Mach Learn
DOI 10.1007/s10994-013-5420-1

Imputation of missing links and attributes in longitudinal social surveys

Vladimir Ouzienko · Zoran Obradovic

Received: 30 July 2012 / Accepted: 25 September 2013

Abstract The predictive analysis of longitudinal social surveys is highly sensitive to the effects of missing data in temporal observations. Such high sensitivity to missing values raises the need for accurate data imputation, because without it a large fraction of collected data could not be used properly. Previous studies focused on the treatment of missing data in longitudinal social networks due to non-respondents and dealt with the problem largely by imputing missing links in isolation or analyzing the imputation effects on network statistics. We propose to account for changing network topology and interdependence between actors' links and attributes to construct a unified approach for imputation of links and attributes in longitudinal social surveys. The new method, based on an exponential random graph model, is evaluated experimentally for five scenarios of missing data models utilizing synthetic and real life datasets with 20 %–60 % of nodes missing. The obtained results outperformed all alternatives, four of which were link imputation methods and two node attribute imputation methods. We further discuss the applicability and scalability of our approach to real life problems and compare our model with the latest advancements in the field. Our findings suggest that the proposed method can be used as a viable imputation tool in longitudinal studies.

Keywords Imputation · Temporal data analysis · Social networks · Exponential random graph models

1 Introduction

Social network surveys have proven to be invaluable tools for social scientists. In such surveys often a group of people from an enclosed social setting (e.g. classroom, village etc.) are

Editors: Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach.

V. Ouzienko (✉) · Z. Obradovic
Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA
e-mail: vladimir.ouzienko@temple.edu

Z. Obradovic
e-mail: zoran.obradovic@temple.edu

Published online: 19 October 2013

asked to identify the people of the same group they think of as a friend. The social network observations which are done over time on the same set of people are called panel surveys and each survey conducted at any given time t is called a wave panel. The collection of wave panels done over a single group of people is called a Longitudinal Survey. In practice, not all respondents always choose to provide answers to such surveys, therefore the social scientists are forced to deal with missing data.

The adverse effects of non-responsive actors in social network surveys have been studied extensively in the past. The general consensus is that missing network information or complete absence of an actor from the network surveys will negatively affect the estimation of network properties (Borgatti and Molina 2003; Gile and Handcock 2006), result in underestimation of the network ties' strength (Burt 1987), and cause poor estimation of graph's diameter (Watts and Strogatz 1998). An alternative take on the problem of non-responsive actors is explored in Gile and Handcock (2010). There the non-responsiveness is viewed as a "Hard To Reach" subset of the population, which skews the sampling process and inference. Such a subset usually consists of members of stigmatized groups that are prohibitively expensive to reach and sample. Hence the Respondent-Driven Sampling was introduced allowing sampling of such groups. In Costenbader and Valente (2003) the effects of missing data on social network centrality measures were studied. The measure of an actor's centrality in the social network is indicator of that actor's influence on his peers (Freeman 1978). Authors conclude that there are very few circumstances under which researchers might be able to use a social network with missing data. A similar study (Kossinets 2006) had shown how various missing data mechanisms (network boundary, survey non-response and vertex degree censoring) dramatically affect estimation of network statistics. The statistical simulation experiments done in Kossinets (2006) suggest that actors' non-response greatly underestimates clustering and assortativity coefficients (Chang et al. 2007) leading to inflated measurement errors. Thus the heightened sensitivity of Social Network Analysis (SNA) to missing data as compared to less structured, non-network datasets raises the importance of accurate imputation. The term "imputation" is defined as replacement of the missing values with substitute values. The purpose of our paper is to introduce such a new and accurate imputation model.

The vast majority of the published work investigates the effects of various imputation techniques on SNA (Huisman and Steglich 2008). Works by Robins et al. (2004) and Koskinen et al. (2010) sought to provide an accurate estimation of network statistics. In Robins et al. (2004), the exponential random graph p^* model (ERGM) (Frank and Strauss 1986) was developed for recovery of the network measurements. It is important to understand the impacts of imputation, which is why later on in Sect. 4.2 we present an analysis on how our approach affects statistics of the imputed networks, but the more thorough treatment of the topic deserves its own paper. In this work we are more interested in comparing the imputation accuracies.

The practical benefits of social network imputation were illustrated in the study of criminal gangs in the city of Los Angeles (Stomakhin et al. 2011). The problem formulated in Stomakhin et al. (2011) is different than the problem addressed in this paper, however, we discuss it to underscore the importance of the research in this field. The problem faced by Los Angeles law enforcement was to identify which gang affiliates were involved in criminal activities such as murder, drive-by shooting etc. There were twenty-nine active gangs in Los Angeles at the time of the study. For the most part, police knew which gangs were involved in the fights. However, in some incidents the only gang affiliation available was the one of the victim. The problem addressed in Stomakhin et al. (2011) was to probabilistically impute the other gang which participated in confrontation.

The task of predicting unobserved links has received much attention in computer science and physics literature. In Clauset et al. (2008) authors propose to infer the hierarchal structure of the static network and use that information to explain or reproduce topological features of the network. Very recent work (Sarkar et al. 2012) proposed to use a deterministic non-parametric approach to predict the future state of a temporal network. The method in Sarkar et al. (2012) relies on matching neighborhood's "datacubes" of the link being predicted to the set of the ones that were already observed, thus inferring the link's probability. The work by Lu and Zhou (2011) provides a very nice summarization of the latest advances in the field, and many of the techniques presented there are used as baselines in our work.

Many data imputation concepts discussed in this paper were covered in great detail by Schafer and Graham (2002) review article. In Schafer and Graham (2002), the imputation fundamentals and historical development of the statistical treatments of missing data were given much attention. It described the various reasons why data might be missing and provided theoretical foundation into two general imputation approaches: the maximum likelihood (Dempster et al. 1977; Handcock and Gile 2007) and multiple imputation (Schafer 1999). The work by Schafer and Graham (2002) was not specific to the problem of social network imputation although it covered some of its aspects. Nevertheless, many important concepts are discussed there.

Preliminaries on the imputation approaches in social networks are introduced in Sect. 2 of this article. A brief review of related models and relevant baselines is presented at Sect. 3. In Sect. 4, our approach which facilitates imputation of missing links and the attribute values in social network surveys is introduced. A summary of experimental evaluations on synthetic and real life problems is provided in Sect. 5. In Sect. 6 the scalability of our approach is discussed. The paper ends with discussion and some directions for future work in Sect. 7.

2 Preliminaries

A popular way to present a temporal sequence of social panel surveys is to use a series of binary adjacency matrices (also called sociomatrices). Each $k \times k$ sociomatrix (where k is a number of actors) at a given observation time t, denoted as A^t, represents the surveyed state of social relationships within the time invariant set of actors. The social link within such matrix from actor i to actor j is denoted by $A^t_{ij} = 1$ and absence of the link is denoted by $A^t_{ij} = 0$. For example, if student i in a wave panel at time t considered student j to be her friend, we denote such a relationship as $A^t_{ij} = 1$. The links could be reciprocal: $A^t_{ij} = 1$ and $A^t_{ji} = 1$, but are not necessarily so. It is assumed the actors shall not have self-referenced relationships, i.e. the main diagonal of the sociomatrix will always contain zeros.

We consider the case of the non-responsiveness by surveyed actors. Due to various reasons certain actors at different observation times might choose to ignore the questionnaire provided to them by social scientists. The non-respondent might choose to ignore the questions because of personal reasons. The panel mortality, where people drop out from the longitudinal survey and cannot be located, is also a possibility. Once the person ignores the questionnaire or drops from the study, she might reappear in the future wave panel(s). For example, if survey panel observations were done at times $t = 1 \cdots 4$ we can have actors who completely ignored all four wave panels, we can also have actors fully participating in all of the surveys. We might encounter the situation when actors have chosen to respond to any combination of the wave panels. In our study we assume real valued attributes (also called features) of the non-respondent actor (alcohol usage score, for example) is also not known. Most importantly, the actor who ignored the wave panel is never completely unobserved,

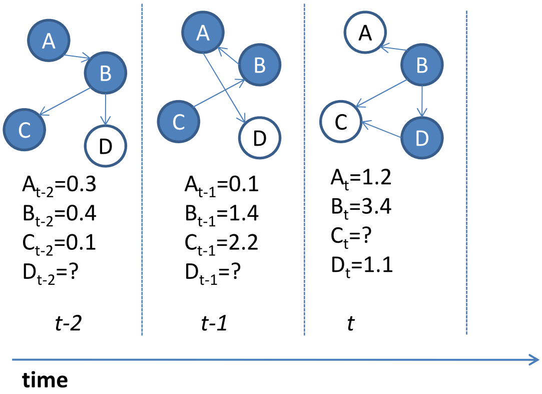


Fig. 1 Representation of the prediction problem addressed in this paper. The actor D at times $t-2$ and $t-1$ and actors A and C at time t were absent from the longitudinal survey (shown as *white nodes*). The prediction task is to guess their outgoing links and real valued features

because the other participants might indicate a friendship link to her. The implication of having more than one non-respondent is that dyadic relationships can be completely unobserved (e.g. two friends had failed to respond to survey). Our approach handles such types of dyadic missingness as well.

More formally, given the sequence of the wave panels surveyed at times $t = 1 \cdots T$ denoted as $A^1 \cdots A^T$, the corresponding actors' attributes denoted as $\mathbf{x}^1 \cdots \mathbf{x}^T$ ($\mathbf{x}^t$ is $1 \times k$ real valued vector, k is the number of actors), and the unobserved actors sets $S^1 \cdots S^T$ ($|S^t| = m^t$, $0 \le m^t < k, \forall t$) our goal is to impute the outgoing links and real valued attributes of the non-respondent actors from the set S^t for each time step t. It should be noted that in this problem setting the social relations can be directed and our proposed model is specifically geared to address such anti-symmetry.

Figure 1 is a schematic representation of the problem addressed in this paper. In this example there were three wave panels done on a group of four actors A, B, C and D at times $t-2$, $t-1$ and t. At each panel wave, actors indicated their social preferences (outgoing social links) and their features were recorded. The only exceptions were absences of actor D at times $t-2$ and $t-1$, and actors A and C at time t, all shown as white nodes in Fig. 1. Note that there are no outgoing links from those actors and we also do not know their features, denoted as a question mark. We need to come up with a good estimate of what outgoing links of those actors would have been, as well as their features.

3 Related work

The treatment of the unobserved links in social networks has been an active area of research for many years. Most of the published work investigated the relationship between the imputation techniques and the introduction of statistical bias and loss of statistical power in a single stationary network (Huisman 2009), or in the context of longitudinal social networks (Snijders 2005). In our paper we rely in the set up of our experiments on one of these works (Huisman and Steglich 2008) because it provided nice comprehensive foundation on how to treat missing data in longitudinal networks. The question of quality of the recovered network statistics is very important. However, in our paper we are more interested in accuracy of the imputation techniques. In this section, we cover some of the most common and recent imputation methods of links and actor's attributes which will serve as baselines in our experiments. We also discuss some recent advances made in the field which cannot be applied directly to our problem but nevertheless bear relevance to our work. Finally, we briefly review the Extended Temporal Exponential Random Graph Model (etERGM), a recently published method by Ouzienko et al. (2011) which we leverage extensively in our approach.

3.1 Link imputation techniques

3.1.1 Reconstruction

A simple but powerful approach to reconstruct links in a single stationary network was proposed by Stork and Richards (1992). This reconstruction method takes advantage of the reciprocity effect very often found in social networks. The imputation procedure works as following: for all ties between respondent and non-respondent actors set the value of unobserved link to the value of the observed reciprocal link: $A^t_{ij\,(\text{imputed})} = A^t_{ji\,(\text{observed})}$. For ties between the non-respondents impute the link with random probability of the observed network density. Here we define the network density as $d = \frac{\sum_{ij}^{k} A^t_{ij}}{k(k-1)}$ where k is number of actors.

3.1.2 Preferential attachment

The Preferential Attachment model proposed by Barabasi and Albert (1999) is based on the assumption that actors with many social links are more likely to be connected to each other. The idea that highly connected vertices are more likely to be connected to each other was also explored in Liben-Nowell and Kleinberg (2003). This technique postulates that the probability of missing actor i having a link to actor j (observed or unobserved) is proportional to indegree r_j of actor j: $P(r_j) = \frac{r_j}{\sum_{i \neq j} r_j}$, i.e., a "popular" actor is more likely to have an incoming link from a missing actor. In the Preferential Attachment model, for each unobserved actor i we randomly draw the outdegree number q_i from the outdegree distribution of the observed network. For the same missing actor we randomly draw, without replacement, and according to probability $P(r_j)$, the q_i number of actors (observed or unobserved). In this step, actors who are popular are more likely to be selected than less popular actors (actors with less incoming links). Finally, we impute the links from actor i to actors which were selected as $A^t_{ij\,(\text{imputed})} = 1$ and $A^t_{ij\,(\text{imputed})} = 0$ to the ones that were not selected.

3.1.3 Constrained random dot product graph

The Constrained Random Dot Product Graph (CRDPG) (Marchette and Priebe 2008) is an imputation technique which models actors as residing in s-dimensional latent space. In this model, the dot product of latent coordinates of two actors yields the probability of the link between the two:

$$p_{ij} = f(\mathbf{x}_i \cdot \mathbf{x}_j + \tilde{\mathbf{y}}_i \cdot \tilde{\mathbf{y}}_j) \tag{1}$$

In (1), f is a simple threshold function

$$f(x) = \begin{cases} 0, & x \leq 0, \\ x, & 0 \leq x \leq 1, \\ 1, & x \geq 1 \end{cases} \tag{2}$$

and $\mathbf{x}_i, \mathbf{x}_j$ are $\mathbf{x} \in \mathbb{R}^d$ latent coordinate vectors of actors i and j in d dimensional latent space. Finally, $\tilde{\mathbf{y}}_i = \mathbf{y}_i * \boldsymbol{\alpha}_i$, where $\mathbf{y}_i$ is actor's i covariates (features) vector and $\boldsymbol{\alpha}_i$ is its associated weights, $*$ is component-wise multiplication. The learning of model parameters is done via the iterative maximum likelihood estimation algorithm described in Marchette and Priebe (2008).

3.1.4 Multiplicative latent factor model

The approach by Hoff models the links' prediction in terms of logistic regression with an extra latent variable (Hoff 2009). In Hoff (2009) the probability of the link is expressed by:

$$\log \text{odds}(y_{i,j} = 1) = \boldsymbol{\beta}'\mathbf{x}_{i,j} + z_{i,j} \tag{3}$$

Here, $\boldsymbol{\beta}$ is the vector of logistic regression coefficients and $\mathbf{x}_{i,j}$ are known predictor variables of the relationship pairs which include the actors' covariates and the presence of the link from j to i. The latent variable $z_{i,j}$ represents patterns in the data unrelated to known predictors. Hoff proposed to use random matrix $\mathbf{Z}$ of latent effects set to deviations of the log-odds from the linear predictor $\boldsymbol{\beta}'\mathbf{x}_{i,j}$. The random matrix $\mathbf{Z}$ is composed of mean matrix $\mathbf{M}$ and noise matrix $\mathbf{E}$: $\mathbf{Z} = \mathbf{M} + \mathbf{E}$. The mean matrix $\mathbf{M}$ of systematic effects is further decomposed using singular value decomposition into lower lank approximation. Such decomposition allows for a better representation of main data patterns and eliminates the lower-order noise. The multiplicative latent factor model is suitable for link imputation by training the model on the observed part of a dataset and using the model parameters to impute the unobserved responses (Hoff 2009). This baseline is different from other link prediction baselines mentioned in this section because it considers both links and actors' features.

3.1.5 Random

We added Random imputation to our baselines more as a sanity check than as a serious predictor. This procedure will randomly fill-in the unobserved portion of the sociomatrix proportional to the density of the observed part.

None of the above-mentioned link imputation techniques consider the real-valued attributes of the observed actors. These methods also do not take advantage of the temporal nature of the longitudinal social survey as they can only impute one stationary network at a time without consideration of other networks in the temporal sequence. Our technique, which we discuss in Sect. 4, will bridge this gap.

3.2 Actor's attribute imputation techniques

In our study we assume the non-respondent actors also fail to provide any other personal information sought by researchers. If the missing information is not available "a priori", then it also has to be imputed. In our paper we will only consider the case of a single real valued attribute per actor per one time step because our goal was to evaluate our approach on a simpler model.

In this section we will discuss two imputation techniques for missing real valued actor attributes which will serve as baselines in our experiments.

3.2.1 Average

The Average method imputes the missing actor's attribute value at each time step as the average of the observed actors in the same survey. This technique is simple and crude, but sometimes simple methods can provide good results.

3.2.2 DynaMMo

The DynaMMo algorithm proposed by Li et al. (2009) is specifically designed to impute the information gaps in multivariate temporal sequence data. The real valued actor attributes in our problem, where each temporal observation is a k-dimensional vector $\mathbf{x}^t$ (k is the number of actors), is in effect such a multivariate temporal sequence which can be imputed by DynaMMo without any modifications. The probabilistic model of DynaMMo consists of two multivariate Gaussian processes. The first process models the transition probabilities between the time steps in the multivariate latent space: $\mathbf{z}_{n+1} = \mathbf{F}\mathbf{z}_n + \omega_n$. The second process describes emission from the latent space to the observed: $\mathbf{x}_n = \mathbf{G}\mathbf{z} + \epsilon_n$. Here $\mathbf{F}$ is transition and $\mathbf{G}$ is observation projections and ω_i, ϵ_i are multivariate Gaussian noises. This model is similar to Linear Dynamical System except it includes an indicator matrix $\mathbf{W}$ of missing values. The joint distribution of observed values $\mathbf{X}_m$, unobserved values $\mathbf{X}_g$ and latent space $\mathbf{Z}$ is expressed as:

$$P(\mathbf{X}_m, \mathbf{X}_g, \mathbf{Z}) = P(\mathbf{z}_1) \cdot \prod_{i=2}^{T} P(\mathbf{z}_i \mid \mathbf{z}_{i-1}) \cdot \prod_{i=1}^{T} P(\mathbf{x}_i \mid \mathbf{z}_i) \tag{4}$$

where T denotes the number of time steps. In (4), the learning of latent parameters $\mathbf{Z}$ is done through iterative coordinate gradient descent optimization procedure. After the model parameters are learned, the imputation of missing values is easily computed from estimation of the latent variables and using the Markov property of the model.

3.3 The extended temporal exponential random graph model

The Extended Temporal Exponential Random Graph Model (etERGM) (Ouzienko et al. 2011) is a decoupled link and attribute prediction model that considers not only prediction of links in temporal networks previously suggested in Hanneke and Xing (2006), but also predicts attributes in such networks. etERGM, however, cannot be used for imputations in its present form. We cover it here because in our approach we include etERGM's prediction models as parts of our iterative solution for imputation of actors' links and attributes. Given the sequence of wave panels $A^1 \cdots A^T$ surveyed at times $t = 1 \cdots T$ and actor attributes $\mathbf{x}^1 \cdots \mathbf{x}^T$, etERGM predicts the network structure A^{T+1} and actor attributes $\mathbf{x}^{T+1}$ at the next unobserved time step $T + 1$. etERGM assumes that all actors have fully participated in surveys at all times $t = 1 \cdots T$.

The etERGM consists of two decoupled models: the link prediction model and the attribute prediction model. The link prediction model is expressed as

$$P\big(A^t \mid A^{t-1}, \mathbf{x}^t, \boldsymbol{\theta}\big) = \frac{1}{Z(A^{t-1}, \mathbf{x}^t, \boldsymbol{\theta})} \exp\big\{\boldsymbol{\theta}' \boldsymbol{\psi}\big(A^t, A^{t-1}, \mathbf{x}^t\big)\big\} \tag{5}$$

The link prediction model (5) defines the transition from A^{t-1} to A^t and incorporates the dependency of A^t over the attributes $\mathbf{x}^t$. In this log-linear model Z is the normalization constant, $\boldsymbol{\psi}$ is a function of $\mathbb{R}^{k\times k} \times \mathbb{R}^{k\times k} \times \mathbb{R}^k \to \mathbb{R}^l$, $\boldsymbol{\psi}(A^t, A^{t-1}, \mathbf{x}^t)$ denotes l-size list of sufficient statistics (k is the number of nodes in the network), which encodes interdependence of actors' links and attributes and $\boldsymbol{\theta}$ is parameter vector. The complete list of statistics used in the link prediction model (Hanneke et al. 2010; Ouzienko et al. 2011) is detailed below:

$$\psi_D\big(A^t, A^{t-1}\big) = \frac{1}{k-1} \sum_{ij}^{k} A_{ij}^t \tag{6}$$

$$\psi_S\big(A^t, A^{t-1}\big) = \frac{1}{k-1}\sum_{ij}^{k}\big[A^t_{ij}A^{t-1}_{ij} + \big(1 - A^t_{ij}\big)\big(1 - A^{t-1}_{ij}\big)\big] \tag{7}$$

$$\psi_R\big(A^t, A^{t-1}\big) = k\frac{[\sum_{ij}^{k} A^t_{ji}A^{t-1}_{ij}]}{[\sum_{ij}^{k} A^{t-1}_{ij}]} \tag{8}$$

$$\psi_T\big(A^t, A^{t-1}\big) = k\frac{\sum_{pqr}^{k} A^t_{pr}A^{t-1}_{pq}A^{t-1}_{qr}}{\sum_{pqr}^{k} A^{t-1}_{pq}A^{t-1}_{qr}} \tag{9}$$

$$\psi_{\text{links}}\big(A^t, \mathbf{x}^t\big) = k\frac{\sum_{i<j}^{k} A^t_{ij}A^t_{ji}\mathbb{I}(|x^t_i - x^t_j| < \sigma)}{\sum_{i<j}^{k} A^t_{ij}A^t_{ji}} \tag{10}$$

Here, ψ_D (6) “density” represents link density of the social network. This statistic governs graph saturation and it naturally models sparsity of social links commonly found in social networks. The ψ_S (7) “stability” measures links’ stability with passing time. The ψ_R (8) “reciprocity” models the reciprocity effect often found in human networks. This is when the feeling of friendship from one person to another is reciprocated in the near future. The transitivity effect, when friend of a friend becomes a friend, is expressed by the “transitivity” statistic ψ_T (9). The “links” statistics ψ_{links} (10) reflects the interdependency between the actors’ links and attributes. It measures the degree to which actors with fully reciprocated links express homophily. To capture the similarity of actors’ attributes the indicator function $\mathbb{I}$ is utilized, which simply counts the actor pairs with similar attributes (defined by the absolute distance and parameter σ).

The node prediction model of etERGM is expressed as:

$$P\big(\mathbf{x}^t \mid \mathbf{x}^{t-1}, A^t, \boldsymbol{\gamma}\big) = \frac{1}{Z(\mathbf{x}^{t-1}, A^t, \boldsymbol{\gamma})}\exp\big\{\boldsymbol{\gamma}'\boldsymbol{\phi}\big(\mathbf{x}^t, \mathbf{x}^{t-1}, A^t\big)\big\}\mathbb{N}\big(\mathbf{x}^t \mid V_0, \Sigma_0\big) \tag{11}$$

It describes the transition of attributes from time $t-1$ to time t, dependent on the network structure A^t at time t. Aside from the variables already mentioned in (5), here $\boldsymbol{\gamma}$ is the vector of model’s weights, $\mathbb{N}$ is the multivariate Gaussian regularization prior, and its mean vector V_0 and covariance matrix Σ_0 are estimated from historical data. The choice for Gaussian as a prior was driven by its smoothening effect on actors’ attributes along the time dimension, diminishing the oscillation of the predicted values. The Gaussian regularizes actors’ attributes so that they stay reasonably within a historically observed range. The node prediction model’s statistics vector consists of three measurements, the “links” statistics from the link prediction model and two additional statistics ϕ_{sim} and ϕ_{dyads}:

$$\phi_{\text{sim}}\big(\mathbf{x}^t, \mathbf{x}^{t-1}\big) = \sum_{i}^{k}\mathbb{I}\big(x^t_i, x^{t-1}_i, \sigma\big) \tag{12}$$

$$\phi_{\text{dyads}}\big(A^t, \mathbf{x}^t\big) = k\frac{\sum_{ij}^{k} A^t_{ij}\mathbb{I}(|x^t_i - x^t_j| < \sigma)}{\sum_{ij}^{k} A^t_{ij}} \tag{13}$$

The “similarity” statistic ϕ_{sim} captures temporal stability of actors’ attributes. This value is large if actors’ attributes do not usually change between the observations, and is small otherwise. $\mathbb{I}$ is the indicator function which counts the actors whose values did not change much during a single time step. Here, the threshold of change is defined by the absolute difference and parameter σ. If $x_i^t = 0.5$, $x_i^{t-1} = 0.4$ and parameter σ is set to 0.2, the ϕ_{sim} is increased by 1. The statistic ϕ_{dyads} expresses the similarity of attributes for the linked actors. It is a ratio of the count of linked pairs which have similar attributes, defined by $\mathbb{I}(x_i^t, x_j^t, \sigma)$, to the total count of all links in the graph.

The use of actors' attributes in both (5) and (11) is consistent with social selection and influence ERGM models detailed in Robins et al. (2001a, 2001b). The more recent work (Snijders et al. 2010) on longitudinal networks, the type of networks studied here, uses the social selection and influence in a stochastic, actor oriented model.

In the next section we show how we have incorporated etERGM's prediction models into our proposed solution for imputation of longitudinal social surveys. etERGM is a natural fit for the problem we are addressing here because it considers the temporal nature of the surveys and interdependence of actors' links and attributes (homophily selection). "Homophily" is often expressed as "birds of a feather flock together", meaning that the actors with similar features are more likely to form a social relationship. While it is a natural fit, the adaptation of etERGM for the imputation task has never been done before. Most importantly we will show how etERGM characteristics allow us to build our own state-of-the-art imputation technique.

4 The proposed ITERGM approach

4.1 Proposed algorithm

The imputation methods we have reviewed so far (Sects. 3.1 and 3.2) can either be applied for link or attribute prediction, or completely ignore the temporal aspect of the surveys. The etERGM model (Sect. 3.3) provides many properties we are looking for in our imputation approach: it encodes the interdependence of actors attributes and links, it considers the time axis in its learning and inference process, and it models directed relations which could be present in social networks. Despite all its characteristics, the etERGM cannot be applied directly to impute attributes or links. Its probability models (5) and (11) can only predict the social network structure at the next unobserved time step given all completely observed previous time steps. Our algorithm, named ITERGM and presented in ALGORITHM 1 is in essence the Expectation Maximization (EM) algorithm over two Markov Chain Monte Carlo (MCMC) inferences. During the Expectation step we draw multiple particles from both link and prediction models (Steps 8–13 and 15–19) of etERGM and use them to impute/update the dataset. During Maximization (Steps 7 and 14) we relearn both models' parameters on the updated data. We repeat these steps until the weights of both models have converged. We choose the iterative solution over a single pass because we want to avoid the dependency of the imputation results on the initialized values. It is unlikely that a single update/imputation pass would reach the point of maximum likelihood. Therefore, we re-learn the model parameters via an iterative approach. More formally, ITERGM method consists of the following steps:

The input of the algorithm (Algorithm 1) is the temporal sequence of the partially observed sociomatrices and actors' attributes. The Steps 1–5 of the algorithm are initializations. In Step 2 we chose "a priori" the DynaMMo algorithm to initialize the missing values of the multivariate temporal sequence of actors' attributes. In Steps 3–5 we apply every imputation technique outlined in Sect. 3.1 to every partially observed sociomatrix and choose the best imputation procedure for initialization of the network's unobserved part. We apply straightforward criteria to select the best initialization technique for links, we choose

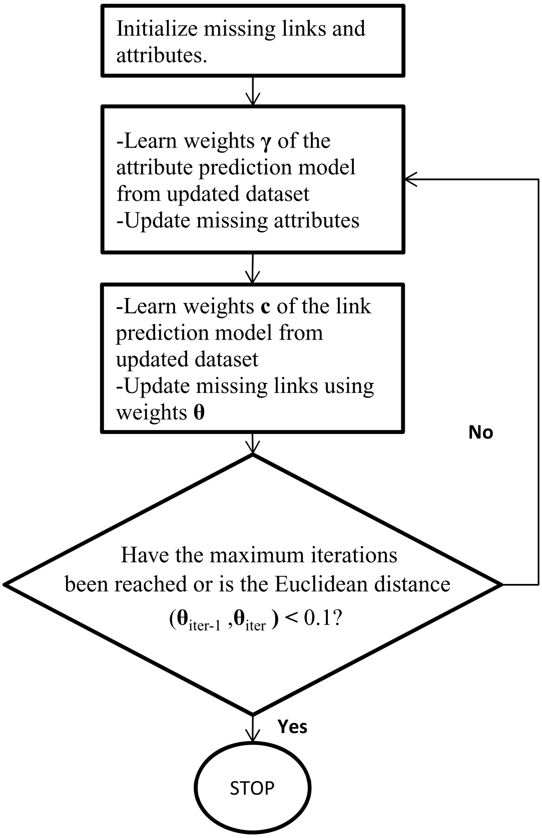
Algorithm 1: ITERGM

```
   Input: The sequence of surveys: $A^{1:T}$, $\mathbf{x}^{1:T}$ where links and attributes, corresponding
          to actor sets $S^{1:T}$ are unobserved: $\mathbf{x}^t(S^t) = \emptyset$ and $A^t(S^t, j) = \emptyset$, $\forall t, j$
   Output: Imputed links score matrices: $\mathbb{S}^{1:T}$. Imputed actors' attributes: $\mathbf{x}^{1:T}_{\text{imputed}}$
1  Initialize iteration counter: $iter = 1$;
2  Apply DynaMMo (Sect. 3.2) to initialize missing values in $\mathbf{x}^{1:T} \to \mathbf{x}^{1:T}_{\text{temporary}}$ ;
3  for $t$ in $1 \cdots T$ do
4      Impute $A^t(S^t, j), \forall j$ with best link imputation technique from Sect. 3.1
       $\to A^t_{\text{temporary}}$
5  end
6  repeat
7      Train etERGM's attribute prediction model (Sect. 3.3) on $A^{1:T}_{\text{temporary}}$, $\mathbf{x}^{1:T}_{\text{temporary}}$ to
       learn weights $\boldsymbol{\gamma}_{\text{iter}}$;
8      for $t$ in $2 \cdots T$ do
9          Sample multiple vectors $\mathbf{x}^t_{\text{inferred}}$ from distribution
           $P(\bar{\mathbf{x}}^t_{\text{inferred}} \mid \mathbf{x}^{t-1}_{\text{temporary}}, A^t_{\text{temporary}}, \boldsymbol{\gamma}_{\text{iter}})$;
10         for missing actor $p$ in $S^t$ do
11             $\mathbf{x}^t_{\text{temporary}}(p) = mean(\mathbf{x}^t_{\text{inferred}}(p))$;
12         end
13     end
14     Train etERGM's link prediction model on $A^{1:T}_{\text{temporary}}$, $\mathbf{x}^{1:T}_{\text{temporary}}$ to learn weights $\boldsymbol{\theta}_{\text{iter}}$;
15     for $t$ in $2 \cdots T$ do
16         Draw multiple networks $A^t_{\text{inferred}}$ from posterior distribution:
           $P(\bar{A}^t_{\text{inferred}} \mid A^{t-1}_{\text{temporary}}, \mathbf{x}^t_{\text{temporary}}, \boldsymbol{\theta}_{\text{iter}})$;
17         Calculate $|S^t| \times k$ score matrix: $\mathbb{S}^t = \sum A^t_{\text{inferred}}(S^t, j), \forall j$;
18         Set $A^t_{\text{temporary}}(S^t, j) = bestcut(\mathbb{S}^t), \forall j$;
19     end
20 until iter < maximum number of iterations;
21 $\mathbf{x}^{1:T}_{\text{imputed}} = \mathbf{x}^{1:T}_{\text{temporary}}$;
```

the algorithm in which imputed density is closest to the density of the observed part. For example, assume that link density of the observed part of network A^t is 0.2. We impute the unobserved part of network A^t by applying every algorithm described in Sect. 3.1 and record the resulting link density of the unobserved part. To initialize links in A^t, we pick the algorithm with computed density of the unobserved part closest to 0.2. The initialization Steps 1–5 are only used to determine initial imputed values for the surveys, and once in the following steps the algorithm runs sufficiently far, it does not matter what the initialized values were. Here we provide Steps 1–5 more as heuristics, but in principle any initialization procedure would work.

At this point, all links and attributes of all networks have been initialized and we begin our iterative approach (Steps 6–20). In Steps 7–13 of the algorithm we apply the etERGM node prediction model to learn its weights and to impute the unobserved attributes by drawing samples from the model over the set of the unobserved actors. Then, in Step 14, we learn the weights of the etERGM link prediction model by training it on the dataset we have just updated with imputed actors' attributes. Knowing the weights, we draw multiple samples

Fig. 2 Schematic diagram of the proposed algorithm ITERGM


Initialize missing links and attributes.
-Learn weights $\boldsymbol{\gamma}$ of the attribute prediction model from updated dataset
-Update missing attributes
-Learn weights **c** of the link prediction model from updated dataset
-Update missing links using weights $\boldsymbol{\theta}$
Have the maximum iterations been reached or is the Euclidean distance ($\boldsymbol{\theta}_{\text{iter-1}}$,$\boldsymbol{\theta}_{\text{iter}}$) < 0.1?
No
Yes
STOP

from the link prediction model and use them to impute the outgoing missing links (Steps 15–19). In Step 20 we check if we reached the number of maximum iterations. Our exit criterion is when there are sufficiently small changes to the weights of the link prediction model between the iterations. If the Euclidean distance between θ_{iter} and $\theta_{\text{iter}-1}$ becomes less than 0.1, then the algorithm stops. If exit criterion is not met, we continue the learning/inference process by constantly updating the dataset over the set of unobserved actors and re-learn the weights of etERGM. Otherwise, the score matrices in Step 17 are our prediction of the imputed links and $\mathbf{x}^{1:T}_{\text{temporary}}$ is our imputed temporal sequence of actors' attributes (Step 21). We present the schematic outline of the ITERGM algorithm in Fig. 2. It is important to note that at each transition we consequently update the missing part of the same dataset (links and attributes) based on the model parameters which were learned at the previous iteration.

The expectation steps of our algorithm deserve closer attention. Computing the expected values of the missing actors' attributes is fairly straightforward. In Step 9, for each survey we sample multiple particles (actors' attributes vectors) based on the weights $\boldsymbol{\gamma}_{\text{iter}}$ learned in the current iteration. We take the mean of the corresponding values of the actors' attribute vector as our prediction of the missing actor's attribute and use that to update our dataset (Steps 10–12). The inference of the links is a bit more involved. Similar to the imputation of actors' attributes we sample multiple sociomatrices for every survey based on the present learned weights $\boldsymbol{\theta}_{\text{iter}}$ of the link prediction model (Step 16). We take the predictions of the imputed values in the form of the score matrices $\mathbb{S}^t$ by adding drawn samples in Step 17. For example, if we sample 100 sociomatrices to predict the network at time step t, then the score matrix $\mathbb{S}^t$ is a sum of all drawn matrices-samples, where the minimal possible score for any entries in $\mathbb{S}^t$ is 0 and the maximum possible score is 100. The score $\mathbb{S}^t_{ij} = 0$ can happen when every matrix in our drawn set predicted absence of a link from i to j. The maximum score $\mathbb{S}^t_{ij} = 100$ is possible when every single matrix from our drawn set predicted a link from i to j. In their present form, the score matrices $\mathbb{S}^t$ cannot be used directly to impute the missing links. We have to convert $\mathbb{S}^t$, which holds the unnormalized integer scores for possible links (larger score denotes higher probability of a link), into binary form, and use that to update the missing part of the network (our sociomatrices are always binary). That is why in Step 18 we apply the *bestcut* procedure to determine the best *threshold* or “cut” to make a binary link imputation matrix suitable to update missing links. The *bestcut* procedure chooses the *threshold* such that the resulting binary matrix is maximizing the probability of the link prediction model in Step 16. This is achieved by moving the *threshold* from the score matrix’s smallest to its largest value. Each time the *threshold* parameter moves, we set all entries in the matrix that are less than *threshold* value to 0 and the remaining entries to 1. Each resulting binary imputation matrix is substituted into the link prediction model and we pick the best matrix, which maximizes the link prediction probability.

4.2 Algorithm convergence

Our algorithm in its essence is a continuous sampling from link and attribute prediction models with iterative updates of model weights. Our exit condition is a sufficient number of iterations, which in practice we limit to 3 or 4. However, we have to ensure that our technique indeed converges. To evaluate convergence of our algorithm we adapted a standard convergence evaluation technique for ERGM models (Snijders 2002). It works as following (we are using link prediction model statistics as an example): (a) take a fully observed network and calculate its real observed statistics ψ_0, (b) randomly remove a given percentage of actors from the network and apply the imputation technique, (c) during imputation draw multiple samples from the model and for each drawn sample calculate network statistics ψ_k, (d) calculate the t-ratio as

$$t_k = \frac{E_\theta(\psi_k) - \psi_0}{SD_\theta(\psi_k)} \tag{14}$$

E_θ is expected value, or simply mean, and SD_θ is standard deviation of the sampled set. In Snijders (2002) it was suggested that $|t_k| \leq 0.1$ is indicative of an excellent convergence, $0.1 < |t_k| \leq 0.2$ is good and $0.2 < |t_k| \leq 0.3$ is fair.

We evaluated the convergence property of our algorithm on the real life dataset *Delinquency* which we describe in the next section. We picked a single transition step from $t = 2$ to $t = 3$ of the dataset and removed 20 % of the actors at random from the network at step $t = 3$. We then ran our imputation technique on the selected transition step and after 3 iterations had collected 1,000 samples of sociomatrices and actors’ attributes. In Table 1 we present the averages of the differences between the true statistics ψ_0 and ϕ_0 and statistics based on the imputed samples, the standard deviation of the differences and corresponding t-ratios. The etERGM statistics $\phi_{\text{links}}, \phi_{\text{sim}}, \phi_{\text{dyads}}, \psi_D, \psi_S, \psi_R, \psi_T$ shown in Table 1 correspond to the measurements of homophily, attributes’ stability, similarity, density, links stability, reciprocity and transitivity (Hanneke and Xing 2006; Ouzienko et al. 2011).

In Table 1 we observe that converging properties are ranging from excellent to poor. However, it should be noted that none of the t-ratios indicate statistical significance. This means that network statistics derived from the imputed data are not significantly different

Table 1 Convergence estimates of the imputation algorithm on time steps $t = 2, 3$ of the real life dataset *Delinquency*

Statistic	Average difference	Standard deviation	t-ratio
ϕ_{links}	0.74	2.12	0.35
ϕ_{sim}	−0.59	2.09	−0.28
ϕ_{dyads}	−0.48	1.20	−0.40
ψ_D	2.02	1.85	1.09
ψ_S	−1.82	1.98	−0.92
ψ_R	0.00	1.01	0.00
ψ_T	0.78	1.15	0.68
ψ_{links}	−0.82	1.55	−0.53

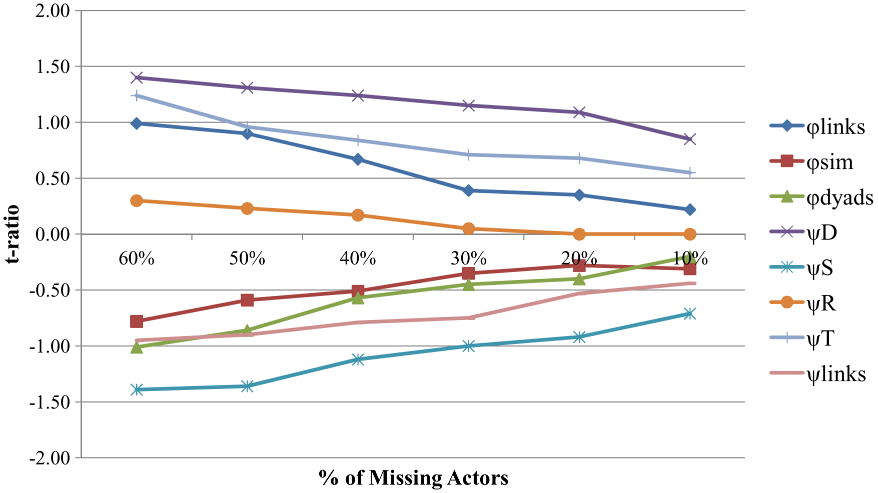


Fig. 3 Comparison of convergence estimates vs. percentage of non-respondent actors in real-life dataset *Delinquency*. On the x-axis is the percentage of actors which were randomly removed from dataset, and on the y-axis is convergence estimate of the model's statistics expressed by t-ratio (14)

than true values. We also examined how convergence measurement relates to the amount of missing data. We repeated the same experiment expressed in Table 1 for various percentages of missing actors which varied from 60 % to 10 % in 10 % decrements for a total of 6 experiments. The results of this series of experiments are presented in Fig. 3. Here we observed the convergent trend of all statistics as the amount of missing data decreased. In other words, the convergence of the model improved as more data was provided to it.

We further examine convergence properties of our algorithm by applying bootstrapping sampling under controlled conditions. In this experiment we generated a synthetic temporal social network with 100 actors and 4 time steps. We then randomly removed 20 % of actors from each time step. Next, we started 15 random walk chains of our algorithm where each chain was initialized with its own set of random values. We recorded the resulting 15 prediction accuracies AUC's and randomly sampled with replacement 10,000 sets of 15 AUC's. For each set we recorded the average, and the histogram of the resulting 10,000 AUC averages is shown in Fig. 4. The 95 % confidence interval of the resulting distribution

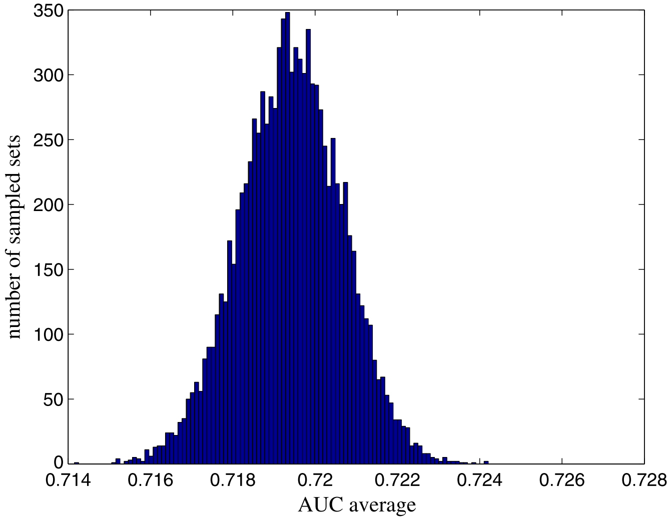


Fig. 4 Histogram of 10,000 bootstrapping samples of AUC's averages based on 15 sampled AUCs with replacement. The original set of 15 AUC's was obtained from the imputation of a synthetic network with 100 actors and 4 time steps. Each algorithm run was initialized with its own random values

is $[0.7169, 0.7219]$ with standard error of 0.0012. Very small variations suggests that our algorithm is convergent.

4.3 Parameter tuning of the random walk

The Gibbs sampling used for estimation of the network parameters can be a slow process (Snijders 2002). The random chain walk is done by randomly flipping a single link of the sociomatrix and then evaluating change in the density of the proposed distribution. After a sufficient number of flips were performed, the model weights were updated via Newton–Raphson optimization step; for more details see the Appendix in Snijders (2002) and also Hanneke and Xing (2006), Hanneke et al. (2010) and Ouzienko et al. (2010). The number of such random flips is a parameter that has to be empirically established. In our experiments this parameter was derived from observing the convergence of the link prediction model weights. If a sufficient number of random steps were performed, meaning the chain had reached stationary distribution, the change in weights $\boldsymbol{\theta}$ demonstrates the model convergence. The Euclidean distance between updated parameters vector $\boldsymbol{\theta}$ should decrease upon each iteration of the outer loop of the Newton–Raphson optimization algorithm. In our experiments the number of random flips was fairly large. For example, when running experiments on the real life dataset *Delinquency* consisting of four sociomatrices of 26 students, we had to perform 1.5 million random flips. When taking such a large number of random steps, the model parameters were always converging, which gave us confidence that random walks were reaching a stationary distribution. The random walk can be sped up by a "big update" technique (Snijders 2002). This is when, instead of a single link, a random block of links is flipped followed by evaluation of distribution density. We customarily applied this technique and set the size of random sampling block of links to ten. It was noted that ERGM models are quite often unstable and can easily become degenerate (Snijders 2002). Degeneracy in sociomatrix sampling is when the random walk completely saturates the matrix in social links, e.g. all nodes become connected with each other, or leaves a matrix without any links. The model degeneracy can be caused by incorrectly instantiated model weights, from which the random walk will never be able to reach stationary distribution. Here, we followed the suggestion from Hanneke and Xing (2006) and randomly initialized the network model parameters $\boldsymbol{\theta}$ in $0 \cdots 1$ range. Such weight initialization strategy worked well, as we never observed model degeneracy in our experiments. We applied similar heuristics to the attribute prediction part of our model (11). On average we sampled 30,000 attribute vectors in experiments on real life dataset *Delinquency*, one third of which were throw away burn-in samples.

5 Experiments

To measure the link prediction accuracy of a temporal network sequence, and in machine learning in general, the AUC was used successfully in the past (Bradley 1997; Dunlavy et al. 2011; Huang and Lin 2009). The AUC is a preferable measurement in the presence of imbalanced datasets such as social networks, where link density is usually low and it is most frequently used in social network literature. The AUC is also equivalent to the Mann–Whitney U test (Hanley and McNeil 1982) and it reflects the probability that a randomly selected true positive is ranked above a randomly selected true negative. Every link imputation algorithm covered in this paper is non-deterministic. Therefore one possible way to measure the link imputation accuracy on a single social network is to compute a score matrix $\mathbb{S}$:

$$\mathbb{S} = \sum_{l} A_l \tag{15}$$

Each run of an imputation algorithm results in the binary $|S| \times k$ subset matrix A_l, which contains only imputed outgoing links. Here, S is the set of actors who did not respond to the survey (did not indicate their outgoing links) and k is the total number of actors in the network. Thus the resulting score matrix $\mathbb{S}$ contains the scores of all imputed links. Using such a matrix we can construct a Receiver Operating Characteristic curve (ROC) by moving the *threshold* parameter in small increments from the matrix $\mathbb{S}$'s smallest to its largest value. Each time we move the *threshold* we create an intermediate binary matrix and set all its entries to 0 if $\mathbb{S}(i, j) < \textit{threshold}, \forall i, j$ and 1 otherwise. Therefore, a binary prediction matrix at the beginning contains all 1's and it contains 0's at the end. While moving the *threshold* parameter we calculate the true positive and false positive rates of imputed links against the true target. True positive rate is the number of correctly imputed links divided by the total count of true links. False positive rate is the number of imputed links which were not in the true target divided by the total count of non-existing links (structural zeros).

To evaluate the accuracy of our approach, we conducted a series of the experiments on synthetic and real life datasets. Two approaches, Missing At Random (MAR) and Missing Not At Random (MNAR), are used to model the non-responses in social network literature (Huisman and Steglich 2008). The former approach assumes there is no underlying hidden structure explaining the missing information, the latter assumes that the missing values are dependent on the actors' attributes or the network topology. For both synthetic and real life datasets we set up our experiments as follows: we randomly remove a predefined percentage of the actors from each wave panel according to MAR or MNAR. We perform repeated imputations ($u = 5$) on the semi-observed dataset by applying the proposed approach and baseline imputation techniques for links and attributes. To compare results we construct the

90 % confidence intervals on both link and attribute imputations according to the "multiple imputation" technique (Schafer 1999). We run our experiments by simulating the removal of the actors according to five missing mechanisms: two MAR and three types of MNAR.

For the first MAR, called "Random", we removed actors at each time step completely at random. At this scenario an actor randomly removed at one time step can potentially reappear at the next time step(s). For the second MAR scenario, called "Absent", an actor was randomly removed completely from all panels. The "Absent" mechanism is useful for modeling of actors who did not respond to a single survey. Such a scenario can occur in a classroom if, for example, a student was out sick for the duration of the study or perhaps persistently ignored a survey. This scenario can be also observed in enclosed medical setting such as a hemodialysis clinic where patients are often absent from their regular environment setting due to hospitalization (DeOreo 1997).

For MNAR, we removed the actors according to probabilities of $\frac{1}{(x_i^t)^2}$, $\frac{1}{(1+\text{indegree})^2}$ and, $\frac{1}{(1+\text{outdegree})^2}$. The first MNAR, which we call "Score", models the absence of actors as being dependent on their real-valued attribute (for example, the actors with higher alcohol consumption score are less likely to respond to survey). The second MNAR, called "Indegree", assumes that the more popular actors are more likely to be survey participants. The third, called "Outdegree", assumes that the socially inactive people are less likely to be willing to answer survey questions. Then, for each missing mechanism we have removed 20 %, 40 % and 60 % of actors from each survey. The 20 %–60 % range of non-responding actors is quite a reasonable assumption, as in the past the response rate of 56 % to a social survey was reported (Berg et al. 2006). Testing our technique against such a dynamic range will give us a fair estimate of its robustness. To summarize, we model the actors' removal at the five types of missingness and three different percentages, to the total of 15 sets of experiments per each dataset and we repeat imputation of each set 5 times.

To assess the imputation accuracy of the actors' attributes we used the Mean Squared Error (MSE) measurement. For the imputation of the links we calculated the Area Under Curve (AUC) measurement on the score matrix of the imputed part of the sociomatrix, discussed previously.

5.1 Synthetic dataset

The purpose of the experiments on the synthetic dataset is to verify the proposed imputation technique under controlled conditions. We generated one synthetic dataset adhering to the Markovian process, where each consecutive social network in the temporal sequence at time t is created from the network of the previous time step $t-1$. We started the generation process by creating a random graph, denoted as A^1, and set $t=1$, then repeated until the needed number of the networks was generated:

- set $t=t+1$
- create a copy of the previous network by setting $A^t = A^{t-1}$
- randomly inverse direction of the 50 % links in A^t
- randomly reassign 10 % of the links in A^t
- randomly pick 20 % of incomplete transitive relationships (link is present from p to q, and q to r, but not from p to r) in A^t, and complete the transitive relationships by adding closure links (from p to r)
- count the number of links added to complete transitive relationships and randomly delete the same number of links from the graph A^t

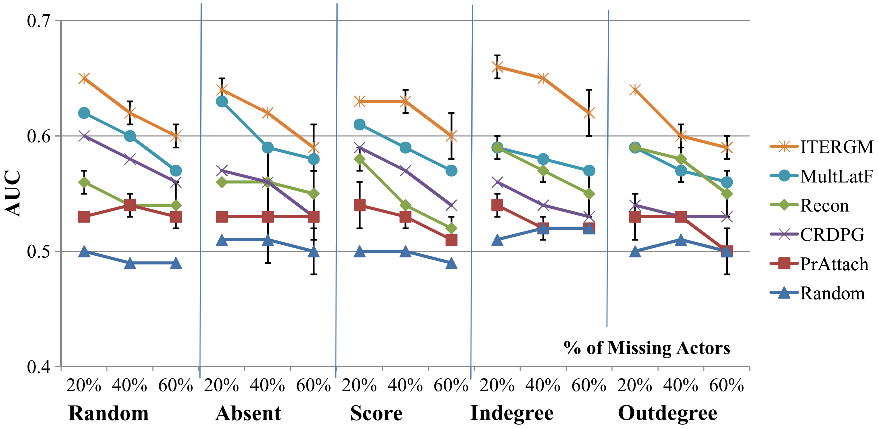


Fig. 5 Links imputation on the synthetic *Dataset1* of 250 actors observed at four time steps (simulating five types of missing mechanisms for 20 %–60 % of missing actors) measured by AUC. The error bars where appropriate show 90 % confidence interval of the result based on five runs of each algorithm

To generate the actors' attributes we used an approach similar to the network generation procedure. We started the generation of the actors' attributes after the networks generation was complete. Similarly to how the sequence of the networks were created we started from the random k-length vector of actors attributes $\mathbf{x}^t$ $(t = 1)$ drawn from the Gaussian distribution with zero mean and one standard deviation. We repeated the following steps until the actors' attributes were populated for all the networks in the sequence:

- set $t = t + 1$
- create a copy of the previous actors' attributes by setting $\mathbf{x}^t = \mathbf{x}^{t-1}$
- randomly select 30 % of actors from $\mathbf{x}^t$ and add to their attribute values a zero mean one standard deviation Gaussian noise
- identify all doubly linked actors pairs in A^t, set the value of one actor in each of these pairs to the value of its doubly linked counterpart plus small random Gaussian noise

We created one synthetic dataset *Dataset1*, simulating a network of 250 actors observed at four time steps, by following our procedure. On average, it took ITERGM four iterations to achieve convergence on this synthetic dataset. The results of the experiments on this dataset are presented in Figs. 5 and 6. In these experiments the ITERGM had the best imputation accuracy of the actors' links and attributes as compared to the baselines every time. We also conducted similar experiments, not presented in this paper, on synthetic networks with 30 and 500 actors. The results were analogous to the ones presented in Figs. 5 and 6. The first thing to note is that the link prediction algorithms "Random" and "Multiple Latent Factors" and attributes prediction algorithm "Average" are deterministic, and this is why their results have zero confidence interval. We did notice that as the size of the network grows the accuracies of the imputation techniques (baselines and ITERGM) were dropping. We attribute this phenomenon to the fact that methods presented here are addressing the imputation globally and do not consider local properties of the communities/clusters that are often found in the large networks. For example, a large social network can span over multiple countries, social classes and ethnicities, which can all have their own customs, traditional behaviors and ways of life. Even though an algorithm can learn weights of large network's statistics (statistics which are based on universally accepted principles like reciprocity, transitivity and homophily), the globally learned weights will not account for variations that are

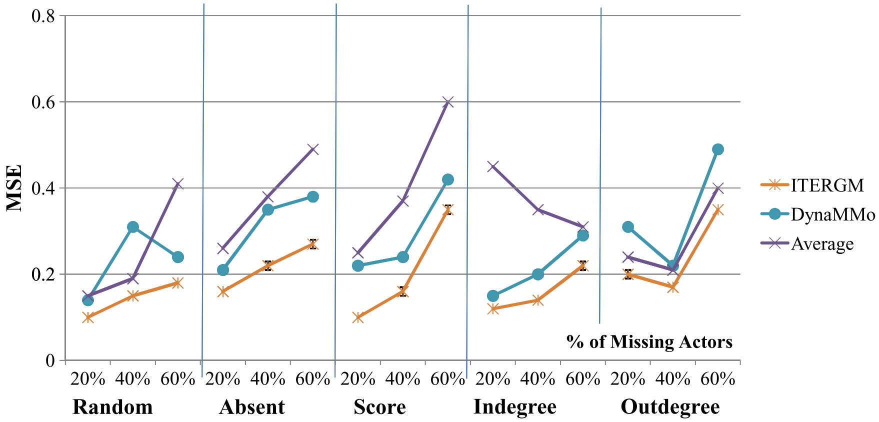


Fig. 6 Features imputation on the synthetic *Dataset1* of 250 actors observed at four time steps (simulating five types of missing mechanisms for 20 %–60 % of missing actors) measured by MSE. The error bars where appropriate show 90 % confidence interval of the result based on five runs of each algorithm

bound to exist within so many various human classes and groups. We posit that any human social network prediction algorithm based on computation of social metrics, including ITERGM, can benefit from identifying local communities and learning model weights on each community separately.

To further characterize our approach, we conducted experiments on eleven synthetic datasets of increasing numbers of actors ranging from 20 to 1,000. All generated datasets were networks observed over four time steps. For each dataset we simulated missing links and attributes by removing 20 % of the actors completely at random (MAR). We applied each of the six link imputation techniques presented in this paper five times on all eleven datasets and calculated the AUC of link imputation accuracy with a 90 % confidence interval (Schafer 1999) for each technique. In general, the imputation accuracy had decreased for all techniques as the number of actors increased (see Fig. 7). However, in all experiments ITERGM was much more accurate than any of the alternative techniques.

5.2 Real life datasets

We conducted an exhaustive set of experiments on two well-known real life datasets. The first dataset, *Delinquency* (Snijders et al. 2009), consists of four temporal observations at intervals of three months between September 2003 and June 2004 of 26 students aged between 11 and 13 in a Dutch school class. At each wave panel researchers asked pupils to identify their 12 best friends. At the same time, researchers recorded the delinquency score, a five point measurement ranging from 1 to 5 as an average number of the delinquent incidents—stealing, vandalism, graffiti, and fighting. The score from 1 to 5 was assigned based on the number of incidents over the last three months: 1 = never, 2 = once, 3 = 2 to 4 times, 4 = 5 to 10 times, 5 = more than 10 times. The second dataset, *Teenagers* (Michell and Amos 1997), is a temporal observation of 50 girls, aged 13, in a school in the West of Scotland over a three year period starting in 1995 and ending in 1997, done in three wave panels. Similarly to *Delinquency*, researchers asked pupils to identify their friends. Also, at each observation the teenagers' alcohol consumption score was compiled. This measurement was also defined on 5 point scale: 1 = none, 2 = once or twice a year, 3 = once a month, 4 = once a week, and 5 = more than once a week. On the *Delinquency*

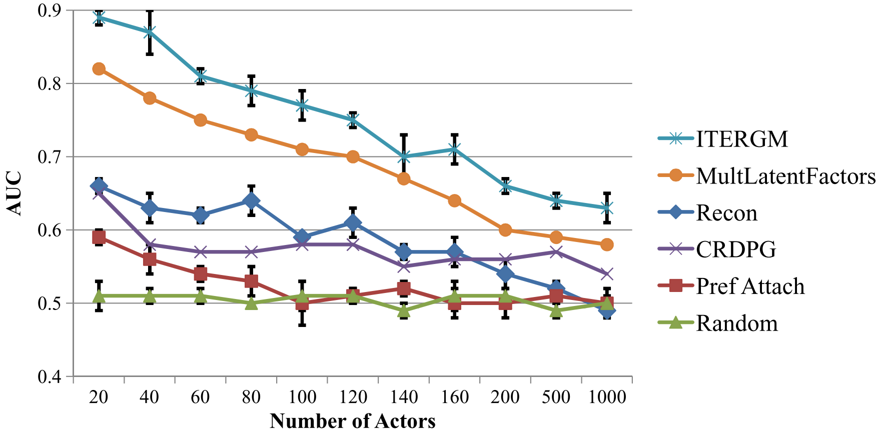


Fig. 7 Comparison of the accuracy of link imputation techniques measured in AUC vs. the number of actors on eleven synthetic datasets of increased size. All datasets consist of 4 time steps and the missing data is modeled by randomly removing 20 % of actors at each time step. The errors bars, where appropriate, show 90 % confidence interval of the result based on five runs of each algorithm

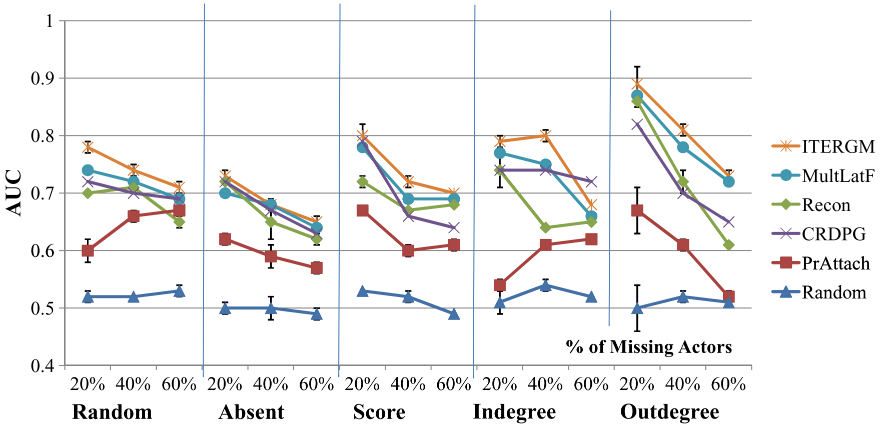


Fig. 8 Links imputation on the *Delinquency* dataset (simulating five types of missing mechanisms for 20 %–60 % of missing actors) measured by AUC. The error bars where appropriate show 90 % confidence interval of the result based on five runs of each algorithm

dataset ITERGM had achieved convergence on average in three iterations, and four iterations on *Teenagers*. We present the results of the experiments on both real life datasets in Figs. 8, 9, 10 and 11. In both real life datasets the ITERGM performed well on the links and attributes' imputation as compared to the baselines. We observed many overlaps in the confidence intervals of the attribute imputation accuracies in the *Delinquency* dataset. However, in many of these instances the confidence intervals of the baseline techniques are rather large whereas the ITERGM is more precise (for example see the experiment of MNAR-score, 60 % missing actors). Results on *Teenagers* were notably better, with less overlaps of the confidence intervals. The CRDPG and "Reconstruction" link imputation algorithms also

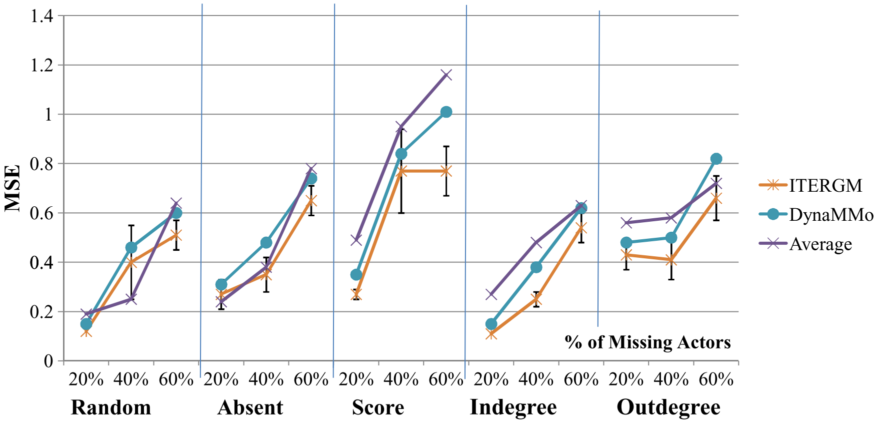


Fig. 9 Features imputation on the *Delinquency* dataset (simulating five types of missing mechanisms for 20 %–60 % of missing actors) measured by MSE. The error bars where appropriate show 90 % confidence interval of the result based on five runs of each algorithm

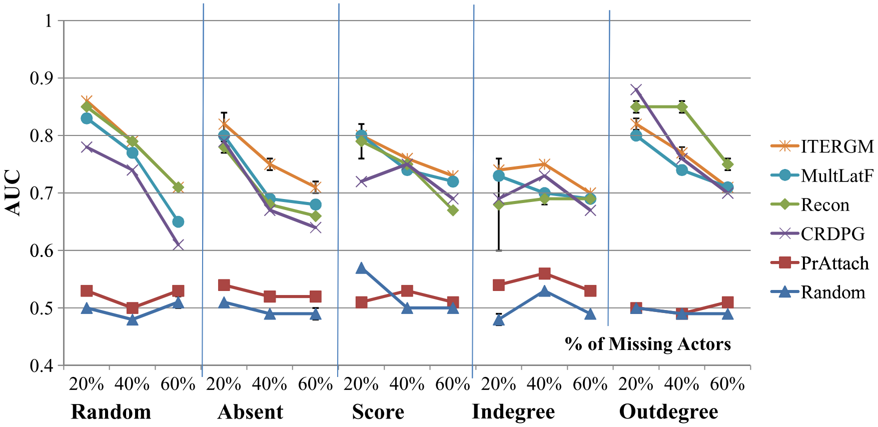


Fig. 10 Links imputation on the *Teenagers* dataset (simulating five types of missing mechanisms for 20 %–60 % of missing actors) measured by AUC. The error bars where appropriate show 90 % confidence interval of the result based on five runs of each algorithm

had good results and in almost all cases were the second best choices. Just as we expected, the "Random" algorithm performed poorly (AUC values are close to 0.5).

The objective of our paper is to introduce a new imputation technique for social network surveys and evaluate its accuracy. Here we are mostly concerned with how various imputation techniques compare in terms of precision. However, it is also interesting to see how most accurate imputation methods reviewed in this paper affect network statistics. To this end we conducted an additional experiment on the real life dataset *Teenagers*. We investigated how imputations affect outdegree

$$\text{outdegree}(i) = \sum_j A_{ij} \tag{16}$$

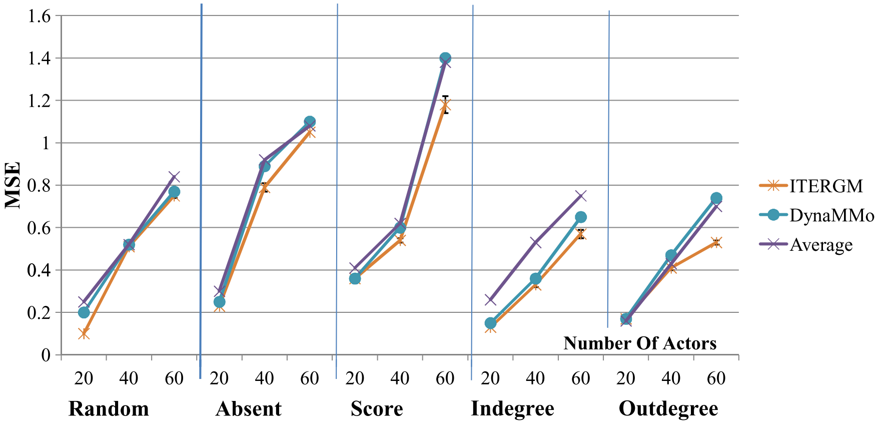


Fig. 11 Features imputation on the *Teenagers* dataset (simulating five types of missing mechanisms for 20 %–60 % of missing actors) measured by MSE. The error bars where appropriate show 90 % confidence interval of the result based on five runs of each algorithm

and reciprocity

$$\text{reciprocity}(i) = \sum_j A_{ij} A_{ji} \tag{17}$$

of the nodes. Outdegree indicates a student's level of social activity. We expect socially active students to have large outdegrees. Reciprocity measures how a student's friendships are reciprocated, with high reciprocity indicating that friendship feelings between a student and her peers are mostly mutual. In our experiment we removed 20 students from the *Teenagers* dataset using a MAR "absent" mechanism. We imputed missing outgoing links using ITERGM, CRDPG and Multiplicative Latent Factors imputation techniques by running each method twenty times and then taking the median of the outdegree and reciprocity statistics for each student at the third wave panel of the dataset. Boxplots shown in Fig. 12 for outdegree and Fig. 13 for reciprocity represent the results of this experiment. The first boxplot in both figures is the true statistics distribution of the students modeled as non-respondents. The following boxplots, two through four of each graph, are corresponding outdegree and reciprocity statistics distributions recovered by three imputation techniques. We note the all three techniques did a great job recovering the distribution of outdegree statistics, as they all managed to correctly predict the median of the distribution and both ITERGM and Multiplicative Latent Factors were fairly accurate in guessing the distribution range. From this we draw the conclusion that our imputation technique ITERGM can accurately guess the social activity level of the missing students. Figure 13 shows that recovery of reciprocity is a harder task, as all three imputation techniques overshoot the median estimate of the reciprocity, although ITERGM does that to a lesser degree. However, the overall recovery of missing students' reciprocity is good to fair, which is encouraging. This and other experiments described in this section suggest that our proposed technique ITERGM can be used in practice by social scientists.

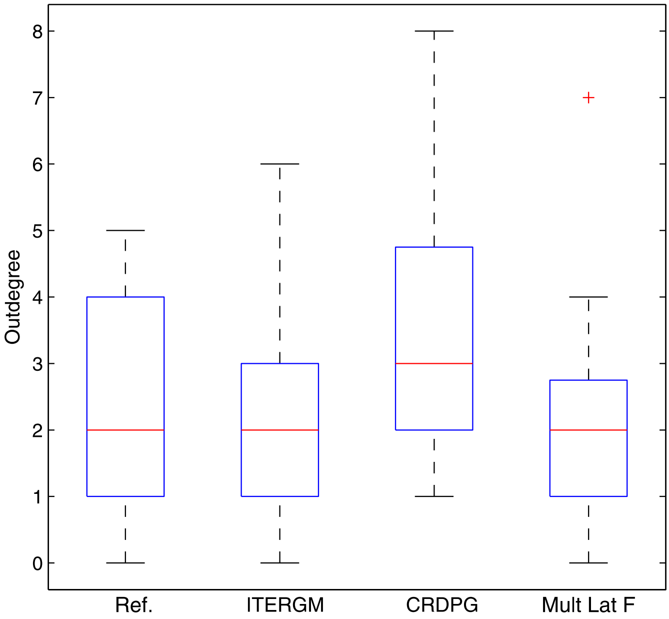


Fig. 12 Boxplots of 20 non-respondent students' outdegrees from the third time step of the real life dataset *Teenagers*. Non-respondents were modeled as missing at random. First boxplot (Ref) is ground truth of observed outdegrees. Last three boxplots are outdegrees recovered by our technique ITERGM and two other baseline models (CRDPG and Multiplicative Latent Factors model)

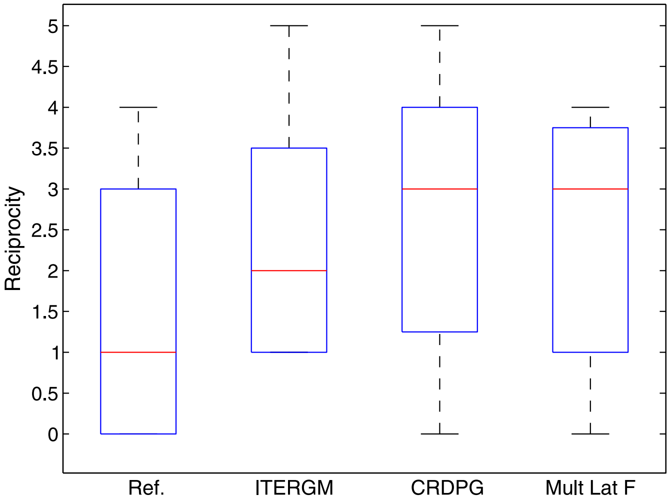


Fig. 13 Boxplots of 20 non-respondent students reciprocity statistics from the third time step of the real life dataset *Teenagers*. Non-respondents were modeled as missing at random. First boxplot (Ref) is ground truth of observed reciprocities. Last three boxplots are reciprocities values recovered by our technique ITERGM and two other baseline models (CRDPG and Multiplicative Latent Factors model)

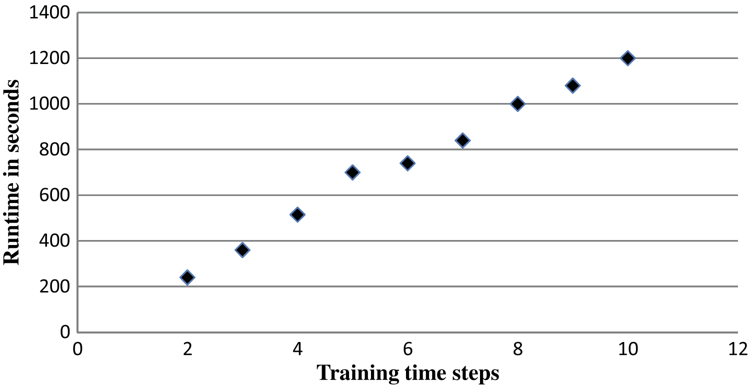


Fig. 14 ITERGM runtime in seconds vs. number of surveys

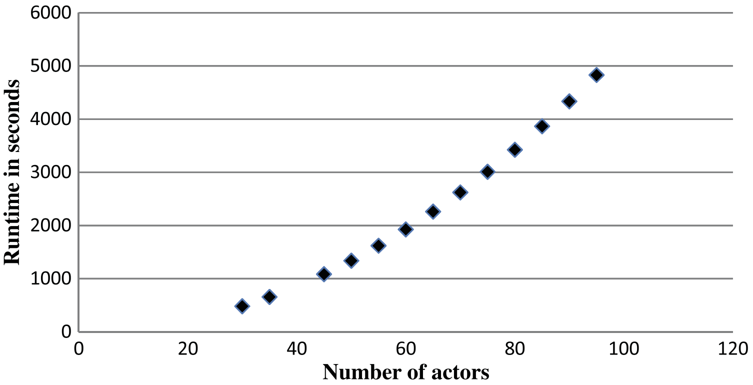


Fig. 15 ITERGM runtime in seconds vs. number of actors

6 Scalability

We investigated the runtime of ITERGM based on two sets of experiments. In one experiment we have created a synthetic dataset with 30 actors and 10 time steps. We ran ITERGM to impute this dataset on a increasing number of time Steps from 2 to 10 and recorded the time in seconds it took the algorithm to run. In Fig. 14 we present the result of this experiment. Here, we clearly observe a linear trend of algorithm runtime in terms of number of survey panels. We conducted a similar experiment on 4 survey panels. This time we held the number of surveys constant but were increasing the number of actors from 30 to 100 in 5 actor increments. We ran the imputation algorithm on the resulting dataset and recorded the time in seconds it took to run. The results of this experiment are shown in Fig. 15. In this experiment, we observed the quadratic term of algorithm runtime in terms of number participating actors. The quadratic scalability in terms of number of actors is not surprising because the algorithm has to consider $k^2 - k$ number of relationships (k is the number of actors in the social networks).

Here we provide formal analysis of ITERGM's runtime behavior. The runtime of ALGORITHM 1 (Sect. 4) can be expressed as:

$$C_1 + \xi\big(C_2 + (T-1)k\iota + C_3 + (T-1)k(k-1)\zeta\big) \tag{18}$$

In (18) we denote the algorithm's initialization Steps 1–5 as constant C_1. We designate variable ξ as the number of iterations needed to achieve convergence, which is basically the number of iterations of the outer loop (Steps 7–19). Within the outer loop we approximate as constants C_2 and C_3 the times required to learn the attribute and link prediction models' weights (Steps 7 and 14 respectively). The expression $(T-1)k\iota$ denotes the runtime of the actors' features sampling. Here we iterate over $T-1$ historical transitions and from each transition step we draw ι samples for all k actors (Steps 8–12). The expression $(T-1)k \times (k-1)\zeta$ denotes sampling of all possible $k(k-1)$ links for each of $T-1$ transitions, where for each link we draw ζ number of samples (Steps 15–19). The limiting behavior of (18) in terms of the number of temporal surveys is linear, or $O(T)$. This is the behavior which was observed in the experiment presented in Fig. 14. However, in terms of the number of actors, the limiting behavior is quadratic, $O(k^2)$, observed in Fig. 15.

The biggest dataset on which we had conducted experiments was a synthetic dataset containing 250 actors, and it is possible to go higher than that. However, it was mentioned in the literature (Snijders et al. 2009) that networks containing more than a few hundred nodes are impractical for real life social surveys. Consider a class which has 500 students. If we ask each student of such class to identify all their friends, we know that the density of such a network will be very low. This is because we would expect each student to have at most twenty or thirty friends and many students will have much fewer friends than that. Human beings are naturally incapable of effectively handling hundreds of close social relationships at the same time. Therefore, the structural 0's in such a network most of the time will indicate that two pupils never had the opportunity to get to know each other rather than a social dislike. Such observation holds true for all large networks. For example, we know for sure that users of Facebook are not aware of the existence of the majority of other users and will never establish links with them. One possible way to scale our approach to the large network is first to apply a community detection algorithm and split the social graph into a collection of subgraphs of manageable size, allowing our algorithm to run in parallel on each subgraph to impute missing data. This opens the possibility of using our method to impute surveys obtained from large electronic datasources such as the one described in Eagle et al. (2009). In this work we do not investigate this possibility, but it is definitely a worthwhile direction for future research.

Recent work (Hanneke et al. 2010) on tERGM, a predecessor of etERGM, had suggested the use of a specially factorized statistics which are expressed in the form:

$$\boldsymbol{\Psi}\big(A^t, A^{t-1}\big) = \sum_{ij} \boldsymbol{\Psi}_{ij}\big(A^t_{ij}, A^{t-1}\big) \tag{19}$$

The use of such statistics avoids the expensive Gibbs sampling process. Instead, it is replaced by direct computation of the expectation step of the learning algorithm. etERGM, which is a cornerstone of our approach, introduces statistics which cannot be factorized into the form specified by (19). These statistics, such as (10), model a homophily selection process often found in social networks. Homophily is an essential part of etERGM, while the factorization method in (19) restricts its expression. This makes it implausible to use such factorization in etERGM and also in our approach.

There are techniques, such as fast approximation, which could speed up the learning and inference process of our approach. However, before applying such techniques one should consider the specifics of collecting the social network surveys. The real life datasets in our paper and other similar real life datasets have one thing in common. It takes months, sometimes years, to collect the temporal social data. Such long collection time is necessary because the social links between humans are relatively stable and do not change that often. Therefore, the prevalent time step granularity of such surveys ranges from months to years. Given the time and cost it takes to conduct the surveys, the imputation accuracy has much greater importance than speed of imputation, as long as the imputation algorithm takes a reasonable time to run. Our proposed approach runs in the order of minutes for the most common temporal social surveys, and it is more than adequate. Hence, we have avoided fast approximation lest the accuracy of the imputation suffer.

7 Discussion and future work

Our proposed imputation technique takes a discrete modeling approach of longitudinal social networks. It only considers discrete time steps when a social network was surveyed by observers. Our approach is different from the continuous-time view of network evolution described in Koskinen and Snijders (2007) and Steglich et al. (2010). Even though models presented in Koskinen and Snijders (2007) and Steglich et al. (2010) cannot be applied directly for data imputation, it is important to discuss the fundamental differences between discrete and continuous-time views of longitudinal networks. In continuous-time models, the network evolution is described as actor-driven and at one microstep at a time. It assumes that network events occur in the infinitely small time intervals where at least one link is added or deleted at a time. The continuous-time view posits that despite the fact that a social network was observed at discrete times there could have been other latent changes which were not seen by the observer. For example, a network was surveyed at times t and $t+1$ and during this time frame a link creation was observed between two actors i and j with similar characteristics (non-drinkers in this example). Such link creation between similar actors will be indisputably labeled as homophily selection by a discrete observer. It is possible however, that during this time frame actor i became a drinker, which prompted actor j to start a therapeutic relationship with him, thus resulting in a social link. Eventually, actor i quit drinking and at the time $t+1$ we observed two similar people in a social relationship. The fact that complex social dynamics took place in-between observations was completely lost. The continuous-time approach provides some assurance that it might capture some of the unseen dynamics by modeling latent observations. In our previous work (Ouzienko et al. 2011) we sought to address questions posed by continuous-time modeling. The main problem of continuous-time models is that it is virtually impossible to go back and verify interactions that might have taken place. It was shown that retrospective studies on the matter are unreliable and prone to error (Bernard et al. 1984). Also, if the longitudinal network is not changing too slowly or too quickly between discrete observations, it provides an additional assurance that a discrete computation model will capture network dynamics adequately (Snijders et al. 2009). The temporal social networks used in this study are indeed of satisfactory quality, as their link change rate is within an acceptable heuristic range (Ouzienko et al. 2011; Snijders et al. 2009).

In Shalizi and Rinaldo (2013) it was suggested that ERGM models in general suffer from undersampling problem. That is, unless the ERGM's model statistics describe a simple dyadic independence (which can be easily factorized akin to (19)) then the model is not useful in making inferences about the larger network from which a studied network sample was drawn. This is because any statistic describing larger motifs than simple dyadic counts, such as "transitivity" statistic ψ_T in (9), cannot have separable increments, and hence the model's projectibility is not attainable. The authors in Shalizi and Rinaldo (2013) recommend instead to treat the unobserved part of the networks as a missing network and suggest the use of an expectation-maximization algorithm. Only one of our statistics, "transitivity", is lacking projectibility power, nevertheless our approach avoids the projectibility issue all together because the learning and inference are done over the whole network graph. The scenario described in Shalizi and Rinaldo assumes that learned model's weights are then applied to the larger graph but in our formulation we do learning and inference over the "same" network structure. Admittedly, the part of our network was not observed but we do know what the complete set of nodes looks like, the only unknown part of the network structure is outgoing links of non-participant nodes and those are randomly initialized at the start of algorithm. Therefore, the model's weights are learned from the whole graph, not part of it, and are never projected to some larger unknown network. To summarize, in our technique we do multiple passes. Within each pass we learn weights from the whole network (its both observed and unobserved parts) and then we apply weights to impute unobserved part. This is done on the same graph until the convergence is achieved and this is what Shalizi and Rinaldo were suggesting to do in their paper.

We demonstrated through the empirical results that our approach can be used as a viable imputation tool for temporal social networks. Our investigation of this technique is not complete because the experiments' results had uncovered new questions which can be addressed in future research. The relationship between the removal technique (MAR or MNAR) used to simulate the unobserved actors and imputation accuracy of ITERGM is not understood. We are currently also expanding our model to handle multivariate and mixed actors' attributes. The disadvantage of our approach is that it cannot infer the first time step. We investigated the possibility of training an ITERGM model on the reversed time sequence of the surveys, so that inference of the first time step would be possible. However, we have encountered the degeneracy of the link prediction model, not uncommon in exponential random graphs (Robins et al. 2006), which we are planning to investigate in our future work.

We believe that our proposed method can be generally applied to any type of network, not only human networks studied here, as long as researchers can derive a set of statistics reasonably describing laws governing the network's evolution. We see the "pluggability" of the network statistics as one of the major advantages of our approach. While we have not attempted to scale our method, we believe it is plausible by way of identifying a large network clusters and then running our model in parallel on each cluster.

Acknowledgements We express special thanks to Dr. Axel Kohlmeyer for his invaluable time and support for this project.

We acknowledge financial support from grant #FA9550-12-1-0406 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA). In addition, this research was supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854.

References

Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The problem of informant accuracy: the validity of retrospective data. *Annual Review of Anthropology*, *13*(1), 495–517.

Borgatti, S., & Molina, J. (2003). Ethical and strategic issues in organizational social network analysis. *The Journal of Applied Behavioral Science*, *39*(3), 337–349.

Bradley, A. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

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

Chang, H., Su, B. B., Zhou, Y. P., & He, D. R. (2007). Assortativity and act degree distribution of some collaboration networks. *Physica A: Statistical Mechanics and Its Applications*, *383*(2), 687–702.

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*(7191), 98–101.

Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, *25*(4), 283–307.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, *39*(1), 1–38.

DeOreo, P. B. (1997). Hemodialysis patient-assessed functional health status predicts continued survival, hospitalization, and dialysis-attendance compliance. *American Journal of Kidney Diseases*, *30*(2), 204–212.

Dunlavy, D. M., Kolda, T. G., & Acar, E. (2011). Temporal link prediction using matrix and tensor factorizations. In *ACM transactions on knowledge discovery from data* (Vol. 5, pp. 1–10).

Eagle, N., Pentland, A., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(36), 1–2.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, *81*(395), 832–842.

Freeman, L. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Gile, K., & Handcock, M. S. (2006). *Model-based assessment of the impact of missing data on inference for networks. css working paper 66*.

Gile, K. J., & Handcock, M. S. (2010). Respondent-driven sampling: an assessment of current methodology. *Sociological Methodology*, *40*(1), 285–327.

Handcock, M. S., & Gile, K. (2007). *Modeling social networks with sampled data or missing data*. Working paper no 75.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29–36.

Hanneke, S., & Xing, E. (2006). Discrete temporal models of social networks. In *Proceedings of the international conference on machine learning workshop on statistical network analysis*, New York: Springer.

Hanneke, S., Fu, W., & Xing, E. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, *4*, 585–605.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, *15*, 261–272.

Huang, Z., & Lin, D. (2009). The time-series link prediction problem with applications in communication surveillance. *Institute for Operations Research and the Management Sciences Journal on Computing*, *21*, 286–303.

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.

Koskinen, J., & Snijders, T. (2007). Bayesian inference for dynamic social network data. *Journal of Statistical Planning and Inference*, *137*(12), 3930–3938.

Koskinen, J. H., Robins, G. L., & Pattison, P. E. (2010). Analysing exponential random graph (p^*) 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*, 247–268.

Li, L., McCann, J., Pollard, N., & Faloutsos, C. (2009). Dynammo: mining and summarization of coevolving sequences with missing values. In *Proc. 15th ACM SIGKDD*.

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the 12th international conference on information and knowledge management*, New York: Assoc. Comput. Mach.

Lu, L., & Zhou, T. (2011). Link prediction in complex networks: a survey. *Physica A: Statistical Mechanics and Its Applications*, *390*(6), 1150–1170.

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

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

Ouzienko, V., Guo, Y., & Obradovic, Z. (2010). Prediction of attributes and links in temporal social networks. In *Proc. euro. conf. artificial intelligence* (pp. 1121–1122).

Ouzienko, V., Guo, Y., & Obradovic, Z. (2011). A decoupled exponential random graph model for prediction of structure and attributes in temporal social networks. *Statistical Analysis and Data Mining*, *4*(5), 470–486.

Robins, G., Elliott, P., & Pattison, P. (2001a). Network models for social selection processes. *Social Networks*, *23*(1), 1–30.

Robins, G., Pattison, P., & Elliott, P. (2001b). Network models for social influence processes. *Psychometrika*, *66*(2), 161–189.

Robins, G., Pattison, P., & Woolcock, J. (2004). Missing data in networks: exponential random graph (p^*) models for networks with non-respondents. *Social Networks*, *26*(3), 257–283.

Robins, G., Snijders, T., Wang, P., & Handcock, M. (2006). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, *29*, 192–215.

Sarkar, P., Chakrabarti, D., & Jordan, M (2012). Nonparametric link prediction in dynamic networks. In *Proceedings of the 29th international conference on machine learning (ICML'12)* (pp. 1687–1694). New York: Omnipress.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*(1), 3.

Schafer, J. L., & Graham, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Shalizi, C. R., & Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *The Annals of Statistics*, *41*(2), 508–535.

Snijders, T. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, *3*(2), 1–40.

Snijders, T. (2005). Models for longitudinal network data. In *Models and methods in social network analysis* (pp. 215–247). New York: Cambridge University Press.

Snijders, T., Steglich, C., & Van de Bunt, G. (2009). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, *32*, 44–60.

Snijders, T., Van de Bunt, G., & Steglich, C. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, *32*(1), 44–60.

Steglich, C., Snijders, T. A. B., & Pearson, M. (2010). Dynamic networks and behavior: separating selection from influence. *Sociological Methodology*, *40*(1), 329–393.

Stomakhin, A., Short, M. B., & Bertozzi, A. L. (2011). Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, *27*(11), 115013.

Stork, D., & Richards, W. (1992). Nonrespondents in communication network studies. *Group & Organization Management*, *17*(2), 193–209.

Van den Berg, G. J., Lindeboom, M., & Dolton, P. J. (2006). Survey non-response and the duration of unemployment. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, *169*(3), 585–604.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, *393*, 440–442.