment geïmputeerd. Eerst wordt een SAOM geschat voor de periode tussen de twee observatiemomenten, waarna de ontbrekende gegevens aan het einde van de periode (op het tweede observatiemoment) worden geïmputeerd met de geschatte SAOM. Hierbij wordt gebruik gemaakt van simulaties uit de *maximum likelihood* schattingsprocedure voor SAOM's, gegeven de waargenomen en geïmputeerde waarden van het eerste observatiemoment, de geobserveerde waarden op het

tweede moment en de geschatte modelparameters. Deze procedure wordt voor elke observatie van het netwerk herhaald, waarbij de geïmputeerde gegevens van eerdere meetmomenten worden gebruikt als startpunt voor de volgende imputatie. Omdat het eerste meetmoment niet op deze manier kan worden gemodelleerd (met een longitudinaal model), worden de ontbrekende gegevens voor dit meetmoment geïmputeerd met een andere procedure. Hiervoor worden twee procedures voorgesteld. In de eerste methode worden de ontbrekende gegevens geïmputeerd met behulp van Bayesiaanse ERGM's, zoals is besproken in hoofdstuk 2. De tweede methode gebruikt een stationair SAOM om een model voor het eerste observatiemoment te schatten waarmee vervolgens door middel van simulaties in de *maximum likelihood* schattingsprocedure van het model de ontbrekende gegevens worden opgevuld. De voordelen van het gebruik van een stationair SAOM zijn tweeledig. Ten eerste is het bij het gebruik van een stationair SAOM niet nodig om twee verschillende netwerkmodelfamilies te combineren voor het imputeren van de volledige dataset (alle meetmomenten). Ten tweede kunnen stationaire SAOM's ook worden gebruikt als er meerdere afhankelijke variabelen zijn, of het nu gaat om afhankelijke actorattributen of meerdere netwerkrelaties (multiplexe netwerken). Nadat de ontbrekende gegevens op alle meetmomenten meerdere keren zijn geïmputeerd, wordt het analysemodel geschat op elke geïmputeerde dataset en worden de resultaten gecombineerd volgens de door Rubin (1987) beschreven procedure. De in dit hoofdstuk voorgestelde procedure wordt gedemonstreerd met behulp van een dataset bestaande uit een vriendschapsnetwerk op een school waarin ontbrekende gegevens zijn gesimuleerd. De analyse laat zien dat de imputatieprocedure resultaten oplevert die zeer dicht in de buurt liggen van de resultaten voor de complete data, en veel beter zijn dan resultaten verkregen met zowel de standaard procedure voor ontbrekende gegevens (Huisman and Steglich, 2008) als met recent voorgestelde alternatieven (Hipp et al., 2015; de la Haye et al., 2017).

De multipele-imputatieprocedure voor ontbrekende gegevens in longitudinale netwerken die is geïntroduceerd in hoofdstuk 4, wordt in hoofdstuk 5 uitgebreid voor de situatie van co-evolutie van netwerken en gedrag. SAOM's worden vaak gebruikt om de afhankelijkheden tussen een netwerk en het gedrag van de actoren te modelleren (bijvoorbeeld, hoe verandert alcoholconsumptie het vriendschapsnetwerk, of hoe beïnvloeden vrienden elkaar in hun alcoholconsumptie; Burk et al., 2012). Het eerder geïntroduceerde algoritme was alleen gericht op ontbrekende gegevens in het netwerk en niet in de gedragsvariabele. In dit hoofdstuk wordt een algoritme gepresenteerd voor de gezamenlijke imputatie van het netwerk en de gedragsvariabele. Hierbij is een extra stap noodzakelijk

voor het imputeren van de ontbrekende gegevens in de gedragsvariabele op het eerste meetmoment. In deze stap worden de ontbrekende gegevens geïmputeerd via een zogenaamde *chained equations* methode (multi-pele imputation by chained equations, MICE; van Buuren, 2012). De geïmputeerde waarden van de gedragsvariabele worden vervolgens gebruikt als startwaarden voor imputatie met een stationair SAOM. Alle SAOM-imputatiemodellen (zowel stationair als longitudinaal) zijn in deze procedure co-evolutiemodellen, dat wil zeggen dat zowel netwerk als gedrag gezamenlijk worden gemodelleerd. Het algoritme wordt gedemonstreerd met behulp van dataset met gegevens over de co-evolutie van vriendschap (netwerk), roken en drinken (gedragsvariabelen) waarin ontbrekende gegevens zijn gesimuleerd. De resultaten van de demonstratie laten zien dat de uitbreiding van de multi-pele-imputatieprocedure tot betrouwbare parameterschatting en bijbehorende standaardfouten leidt.

In het laatste hoofdstuk van dit proefschrift wordt de multi-pele-imputatieprocedure voor longitudinale netwerken uit hoofdstuk 4 op drie manieren verder uitgebreid. De eerste uitbreiding zorgt er voor dat meerdere netwerken met hetzelfde type relatie (d.w.z., meerdere groepen actoren) gezamenlijk kunnen worden geïmputeerd. De tweede dat meerdere netwerken met verschillende typen relaties (d.w.z., multi-pelexe netwerken) kunnen worden geïmputeerd. En de derde uitbreiding is een update van het algoritme naar een volledig Bayesiaanse procedure. De analyse van meerdere groepen maakt het mogelijk om de gegevens uit meer dan één netwerk te gebruiken voor het schatten van het (complexe) netwerkevolutieproces. Dit betekent dat gegevens van meerdere groepen, waarvan wordt aangenomen dat ze homogeen zijn (bijvoorbeeld meerdere schoolklassen), worden gebruikt om één model voor netwerkevolutie te schatten. Hierdoor kan een betrouwbaarder model worden geschat en neemt statistische *power* van de analyse toe. Hierbij wordt gebruik gemaakt van één gezamenlijk imputatiemodel voor alle groepen, wat betere schattingen van de parameters van dit imputatiemodel oplevert, vooral wanneer sommige groepen geen ontbrekende gegevens bevatten. De tweede uitbreiding van het algoritme betreft multi-pelexe netwerken, waarbij alle netwerklagen (de verschillende typen relaties) gezamenlijk worden geïmputeerd, waarbij één co-evoluerend SAOM voor alle netwerklagen wordt geschat. De derde verbetering van het algoritme is de uitbreiding naar een volledig Bayesiaanse imputatieprocedure waardoor betrouwbaardere imputaties worden verkregen waarbij op een correcte manier rekening wordt gehouden met de toegenomen onzekerheid door imputatie. Ook is het mogelijk om met de Bayesiaanse procedure een multilevel-netwerkanalyse uit te voeren, waarbij zowel groep-specifieke willekeurige (*random*) effecten als vaste (*fixed*) effecten over alle netwerken worden geschat. Alle drie de uitbreidingen worden

gedemonstreerd met behulp van een dataset uit een bestaand onderzoek over de co-evolutie van vriendschaps- en hulpnetwerken in een multi-groepsetting (van Rijsewijk et al., 2019), waarbij wordt geïllustreerd hoe convergentieproblemen met het stationaire SAOM-imputatiemodel kunnen worden aangepakt. De resultaten laten zien dat de uitgebreide multi-groep, multipelex, Bayesiaanse procedure goed presteert.

References

- Allison, P. (2001). *Missing data*, volume 136. Sage University Papers Series on Quantitative Applications in the Social Sciences.
- Amati, V., Schoenenberger, F., and Snijders, T. (2015). Estimation of stochastic actor-oriented models for the evolution of networks by generalized method of moments. *Journal de la Société Française de Statistique*, 156(3):140–165.
- Block, P., Stadtfeld, C., and Snijders, T. (2019). Forms of dependence: Comparing saoms and ergms from basic principles. *Sociological Methods & Research*, 48(1):202–239.
- Borgatti, S. and Molina, J. (2003). Ethical and strategic issues in organizational social network analysis. *The Journal of Applied Behavioral Science*, 39(3):337–349.
- Bowman, K. and Shenton, L. (1985). Method of moments. In Kotz, S. and Johnson, N., editors, *Encyclopedia of Statistical Sciences*, volume 5, pages 467–473. Wiley, New York.
- Bright, D., Koskinen, J., and Malm, A. (2019). Illicit network dynamics: The formation and evolution of a drug trafficking network. *Journal of Quantitative Criminology*, 35(2):237–258.
- Burk, W., Van Der Vorst, H., Kerr, M., and Stattin, H. (2012). Alcohol use and friendship dynamics: Selection and socialization in early-, middle-, and late-adolescent peer networks. *Journal of studies on alcohol and drugs*, 73(1):89–98.
- Butts, C. (2008). 4. a relational event framework for social action. *Sociological Methodology*, 38(1):155–200.

- van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman & Hall/CRC, Boca Raton.
- van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.
- Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41 – 55.
- Caimo, A. and Friel, N. (2013). Bayesian model selection for exponential random graph models. *Social Networks*, 35(1):11 – 24.
- Caimo, A. and Friel, N. (2014). Bergm: Bayesian exponential random graphs in R. *Journal of Statistical Software*, 61(2):1–25.
- Caimo, A. and Mira, A. (2015). Efficient computational strategies for doubly intractable problems with applications to Bayesian social networks. *Statistics and Computing*, 25:113–125.
- Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25:283–307.
- de la Haye, K., Embreem, J., Punkay, M., Espelage, D. L., Tucker, J. S., and Green, H. D. (2017). Analytic strategies for longitudinal networks with missing data. *Social Networks*, 50:17–25.
- Dijkstra, J., Kretschmer, T., Pattiselanno, K., Franken, A., Harakeh, Z., Vollebergh, W., and Veenstra, R. (2015). Explaining adolescents’ delinquency and substance use: A test of the maturity gap: The snare study. *Journal of Research in Crime and Delinquency*, 52(5):747–767.
- Ellwardt, L., Labianca, G., and Wittek, R. (2012). Who are the objects of positive and negative gossip at work?: A social network perspective on workplace gossip. *Social Networks*, 34(2):193–205.
- Everitt, R. (2012). Bayesian parameter estimation for latent markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960.
- Fellows, I. and Handcock, M. (2012a). Exponential-family random network models. *arXiv preprint arXiv:1208.0121*.
- Fellows, I. and Handcock, M. (2012b). Exponential-family random network models. *arXiv preprint arXiv:1208.0121*.

- Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81:832–842.
- Fujimoto, K., Wang, P., Ross, M., and Williams, M. (2015). Venue-mediated weak ties in multiplex hiv transmission risk networks among drug-using male sex workers and associates. *American journal of public health*, 105(6):1128–1135.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.
- Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2:1360–1383.
- Gile, K. and Handcock, M. (2006). Model-based assessment of the impact of missing data on inference for networks, CSSS working paper no. 66. *Working Paper Series, University of Washington, Seattle*.
- Handcock, M. and Gile, K. (2007). Modeling social networks with sampled or missing data, CSSS working paper no. 75. *Journal of Statistical Software*.
- Handcock, M. and Gile, K. (2010). Modeling social networks from sampled data. *Annals of Applied Statistics*, 4:5–25.
- Handcock, M., Hunter, D., Butts, C., Goodreau, S., and Morris, M. (2007). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11.
- Hanneke, S., Fu, W., and Xing, E. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605.
- Hipp, J., Wang, C., Butts, C., Jose, R., and Lakon, C. (2015). Research note: The consequences of different methods for handling missing network data in stochastic actor based models. *Social networks*, 41:56–71.
- Huang, F., Zhang, M., and Li, Y. (2019). A comparison study of tie non-response treatments in social networks analysis. *Frontiers in Psychology*, 9.
- Huisman, M. (2009). Imputation of missing network data: some simple procedures. *Journal of Social Structure*, 10:1–29.
- Huisman, M. and Krause, R. (2017). Imputation of missing network data. In Alhajj, R. and J., R., editors, *Encyclopedia of Social Network Analysis and Mining*, pages 707–715. Springer, New York.

- Huisman, M. and Steglich, C. (2008). Treatment of non-response in longitudinal network studies. *Social Networks*, 30:297–308.
- Hunter, D. (2007). Curved exponential family models for social networks. *Social Networks*, 29(2):216–230.
- Hunter, D. and Handcock, M. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15:565–583.
- Hunter, D., Handcock, M., Butts, C., Goodreau, S., and Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29.
- Koskinen, J., Caimo, A., and Lomi, A. (2015). Simultaneous modeling of initial conditions and time heterogeneity in dynamic networks: An application to foreign direct investments. *Network Science*, 3(1):58–77.
- Koskinen, J., Robins, G., and Pattison, P. (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, 7(3):366–384.
- Koskinen, J., Robins, G., Wang, P., and Pattison, P. (2013). Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Social Networks*, 35:514–527.
- Koskinen, J. and Snijders, T. (2007). Bayesian inference for dynamic social network data. *Journal of statistical planning and inference*, 137(12):3930–3938.
- Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28:247–268.
- Krause, R. and Caimo, A. (2019). Multiple imputation for Bayesian exponential random multi-graph models. *International Workshop on Complex Networks*, pages 63–72.
- Krause, R., Huisman, M., and Snijders, T. (2018a). Multiple imputation for longitudinal network data. *Italian Journal of Applied Statistics*, 30:33–57.
- Krause, R., Huisman, M., Steglich, C., and Snijders, T. (2018b). Missing network data a comparison of different imputation methods. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

- Krause, R., Iashina, A., Huisman, M., Steglich, C., and Snijders, T. (2019a). Multiple imputation of missing ties and actor attributes. Dissertation Chapter 5.
- Krause, R., van Rijsewijk, L., Huisman, M., Steglich, C., and Snijders, T. (2019b). Multiple imputation of missing ties and actor attributes. Dissertation Chapter 6.
- Lepkowski, J. (1987). The treatment of wave nonresponse in panel surveys. *THE SURVEY OF INCOME AND PROGRAM PARTICIPATION*, page 88.
- Little, R. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.
- Little, R. and Rubin, D. (1987). Statistical analysis with missing data. *Hoboken, NJ: Wiley*.
- Lusher, D., Koskinen, J., and Robins, G. (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.
- Michell, L. and Amos, A. (1997). Girls, pecking order and smoking. *Social science & medicine*, 44(12):1861–1869.
- Niezink, N., Snijders, T., and van Duijn, M. (2019). No longer discrete: Modeling the dynamics of social networks and continuous behavior. *Sociological Methodology*, page 0081175019842263.
- Padgett, J. and Ansell, C. (1993). Robust action and the rise of the medici, 1400-1434. *American Journal of Sociology*, 98(6):1259–1319.
- Pattison, P. and Wasserman, S. (1999). Logit models and logistic regressions for social networks: Ii. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193.
- Pearson, M. and West, P. (2003). Drifting smoke rings. *Connections*, 25(2):59–76.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- van Rijsewijk, L., Snijders, T., Dijkstra, J., Steglich, C., and Veenstra, R. (2019). The interplay between adolescents' friendships and the exchange of help: A longitudinal multiplex social network study. *Journal of Research on Adolescence*.
- Ripley, R., Boitmanis, K., Snijders, T., and Schoenenberger, F. (2017). Manual for Siena version 3. Technical report, Oxford: University of Oxford, Department of Statistics; Nuffield College.
- Ripley, R., Snijders, T., Bóda, Z., Vörös, A., and Preciado, P. (2019). Manual for Siena version 4.0. Technical report, Oxford: University of Oxford, Department of Statistics; Nuffield College.
- Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 22:400–407.
- Robins, G. (2015). *Doing social network research: Network-based research design for social scientists*. Sage.
- Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph models for social networks. *Social Networks*, 29(2):169–348.
- Robins, G., Pattison, P., and Woolcock, J. (2004). Missing data in networks: exponential random graph (p^*) models for networks with non-respondents. *Social Networks*, 26:257–283.
- Rubin, D. (1976). Inference and missing data. *Biometrika*, 63:581–592.
- Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York.
- Schafer, J. and Graham, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7:147–177.
- Smith, J. and Moody, J. (2013). Structural effects of network sampling coverage i: Nodes missing at random. *Social Networks*, 35:652–668.
- Smith, J., Moody, J., and Morgan, J. (2017). Network sampling coverage ii: The effect of non-random missing data on network measurement. *Social Networks*, 48:78–99.
- Snijders, T. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21:149–172.

- Snijders, T. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, 31(1):361–395.
- Snijders, T. (2005). Models for longitudinal network data. *Models and methods in social network analysis*, 1:215–247.
- Snijders, T. (2017a). Siena algorithms. Technical report, Technical report, University of Groningen, University of Oxford. [http://www . . .](http://www...)
- Snijders, T. (2017b). Stochastic actor-oriented models for network dynamics. *Annual Reviews of Statistics and Its Application*, 4:343–362.
- Snijders, T., Koskinen, J., and Schweinberger, M. (2010a). Maximum likelihood estimation for social network dynamics. *The Annals of Applied Statistics*, 4(2):567.
- Snijders, T., Lomi, A., and Torló, V. (2013). A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. *Social networks*, 35(2):265–276.
- Snijders, T., Pattison, P., Robins, G., and Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36:99–153.
- Snijders, T. and Steglich, C. (2015). Representing micro–macro linkages by actor-based dynamic network models. *Sociological Methods & Research*, 44(2):222–271.
- Snijders, T., Van de Bunt, G., and Steglich, C. (2010b). Introduction to stochastic actor-based models for network dynamics. *Social networks*, 32(1):44–60.
- Stadtfeld, C. and Block, P. (2017). Interactions, actors, and time: Dynamic network actor models for relational events. *Sociological Science*, 4:318–352.
- Stadtfeld, C., Hollway, J., and Block, P. (2017). Dynamic network actor models: Investigating coordination ties through time. *Sociological Methodology*, 47:1–40.
- Stadtfeld, C., Takács, K., and Vörös, A. (2018). The emergence and stability of groups in social networks. *Available at SSRN 3232958*.
- Stadtfeld, C., Takács, K., and Vörös, A. (2018). The emergence and stability of groups in social networks. *SSRN*, <https://ssrn.com/abstract=3232958> or <http://dx.doi.org/10.2139/ssrn.3232958>.

- Steglich, C., Snijders, T., and Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, 40(1):329–393.
- Steglich, C., Snijders, T., and West, P. (2006). Applying Siena. *Methodology*, 2(1):48–56.
- Stork, D. and Richards, W. D. (1992). Nonrespondents in communication network studies. *Group and Organisation Management*, 17:193–209.
- Thiemichen, S., Friel, N., Caimo, A., and Kauermann, G. (2016). Bayesian exponential random graph models with nodal random effects. *Social Networks*, 46:11–28.
- Wang, C., Butts, C., Hipp, J., Jose, R., and Lakon, C. (2016). Multiple imputation for missing edge data: a predictive evaluation method with application to Add Health. *Social Networks*, 45:89–98.
- Wang, P. (2012). Ergm extensions: models for multiple networks and bipartite networks. *Exponential Random Graph Models for Social Networks: Theory, Methods, Applications*, pages 115–129.
- Wang, P., Robins, G., and Pattison, P. (2009). Xpnet: Pnet for multivariate networks.
- Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.
- Wasserman, S. and Pattison, P. (1996). Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika*, 61:401–425.
- Zandberg, T. and Huisman, M. (2019). Missing behavior data in longitudinal network studies: the impact of treatment methods on estimated effect parameters in stochastic actor oriented models. *Social Network Analysis and Mining*, 9(1):8.
- Žnidaršič, A., Doreian, P., and Ferligoj, A. (2012). Absent ties in social networks, their treatments, and blockmodeling outcomes. *Metodološki Zvezki*, 9:119–138.

Acknowledgments

There are many people that have contributed, directly and indirectly, to this Ph.D. thesis. First and foremost, there are my supervisors, Mark Huisman, Christian Steglich, and Tom Snijders. I stormed uncountable times into Mark's office to shower him with questions, crunch over results, have him explain statistics to me, and have him read every chapter hundreds of times. I thank Christian for his patience in explaining networks, being an inspiring teacher, and the teaching opportunities he provided. I thank Tom for having tremendous patience with my lack of formal mathematical education and for many inspiring, thoughtful, and beyond all, educational discussions. Further I thank my co-authors, Alberto Caimo and Anna Iashina for their time, dedication, and inspiration. I would also like to thank the members of my reading committee, Casper Albers, Stef van Buuren, and René Veenstra, for taking the time to read and review my dissertation. I thank my colleagues in the Social Networks cluster (Marijtje, Vera, Jing, Laura, Tomáš, and Nynke) for the thoughtful discussions and invaluable feedback throughout the years. Likewise, I thank all my other colleagues in the ICS, especially in Groningen, for their insides, and for all I have had the pleasure to learn from them in the past four years. Also, I am happy and grateful for having become part of an international family of social network researchers, and I thank especially the Duisterbelt crew for their support, inspiration, and feedback. I am grateful to Mike Rinck, Toon Cillessen, and Bill Burk, for helping me find my way to Groningen, social networks, and thus this dissertation. I want to thank my dungeon crawlers, my friends, and besonders Ferdi, Max, und Niklas, for their support, for many laughs, BBQs, and adventures. A very big and heartfelt salute goes to my fellow Legio Sientis and my paranympths Vera and Tomáš for being by my side. Zuletzt bin ich meiner Familie dankbar. Ohne eure Hilfe, eure kritische, forschende, fördernde Erziehung, finanzielle Unterstützung, Diskussionen, Affenlaute, Humor, Wissen, Weisheit, und kulinarischen Künste wäre ich nicht hier.

About the author

Robert Wilhelm Krause was born in Witten, Germany, on the 2nd of December, 1989. After obtaining his Abitur (A-levels) in 2009 from the Adalbert Stifter Gymnasium in Castrop-Rauxel, Germany, he served in compulsory community service (in lieu of military service) at the Schule am Schwalbenweg, a school for special needs children in Herne, Germany. After his service he commenced his studies at the Ruhr University Bochum in 2010. He obtained a Bachelor of Science in Psychology with a focus on cognitive neuroscience (2013), after which he continued to obtain a Research Master degree in Behavioural Science from Radboud University Nijmegen (2015). During his Bachelors Robert did a six-week internship at the Max-Planck-Institute for Human development, Berlin. In 2015, he started his Ph.D., at the Interuniversity Centre for Social Science Theory and Methodology (ICS) at the Department of Sociology of the University of Groningen. During his Ph.D., he visited the Dublin Institute of Technology, Dublin, Ireland. As of September 2019, Robert works as a researcher at the Institute for Analytical Sociology, Linköping University, Sweden.

ICS dissertation series

The ICS series presents dissertations of the Interuniversity Center for Social Science Theory and Methodology. Each of these studies aims at integrating explicit theory formation with state of the art empirical research or at the development of advanced methods for empirical research. The ICS was founded in 1986 as a cooperative effort of the universities of Groningen and Utrecht. Since 1992, the ICS expanded to the University of Nijmegen and since 2017 to the University of Amsterdam (UvA). Most of the projects are financed by the participating universities or by the Netherlands Organization for Scientific Research (NWO). The international composition of the ICS graduate students is mirrored in the increasing international orientation of the projects and thus of the ICS series itself.

1. C. van Liere. (1990). *Lastige leerlingen. Een empirisch onderzoek naar sociale oorzaken van probleemgedrag op basisscholen*. Amsterdam: Thesis Publishers.
2. Marco H.D. van Leeuwen. (1990). *Bijstand in Amsterdam, ca. 1800–1850. Armenzorg als beheersings- en overlevingsstrategie*. ICS-dissertation, Utrecht.
3. I. Maas. (1990). *Deelname aan podiumkunsten via de podia, de media en actieve beoefening. Substitutie of leereffecten?*. Amsterdam: Thesis Publishers.
4. M.I. Broese van Groenou. (1991). *Gescheiden netwerken. De relaties met vrienden en verwanten na echtscheiding*. Amsterdam: Thesis Publishers.
5. Jan M.M. van den Bos. (1991). *Dutch EC policy making. A model-guided approach to coordination and negotiation*. Amsterdam: Thesis Publishers.
6. Karin Sanders. (1991). *Vrouwelijke pioniers. Vrouwen en mannen met een ‘mannelijke’ hogere beroepsopleiding aan het begin van hun loopbaan*. Amsterdam: Thesis Publishers.
7. Sjerp de Vries. (1991). *Egoism, altruism, and social justice. Theory and experiments on cooperation in social dilemmas*. Amsterdam: Thesis Publishers.
8. Ronald S. Batenburg. (1991). *Automatisering in bedrijf*. Amsterdam: Thesis Publishers.

9. Rudi Wielers. (1991). *Selectie en allocatie op de arbeidsmarkt. Een uitwerking voor de informele en geïnstitutionaliseerde kinderopvang*. Amsterdam: Thesis Publishers.
10. Gert P. Westert. (1991). *Verschillen in ziekenhuisgebruik*. ICS-dissertation, Groningen.
11. Hanneke Hermsen. (1992). *Votes and policy preferences. Equilibria in party systems*. Amsterdam: Thesis Publishers.
12. Cora J.M. Maas. (1992). *Probleemleeringen in het basisonderwijs*. Amsterdam: Thesis Publishers.
13. Ed A.W. Boxman. (1992). *Contacten en carrière. Een empirisch-theoretisch onderzoek naar de relatie tussen sociale netwerken en arbeidsmarktposities*. Amsterdam: Thesis Publishers.
14. Conny G.J. Taes. (1992). *Kijken naar banen. Een onderzoek naar de inschatting van arbeidsmarktkansen bij schoolverlaters uit het middelbaar beroepsopleiding*. Amsterdam: Thesis Publishers.
15. Peter van Roozendaal. (1992). *Cabinets in multi-party democracies. The effect of dominant and central parties on cabinet composition and durability*. Amsterdam: Thesis Publishers.
16. Marcel van Dam. (1992). *Regio zonder regie. Verschillen in en effectiviteit van gemeentelijk arbeidsmarktbeleid*. Amsterdam: Thesis Publishers.
17. Tanja van der Lippe. (1993). *Arbeidsverdeling tussen mannen en vrouwen*. Amsterdam: Thesis Publishers.
18. Marc A. Jacobs. (1993). *Software: Kopen of kopiëren? Een sociaal-wetenschappelijk onderzoek onder PC-gebruikers*. Amsterdam: Thesis Publishers.
19. Peter van der Meer. (1993). *Verdringing op de Nederlandse arbeidsmarkt. Sector- en sekseverschillen*. Amsterdam: Thesis Publishers.
20. Gerbert Kraaykamp. (1993). *Over lezen gesproken. Een studie naar sociale differentiatie in leesgedrag*. Amsterdam: Thesis Publishers.
21. Evelien Zeggelink. (1993). *Strangers into friends. The evolution of friendship networks using an individual oriented modeling approach*. Amsterdam: Thesis Publishers.
22. Jaco Berveling. (1994). *Het stempel op de besluitvorming. Macht, invloed en besluitvorming op twee Amsterdamse beleidsterreinen*. Amsterdam: Thesis Publishers.
23. Wim Bernasco. (1994). *Coupled careers. The effects of spouse's resources on success at work*. Amsterdam: Thesis Publishers.
24. Liset van Dijk. (1994). *Choices in child care. The distribution of child care among mothers, fathers and non-parental care providers*. Amsterdam: Thesis Publishers.
25. Jos de Haan. (1994). *Research groups in Dutch sociology*. Amsterdam: Thesis Publishers.
26. K. Boahene. (1995). *Innovation adoption as a socio-economic process. The case of the Ghanaian cocoa industry*. Amsterdam: Thesis Publishers.
27. Paul E.M. Ligthart. (1995). *Solidarity in economic transactions. An experimental study of framing effects in bargaining and contracting*. Amsterdam: Thesis Publishers.
28. Roger Th. A.J. Leenders. (1995). *Structure and influence. Statistical models for the dynamics of actor attributes, network structure, and their interdependence*. Amsterdam: Thesis Publishers.

29. Beate Völker. (1995). *Should auld acquaintance be forgot...? Institutions of communism, the transition to capitalism and personal networks: The case of East Germany*. Amsterdam: Thesis Publishers.
30. A. Cancrinus-Matthijsse. (1995). *Tussen hulpverlening en ondernemerschap. Beroepsuitoefening en taakopvattingen van openbare apothekers in een aantal West-Europese landen*. Amsterdam: Thesis Publishers.
31. Nardi Steverink. (1996). *Zo lang mogelijk zelfstandig. Naar een verklaring van verschillen in oriëntatie ten aanzien van opname in een verzorgingstehuis onder fysiek kwetsbare ouderen*. Amsterdam: Thesis Publishers.
32. Ellen Lindeman. (1996). *Participatie in vrijwilligerswerk*. Amsterdam: Thesis Publishers.
33. Chris Snijders. (1996). *Trust and commitments*. Amsterdam: Thesis Publishers.
34. Koos Postma. (1996). *Changing prejudice in Hungary. A study on the collapse of state socialism and its impact on prejudice against gypsies and Jews*. Amsterdam: Thesis Publishers.
35. Joeske T. van Busschbach. (1996). *Uit het oog, uit het hart? Stabiliteit en verandering in persoonlijke relaties*. Amsterdam: Thesis Publishers.
36. René Torenvlied. (1996). *Besluiten in uitvoering. Theorieën over beleidsuitvoering modelmatig getoetst op sociale vernieuwing in drie gemeenten*. Amsterdam: Thesis Publishers.
37. Andreas Flache. (1996). *The Double edge of networks. An analysis of the effect of informal networks on cooperation in social dilemmas*. Amsterdam: Thesis Publishers.
38. Kees van Veen. (1997). *Inside an internal labor market: Formal rules, flexibility and career lines in a Dutch manufacturing company*. Amsterdam: Thesis Publishers.
39. Lucienne van Eijk. (1997). *Activity and well-being in the elderly*. Amsterdam: Thesis Publishers.
40. Róbert Gál. (1997). *Unreliability. Contract discipline and contract governance under economic transition*. Amsterdam: Thesis Publishers.
41. Anne-Geerte van de Goor. (1997). *Effects of regulation on disability duration*. ICS-dissertation, Utrecht.
42. Boris Blumberg. (1997). *Das Management von Technologiekooperationen. Partnersuche und Verhandlungen mit dem Partner aus Empirisch-Theoretischer Perspektive*. ICS-dissertation, Utrecht.
43. Marijke von Bergh. (1997). *Loopbanen van oudere werknemers*. Amsterdam: Thesis Publishers.
44. Anna Petra Nieboer. (1997). *Life-events and well-being: A prospective study on changes in well-being of elderly people due to a serious illness event or death of the spouse*. Amsterdam: Thesis Publishers.
45. Jacques Niehof. (1997). *Resources and social reproduction: The effects of cultural and material resources on educational and occupational careers in industrial nations at the end of the twentieth century*. ICS-dissertation, Nijmegen.
46. Ariana Need. (1997). *The kindred vote. Individual and family effects of social class and religion on electoral change in the Netherlands, 1956-1994*. ICS-dissertation, Nijmegen.

47. Jim Allen. (1997). *Sector composition and the effect of education on wages: An international comparison*. Amsterdam: Thesis Publishers.
48. Jack B.F. Hutten. (1998). *Workload and provision of care in general practice. An empirical study of the relation between workload of Dutch general practitioners and the content and quality of their care*. ICS-dissertation, Utrecht.
49. Per B. Kropp. (1998). *Berufserfolg im Transformationsprozeß, Eine theoretisch-empirische Studie über die Gewinner und Verlierer der Wende in Ostdeutschland*. ICS-dissertation, Utrecht.
50. Maarten H.J. Wolbers. (1998). *Diploma-inflatie en verdringing op de arbeidsmarkt. Een studie naar ontwikkelingen in de opbrengsten van diploma's in Nederland*. ICS-dissertation, Nijmegen.
51. Wilma Smeenk. (1998). *Opportunity and marriage. The impact of individual resources and marriage market structure on first marriage timing and partner choice in the Netherlands*. ICS-dissertation, Nijmegen.
52. Marinus Spreen. (1999). *Sampling personal network structures: Statistical inference in ego-graphs*. ICS-dissertation, Groningen.
53. Vincent Buskens. (1999). *Social networks and trust*. ICS-dissertation, Utrecht.
54. Susanne Rijken. (1999). *Educational expansion and status attainment. A cross-national and over-time comparison*. ICS-dissertation, Utrecht.
55. Mérove Gijsberts. (1999). *The legitimization of inequality in state-socialist and market societies, 1987-1996*. ICS-dissertation, Utrecht.
56. Gerhard G. Van de Bunt. (1999). *Friends by choice. An actor-oriented statistical network model for friendship networks through time*. ICS-dissertation, Groningen.
57. Robert Thomson. (1999). *The party mandate: Election pledges and government actions in the Netherlands, 1986-1998*. Amsterdam: Thela Thesis.
58. Corine Baarda. (1999). *Politieke besluiten en boeren beslissingen. Het draagvlak van het mestbeleid tot 2000*. ICS-dissertation, Groningen.
59. Rafael Wittek. (1999). *Interdependence and informal control in organizations*. ICS-dissertation, Groningen.
60. Diane Payne. (1999). *Policy making in the European Union: An analysis of the impact of the reform of the structural funds in Ireland*. ICS-dissertation, Groningen.
61. René Veenstra. (1999). *Leerlingen - klassen - scholen. Prestaties en vorderingen van leerlingen in het voortgezet onderwijs*. Amsterdam: Thela Thesis.
62. Marjolein Achterkamp. (1999). *Influence strategies in collective decision making. A comparison of two models*. ICS-dissertation, Groningen.
63. Peter Mühlau. (2000). *The governance of the employment relation. A relational signaling perspective*. ICS-dissertation, Groningen.
64. Agnes Akkerman. (2000). *Verdeelde vakbeweging en stakingen. Concurrentie om leden*. ICS-dissertation, Groningen.
65. Sandra van Thiel. (2000). *Quangocratization: Trends, causes and consequences*. ICS-dissertation, Utrecht.
66. Rudi Turksema. (2000). *Supply of day care*. ICS-dissertation, Utrecht.
67. Sylvia E. Korupp (2000). *Mothers and the process of social stratification*. ICS-dissertation, Utrecht.

68. Bernard A. Nijstad (2000). *How the group affects the mind: Effects of communication in idea generating groups*. ICS-dissertation, Utrecht.
69. Inge F. de Wolf (2000). *Opleidingsspecialisatie en arbeidsmarktsucces van sociale wetenschappers*. ICS-dissertation, Utrecht.
70. Jan Kratzer (2001). *Communication and performance: An empirical study in innovation teams*. ICS-dissertation, Groningen.
71. Madelon Kroneman (2001). *Healthcare systems and hospital bed use*. ICS/NIVEL-dissertation, Utrecht.
72. Herman van de Werfhorst (2001). *Field of study and social inequality. Four types of educational resources in the process of stratification in the Netherlands*. ICS-dissertation, Nijmegen.
73. Tamás Bartus (2001). *Social capital and earnings inequalities. The role of informal job search in Hungary*. ICS-dissertation, Groningen.
74. Hester Moerbeek (2001). *Friends and foes in the occupational career. The influence of sweet and sour social capital on the labour market*. ICS-dissertation, Nijmegen.
75. Marcel van Assen (2001). *Essays on actor perspectives in exchange networks and social dilemmas*. ICS-dissertation, Groningen.
76. Inge Sieben (2001). *Sibling similarities and social stratification. The impact of family background across countries and cohorts*. ICS-dissertation, Nijmegen.
77. Alinda van Bruggen (2001). *Individual production of social well-being. An exploratory study*. ICS-dissertation, Groningen.
78. Marcel Coenders (2001). *Nationalistic attitudes and ethnic exclusionism in a comparative perspective: An empirical study of attitudes toward the country and ethnic immigrants in 22 countries*. ICS-dissertation, Nijmegen.
79. Marcel Lubbers (2001). *Exclusionistic electorates. Extreme right-wing voting in Western Europe*. ICS-dissertation, Nijmegen.
80. Uwe Matzat (2001). *Social networks and cooperation in electronic communities. A theoretical-empirical analysis of academic communication and internet discussion groups*. ICS-dissertation, Groningen.
81. Jacques P.G. Janssen (2002). *Do opposites attract divorce? Dimensions of mixed marriage and the risk of divorce in the Netherlands*. ICS-dissertation, Nijmegen.
82. Miranda Jansen (2002). *Waardenoriëntaties en partnerrelaties. Een panelstudie naar wederzijdse invloeden*. ICS-dissertation, Utrecht.
83. Anne Rigt Poortman (2002). *Socioeconomic causes and consequences of divorce*. ICS-dissertation, Utrecht.
84. Alexander Gattig (2002). *Intertemporal decision making*. ICS-dissertation, Groningen.
85. Gerrit Rooks (2002). *Contract en conflict: Strategisch management van inkooptransacties*. ICS-dissertation, Utrecht.
86. Károly Takács (2002). *Social networks and intergroup conflict*. ICS-dissertation, Groningen.
87. Thomas Gautschi (2002). *Trust and exchange, effects of temporal embeddedness and network embeddedness on providing and dividing a surplus*. ICS-dissertation, Utrecht.
88. Hilde Bras (2002). *Zeeuwse meiden. Dienen in de levensloop van vrouwen, ca. 1850-1950*. Amsterdam: Aksant Academic Publishers.

89. Merijn Rengers (2002). *Economic lives of artists. Studies into careers and the labour market in the cultural sector*. ICS-dissertation, Utrecht.
90. Annelies Kassenberg (2002). *Wat scholieren bindt. Sociale gemeenschap in scholen*. ICS-dissertation, Groningen.
91. Marc Verboord (2003). *Moet de meester dalen of de leerling klimmen? De invloed van literatuuronderwijs en ouders op het lezen van boeken tussen 1975 en 2000*. ICS-dissertation, Utrecht.
92. Marcel van Egmond (2003). *Rain falls on all of us (but some manage to get more wet than others): Political Context and Electoral Participation*. ICS-dissertation, Nijmegen.
93. Justine Horgan (2003). *High performance human resource management in Ireland and the Netherlands: Adoption and effectiveness*. ICS-dissertation, Groningen.
94. Corine Hoeben (2003). *LETS' be a community. Community in local exchange trading systems*. ICS-dissertation, Groningen.
95. Christian Steglich (2003). *The framing of decision situations. Automatic goal selection and rational goal pursuit*. ICS-dissertation, Groningen.
96. Johan van Wilsem (2003). *Crime and context. The impact of individual, neighborhood, city and country characteristics on victimization*. ICS-dissertation, Nijmegen.
97. Christiaan Monden (2003). *Education, inequality and health. The impact of partners and life course*. ICS-dissertation, Nijmegen.
98. Evelyn Hello (2003). *Educational attainment and ethnic attitudes. How to explain their relationship*. ICS-dissertation, Nijmegen.
99. Marnix Croes en Peter Tammes (2004). *Gif laten wij niet voortbestaan. Een onderzoek naar de overlevingskansen van joden in de Nederlandse gemeenten, 1940-1945*. Amsterdam: Aksant Academic Publishers.
100. Ineke Nagel (2004). *Cultuurdeelname in de levensloop*. ICS-dissertation, Utrecht.
101. Marieke van der Wal (2004). *Competencies to participate in life. Measurement and the impact of school*. ICS-dissertation, Groningen.
102. Vivian Meertens (2004). *Depressive symptoms in the general population: A multifactorial social approach*. ICS-dissertation, Nijmegen.
103. Hanneke Schuurmans (2004). *Promoting well-being in frail elderly people. Theory and intervention*. ICS-dissertation, Groningen.
104. Javier Arregui (2004). *Negotiation in legislative decision-making in the European Union*. ICS-dissertation, Groningen.
105. Tamar Fischer (2004). *Parental divorce, conflict and resources. The effects on children's behaviour problems, socioeconomic attainment, and transitions in the demographic career*. ICS-dissertation, Nijmegen.
106. René Bekkers (2004). *Giving and volunteering in the Netherlands: Sociological and psychological perspectives*. ICS-dissertation, Utrecht.
107. Renée van der Hulst (2004). *Gender differences in workplace authority: An empirical study on social networks*. ICS-dissertation, Groningen.
108. Rita Smaniotto (2004). *'You scratch my back and I scratch yours' versus 'Love thy neighbour'. Two Proximate Mechanisms of Reciprocal Altruism*. ICS-dissertation, Groningen.

109. Maurice Gesthuizen (2004). *The life-course of the low-educated in the Netherlands: Social and economic risks*. ICS-dissertation, Nijmegen.
110. Carlijne Philips (2005). *Vakantiegemeenschappen. Kwalitatief en kwantitatief onderzoek naar gelegenheid- en refreshergemeenschap tijdens de vakantie*. ICS-dissertation, Groningen.
111. Esther de Ruijter (2005). *Household outsourcing*. ICS-dissertation, Utrecht.
112. Frank van Tubergen (2005). *The integration of immigrants in cross-national perspective: Origin, destination, and community effects*. ICS-dissertation, Utrecht.
113. Ferry Koster (2005). *For the time being. Accounting for inconclusive findings concerning the effects of temporary employment relationships on solidary behavior of employees*. ICS-dissertation, Groningen.
114. Carolien Klein Haarhuis (2005). *Promoting anti-corruption reforms. Evaluating the implementation of a World Bank anti-corruption program in seven African countries (1999–2001)*. ICS-dissertation, Utrecht.
115. Martin van der Gaag (2005). *Measurement of individual social capital*. ICS-dissertation, Groningen.
116. Johan Hansen (2005). *Shaping careers of men and women in organizational contexts*. ICS-dissertation, Utrecht.
117. Davide Barrera (2005). *Trust in embedded settings*. ICS-dissertation, Utrecht.
118. Mattijs Lambooi (2005). *Promoting cooperation. Studies into the effects of long-term and short-term rewards on cooperation of employees*. ICS-dissertation, Utrecht.
119. Lotte Vermeij (2006). *What's cooking? Cultural boundaries among Dutch teenagers of different ethnic origins in the context of school*. ICS-dissertation, Utrecht.
120. Mathilde Strating (2006). *Facing the challenge of rheumatoid arthritis. A 13-year prospective study among patients and cross-sectional study among their partners*. ICS-dissertation, Groningen.
121. Jannes de Vries (2006). *Measurement error in family background variables: The bias in the intergenerational transmission of status, cultural consumption, party preference, and religiosity*. ICS-dissertation, Nijmegen.
122. Stefan Thau (2006). *Workplace deviance: Four studies on employee motives and self-regulation*. ICS-dissertation, Groningen.
123. Mirjam Plantinga (2006). *Employee motivation and employee performance in child care. The effects of the introduction of market forces on employees in the Dutch child-care sector*. ICS-dissertation, Groningen.
124. Helga de Valk (2006). *Pathways into adulthood. A comparative study on family life transitions among migrant and Dutch youth*. ICS-dissertation, Utrecht.
125. Henrike Elzen (2006). *Self-management for chronically ill older people*. ICS-dissertation, Groningen.
126. Ayşe Güveli (2007). *New social classes within the service class in the Netherlands and Britain. Adjusting the EGP class schema for the technocrats and the social and cultural specialists*. ICS-dissertation, Nijmegen.
127. Willem-Jan Verhoeven (2007). *Income attainment in post-communist societies*. ICS-dissertation, Utrecht.

128. Marieke Voorpostel (2007). *Sibling support: The exchange of help among brothers and sisters in the Netherlands*. ICS-dissertation, Utrecht.
129. Jacob Dijkstra (2007). *The effects of externalities on partner choice and payoffs in exchange networks*. ICS-dissertation, Groningen.
130. Patricia van Echtelt (2007). *Time-greedy employment relationships: Four studies on the time claims of post-Fordist work*. ICS-dissertation, Groningen.
131. Sonja Vogt (2007). *Heterogeneity in social dilemmas: The case of social support*. ICS-dissertation, Utrecht.
132. Michael Schweinberger (2007). *Statistical methods for studying the evolution of networks and behavior*. ICS-dissertation, Groningen.
133. István Back (2007). *Commitment and evolution: Connecting emotion and reason in long-term relationships*. ICS-dissertation, Groningen.
134. Ruben van Gaalen (2007). *Solidarity and ambivalence in parent-child relationships*. ICS-dissertation, Utrecht.
135. Jan Reitsma (2007). *Religiosity and solidarity - Dimensions and relationships disentangled and tested*. ICS-dissertation, Nijmegen.
136. Jan Kornelis Dijkstra (2007) *Status and affection among (pre)adolescents and their relation with antisocial and prosocial behavior*. ICS-dissertation, Groningen.
137. Wouter van Gils (2007). *Full-time working couples in the Netherlands. Causes and consequences*. ICS-dissertation, Nijmegen.
138. Djamila Schans (2007). *Ethnic diversity in intergenerational solidarity*. ICS-dissertation, Utrecht.
139. Ruud van der Meulen (2007). *Brug over woelig water: Lidmaatschap van sportverenigingen, vriendschappen, kennissenkringen en veralgemeend vertrouwen*. ICS-dissertation, Nijmegen.
140. Andrea Knecht (2008). *Friendship selection and friends' influence. Dynamics of networks and actor attributes in early adolescence*. ICS-dissertation, Utrecht.
141. Ingrid Doorten (2008). *The division of unpaid work in the household: A stubborn pattern?*. ICS-dissertation, Utrecht.
142. Stijn Ruiters (2008). *Association in context and association as context: Causes and consequences of voluntary association involvement*. ICS-dissertation, Nijmegen.
143. Janneke Joly (2008). *People on our minds: When humanized contexts activate social norms*. ICS-dissertation, Groningen.
144. Margreet Frieling (2008). *'Joint production' als motor voor actief burgerschap in de buurt*. ICS-dissertation, Groningen.
145. Ellen Verbakel (2008). *The partner as resource or restriction? Labour market careers of husbands and wives and the consequences for inequality between couples*. ICS-dissertation, Nijmegen.
146. Gijs van Houten (2008). *Beleidsuitvoering in gelaagde stelsels. De doorwerking van aanbevelingen van de Stichting van de Arbeid in het CAO-overleg*. ICS-dissertation, Utrecht.
147. Eva Jaspers (2008). *Intolerance over time. Macro and micro level questions on attitudes towards euthanasia, homosexuality and ethnic minorities*. ICS-dissertation, Nijmegen.

148. Gijs Weijters (2008). *Youth delinquency in Dutch cities and schools: A multilevel approach*. ICS-dissertation, Nijmegen.
149. Jessica Pass (2009). *The self in social rejection*. ICS-dissertation, Groningen.
150. Gerald Mollenhorst (2009). *Networks in contexts. How meeting opportunities affect personal relationships*. ICS-dissertation, Utrecht.
151. Tom van der Meer (2009). *States of freely associating citizens: Comparative studies into the impact of state institutions on social, civic and political participation*. ICS-dissertation, Nijmegen.
152. Manuela Vieth (2009). *Commitments and reciprocity in trust situations. Experimental studies on obligation, indignation, and self-consistency*. ICS-dissertation, Utrecht.
153. Rense Corten (2009). *Co-evolution of social networks and behavior in social dilemmas: Theoretical and empirical perspectives*. ICS-dissertation, Utrecht.
154. Arieke J. Rijken (2009). *Happy families, high fertility? Childbearing choices in the context of family and partner relationships*. ICS-dissertation, Utrecht.
155. Jochem Tolsma (2009). *Ethnic hostility among ethnic majority and minority groups in the Netherlands. An investigation into the impact of social mobility experiences, the local living environment and educational attainment on ethnic hostility*. ICS-dissertation, Nijmegen.
156. Freek Bucx (2009). *Linked lives: Young adults' life course and relations with parents*. ICS-dissertation, Utrecht.
157. Philip Wotschack (2009). *Household governance and time allocation. Four studies on the combination of work and care*. ICS-dissertation, Groningen.
158. Nienke Moor (2009). *Explaining worldwide religious diversity. The relationship between subsistence technologies and ideas about the unknown in pre-industrial and (post-)industrial societies*. ICS-dissertation, Nijmegen.
159. Lieke ten Brummelhuis (2009). *Family matters at work. Depleting and enriching effects of employees' family lives on work outcomes*. ICS-dissertation, Utrecht.
160. Renske Keizer (2010). *Remaining childless. Causes and consequences from a life course perspective*. ICS-dissertation, Utrecht.
161. Miranda Sentse (2010). *Bridging contexts: The interplay between family, child, and peers in explaining problem behavior in early adolescence*. ICS-dissertation, Groningen.
162. Nicole Tieben (2010). *Transitions, tracks and transformations. Social inequality in transitions into, through and out of secondary education in the Netherlands for cohorts born between 1914 and 1985*. ICS-dissertation, Nijmegen.
163. Birgit Pauksztat (2010). *Speaking up in organizations: Four studies on employee voice*. ICS-dissertation, Groningen.
164. Richard Zijdeman (2010). *Status attainment in the Netherlands, 1811-1941. Spatial and temporal variation before and during industrialization*. ICS-dissertation, Utrecht.
165. Rianne Kloosterman (2010). *Social background and children's educational careers. The primary and secondary effects of social background over transitions and over time in the Netherlands*. ICS-dissertation, Nijmegen.
166. Olav Aarts (2010). *Religious diversity and religious involvement. A study of religious markets in Western societies at the end of the twentieth century*. ICS-dissertation, Nijmegen.

167. Stephanie Wiesmann (2010). *24/7 negotiation in couples transition to parenthood*. ICS-dissertation, Utrecht.
168. Borja Martinovic (2010). *Interethnic contacts: A dynamic analysis of interaction between immigrants and natives in Western countries*. ICS-dissertation, Utrecht.
169. Anne Roeters (2010). *Family life under pressure? Parents' paid work and the quantity and quality of parent-child and family time*. ICS-dissertation, Utrecht.
170. Jelle Sijtsema (2010). *Adolescent aggressive behavior: Status and stimulation goals in relation to the peer context*. ICS-dissertation, Groningen.
171. Kees Keizer (2010). *The spreading of disorder*. ICS-dissertation, Groningen.
172. Michael Mäs (2010). *The diversity puzzle. Explaining clustering and polarization of opinions*. ICS-dissertation, Groningen.
173. Marie-Louise Damen (2010). *Cultuurdeelname en CKV. Studies naar effecten van kunsteducatie op de cultuurdeelname van leerlingen tijdens en na het voortgezet onderwijs*. ICS-dissertation, Utrecht.
174. Marieke van de Rakt (2011). *Two generations of crime: The intergenerational transmission of convictions over the life course*. ICS-dissertation, Nijmegen.
175. Willem Huijnk (2011). *Family life and ethnic attitudes. The role of the family for attitudes towards intermarriage and acculturation among minority and majority groups*. ICS-dissertation, Utrecht.
176. Tim Huijts (2011). *Social ties and health in Europe. Individual associations, cross-national variations, and contextual explanations*. ICS-dissertation, Nijmegen.
177. Wouter Steenbeek (2011). *Social and physical disorder. How community, business presence and entrepreneurs influence disorder in Dutch neighborhoods*. ICS-dissertation, Utrecht.
178. Miranda Vervoort (2011). *Living together apart? Ethnic concentration in the neighborhood and ethnic minorities' social contacts and language practices*. ICS-dissertation, Utrecht.
179. Agnieszka Kanas (2011). *The economic performance of immigrants. The role of human and social capital*. ICS-dissertation, Utrecht.
180. Lea Ellwardt (2011). *Gossip in organizations. A social network study*. ICS-dissertation, Groningen.
181. Annemarije Oosterwaal (2011). *The gap between decision and implementation. Decision making, delegation and compliance in governmental and organizational settings*. ICS-dissertation, Utrecht.
182. Natascha Notten (2011). *Parents and the media. Causes and consequences of parental media socialization*. ICS-dissertation, Nijmegen.
183. Tobias Stark (2011). *Integration in schools. A process perspective on students' interethnic attitudes and interpersonal relationships*. ICS-dissertation, Groningen.
184. Giedo Jansen (2011). *Social cleavages and political choices. Large-scale comparisons of social class, religion and voting behavior in Western democracies*. ICS-dissertation, Nijmegen.
185. Ruud van der Horst (2011). *Network effects on treatment results in a closed forensic psychiatric setting*. ICS-dissertation, Groningen.


186. Mark Levels (2011). *Abortion laws in European countries between 1960 and 2010. Legislative developments and their consequences for women's reproductive decision-making*. ICS-dissertation, Nijmegen.
187. Marieke van Londen (2012). *Exclusion of ethnic minorities in the Netherlands. The effects of individual and situational characteristics on opposition to ethnic policy and ethnically mixed neighbourhoods*. ICS-dissertation, Nijmegen.
188. Sigrid M. Mohnen (2012). *Neighborhood context and health: How neighborhood social capital affects individual health*. ICS-dissertation, Utrecht.
189. Asya Zhelyazkova (2012). *Compliance under controversy: Analysis of the transposition of European directives and their provisions*. ICS-dissertation, Utrecht.
190. Valeska Korff (2012). *Between cause and control: Management in a humanitarian organization*. ICS-dissertation, Groningen.
191. Maike Gieling (2012). *Dealing with diversity: Adolescents' support for civil liberties and immigrant rights*. ICS-dissertation, Utrecht.
192. Katya Ivanova (2012). *From parents to partners: The impact of family on romantic relationships in adolescence and emerging adulthood*. ICS-dissertation, Groningen.
193. Jelmer Schalk (2012). *The performance of public corporate actors: Essays on effects of institutional and network embeddedness in supranational, national, and local collaborative contexts*. ICS-dissertation, Utrecht.
194. Alona Labun (2012). *Social networks and informal power in organizations*. ICS-dissertation, Groningen.
195. Michał Bojanowski (2012). *Essays on social network formation in heterogeneous populations: Models, methods, and empirical analyses*. ICS-dissertation, Utrecht.
196. Anca Minescu (2012). *Relative group position and intergroup attitudes in Russia*. ICS-dissertation, Utrecht.
197. Marieke van Schellen (2012). *Marriage and crime over the life course. The criminal careers of convicts and their spouses*. ICS-dissertation, Utrecht.
198. Mieke Maliepaard (2012). *Religious trends and social integration: Muslim minorities in the Netherlands*. ICS-dissertation, Utrecht.
199. Fransje Smits (2012). *Turks and Moroccans in the Low Countries around the year 2000: determinants of religiosity, trend in religiosity and determinants of the trend*. ICS-dissertation, Nijmegen.
200. Roderick Sluiter (2012). *The diffusion of morality policies among Western European countries between 1960 and 2010. A comparison of temporal and spatial diffusion patterns of six morality and eleven non-morality policies*. ICS-dissertation, Nijmegen.
201. Nicoletta Balbo (2012). *Family, friends and fertility*. ICS-dissertation, Groningen.
202. Anke Munniksmä (2013). *Crossing ethnic boundaries: Parental resistance to and consequences of adolescents' cross-ethnic peer relations*. ICS-dissertation, Groningen.
203. Anja Abendroth (2013). *Working women in Europe. How the country, workplace, and family context matter*. ICS-dissertation, Utrecht.
204. Katia Begall (2013). *Occupational hazard? The relationship between working conditions and fertility*. ICS-dissertation, Groningen.
205. Hidde Bekhuis (2013). *The popularity of domestic cultural products: Cross-national differences and the relation to globalization*. ICS-dissertation, Utrecht.

206. Lieselotte Blommaert (2013). *Are Joris and Renske more employable than Rashid and Samira? A study on the prevalence and sources of ethnic discrimination in recruitment in the Netherlands using experimental and survey data.* ICS-dissertation, Utrecht.
207. Wiebke Schulz (2013). *Careers of men and women in the 19th and 20th centuries.* ICS-dissertation, Utrecht.
208. Ozan Aksoy (2013). *Essays on social preferences and beliefs in non-embedded social dilemmas.* ICS-dissertation, Utrecht.
209. Dominik Morbitzer (2013). *Limited farsightedness in network formation.* ICS-dissertation, Utrecht.
210. Thomas de Vroome (2013). *Earning your place: The relation between immigrants' economic and psychological integration in the Netherlands.* ICS-dissertation, Utrecht.
211. Marloes de Lange (2013). *Causes and consequences of employment flexibility among young people. Recent developments in the Netherlands and Europe.* ICS-dissertation, Nijmegen.
212. Roza Meuleman (2014). *Consuming the nation. Domestic cultural consumption: Its stratification and relation with nationalist attitudes.* ICS-dissertation, Utrecht.
213. Esther Havekes (2014). *Putting interethnic attitudes in context. The relationship between neighbourhood characteristics, interethnic attitudes and residential behaviour.* ICS-dissertation, Utrecht.
214. Zoltán Lippényi (2014). *Transitions toward an open society? Intergenerational occupational mobility in Hungary in the 19th and 20th centuries.* ICS-dissertation, Utrecht.
215. Anouk Smeekes (2014). *The presence of the past: Historical rooting of national identity and current group dynamics.* ICS-dissertation, Utrecht.
216. Michael Savelkoul (2014). *Ethnic diversity and social capital. Testing underlying explanations derived from conflict and contact theories in Europe and the United States.* ICS-dissertation, Nijmegen.
217. Martijn Hogerbrugge (2014). *Misfortune and family: How negative events, family ties, and lives are linked.* ICS-dissertation, Utrecht.
218. Gina Potarca (2014). *Modern love. Comparative insights in online dating preferences and assortative mating.* ICS-dissertation, Groningen.
219. Mariska van der Horst (2014). *Gender, aspirations, and achievements: Relating work and family aspirations to occupational outcomes.* ICS-dissertation, Utrecht.
220. Gijs Huitsing (2014). *A social network perspective on bullying.* ICS dissertation, Groningen.
221. Thomas Kowalewski (2015). *Personal growth in organizational contexts.* ICS-dissertation, Groningen.
222. Manu Muñoz-Herrera (2015). *The impact of individual differences on network relations: Social exclusion and inequality in productive exchange and coordination games.* ICS-dissertation, Groningen.
223. Tim Immerzeel (2015). *Voting for a change. The democratic lure of populist radical right parties in voting behavior.* ICS-dissertation, Utrecht.
224. Fernando Nieto Morales (2015). *The control imperative: Studies on reorganization in the public and private sectors.* ICS-dissertation, Groningen.

225. Jellie Sierksma (2015). *Bounded helping: How morality and intergroup relations shape children's reasoning about helping*. ICS-dissertation, Utrecht.
226. Tinka Veldhuis (2015). *Captivated by fear. An evaluation of terrorism detention policy*. ICS-dissertation, Groningen.
227. Miranda Visser (2015). *Loyalty in humanity. Turnover among expatriate humanitarian aid workers*. ICS-dissertation, Groningen.
228. Sarah Westphal (2015). *Are the kids alright? Essays on postdivorce residence arrangements and children's well-being*. ICS-dissertation, Utrecht.
229. Britta Rüschoff (2015). *Peers in careers: Peer relationships in the transition from school to work*. ICS-dissertation, Groningen.
230. Nynke van Miltenburg (2015). *Cooperation under peer sanctioning institutions: Collective decisions, noise, and endogenous implementation*. ICS-dissertation, Utrecht.
231. Antonie Knigge (2015). *Sources of sibling similarity. Status attainment in the Netherlands during modernization*. ICS-dissertation, Utrecht.
232. Sanne Smith (2015). *Ethnic segregation in friendship networks. Studies of its determinants in English, German, Dutch, and Swedish school classes*. ICS-dissertation, Utrecht.
233. Patrick Präg (2015). *Social stratification and health. Four essays on social determinants of health and wellbeing*. ICS-dissertation, Groningen.
234. Wike Been (2015). *European top managers' support for work-life arrangements*. ICS-dissertation, Utrecht.
235. André Grow (2016). *Status differentiation: New insights from agent-based modeling and social network analysis*. ICS-dissertation, Groningen.
236. Jesper Rözer (2016). *Family and personal networks. How a partner and children affect social relationships*. ICS-dissertation, Utrecht.
237. Kim Pattiselanno (2016). *At your own risk: The importance of group dynamics and peer processes in adolescent peer groups for adolescents' involvement in risk behaviors*. ICS-dissertation, Groningen.
238. Vincenz Frey (2016). *Network formation and trust*. ICS-dissertation, Utrecht.
239. Rozemarijn van der Ploeg (2016). *Be a buddy, not a bully? Four studies on social and emotional processes related to bullying, defending, and victimization*. ICS-dissertation, Groningen.
240. Tali Spiegel (2016). *Identity, career trajectories and wellbeing: A closer look at individuals with degenerative eye conditions*. ICS-dissertation, Groningen.
241. Felix Tropf (2016). *Social Science Genetics and Fertility*. ICS-dissertation, Groningen.
242. Sara Geven (2016). *Adolescents problem behavior in school: the role of peer networks*. ICS-dissertation, Utrecht.
243. Josja Rokven (2016). *The victimization-offending relationship from a longitudinal perspective*. ICS-dissertation, Nijmegen.
244. Maja Djundeva (2016). *Healthy ageing in context: Family welfare state and the life course*. ICS-dissertation, Groningen.
245. Mark Visser (2017). *Inequality between older workers and older couples in the Netherlands. A dynamic life course perspective on educational and social class differences in the late career*. ICS-dissertation, Nijmegen.

246. Beau Oldenburg (2017). *Bullying in schools: The role of teachers and classmates*. ICS-dissertation, Groningen.
247. Tatang Muttaqin (2017). *The education divide in Indonesia: Four essays on determinants of unequal access to and quality of education*. ICS-dissertation, Groningen.
248. Margriet van Hek (2017). *Gender inequality in educational attainment and reading performance. A contextual approach*. ICS-dissertation, Nijmegen.
249. Melissa Verhoef (2017). *Work schedules, childcare and well-being. Essays on the associations between modern-day job characteristics, childcare arrangements and the well-being of parents and children*. ICS-dissertation, Utrecht.
250. Timo Septer (2017). *Goal priorities, cognition and conflict: Analyses of cognitive maps concerning organizational changes*. ICS-dissertation, Groningen.
251. Bas Hofstra (2017). *Online Social Networks: Essays on Membership, Privacy, and Structure*. ICS-dissertation, Utrecht.
252. Yassine Khoudja (2018). *Women's labor market participation across ethnic groups: The role of household conditions, gender role attitudes, and religiosity in different national contexts*. ICS-dissertation, Utrecht.
253. Joran Laméris (2018). *Living together in diversity. Whether, why and where ethnic diversity affects social cohesion*. ICS-dissertation, Nijmegen.
254. Maaïke van der Vleuten (2018). *Gendered Choices. Fields of study of adolescents in the Netherlands*. ICS-dissertation, Utrecht.
255. Mala Sondang Silitonga (2018). *Corruption in Indonesia: The impact of institutional change, norms, and networks*. ICS-dissertation, Groningen.
256. Manja Coopmans (2018). *Rituals of the past in the context of the present. The role of Remembrance Day and Liberation Day in Dutch society*. ICS-dissertation, Utrecht.
257. Paul Hindriks (2018). *The Struggle for Power: Attitudes towards the political participation of ethnic minorities*. ICS-dissertation, Utrecht.
258. Nynke Niezink (2018). *Modeling the dynamics of networks and continuous behavior*. ICS-dissertation, Groningen.
259. Simon de Bruijn (2018). *Reaching agreement after divorce and separation. Essays on the effectiveness of parenting plans and divorce mediation*. ICS-dissertation, Utrecht.
260. Susanne van 't Hoff-de Goede (2018). *While you were locked up. An empirical study on the characteristics, social surroundings and wellbeing of partners of prisoners in The Netherlands*. ICS-dissertation, Utrecht.
261. Loes van Rijsewijk (2018). *Antecedents and Consequences of Helping among Adolescents*. ICS-dissertation, Groningen.
262. Mariola Gremmen (2018). *Social network processes and academic functioning. The role of peers in students' school well-being, academic engagement, and academic achievement*. ICS-dissertation, Groningen.
263. Jeanette Renema (2018). *Immigrants' support for welfare spending. The causes and consequences of welfare usage and welfare knowledgeability*. ICS-dissertation, Nijmegen.
264. Suwatin Miharti (2018). *Community health centers in Indonesia in the era of decentralization. The impact of structure, staff composition and management on health outcomes*. ICS-dissertation, Groningen.

265. Chaïm la Roi (2019). *Stigma and stress: Studies on attitudes towards sexual minority orientations and the association between sexual orientation and mental health*. ICS-dissertation, Groningen.
266. Jelle Lössbroek (2019). *Turning grey into gold. Employer-employee interplay in an ageing workforce*. ICS-dissertation, Utrecht.
267. Nikki van Gerwen (2019). *Employee cooperation through training. A multi-method approach*. ICS-dissertation, Utrecht.
268. Paula Thijs (2019). *Trends in cultural conservatism: The role of educational expansion, secularisation, and changing national contexts*. ICS-dissertation, Nijmegen.
269. Renske Verweij (2019). *Understanding childlessness: Unravelling the link with genes and the socio-environment*. ICS-dissertation, Groningen.
270. Niels Blom (2019). *Partner relationship quality under pressing work conditions. Longitudinal and cross-national investigation*. ICS-dissertation, Nijmegen.
271. Müge Simsek (2019). *The dynamics of religion among native and immigrant youth in Western Europe*. ICS-dissertation, Utrecht.
272. Leonie van Breeschoten (2019). *Combining a career and childcare: The use and usefulness of work-family policies in European organizations*. ICS-dissertation, Utrecht.
273. Roos van der Zwan (2019). *The political representation of ethnic minorities and their vote choice*. ICS dissertation, Nijmegen.
274. Ashwin Rambaran (2019). *The classroom as context for bullying: A social network approach*. ICS dissertation, Groningen.
275. Dieko Bakker (2019). *Cooperation and social control: Effects of preferences, institutions, and social structure*. ICS-dissertation, Groningen.
276. Femke van der Werf (2019). *Shadow of a rainbow? National and ethnic belonging in Mauritius*. ICS dissertation, Utrecht.
277. Robert Krause (2019). *Multiple Imputation for Missing Network Data*. ICS-dissertation, Groningen.



A problem for social scientist when conducting empirical research is missing information. Some participants provide incomplete or no (usable) data, leading to lower power of statistical analysis and potentially biased results. This problem is even more severe in network studies, where the connections between people or organizations is the object of the research. Here, missing information from one participant also means missing information for all other members of the network, after all the missing nodes could have indicated a connection to any of the other nodes. In this book multiple imputation methods for missing network data are introduced, implemented, and evaluated. The discussed multiple imputation procedures are based on the two most used statistical network model families in the social sciences, Exponential Random Graph Models (ERGMs) and Stochastic Actor-oriented Models (SAOMs). Examples are given for treating missing network data in cross-sectional, longitudinal, multiplex, and multilevel networks, as well as missing data on nodal attributes. Recommendations for best practices are discussed.

Robert W Krause obtained his Research Master's degree in Behavioural Sciences (2015) at the Radboud University, Nijmegen. The present study was conducted at the Inter-university Center for Social Science Theory and Methodology (ICS) at the University of Groningen.