Imputation of attributes in networked data using Bayesian Autocorrelation Regression Models

1 Introduction

In the cyber domain, large volumes of networked data are being collected, where links (or: edges) can indicate friendship (eg. Facebook), following status (eg. Twitter), sent and received emails, IP-traffic (eg. botnets), financial transactions, and geolocation variables such as zip-code or country codes. Understanding network structure and topology has received attention and is informative for statistical procedures such as clustering [3, 16]. Equally important is the analyses of network attributes (i.e., node covariates), to understand node characteristics and the coordinating role of network topology.

Missing data is feature of much automated or online data collection which can lead to biased estimates ([28, 39]). Outside the network setting, imputation methods often focus on the prediction of missing values in relational datasets consisting of data from randomly selected individuals (eg. [12, 37]). This has fed method-development for imputation of missing data from conditionally independent cases [44]. In a network setting, with observations linked to network structure, complete cases analyses can exacerbate the problem since it is based on a fraction of the information available in the network. When well connected individuals are removed entire clusters drop out with dramatic loss of information [11, 20]. In this setting, we may do far better if we reconnect the clusters by imputation of the missing values.

A previous meta-analysis of network effects [6] provided convincing evidence for homophily, the preference for associating with individuals/actors with similar attributes. If network structure is linked to actor covariate values,

this may help in estimating the attribute of a node surrounded by nodes who have data on that particular attribute. On the other hand, if attribute values can be predicted from network edges, it is likely that the pattern of missing data is also affected by tie structure. The problem of incomplete tie structures has been described as part of the boundary specification problem, arising when the researcher has to decide which actors are relevant to include and which edges do not contribute (reviewed in [23]). Other factors are non-response effects in the data collection procedure (resulting in unobserved parts of the network) and fixed-choice effects (occurring when network actors are asked to nominate a fixed number of friends). Because missingness in networks is often reflected in tie structure, previous studies focusing on missing data have for the most part presented methods for predicting missing edges between actors [19, 22] and different strategies for (sub)network sampling [5, 14].

We are motivated by problems in which the network is observable, and there is important missing (or misleading) data in node attributes. This may appear easier than imputation of edges: from the perspective of parametric Bayesian inference, uncertainty in edges resembles model uncertainty, as interactions must be imputed, and this leads to problems resembling model choice and model averaging; imputation of missing node values conditioned on a network is uncertainty in auxiliary variables and leads to parameter estimation. There is also the joint analysis, using parametric models for joint network and node attributes [42]. For Exponential Random Graph models (ERGMs; [7, 17, 46]) studies have proposed adaptive sampling mechanisms to acquire accurate posterior distributions [9, 11, 22, 48] under missing data. From these models, missing node attributes can be imputed. However, measuring the impact of misspecification, and treating it, is challenging due to the formal intractability of these models. Autocorrelation Regression Models

(ARMs; [2, 33, 41]) provide a more straightforward setting for new statistical methods. In addition, for the probit ARM models for binary node attributes, which we consider below, there must always be fields of missing latent continuous variables. Given these considerations, and the popularity of ARMs, we choose to illustrate misspecification-robust imputation using ARMs.

In a Bayesian setting, where missing data is treated as a latent variable, to be estimated or integrated in a joint posterior distribution along with other parameters of the model, the presence of missing data does not usually impose additional modelling. The conditional distribution of the missing data is determined by the same observation model, model parameters and priors needed for the full-data analysis. We refer to imputation based on the joint posterior of the missing data and model parameters as a (standard) "full Bayes" analysis. In this paper we make two main points: where there is model misspecification, it may be the case that a two stage impute-and-fit approach may be preferred to full Bayes. Our second point is that the effect of model misspecification is reinforced where we have large network "fields" of coupled missing data. In this case the benefit of a two stage approach can be dramatic.

The methods we apply come from the recent literature on Bayesian theory and methods for inference from misspecified models. We demonstrate the usefulness of "cut models" ([21, 29, 36]). Cut models are useful when the model is misspecified and we dont know the correct model, so we cannot immediately fix the problem by model elaboration ([8]). Instead, we try to control the impact of the misspecification by cutting feedback from misspecified model elements.

In the following we define imputation with ARMs and explain how to carry out inference with a cut model. We compare the cut model approach to full Bayes in a misspecified setting, illustrating the difference using a publicly

available social network dataset that is fully observed, and in which we introduce missing data according to different scenario's (snowball/MAR and random/MCAR sampling). Finally, in order to give a simple benchmark for comparison, and underline the robustness of our methods for network data, we compare two naive clustering-based imputation methods (distance- and neighbour based).

2 Methods

2.1 Data

Data were obtained via [removed author name] who collected data from a Facebook-like messaging service from [removed sample reference] and was kind enough to share attribute data (gender and year of study) with us for this project. Data were available from 1899 persons (1118 females) who used the messenger application during 196 days covering the period from [removed date] The data included all users that sent or received at least one message during that period. These longitudinal data were collapsed into covariates; popularity (indegree + outdegree), the day somebody became active using the application, and the day a person reached 75% of the friends in his/her total network. 549 persons received a message but did not respond, and 37 send a message but did not receive a response. A full description of the data is presented in [removed citation]

All data were fully observed, so we introduced missingness by removing some values from the (for the moment) binary *gender* data $y \in \{0,1\}^n$. The data were a $n \times (p+1)$ matrix X with n rows corresponding to the $n = 1899$ people in the study, a column for the intercept, and $p = 7$ columns corresponding to covariates (indegree, five year of study levels, and day active) chosen to inform

gender. A relationship network matrix C was constructed in the following way. For $i, j \in \{1, ..., n\}$ let $a_{i,j}$ denote the number of messages sent from i to j and let $C_{i,j} = \max(a_{i,j}, a_{j,i})$ be the overall network weight for edge $\langle i, j \rangle$.

In the model below, gender y will be the response, with some missing values. We use an ARM to model the relationship between y and the node covariates X in the context of network structure evidenced by C.

2.2 Autocorrelation Regression Model

In the Bayesian ARM for completely observed data [6, 24–27], we let C be a general $n \times n$ matrix of network weights. Let $W_{i,j} = C_{i,j} / \sum_{k=1}^{n} C_{i,k}$ so that $W = [W_{i,j}]_{i=1,...,n}^{j=1,...,n}$ is a row-stochastic version of C. Let X be an $n \times (p+1)$ design matrix of covariates with first column corresponding to the intercept, let $\beta \in R^{p+1}$ be a vector of regression coefficients, I_n the $n \times n$ identity matrix, and $\epsilon \in R^n$ be a vector of n independent network variation differences $\epsilon \sim N(0, I_n \sigma^2)$, with $\sigma = 1$ a variance parameter which can be set equal one in the probit setting of interest. Finally, $\rho \in R$ is the network autocorrelation parameter measuring the network influence. This is positive if attribute values of connected actors tend to converge and negative if those values diverge [6, 27].

The canonical spatial autoregressive model for a real response $z \in R^n$ is

$$z = \rho W z + X\beta + \epsilon \tag{1}$$

or equivalently

$$z = (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1} \epsilon \tag{2}$$

Let $A_\rho = I_n - \rho W$ and $|A_\rho| = \det(A_\rho)$. The likelihood for β and ρ given fully

observed z is

$$p(z|\beta,\rho) \propto |A_\rho| \exp\Big(-\frac{1}{2}(A_\rho z - X\beta)^T(A_\rho z - X\beta)\Big). \tag{3}$$

The log-determinant $\log(|A_\rho|)$ must be evaluated in order to infer ρ. This is non-trivial and we found some schemes were not numerically stable, albeit in rather extreme missingness cases. Of the three $\log(|A_\rho|)$-estimators implemented with [47] (the grid method of Pace and Barry [34], spline approximation using grid points, and a Chebyshev approximation [35]) the grid method proved most reliable, though we had agreement in all but the most extreme cases. The grid method, although robust, can be slow, and would not be used if other faster methods give adequate estimates.

2.2.1 Bayesian Probit ARMs for binary data

The probit-ARM models introduced below are parameter rich. In fact, the number of latent parameters is proportional to the number of response observations. In addition we have missing data. In this setting some form of parameter regularisation is needed. Bayesian network model inference [13, 25] is a coherent regularisation framework. Bayesian implementations of ARM's (e.g. [24]) use Markov Chain Monte Carlo (MCMC) to summarise posterior distribution. In our setting maximum likelihood estimation can result in a downward bias of the network effect parameter ρ when cases are strongly connected [31, 32]. A number of factors contribute to this bias [40]; for example, a network effect can reduce the amount of information gained from each node.

In our data the response variables $y_i, i = 1,...,n$ are binary. Several studies have applied logit or probit Bayesian ARMs to discrete covariate data with, respectively, a dichotomous or multinomial/ordinal outcome [18, 26, 47]. In a probit ARM the binary response y is modeled as a discretisation of an underly-

ing continuous latent field z [1], itself following an ARM as above. The variance is fixed to unity so that the regression parameters β are identifiable [24]. For $i = 1, ..., n$ we model $y_i = \mathbb{I}_{z_i > 0}$, leaving us with parameters z, β and ρ, data y and a posterior distribution

$$\pi(z, \rho, \beta | y) \propto \pi(\rho, \beta) p(z | \rho, \beta) \mathbb{I}_{z \in \mathcal{Z}_y}, \tag{4}$$

where

$$\mathcal{Z}_y = \{z \in R^n : \mathbb{I}y_i = \mathbb{I}z_i > 0 \text{ for each } i = 1, ..., n\}$$

and $p(z|\rho, \beta)$ is given in Equation 3. The prior for $z \in R^n$ is the ARM defined in Equation 3. This can alternatively be thought of as the observation model for the missing data z. We assume independent prior(s) for ρ and β with $\pi(\rho, \beta) = \pi_\rho(\rho)\pi_\beta(\beta)$. For the prior on $\rho \in [-1, 1]$ we take a meta-analytic value based on 183 estimates of ρ [6] encountered in a wide variety of independent data sets. We summarise those data for ρ via a normal distribution with mean $\mu_\rho = 0.36$ and standard deviation $\sigma_\rho = 0.19$ truncated to the interval $[-1, 1]$. Our priors for β are independent near flat normal priors with large variance ($\sigma_\beta = 10^{12}$), $\pi_\beta(\beta) = N(\beta; 0, \sigma_\beta^2 I_{p+1})$.

For completely observed data, functions fitting models of this kind are incorporated in the Spatial Econometrics Toolbox for Matlab. There are some associated R packages (listed by [30]) such as *sarprobit* [47].

2.3 Bayesian inference for missing data

2.3.1 Posterior predictive distribution for missing data

We now consider imputation of missing data in the vector of responses, y. In our case y is a binary vector recording gender. In a Bayesian setting imputation of missing values is formally straightforward. The missing y-entries

are unknown, and treated as parameters alongside z, β and ρ. We outline this "full Bayes" approach in this section. Our point below will be that the full Bayes approach fails where there is model misspecification combined with large amounts of missing data. Indeed it is surprising any approach works in this setting. However, recent developments in Bayes methods for misspecified models, and in particular the use of "cut models" [36], are robust tools for network model parameter inference and offer a way forward.

We assume the following setting. Suppose we are given original data collectively $\tilde{X} = [y, X]$ which contains one column y with some missing entries. Let $y = (y_{obs}, y_{mis})^T$ with $y_{obs} = (y_1, ..., y_{n-q})^T$ observed and $y_{mis} = (y_{n-q+1}, ..., y_n)^T$ missing, so that there are $q \in \{1, ..., n\}$ missing entries in all. It is convenient to sort data in rows so that the missing data are in the last q rows. In order to impute y_{mis} we treat the full observed matrix X as a matrix of covariates and model the relation between y and X using the ARM given in Equation 1.

The posterior distribution conditions on the observed data only,

$$\pi(z, \rho, \beta | y_{obs}) \propto \pi(\rho, \beta) p(z | \rho, \beta) \mathbb{I}_{z \in \mathcal{Z}_{y_{obs}}}, \tag{5}$$

where

$$\mathcal{Z}_{y_{obs}} = \{z \in R^n : \mathbb{I}y_i = \mathbb{I}_{z_i > 0} \text{ for each } i = 1, ..., n - q\},$$

so that the sign condition on $z = (z_1, ..., z_n)$ applies only to those z_i matched with a y_i that is actually observed. The z-values matched with unobserved y-values are informed through their neighbours in the ARM. The posterior predictive distribution for y_{mis} is simulated by simulating $z | y_{obs}$ from the distribution above and setting $z_{mis} = (z_{n-q+1}, ..., z_n)$ and $y_{mis,i} = \mathbb{I}_{z_{mis,i} > 0}$ for $i = 1, ..., q$. In

terms of the posterior in Equation 5, the posterior predictive is

$$P(Y_{mis,i} = 1|y_{obs}) = P(\mathbb{I}_{Z_{mis,i}>0} = 1|y_{obs}) \tag{6}$$

with

$$P(\mathbb{I}_{Z_{mis,i}>0} = 1|y_{obs}) = \int_{z:\mathbb{I}_{z_{mis,i}>0}=1} \pi(z, \rho, \beta|y_{obs}) dz d\beta d\rho.$$

We generate realisations from the marginal distribution $y_{mis}|y_{obs}$ by sampling the joint distribution $z, \rho, \beta \sim \pi(z, \rho, \beta|y_{obs})$ and setting $y_{mis} = \mathbb{I}_{z_{mis}>0}$.

In Bayesian inference for an ARM without missing data, parameter estimates are informed by the whole dataset. When there is missing data the investigator has the opportunity to control the flow of information from the imputed data back to parameters, and this leads to cut models, where parameters are estimated without feedback from imputed missing data. In a full Bayes analysis with missing data, parameters and missing data are coupled, and modelling decisions for missing data impact parameter estimates. If there is no or little model misspecification, the full Bayes approach is likely more effective compared to cut models as all the information available is reliable. Where there is model misspecification, cut models may be far more reliable.

2.3.2 Posterior simulation and estimation for missing data

The full Bayes posterior $(z, \beta, \rho) \sim \pi(z, \rho, \beta|y_{obs})$ is simulated using MCMC as outlined in Algorithm 1. We run Algorithm 1 to generate $(z^{(t)}, \beta^{(t)}, \rho^{(t)})_{t=1,...,T}$ distributed asymptotically in T according to $\pi(z, \beta, \rho|y_{obs})$. For $i \in \{q+1, ..., n\}$ let

$$y_i^{(t)} = \mathbb{I}_{z_i^{(t)}>0}. \tag{7}$$

We estimate the missing data using the marginal posterior mode,

$$\hat{y}_i = \text{mode}(\{y_i^{(t)},\ t = 1,...T\}). \tag{8}$$

To ensure an accurate posterior for ρ, a burn-in period of 1000 plus $T = 25000$ sweeps (where a sweep is one pass over all variables, equal to one loop of Algorithm 1) are used to simulate posterior distributions. The required number of sweeps was determined by targeting an effective sample size (see Table S1) in the thousands.

We use a mixture of Metropolis Hastings (ρ) and Gibbs (z and β) sampling [6, 25]. This is straightforward, but some details of the z-simulation in Algorithm 1 play a role in defining the cut model and it is helpful to be clear that y_{mis} plays no role in the MCMC itself.

2.3.3 Cut model

Cut models treat model misspecification by replacing full Bayesian inference (previous sections) with a form of multiple imputation. Suppose the entire ARM network model is misspecified. For example, suppose the value of ρ is set to an incorrect value. Estimates for parameters such as z_{obs}, which are tightly constrained by their data, may be relatively robust to model misspecification. However, z_{mis}-values are not tied to data and will settle at values consistent with each other, and the miss-specified model. In a full Bayesian setting, these poorly located latent variables feedback to distort z_{obs}, β and ρ estimates. In a cut model, we cut interactions between poorly informed variables z_{mis} and the core parameters z_{obs}, β and ρ. We determine an imputation posterior distribution for the core parameters alone using otherwise standard Bayesian methods. We then use this imputation posterior distribution as "data" to estimate z_{mis}, again, using standard Bayesian methods. This means z_{mis} are informed by

Algorithm 1 Bayesian ARM parameter estimation

MCMC targeting $\pi(z, \rho, \beta | y_{obs})$ in Equation 5.

Suppose at step $t \in \{0, 1, ..., T-1\}$ the current state of the Markov chain is $z^{(t)} = z, \beta^{(t)} = \beta$ and $\rho^{(t)} = \rho$. The state at step $t+1$ is determined in the following way. One update will be one cycle through each element of z, β and ρ.

1. Update $z|\beta, \rho, y_{obs}$: (A) For $i = 1, ..., n$ let $W_{i,:}$ denote the i'th row of W; (1) simulate a new z-value using

$$z_i' \sim N(W_{i,:}z + X_{i,:}\beta, 1 | y_i = \mathbb{I}_{z_i' > 0})$$

   if $i \leq n - q$ (note that $W_{i,i} = 0$ so the mean does not depend on z_i) and

$$z_i' \sim N(W_{i,:}z + X_{i,:}\beta, 1)$$

   if $i > n - q$ and then (B) set $z_i \leftarrow z_i'$ (ie, before moving onto the next i). Denote by z' the updated z-vector.

2. Update $\beta|z', \rho, y_{obs}$: the conditional probability density of β is normal, so simulate

$$\begin{aligned}
\beta' &\sim N(\mu_\beta^*, \Sigma_\beta^*) \\
\mu_\beta^* &= (X^T X + \Sigma_\beta^{-1})(X^T A_\rho z' + \Sigma_\beta^{-1}\mu_\beta) \\
\Sigma_\beta^* &= (X^T X + \Sigma_\beta^{-1})^{-1} \\
A_\rho &= (I_n - \rho W)
\end{aligned}$$

   In our case $\mu_\beta = 0$ and $\Sigma_\beta = \sigma_\beta^2 I_{p+1}$ with large σ_β^2, so these distributions simplify. Notice that μ_β^* is calculated using the new z'-values inherited from the z-update above. Denote by β' the updated β-vector.

3. update $\rho|z', \beta', y_{obs}$: the conditional density of ρ depends on ρ through $|A_\rho|$, so Gibbs sampling is infeasible. We use Metropolis Hastings with a simple random walk proposal

$$\tilde{\rho} = \rho + uR, \quad R \sim N(0, 1) \tag{9}$$

   where u is the tuning parameter, chosen by monitoring the acceptance rates for this step, and acceptance probability

$$\alpha(\tilde{\rho}|\rho) = \min\left\{ 1, \frac{\pi_\rho(\tilde{\rho}) p(z'|\beta', \tilde{\rho})}{\pi_\rho(\rho) p(z'|\beta', \rho)} \right\}.$$

   With probability α set $\rho' = \rho$ and otherwise set $\rho' = \tilde{\rho}$.

The new state is $z^{(t+1)} = z', \beta^{(t+1)} = \beta'$ and $\rho^{(t+1)} = \rho'$.

more reliable z_{obs}, β and ρ values.

Denote by $\pi_{cut}(z, \rho, \beta | y_{obs})$ the full distribution determined by the cut model. This will have the form

$$\pi_{cut}(z, \rho, \beta | y_{obs}) = p_{cut,mis}(z_{mis} | z_{obs}, \beta, \rho) \pi_{cut,obs}(z_{obs}, \rho, \beta | y_{obs}), \qquad z \in \mathcal{Z}_{y_{obs}}, \tag{10}$$

where $z = (z_{obs}, z_{mis})$ as above, and the distributions on the right hand side are defined below. In cut model MCMC, Algorithm 2, we use MCMC to simulate

$$(z_{obs}^{(t)}, \beta^{(t)}, \rho^{(t)}) \sim \pi_{cut,obs}(z_{obs}, \rho, \beta | y_{obs})$$

and then simulate

$$z_{mis}^{(t)} \sim p_{cut,mis}(z_{mis}^{(t)} | z_{obs}^{(t)}, \beta^{(t)}, \rho^{(t)}),$$

setting $z^{(t)} = (z_{obs}^{(t)}, z_{mis}^{(t)})$, for $t = 1, ..., T$. Estimation of $\hat{y}_{mis,i}$ and further analysis is then unchanged from the full Bayes case in Section 2.3.2.

We now give details for $p_{cut,mis}(z_{mis} | z_{obs}, \beta, \rho)$ and $\pi_{cut,obs}(z_{obs}, \rho, \beta | y_{obs})$. Group the model elements according to the way they are linked to observed or missing data, dividing the ARM equations into blocks corresponding to connections between observed pairs of nodes, missing pairs of nodes, and missing and observed pairs of nodes. The $n \times n$ network weight matrix W is given in terms of its blocks as

$$W = \begin{bmatrix} W_{[obs,obs]} & W_{[obs,mis]} \\ W_{[mis,obs]} & W_{[mis,mis]} \end{bmatrix} \tag{11}$$

where the blocks have dimension

$$\dim W = \begin{bmatrix} (n-q) \times (n-q) & (n-q) \times q \\ q \times (n-q) & q \times q \end{bmatrix}.$$

Let $\mathbb{O}$ be an $(n-q)\times q$ matrix of zeros. Define a new cut matrix W_{cut} by removing feedback from missing to observed,

$$W^{cut} = \begin{bmatrix} W_{[obs,obs]} & \mathbb{O} \\ W_{[mis,obs]} & W_{[mis,mis]} \end{bmatrix} \tag{12}$$

We block covariates similarly. Let

$$\mathrm{X} = \begin{bmatrix} X_{obs} \\ X_{mis} \end{bmatrix} \qquad \text{with dimensions: } \begin{bmatrix} (n-q)\times p \\ q \times p \end{bmatrix}, \tag{13}$$

Substituting W^{cut} for W in Equation 1 we have a new cut ARM,

$$z_{obs} = \rho W_{[obs,obs]} z_{obs} + X_{obs}\beta + \epsilon_{obs} \tag{14}$$

$$z_{mis} = \rho W_{[mis,obs]} z_{obs} + \rho W_{[mis,mis]} z_{mis} + X_{mis}\beta + \epsilon_{mis} \tag{15}$$

where $\beta = (\beta_1, ..., \beta_p)^T$, $\epsilon_{obs} \sim N(0, \sigma^2 I_{n-q})$ and $\epsilon_{mis} \sim N(0, \sigma^2 I_q)$. The cut distribution for the missing data is determined from Equation 4. Let $V = [X_{mis}, \rho W_{[mis,obs]}]$ (so V is a $q \times (n+1+p-q)$ matrix) and let $\theta = (\beta^T, z_{obs}^T)^T$ (a $(n+1+p-q)\times 1$ vector). Let $A_\rho^{(mis)} = I_q - \rho W_{[mis,mis]}$. The cut prediction distribution $p_{cut,mis}$ in Equation 10 is

$$p_{cut,mis}(z_{mis}|z_{obs}, \beta, \rho) \propto |A_\rho^{mis}| \exp\Big(-\frac{1}{2}(A_\rho^{mis} z_{obs} - V\theta)^T (A_\rho^{mis} z_{obs} - V\theta)\Big).$$

The cut posterior distribution $\pi_{cut,obs}$ on the RHS of Equation 10 is

$$\pi_{cut}(z_{obs}, \rho, \beta|y_{obs}) \propto \pi(\rho,\beta) p_{cut,obs}(z_{obs}|\rho,\beta) \mathbb{I}_{z_{obs} \in Z_{y_{obs},obs}},$$

where

$$\mathcal{Z}_{y_{obs},obs} = \{z_{obs} \in R^{n-q} : \mathbb{I}y_i = \mathbb{I}_{z_i>0} \text{ for each } i = 1,...,n-q\},$$

and likelihood from Equation 14,

$$p_{cut,obs}(z_{obs}|\rho,\beta) \propto |A_\rho^{obs}| \exp\Big(-\frac{1}{2}(A_\rho^{obs}z_{obs} - X_{obs}\beta)^T(A_\rho^{obs}z_{obs} - X_{obs}\beta)\Big),$$

where now $A_\rho^{obs} = I_{n-q} - \rho W_{[obs,obs]}$.

Cut models may be helpful with high missingness as even slight model miss-specification can bias full Bayes estimates badly. Cut models can be characterised as Bayesian multiple imputation. Multiple imputation has two stages, an imputation stage, in which multiple copies of the missing data are imputed, followed by an analysis stage, in which in which a model is fit to the imputed and observed data and parameters estimated. In our setting there is some flexibility in what we identify as missing data, and what we call a parameter. We use this flexibility to get robustness to model misspecification. Recall that y_{obs} is the observed data and β, ρ, z_{obs}, and z_{mis} and y_{mis}, are all unknown. Algorithm 2 does multiple imputation of "missing data" β, ρ, z_{obs} followed by estimation of the parameters z_{mis} and y_{mis}. When there is no model misspecification this cut model is consistent for β and ρ estimation, like full Bayes. However in that (well-specified) case, cut models tend to give estimates with less precision, as (desirable) information spread through the network via missing data is lost.

2.3.4 Experiments

From the original fully observed attribute data we created two types of missing data scenario's: Missing Completely at Random (MCAR) and Missing at Random (MAR). In MCAR, the missingness property is unrelated to the missing

Algorithm 2 Cut model ARM parameter estimation

MCMC targeting $\pi_{cut}(z, \rho, \beta | y_{obs})$ in Equation 10.

Suppose at step $t \in \{0, 1, ..., T-1\}$ the current state of the Markov chain is $z_{obs}^{(t)} = z_{obs}, z_{mis}^{(t)} = z_{mis}, \beta^{(t)} = \beta$ and $\rho^{(t)} = \rho$. The state at step $t+1$ is determined in the following way. One update will be one cycle through each element of z, β and ρ.

1. Update $z_{obs} | \beta, \rho, y_{obs}$: (A) For $i = 1, ..., n-q$ let $W_{i,:}^{cut}$ denote the i'th row of W^{cut}; (A) simulate a new z-value using

$$z'_{obs,i} \sim N(W_{i,:}^{cut} z_{obs} + X_{obs,i,:}\beta, 1 | y_{obs,i} = \mathbb{I}_{z'_{obs,i} > 0}).$$

   and then (B) set $z_{obs,i} \leftarrow z'_{obs,i}$ (ie, before moving onto the next i). Denote by z'_{obs} the updated z-vector.

2. Update $\beta | z'_{obs}, \rho, y_{obs}$: simulate

$$\begin{aligned}
\beta' &\sim N(\mu_\beta^*, \Sigma_\beta^*) \\
\mu_\beta^* &= (X_{obs}^T X_{obs} + \Sigma_\beta^{-1})(X_{obs}^T A_\rho^{obs} z'_{obs} + \Sigma_\beta^{-1} \mu_\beta) \\
\Sigma_\beta^* &= (X_{obs}^T X_{obs} + \Sigma_\beta^{-1})^{-1} \\
A_\rho &= (I_{n-q} - \rho W_{[obs,obs]})
\end{aligned}$$

   Denote by β' the updated β-vector.

3. update $\rho | z', \beta', y_{obs}$: We use Metropolis Hastings with a simple random walk proposal

$$\tilde{\rho} = \rho + uR, \quad R \sim N(0, 1) \tag{16}$$

   where u is the tuning parameter, chosen by monitoring the acceptance rates for this step, and acceptance probability

$$\alpha(\tilde{\rho} | \rho) = \min \left\{ 1, \frac{\pi_\rho(\tilde{\rho}) p_{cut}(z'_{obs} | \beta', \tilde{\rho})}{\pi_\rho(\rho) p_{cut}(z'_{obs} | \beta', \rho)} \right\}.$$

   With probability α set $\rho' = \rho$ and otherwise set $\rho' = \tilde{\rho}$.

The new state is $z_{obs}^{(t+1)} = z'_{obs}, \beta^{(t+1)} = \beta'$ and $\rho^{(t+1)} = \rho'$.

4. Update $z_{mis} | z_{obs}^{(t+1)} \beta^{(t+1)}, \rho^{(t+1)}$: Simulate $\epsilon_{mis}^{(t+1)} \sim N(0, \sigma^2 I_q)$ and set

$$z_{mis}^{(t+1)} \leftarrow \rho W_{[mis,obs]} z_{obs}^{(t+1)} + \rho^{(t+1)} W_{[mis,mis]} z_{mis}^{(t)} + X_{mis}\beta^{(t+1)} + \epsilon_{mis}^{(t+1)}$$

   Note that since Steps 1-3 do not depend on $z_{mis}^{(t)}$, step 4 can be implemented in post-processing on the MCMC-output chain.

value itself or other attribute data. In the MAR data, the probability of being missing is the same only within groups defined by the observed data. On a network, missingness may be correlated by the network in the same way as any other node attribute. If there is a network effect on gender, there is may well be a network effect on missing gender.

In our experiments, four scenarios with 10%, 25%, 50% and 75% missing gender values (this is $q = 190, 475, 950, 1424$ missing node gender values out of $n = 1899$ in all) were created and compared with a baseline analysis of the fully observed data. For the MCAR setting, the individuals selected for imputation were randomly selected with the *sample* function in R. To mimic MAR, snowball sampling was used. We chose the percentage missing node values in the snowball-sampling so that the number of "informative edges" in the snowball sampling matched the number of informative edges in the corresponding random/MCAR sampling. An edge is "informative" if both the two gender node-values adjacent the edge are not missing data. We refer to non-informative edges as "missing". We used the number of missing edges as a rough measure of the amount of network information in the data. Operationally, we select m seed nodes and their direct neighbours (using *LSMI* from the *snowball* R-package) and remove their gender data. The number of nodes q removed in our MAR setup was determined by removing data at seed nodes and their neighbours until the target number of missing edges was reached. A smaller number of snowball-sampled nodes gives the same number of missing edges as a larger number of random/MCAR nodes. The figures are reported in Table S2). For example, removing data on 75% of nodes chosen completely at random gives the same number of missing edges as removing data on 37.8% of nodes chosen by snowball sampling (in one realisation of the missing-data snowball process). For further discussion of the Snowball/MAR missing-data

process see Section 5.3.

We analyse each of the four MCAR missing-data sets and each of the four MAR data sets twice, first using the (standard) full Bayes machinery of Section 2.3.2 and then second using the cut model setup of Section 2.3.3. This leads to four data/inference pairs of MCAR analyses and four pairs of MAR analyses. There is also a single baseline Bayesian analysis made with no missing data and the same observation model and priors common to all analyses.

2.3.5 Performance evaluation

Parameter estimates (always posterior mean unless indicated) obtained using a full Bayes analysis on the complete data can for our purpose be treated as the truth, since we are interested in methods which continue to recover the full-data parameter values and predict missing data well as the percentage of missing data becomes large. Different fitting procedures were evaluated by comparing parameter estimates and standard deviations. A method is successful (on this first criterion) if parameter estimates do not change significantly as we increase the proportion of missing data. We will see that the full Bayes analysis fails very badly on this score (due to model misspecification) but a cut model approach is much more reliable, out to even very large proportions of missing data.

Our second criterion is predictive performance on withheld data. Since we generate test missing data by withholding completely random- and snowball-sampled data, the withheld data for performance evaluation is the missing data for the original analysis. We run MCMC for each data/inference pair and use the sampled parameters to estimate parameters using the posterior mean for $\hat{\beta}$ and $\hat{\rho}$ and the posterior mode for $\hat{y}_{mis}$ (ie, Equations 7 and 8). We report the percentage of misclassified missing observations, $\sum_i \mathbb{I}(\hat{y}_{mis,i} \neq y_{true,i})/q$ not equal to its true withheld value, $y_{true,i}$ say, for each data/inference pair.

A good inference method should be well calibrated, that is $E(Y_{mis,i}|\hat{p}_{mis,i}) = \hat{p}_{mis,i}$, so predictions have the correct level of confidence. The Brier score is sensitive to calibration (and other things, see [38, 43]). The Brier Score B is given by

$$B = \frac{1}{q}\sum_{i=1}^{q}(\hat{p}_{mis,i} - y_{true,i})^2,$$

where $\hat{p}_{mis,i} = \sum_{t=1}^{T} y_{mis,i}^{(t)}/T$ is our Monte Carlo estimate of $P(Y_{mis,i} = 1|y_{obs})$ in Equation 6, and $y_{true,i}$ is the true value of the missing (in fact withheld) data. Smaller values of B indicate better-calibrated prediction. The misclassification rate and Brier score take values between 0 and 1. For reference, the ratio of males to females is approximately 6 : 4 in these data, so ignoring network data, taking $\hat{p}_{mis,i} \simeq 0.6$ and simply assigning values to missing data independently at random in these proportions gives (approximately) a misclassification rate of 0.48 and a Brier score of 0.24. This procedure is actually perfectly calibrated (but lacking in resolution) so 0.24 should be thought of as a fair score.

2.4 Model-free network-based prediction method

Given that model-misspecification is at the root of the difference between our cut model and full Bayes analysis, it is of interest to see how a straightforward model-free method performs. We tried a number of methods which we do not report as they all gave poor performance. We report a K-nearest-neighbour scheme which also gives poor performance, but seemed at least a priori a sensible attempt.

For each node $i = n-q+1,...,n$ with missing gender, we have covariates X_i. For each $j = 1,...n-q$ corresponding to an observed node, let $D_{i,j} = |X_i - X_j|$ be the Euclidean covariate distance. We took the K-nearest neighbours of i in this covariate distance and predicted the value of $y_{mis,i}$ using the majority gender

in this K-nearest-neighbour set. The value of K was chosen by applying the method to the fully observed part of the data and choosing K to minimise the misclassification rate on that data.

3 Results

We illustrate our methods by recovering missing gender data (recall y_i is this binary gender variable). We begin with a brief summary of gender marked imbalance in the data. Table 1 describes covariates across gender. Females and males are equally popular, but females have a significantly higher in-degree (Mean = 12.47, SD = 16.15), indicating they receive more messages then males (Mean = 9.45, SD = 14.54). Males were more likely in a higher study year (Mean = 2.5, SD = 1.37) compared to females (Mean = 2.24, SD = 1.19). We selected the significant variables (indegree, year of study, and day active) as predictors for the imputation.

Table 1: Covariates descriptives for Males (N = 1118) and Females (N = 781)

Covariate	$mean_M$ (sd)	$mean_F$ (sd)	$t(df)$	p
Outdegree	10.92 (23.79)	10.36 (18.80)	.571 (1868.3)	.568
Indegree	9.45 (14.54)	12.47 (16.15)	-4.178 (1563.2)	<.001
Popularity	20.36 (37.01)	22.82 (33.47)	-1.509 (1776.4)	.131
Year of study	2.50 (1.37)	2.24 (1.19)	4.36 (1805.8)	<.001
Average characters	68.09 (88.16)	63.50 (75.72)	1.215 (1818.1)	.224
Day user became active	30.43 (29.22)	37.54 (35.99)	-4.568 (1449)	<.001
Day user contacted 75% of his/her friends	42.67 (33.74)	48.66 (38.21)	-3.526 (1541.7)	<.001

Descriptives obtained from the Facebook data used in this study, where the mean values are compared between males and females. Columns are independent-sample T-test statistics (t) comparing means with p value $< .05$ indicating a significant effect, given a standard deviation (sd) and significance threshold depending on the stated degrees of freedom (df).

In Tables 2 and 3 we present the main results of our fitting data with random/MCAR and snowball/MAR missingness respectively. MCMC conver-

gence was checked. Specimen traces, presented in the Supplement in Section 5.1, showed negligible burn-in and very good mixing. Effective sample sizes, reported in Section 5.2, are all over 10000. The results in the tables are representative. We replicated the parameter estimation results in two different missing-data subsets of the same size (see Supplementary Material Tables S4, S5) so we can be confident the results we present are representative, and not an artifact of one specific realisation of the missing data process.

To sum up the results briefly, parameter estimates in Tables 2 and 3 are far more stable (that is, they match parameter values in the complete-data analysis) when we use the cut model. Parameter estimates remain approximately constant across rows of the cut model analysis (top half of each table) while they shrink towards zero as we scan across columns in the full Bayes analysis (bottom half of each table). This is what we would expect in a misspecified setting. The cut model protects parameter estimates from distortion due to model misspecification in the missing data. These improved parameter estimates then give better prediction when applied to the missing data. Interestingly, the key network parameter, ρ, is negative and significant: the residual network effect on gender, after accounting for our covariates, is anti-correlated. The significance of this effect was lost in the full Bayes analysis at high levels of missingness, but detected by the cut model.

Turning to our other criteria, the cut model has significantly smaller misclassification rate when the missingness is MCAR (Table 2) but only roughly equal misclassification rate when missingness is MAR (Table 3). The Brier scores are similar. Further investigation showed that at the highest levels of missingness, full Bayes is essentially predicting the missing binary gender data using the constant gender ratio, as there is little other information left in the data.

We repeated the analysis using snowball-sampling without edge-correction. This led to data with almost no informative edges. The levels of missingness are extreme, and we think network analysis is no-longer sensible. We present these results in Section 5.4) for completeness.

The model-free comparison lead to an optimal value of $K = 21$, so we assign each missing gender value by taking the modal gender of the 21 nearest gendered individuals. We tried this approach on the random/MCAR data with 10% missingness. The misclassification rate was 0.39, to be compared with the values 0.38 (cut-model) and 0.42 (Bayes) taken from Table 2.

Our KNN method could be improved, for example by weighting covariates in the distance measure. However, our parameteric models give parameter estimates which are useful for interpretation, and absent in a model free approach, and they could also be improved. Our purpose here is to show that the although the network based parametric model is misspecified, inference outcomes can be improved by changing the inference procedure, and not necessarily improving the model.

4 Discussion

The aim of this paper was to present a misspecification-robust imputation procedure for attributes in networked data via autocorrelation regression models. We presented a cut model, where there is no feedback from the imputed data to parameter estimation, and a full Bayes approach, where feedback exists. These models were applied to multiple scenarios with increasing missingness.

When snowball/MAR sampling was used to remove the same number of observations, the amount of edges (or friendship connections) lost was significantly different, following a convincingly negative exponential distribution. We have shown that model parameters are diversely affected with different

Table 2: Comparison of posterior mean parameter estimates for *Gender* under a cut and full Bayes imputation model with random missingness/MCAR-sampling.

	0% missing	10% missing	25% missing	50% missing	75% missing
Parameters		CM	CM	CM	CM
Intercept	-.591(.073)	-.599(.075)	-.566(.082)	-.564(.098)	-.461(.133)
Indegree	.010(.002)	.011(.002)	.012(.002)	-.014(.003)	.013(.004)
Year of study 2	-.031(.079)	-.007(.082)	-.042(.090)	-.059(.111)	-.135(.157)
Year of study 3	-.043(.086)	-.021(.092)	-.095(.102)	-.050(.124)	.048(.170)
Year of study 4	-.087(.098)	-.044(.103)	-.098(.111)	-.077(.134)	-.152(.189)
Year of study 5	-.226(.155)	-.208(.160)	-.383(.188)	-.373(.229)	-.393(.296)
Year of study 6	-.754(.261)	-.730(.270)	-1.399(.443)	1.235(.461)	-3.737(2.611)
Day active	.005(.001)	-.005(.001)	.005(.001)	.006(.001)	.004(.002)
ρ	-.507(.045)	-.517(.050)	-.530(.059)	-.576(.090)	-.583(.143)
Misclassification rate	.331	.381	.332	.285	.220
Brier score		.267	.255	.262	.264
Parameters		FBM	FBM	FBM	FBM
Intercept		-.579(.072)	-.426(.069)	-.273(.066)	-.099(.064)
Indegree		.009(.002)	.006(.002)	.007(.002)	.003(.002)
Year of study 2		.040(.078)	-.040(.077)	-.084(.077)	-.066(.075)
Year of study 3		.035(.086)	-.078(.085)	-.058(.083)	.031(.081)
Year of study 4		.037(.098)	-.026(.095)	-.039(.093)	-.095(.092)
Year of study 5		-.148(.155)	-.265(.153)	-.405(.153)	-.075(.145)
Year of study 6		-.681(.252)	-.648(.236)	-.696(.227)	-.455(.209)
Day active		.005(.001)	.005(.001)	.004(.001)	.001(.001)
ρ		-.416(.047)	-.323(.047)	-.171(.047)	-.036(.046)
Misclassification rate		.416	.399	.418	.408
Brier score		.259	.243	.246	.247

Estimates based on 25000 MCMC steps with burn-in 1000, flat prior for β, CM = Cut Model, FBM = Full Bayes Model and "Bayes" is simply the full analysis of all data. Covariate "Year of Study" is treated as a categorical variable with first-year baseline. misclassification rate of 0% missing data is computed using leave-one-out prediction.

Table 3: Comparison of posterior mean parameter estimates for *Gender* under a cut and full Bayes imputation model with snowball/MAR sampling based missingness, conditioned on the number of missing edges.

	0% missing	2% missing	9% missing	19% missing	38% missing
Parameters		CM	CM	CM	CM
Intercept	-.591(.073)	-.599(.073)	-.586(.073)	-.557(.076)	-.650(.088)
Indegree	.010(.002)	.010(.002)	.007(.003)	.011(.004)	.001(.006)
Year of study 2	-.031(.079)	-.023(.079)	-.041(.082)	-.049(.086)	-.051(.098)
Year of study 3	-.043(.086)	-.055(.088)	-.063(.091)	-.065(.094)	-.049(.109)
Year of study 4	-.087(.098)	-.127(.098)	-.142(.101)	-.146(.106)	-.066(.121)
Year of study 5	-.226(.155)	-.217(.157)	-.229(.158)	-.170(.171)	-.256(.189)
Year of study 6	-.754(.261)	-.898(.284)	-.952(.287)	-.991(.312)	-.905(.3970)
Day active	.005(.001)	.006(.001)	.005(.001)	.006(.001)	.007(.001)
ρ	-.507(.045)	-.515(.048)	-.493(.056)	-.367(.069)	-.178(.092)
misclassification rate	.331	.500	.567	.434	.528
Brier score		.309	.294	.265	.304
Parameters		FBM	FBM	FBM	FBM
Intercept		-.546(.071)	-.485(.069)	-.502(.067)	-.483(.066)
Indegree		.008(.002)	.007(.002)	.009(.002)	.008(.002)
Year of study 2		-.002(.078)	-.050(.077)	.014(.077)	-.072(.077)
Year of study 3		-.027(.086)	-.092(.085)	.017(.084)	-.059(.084)
Year of study 4		-.111(.098)	-.184(.097)	-.098(.095)	.003(.094)
Year of study 5		-.214(.155)	-.199(.152)	-.030(.150)	-.137(.150)
Year of study 6		-.692(.249)	-.743(.247)	-.535(.229)	-.266(.214)
Day active		.005(.001)	.005(.001)	.005(.001)	.005(.001)
ρ		-.443(.051)	-.306(.058)	-.183(.056)	-.097(.054)
misclassification rate		.529	.561	.439	.511
Brier score		.305	.283	.260	.268

Estimates based on 25000 draws and 1000 burn-in, m = 10, flat prior for β, CM = cut model, FBM = Full Bayes model. Percentage missing nodes do not match random/MCAR levels as percentage missing edges were matched. See Table S2

types of missingness and the network element exacerbates the consequences of missing data, especially when entire clusters drop out. Notably, the combination of increased missing data and model misspecification deteriorated imputation performance. Both methods (i.e. cut model and full Bayes) fail at prediction when the missingness falls in clusters, as in the snowball-sampled case. However cut model parameter estimates remain stable even in this case, where full Bayes estimates shrink to zero and loose significance.

If there were no model misspecification (for example, if we carried out this analysis on synthetic data, with parameters sampled from the prior, and data from the observation model) then straightforward Bayesian inference will work. Moreover, because the cut model is discarding information in cutting feedback from missing data, its returns parameter estimates with greater associated variance. In this case straightforward Bayesian inference will give correctly calibrated answers with greater confidence and precision in the well-specified setting. This lower-variance aspect is already visible in Tables 2 and 3, in the misspecified case, where we see the Bayes estimates have associated errors which are slightly, but uniformly, smaller than the corresponding cut model errors.

We shortly discussed a naive model-free but network-informed alternative with a misclassification rate between the presented models. Matching procedures are less robust compared to models where stable parameters are obtained from which multiple observations can be imputed. Yet, given the quality of spectral clustering literature, the outcome of our experiment postulates machine learning alternatives become competitive with existing methods. Perhaps dimensionality reduction and clustering approaches, recently claimed to outperform multiple imputation [15], can be useful in networked data, but other new methods employing "learning rates" ([10]) might also be consid-

ered. One open problem in network based predictive mean matching is proper donor selection. Our proposal was to combine network structure with covariate distance, but more work is required to test different scenario's, covariates types, and network structures, specifically if small clusters exist.

This study assumed fully observed tie-structure in the network. Edge weights depended on collapsing data from 196 days, which seemed a solid solution to provide an indication whether a relationship existed (compared to cross-sectional designs) but does not completely rule out that other edges may come to existence (or break) in the future. A potential misrepresentation of the network by missing edges could result in an over-or underestimation of ρ, depending on whether covariates diverse or converge with edges. Several studies have suggested imputation-like models to complete tie structure in networks (e.g. [19]), and one strategy in future studies could be to first complete the tie-structure, towards optimal estimation of ρ.

Another issue in Bayesian imputation of categorical y is unbalance, which changes the shape of the distribution of latent proxy z. Although z is drawn from a truncated normal, its shape fully depends on the distribution of y. The gender variable used in this study was not severely imbalanced yet severe imbalance may lead to problems in parameter estimation if z is highly skewed. non-linearity could be dealt with by applying non-linear Bayesian regression, but doing this in the context of imputation with ARMs is yet uncharted.

The main conclusion from our model comparison is as follows: researchers faced with missing data are advised to use the cut model, which is likely more able to retain accurate posterior distributions, especially when the pattern of missingness correlates with network edges and with high proportions of missing data > 25%. Parameter estimates in the full Bayes approach suffer from model misspecification and strongly depend on the starting values given to

each observation. Perhaps more study into different ways to determine starting values combined with full Bayes in missing data can make full Bayes more applicable in this context. The outcome that cut models outperform full Bayes is convenient as current imputation methods in independent observations are abundantly cut models [4]. As mentioned, data completeness and veracity are a major issue for any analyst, especially in the cyber domain. There are numerous cases where online identities are copied, faked, or profiles use false information to misguide other users (e.g. online grooming). By applying ARM imputation, attribute data of all observations in a network can be completed by incorporating all the observed information. When attribute data is observed, the model may be useful in analyzing residuals acquired by comparing imputation output to observed data, as a form of robust anomaly detection.

References

[1] James H Albert and Siddhartha Chib, *Bayesian analysis of binary and polychotomous response data*, Journal of the American Statistical Association **88** (1993), no. 422, 669–679.

[2] Luc Anselin, *Spatial econometrics: methods and models*, Vol. 4, Springer Science & Business Media, 2013.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: theory and experiment **2008** (2008), no. 10, P10008.

[4] S van Buuren and Karin Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in r*, Journal of statistical software (2010), 1–68.

[5] Kayla de la Haye, Joshua Embree, Marc Punkay, Dorothy L Espelage, Joan S Tucker, and Harold D Green Jr, *Analytic strategies for longitudinal networks with missing data*, Social networks **50** (2017), 17–25.

[6] Dino Dittrich, Roger Th AJ Leenders, and Joris Mulder, *Bayesian estimation of the network autocorrelation model*, Social Networks **48** (2017), 213–236.

[7] Ove Frank and David Strauss, *Markov graphs*, Journal of the American Statistical Association **81** (1986), no. 395, 832–842.

[8] Andrew Gelman, *Parameterization and bayesian modeling*, Journal of the American Statistical Association **99** (2004), no. 466, 537–545.

[9] Krista J Gile and Mark S Handcock, *Analysis of networks with missing data with application to the national longitudinal study of adolescent health*, Journal of the Royal Statistical Society: Series C (Applied Statistics) **66** (2017), no. 3, 501–519.

[10] Peter Grünwald, Thijs Van Ommen, et al., *Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it*, Bayesian Analysis **12** (2017), no. 4, 1069–1103.

[11] MS Hancock and K Gile, *Modeling social networks with sampled or missing data*, Seattle: Center for Statistics and the Social Sciences, University of Washington (2007).

[12] Ofer Harel and Xiao-Hua Zhou, *Multiple imputation: review of theory, implementation and software*, Statistics in medicine **26** (2007), no. 16, 3057–3077.

[13] Leslie W Hepple, *Bayesian techniques in spatial and network econometrics: 2. computational methods and algorithms*, Environment and Planning A **27** (1995), no. 4, 615–644.

[14] John R Hipp, Cheng Wang, Carter T Butts, Rupa Jose, and Cynthia M Lakon, *Research note: The consequences of different methods for handling missing network data in stochastic actor based models*, Social networks **41** (2015), 56–71.

[15] Domonique W Hodge, Sandra E Safo, and Qi Long, *Multiple imputation using dimension reduction techniques for high-dimensional data*, arXiv preprint arXiv:1905.05274 (2019).

[16] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, *Stochastic blockmodels: First steps*, Social Networks **5** (1983), no. 2, 109–137.

[17] Paul W Holland and Samuel Leinhardt, *An exponential family of probability distributions for directed graphs*, Journal of the American Statistical Association **76** (1981), no. 373, 33–50.

[18] Garth Holloway, Bhavani Shankar, and Sanzidur Rahmanb, *Bayesian spatial probit estimation: a primer and an application to hyv rice adoption*, Agricultural Economics **27** (2002), no. 3, 383–402.

[19] Mark Huisman, *Imputation of missing network data: some simple procedures*, Journal of Social Structure **10** (2009), no. 1, 1–29.

[20] Mark Huisman and Christian Steglich, *Treatment of non-response in longitudinal network studies*, Social networks **30** (2008), no. 4, 297–308.

[21] Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert, *Better together? statistical learning in models made of modules*, arXiv preprint arXiv:1708.08719 (2017).

[22] Johan H Koskinen, Garry L Robins, Peng Wang, and Philippa E Pattison, *Bayesian analysis for partially observed network data, missing ties, attributes and actors*, Social Networks **35** (2013), no. 4, 514–527.

[23] Gueorgi Kossinets, *Effects of missing data in social networks*, Social Networks **28** (2006), no. 3, 247–268.

[24] James LeSage, *Applied econometrics using matlab*, Manuscript, Dept. of Economics, University of Toronto (1999), 154–159.

[25] James P LeSage, *Bayesian estimation of spatial autoregressive models*, International Regional Science Review **20** (1997), no. 1-2, 113–129.

[26] James P LeSage, R Kelley Pace, Nina Lam, Richard Campanella, and Xingjian Liu, *New orleans business recovery in the aftermath of hurricane katrina*, Journal of the Royal Statistical Society: Series A (Statistics in Society) **174** (2011), no. 4, 1007–1027.

[27] James P LeSage and R K Pace, *Introduction to spacial econometrics*, New York: Taylor & Francis Group: Chapman & Hall, 2009.

[28] Roderick JA Little and Donald B Rubin, *Statistical analysis with missing data*, Vol. 793, John Wiley & Sons, 2019.

[29] David Lunn, Nicky Best, David Spiegelhalter, Gordon Graham, and Beat Neuenschwander, *Combining mcmc with âsequentialâpkpd modelling*, Journal of Pharmacokinetics and Pharmacodynamics **36** (2009), no. 1, 19.

[30] Davide Martinetti and Ghislain Geniaux, *Probitspatial r package: Fast and accurate spatial probit estimations*, 22. international conference on computational statistics (compstat), 2016, pp. np.

[31] Mark S Mizruchi and Eric J Neuman, *The effect of density on the level of bias in the network autocorrelation model*, Social Networks **30** (2008), no. 3, 190–200.

[32] Eric J Neuman and Mark S Mizruchi, *Structure and bias in the network autocorrelation model*, Social Networks **32** (2010), no. 4, 290–300.

[33] Keith Ord, *Estimation methods for models of spatial interaction*, Journal of the American Statistical Association **70** (1975), no. 349, 120–126.

[34] R Kelley Pace and Ronald Barry, *Quick computation of spatial autoregressive estimators*, Geographical Analysis **29** (1997), no. 3, 232–247.

[35] R Kelley Pace and James P LeSage, *Chebyshev approximation of log-determinants of spatial weight matrices*, Computational Statistics & Data Analysis **45** (2004), no. 2, 179–196.

[36] Martyn Plummer, *Cuts in bayesian graphical models*, Statistics and Computing **25** (2015), no. 1, 37–43.

[37] Panteha Hayati Rezvan, Katherine J Lee, and Julie A Simpson, *The rise of multiple imputation: a review of the reporting and implementation of the method in medical research*, BMC medical research methodology **15** (2015), no. 1, 30.

[38] Frederick Sanders, *On subjective probability forecasting*, Journal of Applied Meteorology **2** (1963), no. 2, 191–201.

[39] Joseph L Schafer and John W Graham, *Missing data: our view of the state of the art.*, Psychological Methods **7** (2002), no. 2, 147.

[40] Tony E Smith, *Estimation bias in spatial models with strongly connected weight matrices*, Geographical Analysis **41** (2009), no. 3, 307–332.

[41] Tony E Smith and James P LeSage, *A bayesian probit model with spatial dependencies*, Spatial and spatiotemporal econometrics, 2004, pp. 127–160.

[42] Tom AB Snijders, *Statistical models for social networks*, Annual Review of Sociology **37** (2011).

[43] David B Stephenson, Caio AS Coelho, and Ian T Jolliffe, *Two extra components in the brier score decomposition*, Weather and Forecasting **23** (2008), no. 4, 752–757.

[44] Stef Van Buuren, *Flexible imputation of missing data*, CRC press, 2012.

[45] Dootika Vats, James M Flegal, and Galin L Jones, *Multivariate output analysis for markov chain monte carlo*, arXiv preprint arXiv:1512.07713 (2015).

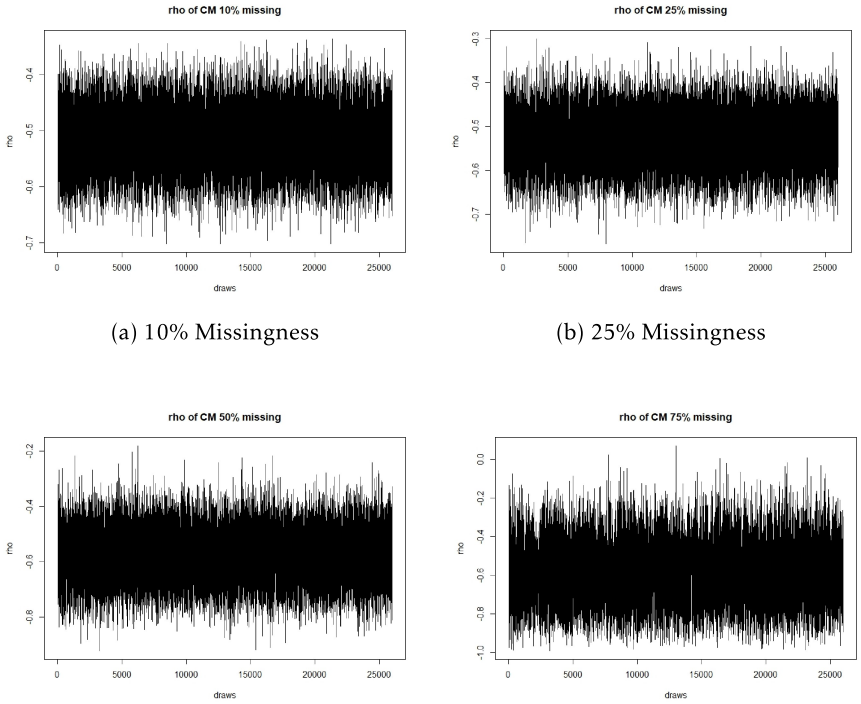
[46] Stanley Wasserman and Philippa Pattison, *Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp*, Psychometrika **61** (1996), no. 3, 401–425.

[47] Stefan Wilhelm and Miguel Godinho de Matos, *Estimating spatial probit models in r.*, R Journal **5** (2013), no. 1.

[48] Aaron Zimmerman, Tyler McCormick, Ali Shojaie, and Hedwig Lee, *Improving attribute prediction through network-augmented attribute prediction* (2015).

5 Supplementary Material

5.1 MCMC output traces for selected parameters



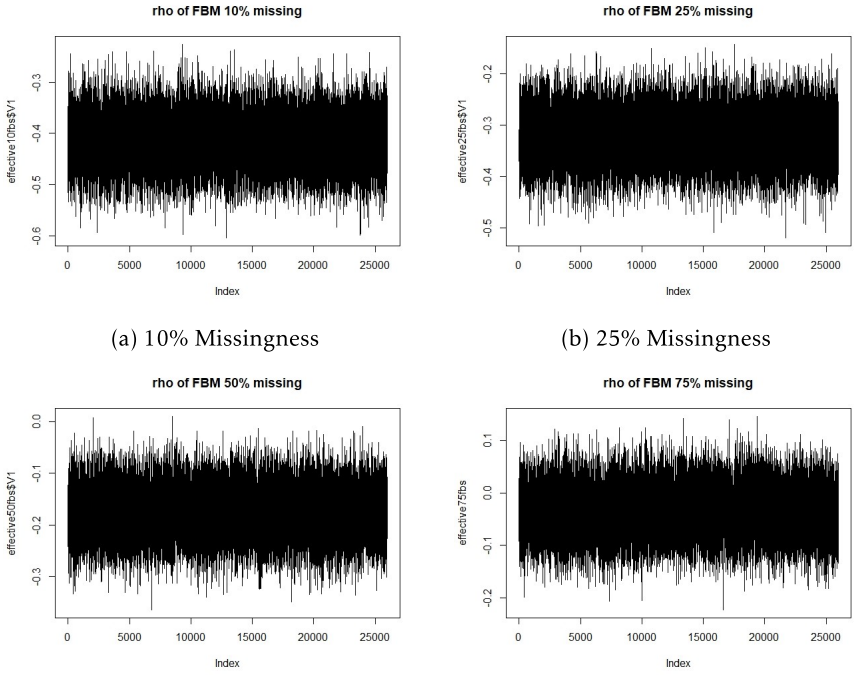
(a) 10% Missingness

(b) 25% Missingness

(c) 50% Missingness

(d) 75% Missingness

Figure S1: Network parameter MCMC traces from the random-missingness/MCAR sampling procedure in the cut model; burn in period of 1000 draws followed by 25000 draws.



(a) 10% Missingness

(b) 25% Missingness

(c) 50% Missingness

(d) 75% Missingness

Figure S2: Network parameter MCMC traces from the random-missingness/MCAR sampling procedure in the full Bayes model; burn in period of 1000 draws followed by 25000 draws.

5.2 Effective sample size of ρ in the snowball/MAR sampling scenario.

In this section we give the effective sample size (ESS) for the slowest mixing parameter, ρ, in the worst missing-data process (Snowball, with no edge matching). This is example is representative. The ESS values of the β-parameters were similar or better. To ensure robustness of parameter estimation, we used *MultiESS* from the *mcmcse* package to estimate the effective sample and observed (see Table S1) that the number of missing observations influenced the effective sample size of our Markov Chain [45].

Table S1: Effective sample size of ρ in the snowball/MAR sampling scenario

%missing	Cut model	full Bayes
10%	21352.32	23008.53
25%	18658.36	24926.45
50%	26007.73	21691.17
75%	20346.38	24804.03

This table presents the effective sample size estimates of network parameter ρ, where the ρ estimates from the fully observed model are compared with the ρ estimates from the missing data.

5.3 Snowball sampling with edge conditioning

We give some further details of the Snowball/MAR sampling scheme, and in particular the edge-matching. The use of different sampling techniques to select nodes for the imputation analyses influenced model estimation. Snowball sampling tends to remove data from well-connected actors. This leads to large numbers of missing (ie non-informative) edges. When data are missing completely at random over the network, network information is retained even at very high levels of missingness, as much as 75%. A straightforward application of snowball sampling (next section of supplement) at a fixed percentage missing node values leads to extreme low levels of informative edges, so that little network information remains and there is little point in making a network analysis. The correspondence is shown in Table S2 and in Figure S7. In our Snowball/MAR missing data process we therefore match a fixed percentage missing edges. The data are missing in clusters in contrast to the random scatter generated by the random/MCAR missing-data process.

Table S2: Amount of remaining edges between different sampling techniques

%nodes missing	random	snowball	edge matched snowball (%nodes)
10%	22220	13708	23072 (1.80)
25%	15002	4640	15296 (8.60)
50%	6946	590	6944 (19.2)
75%	1814	74	1802 (37.8)

This Table gives the number of remaining edges under different sampling methods. W starts complete with 27676 edges and we introduce 10, 25, 50, and 75% missingness by removing nodes. The last column presents the amount of remaining edges if matched on edges-amount, and the percentage of missing nodes that scenario corresponds to. For example, in the 50% nodes missing scenario, random sampling left 6946 edges, and if snowball sampling to prune a similar amount of edges, this results in 19.2% missing nodes.

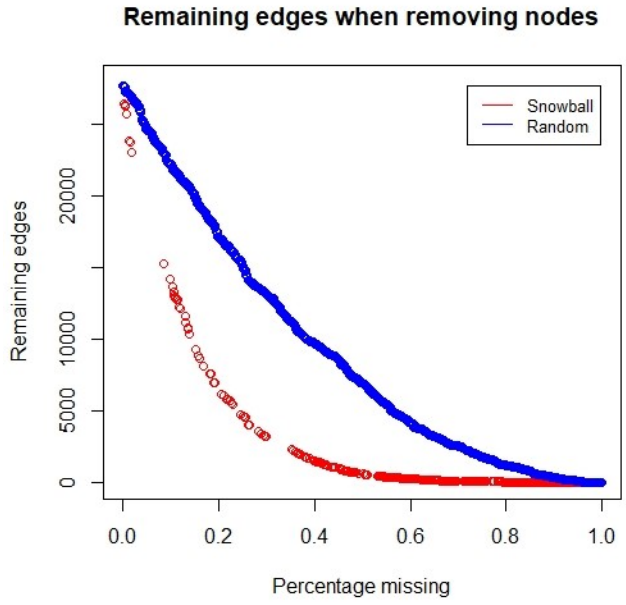


Figure S7: This Figure shows the number of informative edges when using different sampling methods to select nodes to remove attribute data. When selecting nodes randomly, the number of edges lost drops relatively gently, while snowball sampling drops more rapidly. Beyond about 60% node-missingness the number of informative edges in snowball sampling is very small as only isolated nodes remain.

5.4 Snowball sampling without matching missing-edge counts

In snowball sampling with out matching missing-edges, we matched missingness levels of gender values in the MCAR setup approximately. We adjusted the number of seeds ($m = 20, 44, 159, and 457$) to create four datasets with $q = 196, 401, 952$, and $q = 1424$ persons with a missing value in y. This roughly matched the percentage missing nodes (10,25,50,75) in the MCAR analysis. However, since data on better-connected individuals are more likely to be removed by snowball sampling than data on individuals with fewer connections, this quickly leads to a scenario with insufficient data for interesting or meaningful analysis, so in the main paper we report results for scenarios where we matched the number of edges with missing node data at both ends in MAR and MCAR.

Model estimates are provided in Table S3. This table can be thought of as an extension of the informative-edge-matched Table 3 adding columns on the right side of the table at higher levels of missingness. Both methods are predicting very poorly as network based inference becomes irrelevant. The cut model picks up the significant negative network parameter $\rho < 0$ out to the greatest levels of missingness.

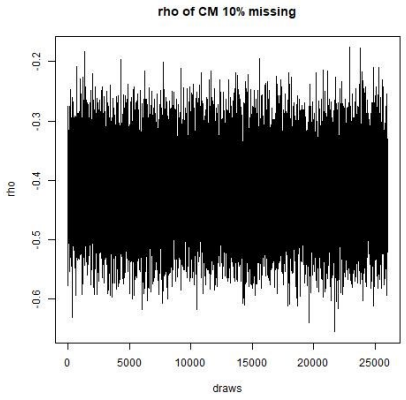
5.5 Replication outcomes

For random/MCAR missingness and snowball/MAR sampling we replicated the parameter estimation results by using different seeds as input for the sampling methods to select which nodes had their data removed (which nodes selected for imputation).

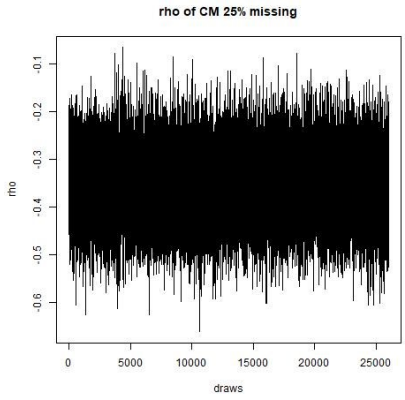
Table S3: Comparison of parameter estimates for *Gender* under a cut and full Bayes imputation model with snowball/MAR sampling based missingness.

	0% missing	10% missing	25% missing	50% missing	75% missing
Parameters		CM	CM	CM	CM
Intercept	-.591(.073)	-.600(.074)	-.569(.077)	-.668(.098)	-1.022(.150)
Indegree	.010(.002)	.006(.003)	.009(.004)	.029(.011)	.176(.036)
Year of study 2	-.031(.079)	-.050(.083)	-.046(.087)	-.039(.112)	.090(.162)
Year of study 3	-.043(.086)	-.057(.091)	-.046(.096)	-.097(.121)	.011(.171)
Year of study 4	-.087(.098)	-.150(.102)	-.143(.109)	-.136(.133)	-.118(.188)
Year of study 5	-.226(.155)	-.215(.158)	-.143(.172)	-.266(.194)	-.142(.276)
Year of study 6	-.754(.261)	-.967(.299)	-.898(.309)	-.831(.418)	-13.585(8.140)
Day active	.005(.001)	.005(.001)	.006(.001)	.006(.001)	.007(.002)
ρ	-.507(.045)	-.489(.058)	-.354(.073)	-.097(.117)	-.037(.206)
Misclassification rate	.331	.622	.484	.428	.467
Brier score		.312	.268	.275	.369
Parameters		FBM	FBM	FBM	FBM
Intercept		-.507(.069)	-.482(.066)	-.378(.066)	-.314(.065)
Indegree		.005(.002)	.009(.002)	.010(.002)	.009(.002)
Year of study 2		-.047(.077)	-.045(.076)	-.004(.076)	.021(.076)
Year of study 3		-.016(.084)	-.009(.083)	-.100(.083)	.066(.082)
Year of study 4		-.150(.096)	-.145(.094)	-.131(.094)	.022(.091)
Year of study 5		-.184(.152)	-.057(.150)	-.207(.149)	-.126(.148)
Year of study 6		-.724(.246)	-.551(.227)	.015(.203)	-.492(.216)
Day active		.005(.001)	.005(.001)	.004(.001)	.003(.001)
ρ		-.305(.056)	-.076(.061)	-.068(.053)	.015(.049)
Misclassification rate		.617	.461	.441	.415
Brier score		.296	.260	.256	.254

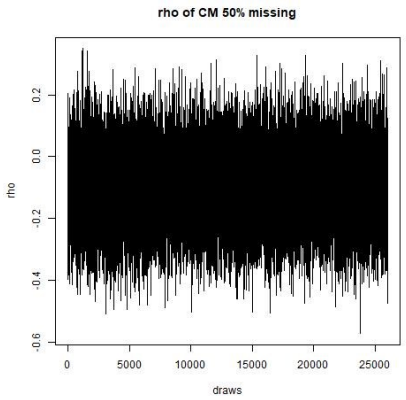
Estimates based on 25000 draws and 1000 burn-in, flat prior for β, CM = cut model, FBM = Full Bayes model.



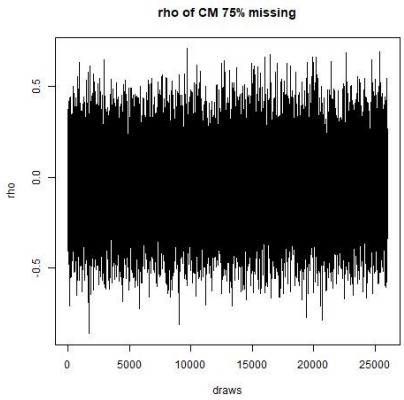
(a) 10% Missingness



(b) 25% Missingness

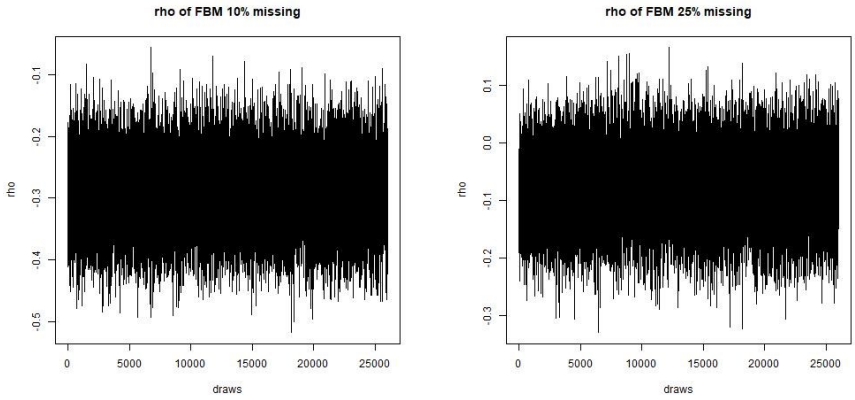


(c) 50% Missingness

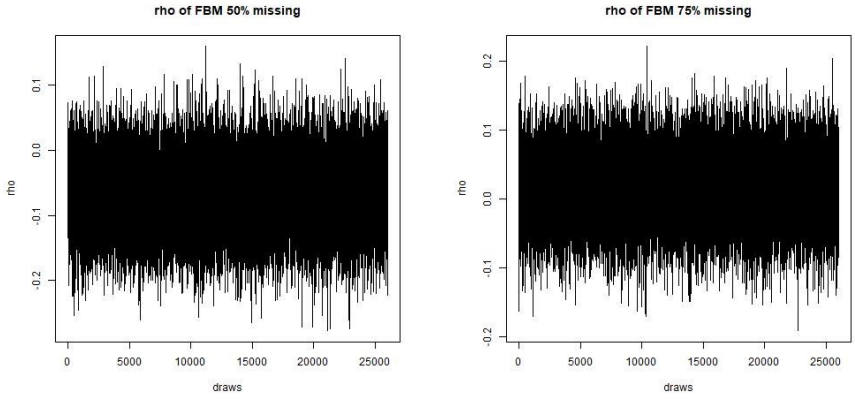


(d) 75% Missingness

Figure S3: Network parameter draws from the sampling procedure on the snowball/MAR sampled missingness, in the cut model; burn in period of 1000 draws followed by 25000 draws.

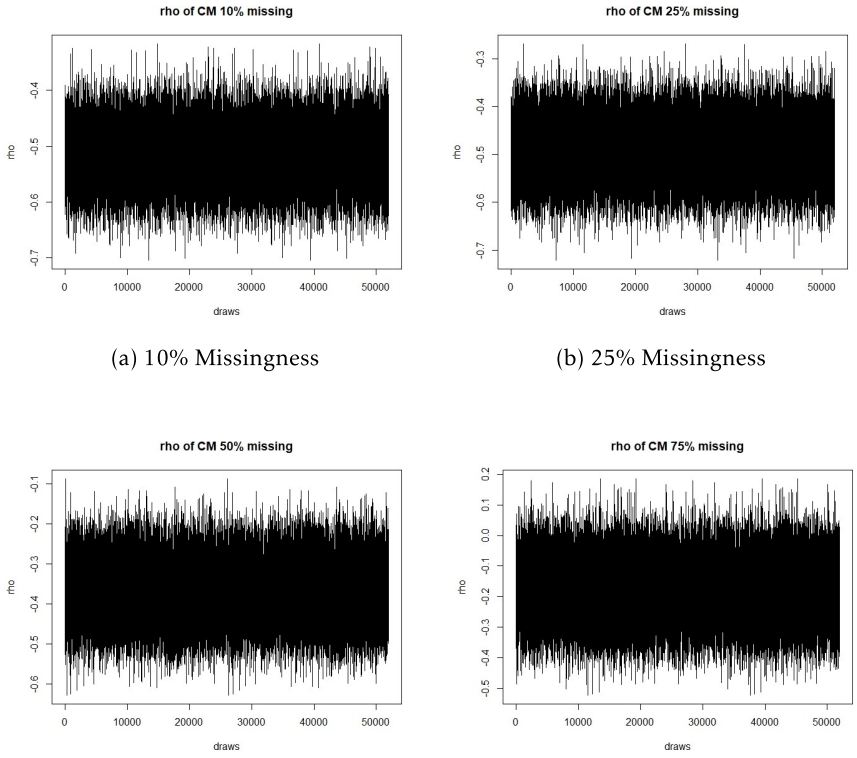


(a) 10% Missingness (b) 25% Missingness



(c) 50% Missingness (d) 75% Missingness

Figure S4: Network parameter draws from the sampling procedure on the snowball/MAR sampled missingness, in the full Bayes; burn in period of 1000 draws followed by 25000 draws.



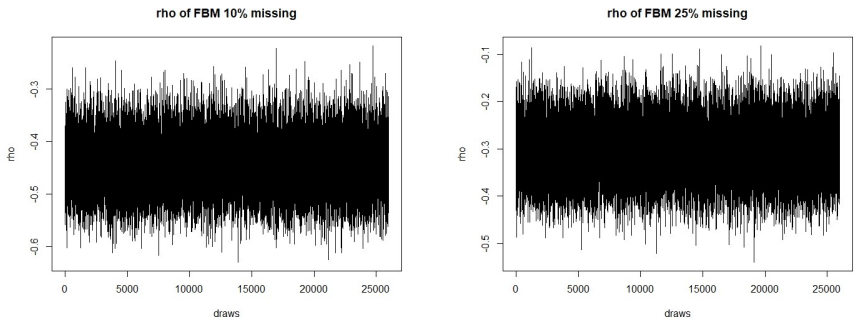
(a) 10% Missingness

(b) 25% Missingness

(c) 50% Missingness

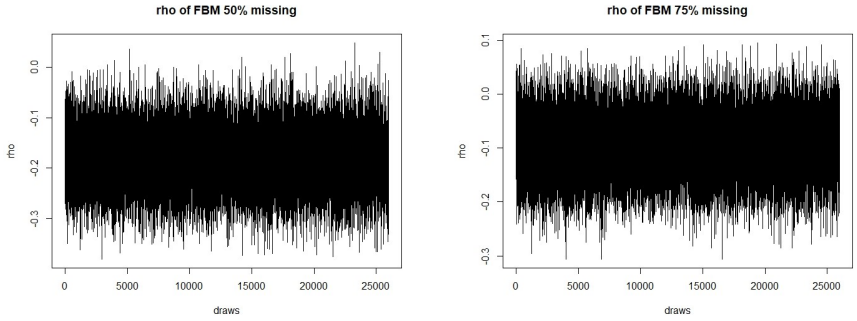
(d) 75% Missingness

Figure S5: Network parameter draws from the sampling procedure on the snowball/MAR sampled missingness, matching the number of missing edges, in the cut model; burn in period of 1000 draws followed by 25000 draws..



(a) 10% Missingness

(b) 25% Missingness



(c) 50% Missingness

(d) 75% Missingness

Figure S6: Network parameter draws from the sampling procedure on the snowball/MAR sampled missingness, matching the number of missing edges, in full Bayes; burn in period of 1000 draws followed by 25000 draws.

Table S4: Parameter estimates from replication with 10% random missingness

	replication 1	replication 2
Parameters	CM	CM
Intercept	-.642(.074)	-.567(.074)
Indegree	.013(.002)	.009(.002)
Year of study 2	-.019(.084)	-.050(.082)
Year of study 3	-.039(.091)	-.024(.091)
Year of study 4	-.095(.104)	-.054(.101)
Year of study 5	-.178(.162)	-.192(.164)
Year of study 6	-.732(.264)	-.743(.267)
Day active	.006(.001)	.005(.001)
ρ	-.533(.050)	-.515(.051)
N missing	185	185
Parameters	FBM	FBM
Intercept	-.564(.071)	-.537(.070)
Indegree	.012(.002)	.010(.002)
Year of study 2	.008(.077)	-.029(.078)
Year of study 3	-.001(.086)	-.030(.086)
Year of study 4	-.125(.097)	-.038(.096)
Year of study 5	-.228(.155)	-.151(.153)
Year of study 6	-.872(.265)	-.679(.247)
Day active	.005(.001)	.005(.001)
ρ	-.439(.045)	-.450(.046)
N missing	185	185

This table presents the parameter estimates from the two replication studies where gender of 10% of the observations was set to missing. Observations were selected via snowball/MAR sampling with a different seed. Estimates are based on 10000 draws with a burn-in period of 500. The number of missing observations fluctuates due to the initial number of seed persons for the snowball/MAR sampling.

Table S5: Parameter estimates from replication with different snowball/MAR sampled subsets (10% missingness)

	replication 1	replication 2
Parameters	CM	CM
Intercept	-.593(.074)	-.483(.074)
Indegree	.011(.003)	.017(.003)
Year of study 2	-.038(.083)	-.052(.082)
Year of study 3	-.056(.091)	-.160(.089)
Year of study 4	-.092(.101)	-.174(.101)
Year of study 5	-.277(.161)	-.297(.161)
Year of study 6	-.985(.301)	-.886(.276)
Day active	.005(.001)	.005(.001)
ρ	-.452(.058)	-.483(.054)
N missing	234	178
Parameters	FBM	FBM
Intercept	-.505(.069)	-.390(.070)
Indegree	.007(.002)	.011(.002)
Year of study 2	.008(.077)	-.102(.077)
Year of study 3	-.005(.085)	-.220(.087)
Year of study 4	-.074(.096)	-.216(.097)
Year of study 5	-.223(.150)	-.318(.153)
Year of study 6	-.698(.236)	-.624(.231)
Day active	.004(.001)	.005(.001)
ρ	-.298(.061)	-.357(.054)
N missing	234	178

This table presents the parameter estimates from the two replication studies where gender of 10% of the observations was set to missing. Observations were selected via snowball/MAR sampling with a different seed. Estimates are based on 10000 draws with a burn-in period of 500. The number of missing observations fluctuates due to the initial number of seed persons for the snowball sampling ($N = 14$ for *set.seed(3030)* and $N = 20$ for *set.seed(6060)*).