Journal of Data Science 0 (0), 1–20
??? 2022

DOI: 10.6339/22-JDS1045
Data Science Reviews

Comparison of Methods for Imputing Social Network Data

Ziqian Xu[1,*], Jiarui Hai[2], Yutong Yang[3], and Zhiyong Zhang[1]

[1]*Department of Psychology, 390 Corbett Family Hall, Notre Dame, IN 46556, University of Notre Dame, United States*

[2]*Department of Hydraulic Engineering, 30 Shuangqing Rd, Haidian District, Beijing 100000, Tsinghua University, China*

[3]*School of Statistics, Renmin Daxue N Rd, Haidian District, Beijing 100000, Renmin University of China, China*

Abstract

Social network data often contain missing values because of the sensitive nature of the information collected and the dependency among the network actors. As a response, network imputation methods including simple ones constructed from network structural characteristics and more complicated model-based ones have been developed. Although past studies have explored the influence of missing data on social networks and the effectiveness of imputation procedures in many missing data conditions, the current study aims to evaluate a more extensive set of eight network imputation techniques (i.e., null-tie, Reconstruction, Preferential Attachment, Constrained Random Dot Product Graph, Multiple Imputation by Bayesian Exponential Random Graph Models or BERGMs, k-Nearest Neighbors, Random Forest, and Multiple Imputation by Chained Equations) under more practical conditions through comprehensive simulation. A factorial design for missing data conditions is adopted with factors including missing data types, missing data mechanisms, and missing data proportions, which are applied to generated social networks with varying numbers of actors based on 4 different sets of coefficients in ERGMs. Results show that the effectiveness of imputation methods differs by missing data types, missing data mechanisms, the evaluation criteria used, and the complexity of the social networks. More complex methods such as the BERGMs have consistently good performances in recovering missing edges that should have been present. While simpler methods like Reconstruction work better in recovering network statistics when the missing proportion of present edges is low, the BERGMs work better when more present edges are missing. The BERGMs also work well in recovering ERGM coefficients when the networks are complex and the missing data type is actor non-response. In conclusion, researchers analyzing social networks with incomplete data should identify the network structures of interest and the potential missing data types before selecting appropriate imputation methods.

Keywords *Bayesian ERGM; ERGM; missing data; multiple imputation*

1 Introduction

Social networks are useful for analyzing social structures, which, in turn, can help us understand how social interactions are associated with behaviors and changes. Indeed, social network data on friendships have been frequently used to explore friends' influence on behaviors such as suicide

[*] Corresponding author. Email: zxu9@nd.edu.

Received December 21, 2021; Accepted March 30, 2022

(Massing-Schaffer et al., 2020) and smoking (Liu et al., 2021). Social network analysis was also applied to understanding how knowledge transfers in workplaces (Yang et al., 2019). When not restricting network nodes as individuals, networks can shed light on measurement theories when viewing measurement items as nodes and item relationships as ties (Epskamp et al., 2018). Additionally, with the development of technology, social network analysis can further inform fields like social media analytics and information sciences. In fact, network analysis has been applied to various online media platforms such as Twitter (Himelboim et al., 2017) and Facebook (Akhtar et al., 2013), and the same idea can be transferred to citation network analysis as a branch of information sciences (Otte and Rousseau, 2002).

While social network analysis is useful for many fields, its validity is often threatened by the presence of missing data. The issue of missing data is especially prominent when data are collected through questionnaires because respondents often omit questions due to the broad extent and the sensitive nature of the information acquired. Besides, network data naturally have dependency among actors, meaning that missing information from one respondent results in the information about other respondents more likely to miss as well, making network analysis more susceptible to missing data than non-network data (Huisman, 2009).

Studies investigating the effect of missing data on social networks have shown that different kinds of missing network data can lead to biased estimations of network structural statistics to different extents (Kossinets, 2006; Smith and Moody, 2013; Smith et al., 2017). While actor non-responses or missing all information from certain respondents can lead to under-estimations of some network statistics such as assortativity, missing data due to constraining the number of neighbors that one respondent can nominate results in over-estimations of such statistics (Kossinets, 2006). Further, the performance of network statistics against missing data depends on the statistics themselves as well. For example, network characteristics such as transitivity and closeness are relatively robust against missing data, whereas distances and betweenness are more sensitive (Smith and Moody, 2013). Additionally, central actors are likely to bias the network statistics more than the less central actors when they are missing (Smith et al., 2017). Such characteristics of missing data in social networks make them difficult to analyze, which, in turn, emphasizes the importance of imputation procedures. One advantage of the imputation methods is that they do not need to assume a specific model and once the missing data are imputed, any complete data methods can be readily applied to analyze the social networks.

Researchers have proposed various methods for treating missing data in social networks. Besides the simplest procedures such as listwise deletion and null-tie imputation (i.e., replacing missing values with 0's), methods relying on network structural traits such as Reconstruction (Stork and Richards, 1992), Preferential Attachment (Barabási and Albert, 1999), and Constrained Random Dot Product Graph (Marchette and Priebe, 2008) can be used to impute missing values through rather simple steps. More complicated methods such as imputation using Bayesian Exponential Random Graph Model (BERGM) further fill in missing values in networks using model-based approaches (Caimo and Friel, 2011), which can consider both structural characteristics of a network and the associated covariates. On the other side, existing methods such as Multiple Imputation by Chained Equations (Van Buuren and Groothuis-Oudshoorn, 2011), k-Nearest Neighbors, and Random Forest can be used to impute binary data without assuming a specific model for social networks. Based on the performance of these methods on binary data, they may also be useful in binary social network data imputation through data manipulation.

Researchers have also explored the performance of social network missing data imputation procedures from various aspects. Huisman (2009) compared the effectiveness of several simple procedures including Reconstruction and Preferential Attachment under different missing data mechanisms (i.e., Missing Completely at Random and missingness related to covariates or network characteristics) for different missing data types including actor and tie (i.e., partial information from respondents is available) non-responses, and found that while the imputation procedures performed better than ignoring missing data in restoring network statistics, they still led to biased estimations. Žnidaršič et al. (2017) evaluated several imputation strategies including Reconstruction, 3 Nearest Neighbors, null-tie imputation, and mean imputation on actor non-responses when applied to network blockmodeling, and found that imputation using the median of 3 neighbors performed better than the rest of the methods. Krause et al. (2020) tested imputation strategies including null-tie imputation, Reconstruction, and multiple imputation by BERGM for simulated binary social networks, and found that for actor non-responses, simpler methods such as Reconstruction and the simple BERGM worked well in reconstructing missing edges for small data sets whereas the complex BERGM worked better for larger data sets. Additionally, Smith et al. (2022) compared listwise deletion, Reconstruction, null-tie imputation, reciprocity-based probabilistic imputation, and imputation by ERGM on a range of empirical social network and concluded that excluding listwise deleting, performance of imputation methods differed by which network characteristic the researchers were interested in. Further, de la Haye et al. (2017) proposed an analytic strategy for longitudinal social network data with missing values, suggesting that subsetting to include participants who had complete data in any two consecutive waves would generate a representative sample of the original study sample.

Although the above studies have investigated and compared the performance of several imputation methods on missing network data, the current study aims to provide a more comprehensive evaluation and comparison of many methods. First, the existing studies are still limited. For example, although Huisman (2009) explored multiple missing data mechanisms, types, and proportions, only simple methods such as Reconstruction were used. While Krause et al. (2020) evaluated the effectiveness of more complex imputation methods using BERGMs and ERGMs, they only compared them with null-tie imputation and Reconstruction, and only considered actor non-response data. Similarly, Žnidaršič et al. (2017) only explored imputation methods for actor non-response and did not consider tie non-response or a mix between tie and actor non-responses that is more likely to occur in empirical studies. Additionally, to our knowledge, no studies have evaluated imputation methods such as MICE and Random Forest that are normally applied to binary data in the context of network data imputation. Therefore, in the current study, we will conduct a comprehensive comparison of many cross-sectional social network imputation methods considering different missing data conditions, network sizes, and network complexities. The study will provide a fuller picture of the performance of different methods for dealing with missing data in social network analysis.

The rest of the paper is organized as following. First, we will review the imputation methods evaluated in our study including null-tie imputation, Reconstruction, Preferential Attachment, Constrained Random Dot Product Graph, Multiple Imputation by BERGM, Multiple Imputation by Chained Equations, k-Nearest Neighbors, and Random Forest. Then, we will compare these methods through a comprehensive simulation study and identify the best method under each condition. Finally, we will discuss and make recommendations on when to use each imputation method based on the simulation results.

2 Review of Imputation Methods

A social network can be stored in an adjacency matrix A, where each row or column represents a unique actor, and the value of the tie from actor i to actor j is denoted as a_{ij} ($i = 1, \ldots, n$, $j = 1, \ldots, n, i \neq j$). This study will focus on cross-sectional, binary, and directed social networks without self-loops, so A is binary but not necessarily symmetric. Throughout the paper, we will denote $a_{ij} = 1$ ($i \neq j$) as a present edge (or tie) and $a_{ij} = 0$ ($i \neq j$) as an absent edge (or tie). For convenience, we set $a_{ii} = 0$ for the non-existence of self-loops. If a_{ij} (either 1 or 0) is missing, we call it a missing edge (or tie), which should be distinguished from absent edges, meaning the lack of a relationship between two actors.

2.1 Null-Tie Imputation

The method of null-tie imputation simply replaces missing edges with the value of 0. Thus, missing edges are imputed as absent edges. Null-tie imputation is the same as imputing missing values by the unconditional mean in social network data when the network is sparse (Žnidaršič et al., 2012). In our simulation study, we include null-tie imputation as the baseline method. Generally speaking, the more complicated imputation methods introduced below are expected to have better performance than the null-tie imputation method.

2.2 Reconstruction

The reconstruction (RE) method (Stork and Richards, 1992) imputes missing edges of a network based on reciprocity. Imputation by RE is conducted using one of the following two ways (Huisman and Steglich, 2008).

1. If an edge a_{ij} ($i \neq j$) is missing, its value is replaced with a_{ji} when a_{ji} is known.

2. If both a_{ij} and a_{ji} ($i \neq j$) are missing between the two actors i and j, the edge is randomly imputed based on the observed edge density of the network.

While reciprocity is a natural assumption in undirected networks (i.e., A is symmetric), edges are not necessarily reciprocated in directed networks. We can expect the RE method to work better for directed networks with high reciprocity compared to those with low reciprocity, and the RE method is expected to lead to higher reciprocity in the imputed networks.

2.3 Preferential Attachment

The preferential attachment (PA) method (Barabási and Albert, 1999) imputes missing edges based on how connected each actor is. This method assumes that the probability of $a_{ij} = 1$ ($i \neq j$) is proportional to the indegree of actor j. (Huisman and Steglich, 2008; Huisman, 2009). Further, the total number of outgoing edges for actor i is determined using the outdegree distribution assumed by the observed network. The process of preferential attachment can be described as the following.

1. From the outdegree distribution of an observed network, an outdegree O_i is drawn for an actor i with any number of missing outgoing edges.

2. The number of outgoing edges that should be imputed as present for actor i can now be calculated as difference between O_i and the observed number of outgoing edges $O_i^{'}$ of actor i, or $O_i - O_i^{'}$. If this number is non-positive, then no edges need to be imputed.

3. The probability of $a_{ij} = 1$ is proportional to the indegree of actor j. A total of $O_i - O_i^{'}$ outgoing edges from actor i at places when actor i have missing edges are imputed as present

based on the indegree probability density distribution of all potential actors j. The remaining missing edges from actor i are imputed as absent.

2.4 Constrained Random Dot Product Graph

Constrained Random Dot Product Graph (DP) models actors in an s-dimensional latent space, and the missing edges are imputed by the dot product of each pair of actors' latent position vectors (Marchette and Priebe, 2008; Ouzienko and Obradovic, 2014). The imputation process can be described as the following.

1. The s-dimensional latent positions of actors i and j are identified as size s vectors x_i and x_j. Because we are using directed social networks, there are two sets of positions: x_i^o and x_j^o for outgoing ties, and x_i^i and x_j^i for incoming ties.
2. The probability of $a_{ij} = 1$ $(i \neq j)$ can be identified by $P(a_{ij} = 1) = f(x_i^o \cdot x_j^i)$.
3. The function f is characterized by

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \leqslant x \leqslant 1, \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

The DP method can be viewed as conducting a singular value decomposition on the adjacency matrix. Suppose that the adjacency matrix M has the singular value decomposition UDV^T, then matrix U contains singular vectors for x^o and V contains singular values for x^i (Marchette and Priebe, 2008) because U contains the eigenvectors for MM^T which are outdegree-dominant and V contains the eigenvectors for M^TM which are indegree-dominant.

Unlike the RE and PA methods, the DP method can be extended to include covariates. However, including covariates makes it necessary to use maximum likelihood to estimate the latent positions, and the closed form singular value decomposition solution is no longer valid. Finding latent positions along with covariates are computationally challenging (Marchette and Priebe, 2008) and the implementation is complicated. This is in contrast to the methods introduced below where adding covariates do not significantly increase the computational complexity of the imputation compared to adding network structural features. Therefore, for the DP method, we will use the form without covariates.

2.5 Multiple Imputation by Bayesian ERGM

An Exponential Random Graph Model (ERGM) is a probabilistic way to model network data that views an observed network (y) as a realization of a graph (Y). As noted by Caimo and Friel (2011), the distribution of a set of networks can be described as a function of network statistics,

$$\pi(y|\theta) = \frac{\exp\left[\theta^t s(y)\right]}{z(\theta)}, \tag{2}$$

where $s(y)$ is a vector of sufficient network statistics, θ is the corresponding coefficients to the network statistics, and $z(\theta)$ is a normalizing constant equal to the sum of all statistics in all realizations of the graph. The network statistics can be structural ones like reciprocity and triangles in the networks, and it can also be related to nodal or dyad covariates.

In ERGM, the ultimate goal is to estimate θ, but it can be difficult statistically. Traditionally, maximum pseudo-likelihood estimation is used for such estimations, however, this method only uses local information and can be inefficient (Caimo and Friel, 2011). Instead, a Bayesian approach (Bayesian ERGM or BERGM) that draws the estimates from the full conditional distributions through a Gibbs sampler in a MCMC scheme using the exchange algorithm is possible and is also more effective (Caimo et al., 2021b). The Bayesian approach can also handle missing data in social networks. Koskinen et al. (2010) proposed a data augmentation scheme using Bayesian inference through a linked importance sampler auxiliary Metropolis-Hastings algorithm to deal with the missing data problem in a model-based way. In the current study, we implemented two different BERGM imputation algorithms each with 30 imputations, which will be introduced in section 3.

2.6 Multiple Imputation by Chained Equations

Multiple Imputation by Chained Equations (MICE) uses the fully conditional specification as described by Van Buuren and Groothuis-Oudshoorn (2011) to impute missing values. MICE can account for different data types and iteratively impute missing data one by one. While MICE can be used to impute different data types, here we focus on binary data since this form is what is present in the social networks of our interest. For imputation of binary data, the model that MICE uses is logistic regression. While MICE was not originally developed for social network data, it has been widely applied to other types of data including health record network data (Chang et al., 2020) successfully. Thus, we want to evaluate its effectiveness for social network data imputation.

In implementation, a network adjacency matrix is reconstructed similar to an edge list matrix with three columns: for an edge a_{ij}, one column denotes i, one column denotes j, and the last column denotes the value of a_{ij}. Other attributes such as the difference in covariate levels across two neighboring actors connected to an edge can also be added. For simple analysis, 5 imputations have been suggested (Little and Rubin, 1987). For more complex models, typically more imputations are needed. In this study, we use 30 imputations.

2.7 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a non-parametric supervised machine learning algorithm useful for prediction and classification tasks initially developed by Fix and Hodges (1989). The main ideas of the KNN algorithm for classification can be simplified as the following.

1. Initialize K as the number of neighbors the algorithm will use. In our simulation study, $K = 3$ is used.

2. Calculate the Euclidean distances between the target to be classified and each existing point in the present data. The number of attributes that the target has will determine the number of dimensions the data will rest in.

3. After sorting distances from each data point to the target, the K closest data points or neighbors to the target will be chosen, and the classification for the target will be the majority vote of all K neighbors' classes.

Similar to MICE, KNN is conducted on the modified form of social network data. The KNN algorithm has been applied to data imputation, and researchers have shown that the KNN imputation method outperformed simple imputation methods such as mean imputation, median imputation, and linear regression imputation (Jadhav et al., 2019). Žnidaršič et al. (2017) also used the KNN method in imputing social networks. Our procedure differs from what Žnidaršič et al. (2017) did, as Žnidaršič et al. (2017) viewed actors as neighbors whereas we view edges with similar attributes as neighbors.

2.8 Random Forest

Random Forest (RF) is another commonly used supervised machine learning algorithm. Initially developed by Ho (1995), RF is an ensemble method that utilizes multiple decision trees. A decision tree branches off from the root with each split on the nodes representing an addition of a feature that can split the data. For example, the tree can be split using the logic of whether the value of a covariate is greater than a cutoff. Then, the observations can be grouped into two sets based on this question. Decision trees fit classification tasks, which predicting the missing ties in social networks as present or absent is an example of.

RF takes multiple decision trees, and combines their classification results. While bagging using majority votes was the original algorithm used to aggregate results from trees, a more commonly used strategy nowadays is to select a fixed number of features from each tree to enhance the effectiveness and accuracy of prediction. In our study, 500 trees are used in the RF algorithm.

RF, similar to KNN, has also been applied to data imputation (Pantanowitz and Marwala, 2009). RF could also be a good choice to impute social network data despite the current sparsity of such usage because of RF's generally good prediction accuracy (Bernard et al., 2009) and its application to other types of data such as numeric data (Pantanowitz and Marwala, 2009) and metabolomics data (Kokla et al., 2019). Similar to MICE and KNN, RF is used on the modified form of social network data.

3 Simulation Study

We simulate cross-sectional directed binary networks with missing data for imputation in R (R Core Team, 2022). Networks simulated in this study all have three nodal covariates, A, B, C, associated with them. A is a continuous variable, normally distributed with the mean of 20 and the standard deviation of 3; B is a binary variable with a 60% probability to be 1 and 40% probability to be 2; C is an ordinal variable that has possible values from 1 to 5, with respective probabilities of 0.1, 0.3, 0.3, 0.15, and 0.15. We first simulate complete network data with varying numbers of nodes using ERGM models based on different model coefficients, and then create missing data from the complete networks. One hundred replications of each of the networks are generated.

3.1 Simulation of Complete Data

ERGM models are used to simulate complete network data with directed edges using the "ergm" package in R (Handcock et al., 2021). The following sample sizes and network statistics are considered. The complete data will serve as the baseline for comparison when evaluating the performance of different imputation methods.

3.1.1 Sample Size

Two sample sizes are used in the simulation of complete data. In social science studies, sample sizes of social networks are often small (Krause et al., 2020; Kc et al., 2019). Here, we simulate small networks with 50 nodes and also large networks with 100 nodes to explore the effect of sample size.

Table 1: ERGM coefficients used for data simulation.

simulation	edges	mutual	gwesp	nodematchB	nodematchC	nodecovA	nodefactorB	nodecovC
1	−2	2	0	0	0	0	0	0
2	−3	2	0	0.8	0.8	0	0	0
3	−8	2	0	0.8	0.8	0.1	0.1	0.1
4	−8	2	0.15	0.8	0.8	0.1	0.1	0.1

3.1.2 ERGM Terms

To consider social networks with different structural variations, we use 4 sets of ERGM coefficients (referred to as Simulations 1 to 4) during complete data simulation. These coefficients are based on the following ERGM terms.

- Edges: The edges term in the "ergm" package counts how many edges are present in a network. This term acts like the intercept term in a regression model and is related to the density of a social network.
- Mutual: We use the mutual term which counts the number of pairs of nodes with present edges in both directions, which is related to the reciprocity of a network.
- Gwesp: The gwesp term stands for "geometrically weighted edgewise shared partner" distribution and it is a measure of triad closure and clustering. This term counts the number of outgoing shared partners. A fixed decay parameter of 0.5 is used for this term, which controls the weight given to additional shared partners.
- Nodematch: We use uniform homophily on covariates B and C. This term counts the number of edges such that the levels of the covariate at the two neighboring actors are the same. We denote the nodematch terms for covariates B and C as nodematchB and nodematchC in this paper.
- Nodecov: This term specifies the main effect of a continuous covariate. Here we denote the term on covariates A and C as nodecovA and nodecovC. The terms sums levels of covariate for neighboring actors.
- Nodefactor: The nodefactor term is similar to nodecov but is for categorical data. Here we included a nodefactor term for covariate B and denote it as nodefactorB. The terms sums the number of times an actor with the given attribute occurs in the edges.

The 4 sets of ERGM coefficients for the selected terms are presented in Table 1. Based on the simulation design, networks from Simulations 1 to 4 are incrementally more complex.

3.2 Simulation of Missing Data

A 3 × 3 × 4 design is used to simulate missing data for networks of each size. The three factors used are missing data types, proportions, and mechanisms.

3.2.1 Missing Data Type

Two types of missing data are distinguished in social network data: tie non-response and actor non-response (Huisman and Steglich, 2008). Tie non-response occurs when a participant's response is only observed on some items but not on all items, meaning that only some outgoing edges of an actor are missing. Actor non-response occurs when a case is completely missing, meaning that all outgoing edges of an actor are not observed in networks. We consider three missing data types: actor non-response, tie non-response, and a mixed type consisting of half actor non-responses and half tie non-responses.

3.2.2 Missing Data proportion

We consider three missing data proportions: 10%, 20%, and 30% missing. Self-links (a_{ii}'s) are excluded as candidates for missing data.

3.2.3 Missing Data Mechanisms

According to Rubin (1976), there are three different missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Under the MCAR mechanism, each individual edge (or actor) in a social network is missing independently of all the data. Under the MAR mechanism, the missingness of data is independent of the missing edge (or actor) itself, but is dependent on other observed information. Under the MNAR mechanism, missingness is related to the unknown missing values themselves.

In the current study, four conditions of missing data mechanisms are considered. We construct one MCAR, one MAR, and two MNAR (MNAR1 and MNAR2) conditions. Missing data in the MCAR condition is generated randomly. The MAR condition is based on the binary covariate, B, following the rule that 80% of missing outgoing edges come from actors with B values of 1, and the rest 20% came from actors with B values of 2. The MNAR conditions differ based on missing data types. For actor non-response, the two MNAR conditions are based on outdegree (MNAR1) and indegree (MNAR2) such that outgoing edges from actors with the lowest respective degrees are missing. For tie non-response, the two MNAR conditions are based on the percentage of edges with the value of 1 (present edges) such that the proportion of present edges missing is larger in the MNAR1 condition than in the MNAR2 condition. In the simulated datasets, the MNAR1 condition has on average 28% of missing ties being present ties, whereas the MNAR2 condition has an average of 10% missing present ties. We also creates a complex pattern in the MNAR1 condition such that as the missing data proportion increases, the proportion of missing present edges out of all missing edges slightly decreases. In contrast, in the MNAR2 condition, the missing present edges proportion increases with missing data proportion. This would facilitate the understanding of the effect of missing data proportion and missing present edge proportion on imputation.

3.3 Imputation

The methods in section 2 are implemented in R (R Core Team, 2022) using the packages “igraph” (Csardi and Nepusz, 2022), “Bergm” (Caimo et al., 2021a), “class” (Ripley and Venables, 2022), “mice” (van Buuren et al., 2021), and “randomForest” (Liaw and Wiener, 2022). Since BERGMs can have different complexities based on what terms are entered in the models, we use two models, one with terms including edges, mutual, nodecovA, and nodematchB (BERGM1), and the other with all terms used in data simulation (BERGM2). As Krause et al. (2020) stated, in general, multiple imputation by BERGM should use models at least as complex as the generation and analysis models. However, it is practically difficult to obtain complete information on the mechanism and structure of the data in empirical studies, and often, the imputation models used are not specified in the same way as how the data would behave. Thus, we use two models, one not assuming perfect knowledge on the data and the other assuming perfect knowledge on the data for BERGM. Comparing the two BERGMs can show the difference in imputation performance between correctly and incorrectly specified models.

For MICE, KNN, and RF which make use of the modified form of social network data that is not in the adjacency matrix format, variables including indegree and outdegree of the actor where the tie is going out from, the differences in levels of covariate A and B across the tie, as well as whether the reciprocated tie is present in the network are used as predictors for imputation. Not all information about the network is included so that the imputation can depart from the simulation process, which is what researchers often encounter in real life.

For methods that use a cutoff value in imputation such as KNN and RF, we set the cutoff to be the observed network density. For multiple imputations, we analyze an aggregated network constructed using majority vote from the 30 imputations. For a specific missing edge, if the proportion of imputed networks that deem it as present exceeds the observed network density, then in the aggregated network, the edge is imputed as present. This aggregation method is introduced by Wang et al. (2016) and used in Krause et al. (2020).

3.4 Evaluation Criteria

The following criteria are used to compare the performance of the different imputation methods.

3.4.1 Link Reconstruction

Link reconstruction is used by Wang et al. (2016) and Krause et al. (2020) as a criterion for evaluating social network imputation. This criterion looks at how well the present and absent edges are imputed. For single imputations, the percentage of correctly imputed present edges and the percentage of correctly imputed absent edges are evaluated. For multiple imputations, an aggregated network as described in section 3.3 is used to compute the percentage of correctly imputed present and absent edges.

3.4.2 Degree Correlations

While the other criteria evaluated emphasize on an imputed network as a whole, indegree correlation and outdegree correlation can further help us understand how each imputation method performs in terms of recovering each actor's characteristics. Here, indegree refers to the total number of incoming edges for each actor and outdegree refers to the total number of outgoing edges for each actor. Correlations of these two degrees for all actors between the imputed networks and the complete networks are obtained for comparison.

3.4.3 Network Statistics of Imputed Networks

We also calculate six network statistics on the imputed networks as the following.

- Density: density is calculated as the proportion of present edges over the number of all possible edges.
- Reciprocity: we calculate the edgewise reciprocity, which is the proportion of edges that is reciprocated.
- Gwesp: we calculated the gwesp statistics as the number of outgoing shared partners over all possible outgoing shared partners.
- Homophily of covariate A (homophilyA): We use Moran's I as a measure of homophily for the continuous covariate A. Moran's I calculates the auto-correlation in covariate A based on a actors' locations which can be seen as their positions in the social networks.

- Homophily of covariate B (homophilyB): We calculate the number of in-group ties for covariate B over all ties as a measure of homophily for the binary variable B.
- Homophily of covariate C (homophilyC): We calculate Moran's I for covariate C in a similar fashion to covariate A.

While density and reciprocity are calculated using the "sna" package in R (Butts, 2020), gwesp and homophilyB are calculated via the "ergm" package (Handcock et al., 2021). Moran's I for homophilyA and homophilyC are calculated using the "ape" package (Paradis et al., 2022).

Network statistics are evaluated through the normalized root mean square errors (NRMSEs) from the corresponding statistics of complete networks. The calculation of RMSEs follows the equation below,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(s_i^{impute} - s_i^{complete})^2}{n}}, \tag{3}$$

where $s_i^{complete}$ is the statistics of the complete network, s_i^{impute} is the statistics of the imputed network, and n is the total number of networks in each category being compared. NRMSEs are then calculated as $RMSE/\bar{s_i}^{complete}$ where $\bar{s_i}^{complete}$ is the mean of the corresponding complete network statistics.

3.4.4 ERGM Coefficients of Imputed Networks

To study how well the imputed networks restore the ERGM coefficients of the complete data, we estimate ERGM coefficients on the eight terms used to generate complete data. These coefficients from imputed networks are compared to complete network coefficients. Similar to the network statistics, NRMSEs are used to compare ERGM coefficients. For all the evaluation criteria, we obtain aggregated mean values of the criteria for each imputation method across the 100 replications. While it is suggested that the imputation models for multiple imputations to be at least as complex as the analysis and simulation models (Krause et al., 2020), which could be possible in a simulation study, it is impractical to assume that the mechanisms that produced data are known in empirical studies. However, it would also be biased to make the analysis model different from the simulation model. Thus, we choose to use the same ERGM in analysis as in the simulation despite that some multiple imputation methods (i.e., BERGM1 and MICE) are purposefully not provided with all available information about the datasets. We think this serves as an illustration of how imputations would work in empirical studies, and analyzing the imputation results using the simulation model could reveal how real-life approaches to imputations may bias the underlying data generation mechanisms if known.

In addition to estimating ERGM coefficients of the imputed networks, we also directly estimate the ERGM coefficients on the networks with missing data. Given that ERGMs can handle missing data internally when estimating the coefficients, we expect coefficient estimates without imputation methods to serve as a comparison on how good the estimates can possibly be. Because directly estimating ERGMs is synonymous to using the simulation model as the imputation model, we expect the directly estimated ERGM coefficients to be closer to true values than those of the imputed networks. Further, unlike in the BERGM2 method, directly estimating ERGMs does not require us to extract certain number of networks based on the estimated coefficients, and then re-estimating coefficients based on the sampled networks. Thus, we also expect the directly estimated ERGM coefficients to be closer to true values than those imputed by the BERGM2 method which also utilizes the simulation model as the imputation model.

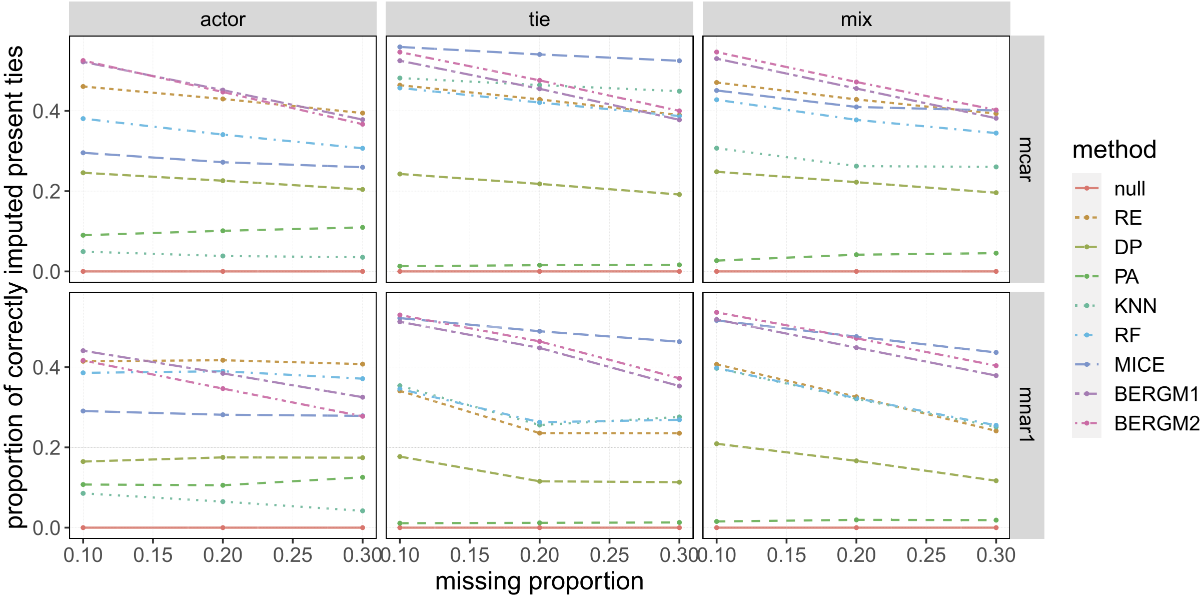

Figure 1: Proportion of correctly imputed present ties by method.

3.5 Results

3.5.1 Overview of Complete Data

The mean estimated EGRM coefficients from the simulated complete networks were close to what we specified in the models as shown in Table 1 in the supplementary materials. The estimated network statistics with different sample size are shown in Table 2 in the supplementary materials. Overall, the estimated statistics were similar for different sample sizes except the Simulation 4 with the most complex model. The same sets of ERGM coefficients produced networks with larger density when sample size was large compared to when sample size was small.

3.5.2 Link Reconstruction

The relative link reconstruction performance of the imputation methods only differed negligibly across the 4 sets of simulations and across sample sizes, whereas differences in which methods performed better were more pronounced across missing data types, proportions, and mechanisms. Therefore, we aggregated link reconstruction performance of each imputation method to levels of missing data types, proportions, and mechanisms only. Patterns across the MCAR, MAR, and MNAR2 mechanisms were also similar, so we only showed results for MCAR and MNAR1. Figure 1 shows the performance of different imputation methods on imputing present ties. In general, the proportion of correctly imputed present ties decreased with increasing missing data proportions. Null-tie imputation had the worst performance on recovering present ties in all conditions. Both BERGMs and RE worked well for actor non-response whereas BERGMs generally worked better for tie non-response and mix non-response and MICE generally worked better for tie non-response. The simpler BERGM1 algorithm worked better than the more complex BERGM2 algorithm under actor non-response, whereas BERGM2 worked better under tie and mix non-responses. Figure 2 shows the performance of different imputation methods on imputing absent ties. As expected, null-tie imputation worked best as it imputed all missing ties as

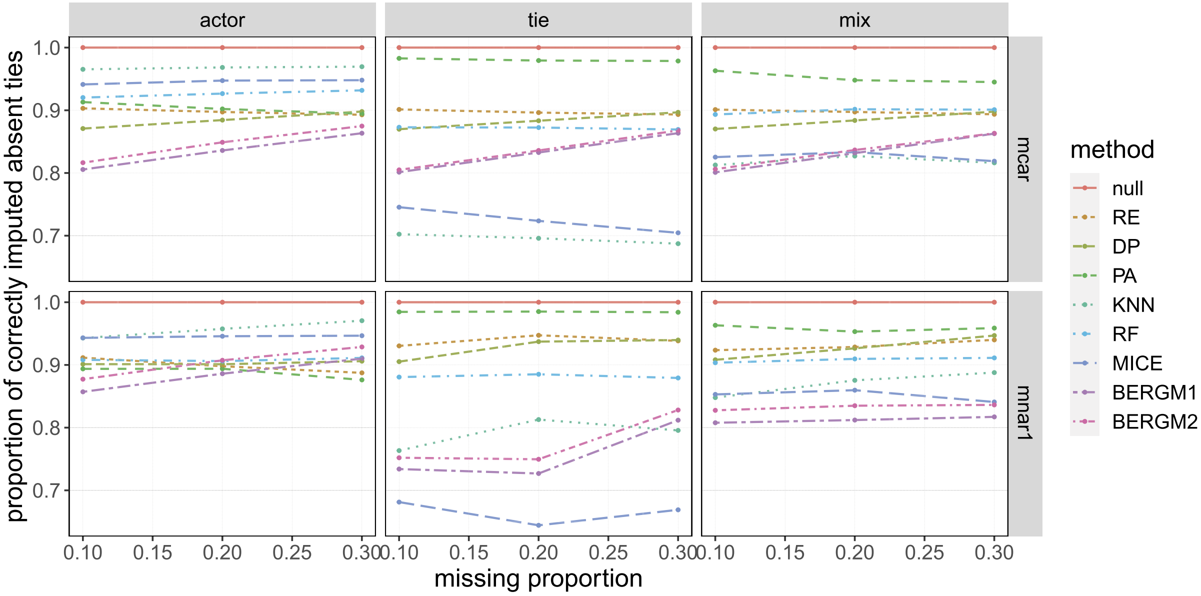

Figure 2: Proportion of correctly imputed absent ties by method.

absent ties. For imputing absent ties, MICE and KNN performed well under actor non-response, whereas PA worked well under tie and mix non-responses.

3.5.3 Degree Correlations

Indegree and outdegree correlations among the actors between the imputed networks and the complete networks were similar across simulations and sample sizes, so we aggregated results over these conditions. Further, similar to link reconstruction results, patterns across the MCAR, MAR, and MNAR2 missing mechanisms were similar, so we only showed the MCAR and MNAR1 mechanisms in Figure 3 and Figure 4, which contain the results for indegree and outdegree correlations correspondingly. Under missing data mechanisms other than MNAR1, null-tie imputation actually had the highest indegree correlation. Under the MNAR1 missing mechanism with tie and mix non-responses, other methods such as BERGM2 and MICE had better performance. In these conditions, the proportion of missing present edges was higher, thus, more complex algorithms worked better. RE had the highest outdegree correlations under actor and mix non-responses with the MCAR, MAR, and MNAR2 mechanism, whereas null-tie and MICE had the highest outdegree correlations under actor and mix non-responses when the mechanism was MNAR1. However, the BERGMs, especially the more complex BERGM2 algorithm worked best under the tie non-response MNAR1 condition. Again, the tie non-response MNAR1 condition had more missing present ties than other conditions, and more complex methods worked better in such a situation.

3.5.4 Network Statistics of Imputed Networks

In general, NRMSEs of the network statistics increased with increasing missing data proportions, and relative performance patterns of different methods were similar across sample sizes. We investigated if the aggregated NRMSEs by sample sizes were different across imputation

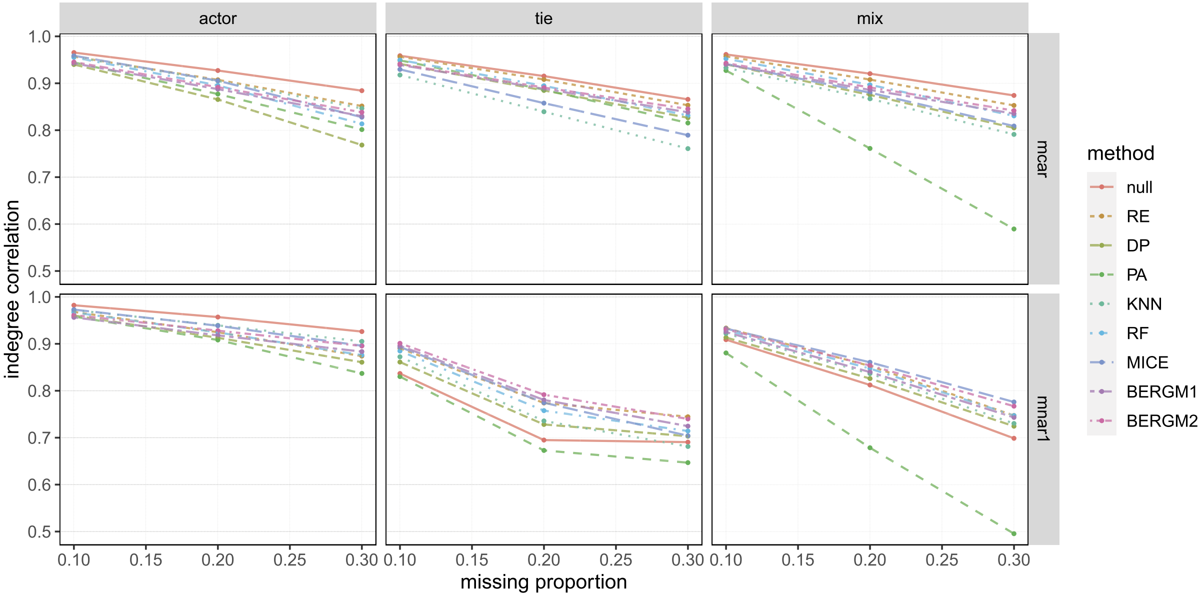

Figure 3: Indegree correlations.

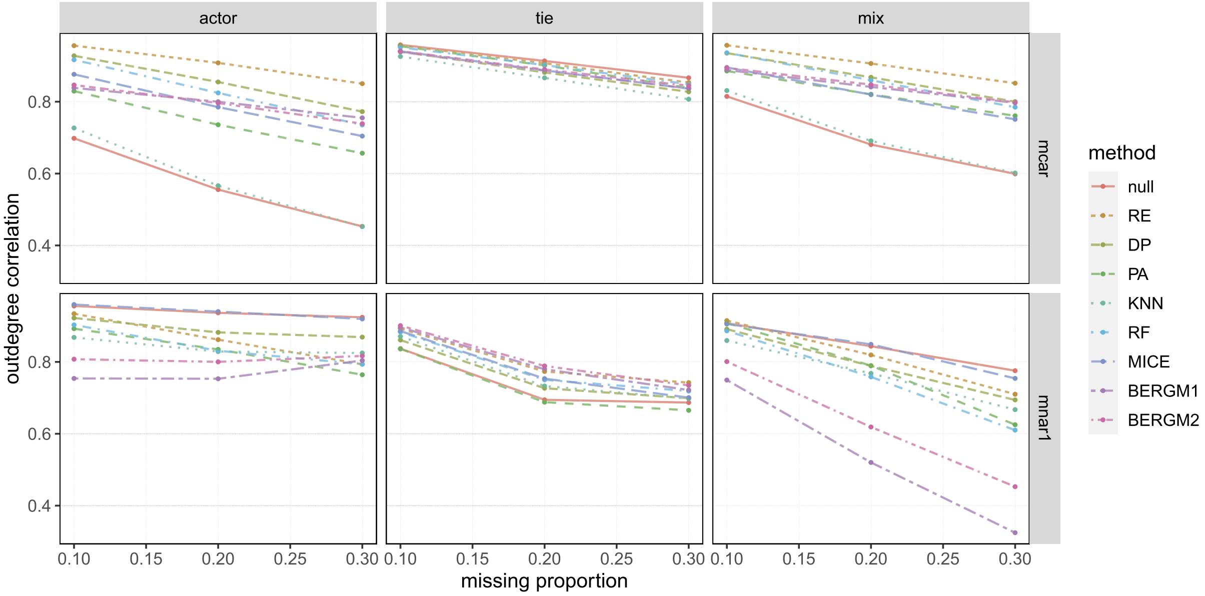

Figure 4: Outdegree correlations.

methods. The results (shown in Table 3 in the supplementary materials) indicated that while reciprocity and the homophily of covariate A were generally imputed slightly better when sample size was small, the homophily of covariates B and C and gwesp tended to be imputed better when sample size was large. NRMSEs of density were not very different by sample sizes.

Because patterns were similar across sample sizes, we used the sample size of 100 as an illustration to further analyze the effect of missing data conditions on network statistics. In Figures 1 to 6 of the supplementary materials, we showed heatmaps of NRMSEs of network

Table 2: Best method for each network statistics in Simulation 4.

mechanism	type	density	reciprocity	gwesp	homophilyA	homophilyB	homophilyC
mcar	actor	RE	DP	RE	RE	null	RE
mcar	tie	RE	MICE	RE	RE	null	null
mcar	mix	RE	BERGM2	RF	BERGM2	null	RE
mar	actor	RE	DP	RE	RE	null	RE
mar	tie	RE	MICE	RE	BERGM2	null	RE
mar	mix	RE	DP	RF	RE	null	null
mnar1	actor	DP	DP	null	DP	null	null
mnar1	tie	MICE	BERGM2	BERGM2	BERGM2	RE	RE
mnar1	mix	MICE	BERGM2	BERGM2	BERGM1	RE	BERGM1
mnar2	actor	RE	null	DP	DP	null	BERGM1
mnar2	tie	DP	KNN	RE	null	null	null
mnar2	mix	DP	DP	RE	DP	null	null

statistics nested by simulations, missing data types, mechanisms, and proportions when sample size was 100, similar to what Krause et al. (2020) presented. We also summarized the relative performance of imputation methods for each network statistics in section 1 of the supplementary materials.

We aggregated network statistics NRMSEs by simulation numbers as well as missing data mechanisms and types to investigate which method was best in each condition. For simplicity, we showed results only for Simulation 4 here (Table 2). In fact, the best performing imputation methods tended to be consistent across different simulations for network statistics. The full results were included in the supplementary materials (Table 4). While simpler methods such as null-tie imputation and RE worked well for conditions including most of MCAR, MAR, and MNAR2, where the proportion of missing present edges was lower, the BERGMs worked well for the MNAR1 tie or mix non-response conditions, where the missing proportion of present edges was higher in comparison.

3.5.5 ERGM Coefficients of Imputed Networks

ERGM coefficients estimates also had similar patterns across sample sizes. We aggregated NRMSEs across sample sizes to see if sample sizes played a role in imputation. As shown in the supplementary materials (Table 5), most imputation methods generally had better NRMSEs in recovering ERGM coefficients when sample size was large. In section 2 and Figures 7 to 14 of the supplementary materials, we analyzed ERGM coefficients of networks with 100 actors in more details.

We also aggregated NRMSEs of the ERGM coefficients by simulation, missing data mechanisms and types to investigate which method was best in each condition. As for the network statistics, we showed results only for simulation 4 here (Table 3), and the full results were shown in the supplementary materials (Table 6). The BERGMs worked best for most coefficients under actor non-response conditions, whereas simpler methods such as RE worked better for tie and mix non-response conditions. For ERGM coefficients in the actor non-response conditions, more complex methods such as the BERGM2 algorithm worked better when the networks simulated were more complicated, whereas in simpler networks, methods such as RE had better performance.

Table 3: Best method for each ERGM coefficient in Simulation 4.

mechanism	type	edges	mutual	gwesp	nodematchB	nodematchC	nodecovA	nodefactorB	nodecovC
mcar	actor	DP	DP	BERGM1	DP	BERGM2	DP	RE	BERGM2
mcar	tie	RF	MICE	RF	PA	null	null	null	null
mcar	mix	RF	BERGM2	RF	null	null	DP	RE	DP
mar	actor	DP	DP	BERGM1	BERGM2	BERGM2	BERGM2	RE	DP
mar	tie	DP	DP	RF	null	null	null	RE	null
mar	mix	DP	DP	RF	null	null	BERGM2	RE	DP
mnar1	actor	null	DP	BERGM1	BERGM2	BERGM2	BERGM2	MICE	BERGM2
mnar1	tie	RE	BERGM1	RF	RF	PA	RF	BERGM1	MICE
mnar1	mix	MICE	BERGM2	MICE	PA	null	DP	MICE	MICE
mnar2	actor	DP	DP	BERGM1	BERGM2	BERGM2	BERGM2	RE	MICE
mnar2	tie	null	MICE	null	null	null	null	null	null
mnar2	mix	DP	null	BERGM1	null	null	null	RE	null

4 Discussion

Through a comprehensive simulation study, we find that the performance of social network imputation methods depend on the purposes of analysis (evaluation criteria), the networks themselves, and the missing data conditions. While none of the imputation methods we evaluated (i.e. null-tie, RE, PA, DP, BERGMs, KNN, MICE, and RF) can correctly impute present ties very well, their performance on imputing absent ties are much better. In imputing missing present ties, BERGMs are most stable, whereas simpler methods such as null-tie imputation and RE perform better in imputing missing absent ties. Furthermore, the degree correlations are mostly above 0.8, suggesting that the imputed networks could be reasonable for studying associations.

Relative performance of different imputation methods in recovering network statistics and ERGM coefficients are different across missing data mechanisms, missing data types, and the complexity of the simulated networks. Such patterns tend to be consistent across missing data proportions and network sample sizes. Overall, imputation biases on network statistics including reciprocity and homophily for the continuous covariate are lower if network size is smaller, whereas the homophily of the binary and ordinal covariates as well as gwesp are better estimated when the sample size is larger. Imputation biases on the estimated ERGM coefficients tend to be lower if network size is larger. Performances in recovering network statistics and ERGM coefficients deteriorate as missing proportion increases in all missing data conditions except for when data are missing according to tie non-response and the percent of missing present edges decreases with the increasing missing proportion. In this case, the performance in recovering the constructs of interest decreases and then increases in a subset of conditions for imputation methods such as MICE and RF. This corresponds to how the specific missing data condition is constructed: under this condition, the proportion of missing present edges out of all missing edges decreases as the total number of missing edges increase. When the missing data proportion is small, the number of total missing edges might dominate the pattern whereas when the missing data proportion is large, the proportion of missing present edges begins to dominate. The performance discrepancies among imputation methods also increase with increasing missing proportions. Therefore, when the percentage of missing data is high, choosing a competent imputation method is especially vital.

As mentioned above, imputation methods with better performances differ across missing data mechanisms and missing data types when recovering network statistics or ERGM coefficients. In a relatively complex network, the simpler methods such as RE and null-tie imputation work better in recovering network statistics when the proportion of missing present ties is low, but when the proportion of missing present ties becomes large, the more complex BERGMs work better. The relatively good performance of RE compared to the methods such as PA is consistent with previous literature (Huisman, 2009). For ERGM coefficients, the model-based BERGMs perform consistently best for terms such as gwesp, nodematch, and nodecov when the missing data type is actor non-response. When the missing data type is tie or mix non-response, the simpler methods perform better. This pattern could be explained by the fact that in the tie non-response and mix non-response conditions, the data is less obfuscated than in actor non-response. Thus, more complex BERGMs are not needed. It should be noted that the more complex BERGM with all terms also work better than the simpler BERGM with selected terms in many actor non-response conditions. The satisfactory performance of BERGM echos what Krause et al. (2020) found for the actor non-response networks and the link reconstruction. One thing we find interesting is that in some situations, null-tie imputation appears to bias the target measures the least, suggesting that sophisticated imputations may not be needed when the missing data patterns are simple.

Based on the simulation results, network statistics can be recovered using simpler methods such as RE, DP, and MICE if the missing proportion of present edges is not very large. Otherwise, BERGM is a better option. ERGM coefficients can be recovered using BERGM if the analysis model is complex and the missing data type is actor non-response. RF, PA, and KNN are not recommended in general because they do not outperform other imputation methods in most cases. In certain situations, methods such as RE and null-tie imputation are also competent. Specific methods that work best in each evaluated situation are summarized in Table 2 and 3.

Several places of our study can be improved in the future. First, while we included nodal covariates for the simulated social networks, we did not include covariates associated with edges or dyads. Thus, some of the network characteristics may not be captured when using methods such as BERGM that could have captured more complex network structures. Second, our current study focused on cross-sectional network data, but longitudinal data are equally important for social sciences as the use of repeated measures steadily increases. Third, other models to simulate network data exist such as the latent space model (Hoff et al., 2002) and the stochastic actor-oriented model (Snijders, 1996) which we did not consider in the current study. Fourth, in practice, the data on covariates could be missing too, but here we assumed complete covariates because our goal is to evaluate the imputation of networks themselves. Fifth, although we identified methods to use for each missing data mechanism and missing data type, it is impossible to identify the missing data mechanism underlying a given dataset in empirical studies, posing a threat to the application of imputation methods. Finally, while the BERGMs using the simulation model as the imputation model work better than the BERGMs with simpler terms, it is unlikely that in practice, all information of a sample is known. The discrepancies between simulation studies and empirical studies can be explored in the future.

In conclusion, the effectiveness of network imputation methods depends on multiple factors including missing data mechanisms, missing data types, target evaluation criteria, complexity of the networks, and imputation methods themselves. BERGMs, ideally with correctly specified terms, work better when the network is more complex or when the missing data pattern is more complicated. In other cases, simpler methods like RE and DP could suffice.

Supplementary Material

- supplement.pdf: Supplementary analyses, tables, and figures mentioned in the paper.

- code: Code used in this study. This folder contains a README.txt file which explains how the code can be used.

Acknowledgements

We thank the editor and the anonymous reviewers for their comments that have significantly improved the study.

Funding

The research was supported by a grant from the US Department of Education (R305D210023).

References

Akhtar N, Javed H, Sengar G (2013). Analysis of facebook social network. In: *2013 5th International Conference and Computational Intelligence and Communication Networks*, 451–454. IEEE.

Barabási AL, Albert R (1999). Emergence of scaling in random networks. *Science*, 286(5439): 509–512.

Bernard S, Heutte L, Adam S (2009). Influence of hyperparameters on random forest accuracy. In: *International Workshop on Multiple Classifier Systems*, volume 5519, 171–180. Springer.

Butts CT (2020). sna: Tools for social network analysis. In: R package version 2.6. https://cran.r-project.org/web/packages/sna.

Caimo A, Bouranis L, Krause R, Friel N (2021a). Bergm: Bayesian exponential random graph models. R package version 5.0.3. https://cran.r-project.org/web/packages/Bergm/.

Caimo A, Bouranis L, Krause R, Friel N (2021b). Statistical network analysis with bergm. arXiv preprint https://arxiv.org/abs/2104.02444.

Caimo A, Friel N (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1): 41–55.

Chang C, Deng Y, Jiang X, Long Q (2020). Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature Communications*, 11(1): 1–11.

Csardi G, Nepusz T (2022). igraph: Network analysis and visualization. R package version 1.2.11. https://cran.r-project.org/web/packages/igraph/.

de la Haye K, Embree J, Punkay M, Espelage DL, Tucker JS, Green Jr HD (2017). Analytic strategies for longitudinal networks with missing data. *Social Networks*, 50: 17–25.

Epskamp S, Borsboom D, Fried EI (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1): 195–212.

Fix E, Hodges JL (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3): 238–247.

Handcock MS, Hunter DR, Butts CT, Goodreau SM, Krivitsky PN, Morris M (2021). ergm: Fit, simulate and diagnose exponential-family models for networks. https://cran.r-project.org/web/packages/ergm/.

Himelboim I, Smith MA, Rainie L, Shneiderman B, Espina C (2017). Classifying twitter topic-networks using social network analysis. *Social Media + Society*, 3(1): 1–13.

Ho TK (1995). Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, 278–282. IEEE.

Hoff PD, Raftery AE, Handcock MS (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098.

Huisman M (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, 10(1): 1–29.

Huisman M, Steglich C (2008). Treatment of non-response in longitudinal network studies. *Social Networks*, 30(4): 297–308.

Jadhav A, Pramod D, Ramanathan K (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10): 913–933.

Kc B, Morais DB, Smith JW, Peterson M, Seekamp E (2019). Using social network analysis to understand trust, reciprocity, and togetherness in wildlife tourism microentrepreneurship. *Journal of Hospitality & Tourism Research*, 43(8): 1176–1198.

Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K (2019). Random forest-based imputation outperforms other methods for imputing lc-ms metabolomics data: a comparative study. *BMC Bioinformatics*, 20(1): 1–11.

Koskinen JH, Robins GL, Pattison PE (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. *Statistical Methodology*, 7(3): 366–384.

Kossinets G (2006). Effects of missing data in social networks. *Social Networks*, 28(3): 247–268.

Krause RW, Huisman M, Steglich C, Snijders T (2020). Missing data in cross-sectional networks–an extensive comparison of missing data treatment methods. *Social Networks*, 62: 99–112.

Liaw A, Wiener M (2022). randomforest: Breiman and cutler's random forests for classification and regression. R package version 4.7-1. https://cran.r-project.org/web/packages/randomForest/.

Little RJ, Rubin DB (1987). *Statistical Analysis With Missing Data*. John Wiley & Sons.

Liu H, Jin IH, Zhang Z, Yuan Y (2021). Social network mediation analysis: A latent space approach. *Psychometrika*, 86(1): 272–298.

Marchette DJ, Priebe CE (2008). Predicting unobserved links in incompletely observed networks. *Computational Statistics & Data Analysis*, 52(3): 1373–1386.

Massing-Schaffer M, Nesi J, Telzer EH, Lindquist KA, Prinstein MJ (2020). Adolescent peer experiences and prospective suicidal ideation: the protective role of online-only friendships. *Journal of Clinical Child & Adolescent Psychology*, 51(1): 1–12.

Otte E, Rousseau R (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6): 441–453.

Ouzienko V, Obradovic Z (2014). Imputation of missing links and attributes in longitudinal social surveys. *Machine Learning*, 95(3): 329–356.

Pantanowitz A, Marwala T (2009). Missing data imputation through the use of the random forest algorithm. In: *Advances in Computational Intelligence*, volume 116, 53–62. Springer.

Paradis E, Blomberg S, Bolker B, Brown J, Claude J, Cuong HS, et al. (2022). ape: Analyses of phylogenetics and evolution. R package version 5.6-2. https://cran.r-project.org/web/packages/ape.

R Core Team (2022). *R: A language and environment for statistical computing*. https://www.R-project.org/.

Ripley B, Venables W (2022). class: Functions for classification. R package version 7.3-20. https://cran.r-project.org/web/packages/class/.

Rubin DB (1976). Inference and missing data. *Biometrika*, 63(3): 581–592.

Smith JA, Moody J (2013). Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks*, 35(4): 652–668.

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

Smith JA, Morgan JH, Moody J (2022). Network sampling coverage III: Imputation of missing network data under different network and missing data conditions. *Social Networks*, 68: 148–178.

Snijders TA (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21(1–2): 149–172.

Stork D, Richards WD (1992). Nonrespondents in communication network studies: Problems and possibilities. *Group & Organization Management*, 17(2): 193–209.

Van Buuren S, Groothuis-Oudshoorn K (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3): 1–67.

van Buuren S, Groothuis-Oudshoorn K, Robitzsch A, Vink G, Doove L, Jolani S, et al. (2021). mice: Multivariate imputation by chained equations. R package version 3.14.0. https://cran.r-project.org/web/packages/mice/.

Wang C, Butts CT, Hipp JR, Jose R, Lakon CM (2016). Multiple imputation for missing edge data: A predictive evaluation method with application to add health. *Social Networks*, 45: 89–98.

Yang CL, Yuan CW, Wang HC (2019). When knowledge network is social network: Understanding collaborative knowledge transfer in workplace. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–23.

Žnidaršič A, Doreian P, Ferligoj A (2012). Absent ties in social networks, their treatments, and blockmodeling outcomes. *Advances in Methodology and Statistics*, 9(2): 119–138.

Žnidaršič A, Ferligoj A, Doreian P (2017). Actor non-response in valued social networks: The impact of different non-response treatments on the stability of blockmodels. *Social Networks*, 48: 46–56.