Imputation of Missing Links and Attributes in Longitudinal Social Surveys

Vladimir Ouzienko and Zoran Obradovic
Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, PA 19122, USA
zoran.obradovic@temple.edu

***Abstract*—We propose a unified approach for imputation of the links and attributes in longitudinal social surveys which accounts for changing network topology and interdependence between the actor's links and attributes. The previous studies on the treatment of non-respondents in longitudinal social networks were mostly concerned with imputation of the missing links only or imputation effects on the networks statistics. For this study we conduct a set of experiments on synthetic and real life datasets with 20%-60% of nodes missing under four mechanisms. The obtained results were better than when using alternative methods which suggest that our method can be used as a viable imputation tool.**

***Keywords*-imputation, temporal data analysis, social networks, exponential random graph models**

I. 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.) is 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. 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 were 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 effect the estimation of network properties [1] and underestimation of the network ties strength [2]. The heightened sensitivity of SNA to the missing data as compared to the less structured, non-network datasets raises the importance of accurate imputation models. 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 [3]. Study of such effects is a separate research topic which is not addressed here because we are more interested in comparing the imputation accuracies. Nevertheless, it is important to understand the impacts of imputation, which is why later on in Section IV-C we present an analysis on how our approach affects statistics of the imputed networks.

Preliminaries on the imputation approaches in social networks are introduced in Section II of this article. A brief review of related models and relevant baselines is presented at Section III. In Section IV, 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 Section V. In Section VI the scalability of our approach is discussed. The paper ends with discussion and some directions for future work in Section VII.

II. Preliminaries

A popular way to present a temporal sequence of the social panel surveys is to use a series of binary adjacency matrices (also called sociomatrices) [4]. Each $k \times k$ sociomatrix (where k is a number of actors) at a given observation time t, denoted as N^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 $N^t_{ij} = 1$ and absence of the link is denoted by $N^t_{ij} = 0$. For example, if student i at wave panel at time t considered student j to be her friend, we denote such a relationship as $N^t_{ij} = 1$. The links could be reciprocal: $N^t_{ij} = 1$ and $N^t_{ji} = 1$, 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 the 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 $t = 1 \ldots 4$ 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 the real valued attribute 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, because the other participants might indicate a friendship link to her.

More formally, given the sequence of the wave panels surveyed at times $t = 1 \dots T$ denoted as $N^1 \dots N^T$, the corresponding actors attributes denoted as $\mathbf{x}^1 \dots \mathbf{x}^T$ ($\mathbf{x}^t$ is $1 \times k$ real valued vector, k is the number of actors), and the unobserved actors sets $S^1 \dots S^T$ ($|S^t| = m^t$, $0 \leq m^t < k \quad \forall t$) 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.

III. 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 [5], or in the context of longitudinal social networks [6]. In our paper we rely in the set up of our experiments on one of these works [3] because it provided a nice comprehensive foundation on how to treat a 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 will also briefly review the Extended Temporal Exponential Random Graph Model (etERGM), a recently published method by [7] which we leverage extensively in our approach.

A. Link Imputation Techniques

1) Reconstruction.: A simple but powerful approach to reconstruct links in a single stationary network was proposed at [8]. 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 impute the unobserved link as opposite to the one that was observed: $N^t_{ij(imputed)} = N^t_{ji(observed)}$. For ties between the non-respondents impute the link with random probability of the observed network density. Here we define the network density as $d = \frac{\sum_{ij}^k N^t_{ij}}{k(k-1)}$

2) Preferential Attachment.: The Preferential Attachment technique proposed at [9] is based on the assumption that actors with many social links are more likely to be connected to each other. 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 procedure, for each unobserved actor i we randomly draw the outdegree number q_i from 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 $N^t_{ij(imputed)} = 1$ and $N^t_{ij(imputed)} = 0$ to the ones that were not selected.

3) Constrained Random Dot Product Graph.: The Constrained Random Dot Product Graph (CRDPG) [10] is an imputation technique which models actors as residing in s-dimensional latent space. In this model the dot product of actors pair latent coordinates yields the probability of the link between the two.

4) 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 according to the random probability of the density d 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 Section IV, will bridge this gap.

B. 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 each actor per one time step because our goal was to evaluate our approach on a simpler model. Also, the multivariate datasets are somewhat hard to obtain. In this subsection we will discuss two imputation techniques for missing real valued actor attributes which will serve as baselines in our experiments.

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.

2) DynaMMo.: The DynaMMo algorithm proposed at [11] 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 multivariate variable $\mathbf{x^t}$ (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. First process models the transition probabilities between the time steps in the multivariate latent space. The second process describes emission from the latent space to the observed.

C. The Extended Temporal Exponential Random Graph Model

The Extended Temporal Exponential Random Graph Model (etERGM) [7] is a decoupled link and attribute prediction model that considers not only prediction of links in temporal networks previously suggested at [12], but it 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 $N^1 \dots N^T$ surveyed at times $t = 1 \dots T$ and actor attributes $\mathbf{x}^1 \dots \mathbf{x}^T$, etERGM predicts the network structure N^{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 $t = 1 \dots T$.

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

$$P(N^t | N^{t-1}, \boldsymbol{x}^t, \boldsymbol{\theta}) = \\ \frac{1}{Z(N^{t-1}, \boldsymbol{x}^t, \boldsymbol{\theta})} exp\{\boldsymbol{\theta}^{'} \boldsymbol{\psi}(N^t, N^{t-1}, \boldsymbol{x}^t)\} \quad (1)$$

The link prediction model (1) defines the transition from N^{t-1} to N^t and incorporates the dependency of N^t over the attributes $\boldsymbol{x}^t$. In this log-linear model Z is the normalization constant, $\boldsymbol{\psi}$ is a function of $\mathbb{R}_{k \times k} \times \mathbb{R}_{k \times k} \times \mathbb{R}_k \rightarrow \mathbb{R}^l$, $\boldsymbol{\psi}(N^t, N^{t-1}, \boldsymbol{x}^t)$ denotes l-size list of sufficient statistics, which encode interdependence of actors' links and attributes and $\boldsymbol{\theta}$ is parameter vector. The complete list of statistics used in link prediction model is detailed in [7], but we present one of these statistics as an example:

$$\psi_L(N^t, \boldsymbol{x}^t) = k \frac{\sum_{i<j}^k N^t_{ij} N^t_{ji} \mathbb{I}(|x^t_i - x^t_j| < \sigma)}{\sum_{i<j}^k N^t_{ij} N^t_{ji}} \quad (2)$$

Here, ψ_L (2) 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 :

$$P(\boldsymbol{x}^t | \boldsymbol{x}^{t-1}, N^t, \boldsymbol{\gamma}) = \\ \frac{1}{Z(\boldsymbol{x}^{t-1}, N^t, \boldsymbol{\gamma})} exp\{\boldsymbol{\gamma}^{'} \boldsymbol{\psi}(\boldsymbol{x}^t, \boldsymbol{x}^{t-1}, N^t)\} \mathbb{N}(\boldsymbol{x}^t | V_0, \Sigma_0) \quad (3)$$

It describes the transition of attributes from time $t-1$ to time t, dependent on the network structure N^t at time t.

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). 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.

IV. The Proposed ITERGM Approach

Before we discuss our approach we should explain the Area Under the Curve (AUC) measure and how it is computed to measure link prediction accuracy. Readers familiar with the use of AUC for link prediction in social networks can safely skip next paragraph.

A. AUC Measure of Link Prediction Accuracy

In general, to measure the link prediction accuracy of a temporal network sequence the AUC was used successfully in the past ([13],[14]). The AUC is a preferable measurement in the presence of imbalanced datasets such as social networks where link density is usually low. 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 $\mathbb{S}$:

$$\mathbb{S} = \sum_l N_l \quad (4)$$

Each run of an imputation algorithm results in the binary $|S| \times k$ subset matrix N_l, which contains only imputed outgoing links. Here, S is the set of actors who did not respond to survey (did not indicate their outgoing links) and k is a total number of actors in the network. Thus the resulting score matrix $\mathbb{S}$ contains the probabilities scores of all imputed links. Using such a matrix we can construct a Receiver Optimization Curve (ROC) by moving the *threshold* parameter in small increments from the matrix's $\mathbb{S}$ 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 $\mathbb{S}(i,j) < threshold \; \forall i,j$ and 1 otherwise. Therefore, a binary prediction matrix at the beginning contains all 1s and it contains 0s 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 number of correctly imputed links divided by the total count of true links. False positive rate is number of imputed links which were not in the true target divided by the total count of non-existing links (structural zeros). We construct ROC by using the x-axis for the false positive rate and the y-axis for the true positive. We calculate AUC, bounded between 0 and 1, based on the constructed curve. A perfect imputation algorithm will have AUC=1, and random algorithm will have AUC=0.5. A better predictor always have larger AUC.

B. Proposed Algorithm-ITERGM

The imputation methods we have reviewed so far (Sections III-A and III-B) can either be applied for link or attribute prediction, or completely ignore the temporal aspect of the surveys. The etERGM model (Section III-C) provides many properties we are looking for in our imputation approach: it encodes the interdependence of actors attributes and links, it also considers the time axis in its learning and inference process. Despite all its characteristics, the etERGM cannot be applied directly to impute attributes or links. Its probability models (1) and (3) can only predict the social network structure at the next unobserved time step given all completely observed previous time steps. Our algorithm, named ITERGM, is in essence the Expectation Maximization (EM) algorithm over two Markov Chain Monte Carlo (MCMC) inferences. During Expectation step we draw multiple particles from both link and prediction models (Steps 4 and 6) of etERGM and in interlocking fashion use them to impute/update the dataset. During Maximization (Steps 3 and 5) 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:

Algorithm 1 ITERGM

```
    Input: The sequence of surveys: $N^{1:T}$, $\mathbf{x}^{1:T}$ where links and attributes, corresponding to actor sets $S^{1:T}$ are unobserved: $\mathbf{x}^t(S^t) = \varnothing$ and $N^t(S^t, j) = \varnothing \;\; \forall t, j$
    Output: Imputed links score matrices: $\mathbb{S}^{1:T}$. Imputed actors' attributes: $\mathbf{x}^{1:T}_{imputed}$
 1: Initialize iteration counter: $iter = 1$
 2: Apply DynaMMo (Section III-B) to initialize missing values in $\mathbf{x}^{1:T} \rightarrow \mathbf{x}^{1:T}_{temporary}$
 3: for $t$ in $1 \dots T$ do
 4:     Impute $N^t(S^t, j), \forall j$ with best link imputation technique from Section III-A $\rightarrow N^t_{temporary}$
 5: end for
 6: Train etERGM's attribute prediction model (Section III-C) on $N^{1:T}_{temporary}$, $\mathbf{x}^{1:T}_{temporary}$ to learn weights $\boldsymbol{\gamma}_{iter}$
 7: for $t$ in $2 \dots T$ do
 8:     Sample multiple vectors $\mathbf{x}^t_{inferred}$ from distribution $P(\bar{\mathbf{x}}^t_{inferred}|\mathbf{x}^{t-1}_{temporary}, N^t_{temporary}, \boldsymbol{\gamma}_{iter})$
 9:     for all missing actor $p$ in $S^t$ do
10:         $\mathbf{x}^t_{temporary}(p) = mean(\mathbf{x}^t_{inferred}(p))$
11:     end for
12: end for
13: Train etERGM's link prediction model on $N^{1:T}_{temporary}$, $\mathbf{x}^{1:T}_{temporary}$ to learn weights $\boldsymbol{\theta}_{iter}$.
14: for $t$ in $2 \dots T$ do
15:     Draw multiple networks $N^t_{inferred}$ from posterior distribution: $P(\bar{N}^t_{inferred}|N^{t-1}_{temporary}, \mathbf{x}^t_{temporary}, \boldsymbol{\theta}_{iter})$.
16:     Calculate $|S^t| \times k$ score matrix: $\mathbb{S}^t = \sum N^t_{inferred}(S^t, j), \forall j$
17:     Set $N^t_{temporary}(S^t, j) = bestcut(\mathbb{S}^t), \forall j$
18: end for
19: if $iter >$ maximum number of iterations then
20:     $\mathbf{x}^{1:T}_{imputed} = \mathbf{x}^{1:T}_{temporary}$
21:     return
22: else
23:     $iter = iter + 1$
24:     go to Step 6
25: end if
```

The input of the algorithm 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 Section III-A 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 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 N^t is 0.2. We impute the unobserved part of network N^t by applying every algorithm described in Section III-A and record the resulting link density of the unobserved part. To initialize links in N^t we pick the algorithm with computed density of the unobserved part closest to 0.2.

At this point, all links and attributes of all networks have been initialized and we begin our iterative approach. In Steps 6-12 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 13, we learn the weights of the etERGM link prediction model by training it on the dataset we have just updated with imputed actors' attributes in interlocking fashion. Knowing the weights, we draw multiple samples from the link prediction model and use them to impute the outgoing missing links (Steps 14-18). In Step 19 we check if we reached number of maximum iterations. If the number of maximum iterations is not reached, we continue the learning/inference process constantly updating the dataset over the set of unobserved actors in interlocking fashion and re-learn the weights of etERGM. Otherwise, the score matrices in Step 16 are our prediction of the imputed links and $\mathbf{x}^{1:T}_{temporary}$ is our imputed temporal sequence of actors' attributes. 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 8, for each survey we sample multiple particles (actors' attributes vectors) based on the weights $\boldsymbol{\gamma}_{iter}$ 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 9-11). The inference of the links is a little bit more involved. Similarly to the imputation of actors' attributes we sample multiple sociomatrices for every survey based on the present learned weights $\boldsymbol{\theta}_{iter}$ of the link prediction model (Steps 14-15). We take the predictions of the imputed values in the form of the score matrices $\mathbb{S}^t$ by adding drawn samples in Step 16. 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 hold the relative probability scores of possible links, into binary form and use that to update the missing part of network (sociomatrices are always binary). That is why in Step 17 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 15. This is achieved by moving the *threshold* from the score matrix's smallest to its largest value. Each resulting binary imputation matrix is substituted into a link prediction model and we pick the best matrix ("cut") which maximizes the link prediction probability.

C. 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 [15]. It works as following: 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

$$t_k = \frac{E_\theta(\psi_k) - \psi_0}{SD_\theta(\psi_k)} \qquad (5)$$

In [15] 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.

Table I
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
ψ_{links}	0.74	2.12	0.35
ψ_{sim}	-0.59	2.09	-0.28
ψ_{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

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 I we present the averages of the differences between the true statistics ψ_0 and statistics based on the imputed samples, the standard deviation of the differences and corresponding t-ratios. The etERGM statistics $\psi_{links}, \psi_{sim}, \psi_{dyads}, \psi_D, \psi_S, \psi_R, \psi_T$ shown in Table I correspond to the measurements of homophily, attributes' stability, similarity, density, links stability, reciprocity and transitivity [12],[7].

In Table I 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 than true values.

V. EXPERIMENTS

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 [3]. 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 [16]. We run our experiments by simulating the removal of the actors according to four missing mechanism: one MAR and three types of MNAR. For MAR, we removed actors at each time step completely at random. For MNAR, we removed the actors according to probabilities of $\frac{1}{(x_i^t)^2}$, $\frac{1}{(1+indegree)^2}$ and , $\frac{1}{(1+outdegree.)^2}$. The first MNAR, that 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 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. To summarize, we model the actors' removal at the four types of missingness and three different percentages, to the total of 12 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 in the previous section. A perfect imputation algorithm will have AUC=1 and a random algorithm will have AUC=0.5, the larger AUC value indicates the better algorithm.

A. Synthetic Dataset

The purpose of the experiments on the synthetic dataset was 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 $t - 1$. The transformation steps from $t - 1$ to t consisted of random link inversions, random link reassignment and completion of the transitive relations. We repeated these transformation steps until the desired number of synthetic networks was generated. To generate the actors' attributes we used an approach similar to the network generation procedure. Here, the transformation steps of actor's attributes from $t - 1$ to t consisted of addition of Gaussian random noise and equalizing attribute values of doubly linked actors.

To 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 five link imputation techniques presented in this paper on all eleven datasets and calculated the AUC of link imputation accuracy for each technique. In general, the imputation accuracy had decreased for all techniques as the number of actors increased (see Figure 1). However, in all experiments ITERGM was much more accurate than any of the alternative techniques.

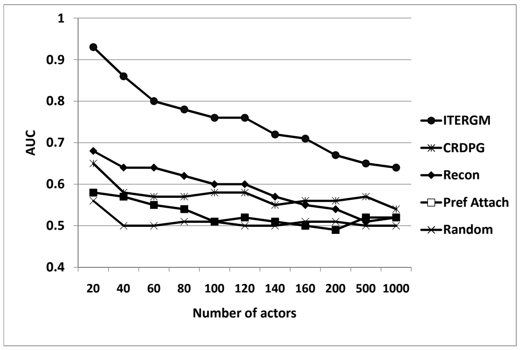

Figure 1. Comparison of the accuracy of link imputation techniques measured in AUC vs. the number of actors on eleven synthetic datasets of increased size. All datasets consist of 4 time steps and the missing data is modeled by randomly removing 20% of actors at each time step.

B. Real Life Datasets

We conducted exhaustive set of experiments on two well known real life datasets. The first dataset, *Delinquency* [17], consists of four temporal observations of 26 students 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 of the delinquent incidents. The second dataset, *Teenagers* [18], is a temporal observation of 50 teenagers 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 from 1=none to 5=more than once a week. On the *Delinquency* 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 Tables II and III. 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

Table II

Links and attributes imputation on the *Delinquency* dataset (simulating four types of missing mechanisms for 20%-60% of missing actors): 90% confidence interval of the AUC and MSE. **Bold** denotes the best result, the underline represents the overlap of the confidence interval with the best result.

Type	%	Link Imputation AUC					Attributes Imputation MSE		
		Random	PrAttach.	Reconstr.	CRDPG	ITERGM	Average	DynaMMo	ITERGM
random	20	0.52±0.01	0.60±0.02	0.70±0.00	0.72±0.00	**0.78±0.01**	0.19±0.00	0.15±0.00	**0.12±0.00**
	40	0.52±0.00	0.66±0.01	0.71±0.01	0.70±0.01	**0.74±0.01**	**0.25±0.00**	0.46±0.09	0.40±0.15
	60	0.53±0.01	0.67±0.00	0.65±0.01	0.69±0.01	**0.71±0.01**	0.64±0.00	0.60±0.05	**0.51±0.06**
score	20	0.53±0.00	0.67±0.00	0.72±0.01	0.79±0.02	**0.80±0.02**	0.49±0.00	0.35±0.25	**0.27±0.02**
	40	0.52±0.01	0.60±0.01	0.67±0.00	0.66±0.01	**0.72±0.01**	0.95±0.00	0.84±0.45	**0.77±0.17**
	60	0.49±0.00	0.61±0.01	0.68±0.00	0.64±0.02	**0.70±0.00**	1.16±0.00	1.01±0.54	**0.77±0.10**
indegree	20	0.51±0.02	0.54±0.01	0.74±0.03	0.74±0.03	**0.79±0.01**	0.27±0.00	0.15±0.01	**0.11±0.00**
	40	0.54±0.01	0.61±0.00	0.64±0.00	0.74±0.00	**0.80±0.01**	0.48±0.00	0.38±0.03	**0.25±0.03**
	60	0.52±0.00	0.62±0.00	0.65±0.00	**0.72±0.00**	0.68±0.00	0.63±0.00	0.62±0.04	**0.54±0.06**
outdegree	20	0.50±0.04	0.67±0.04	0.86±0.01	0.82±0.06	**0.89±0.03**	0.56±0.00	0.48±0.15	**0.43±0.06**
	40	0.52±0.01	0.61±0.01	0.72±0.02	0.70±0.00	**0.81±0.01**	0.58±0.00	0.50±0.08	**0.41±0.08**
	60	0.51±0.00	0.52±0.01	0.61±0.00	0.65±0.01	**0.73±0.01**	0.72±0.00	0.82±0.14	**0.66±0.09**

Table III

Links and attributes imputation on the *Teenagers* dataset (simulating four types of missing mechanisms for 20%-60% of missing actors): 90% confidence interval of AUC and MSE. **Bold** denotes the best result, the underline represents the overlap of the confidence interval with the best result.

Type	%	Link Imputation AUC					Attributes Imputation MSE		
		Random	PrAttach.	Reconstr.	CRDPG	ITERGM	Average	DynaMMo	ITERGM
random	20	0.50±0.00	0.53±0.00	0.85±0.00	0.78±0.00	**0.86±0.00**	0.25±0.00	0.11±0.00	**0.10±0.00**
	40	0.48±0.00	0.50±0.00	0.71±0.00	0.61±0.00	**0.71±0.00**	0.52±0.00	0.52±0.01	**0.51±0.00**
	60	0.51±0.01	0.53±0.00	0.79±0.00	0.74±0.00	**0.79±0.00**	0.84±0.00	0.77±0.01	**0.75±0.01**
score	20	0.57±0.00	0.51±0.00	0.79±0.03	0.72±0.00	**0.80±0.00**	0.41±0.00	**0.36±0.00**	0.36±0.02
	40	0.50±0.00	0.53±0.00	0.75±0.00	0.75±0.02	**0.76±0.00**	0.62±0.00	0.60±0.01	**0.54±0.01**
	60	0.50±0.00	0.51±0.00	0.67±0.00	0.69±0.00	**0.73±0.00**	1.38±0.00	1.40±0.03	**1.18±0.04**
indegree	20	0.48±0.01	0.54±0.00	0.68±0.08	0.69±0.04	**0.74±0.00**	0.26±0.00	0.15±0.00	**0.13±0.00**
	40	0.53±0.00	0.56±0.00	0.69±0.01	0.73±0.00	**0.75±0.00**	0.53±0.00	0.36±0.00	**0.33±0.01**
	60	0.49±0.00	0.53±0.00	0.69±0.00	0.67±0.01	**0.70±0.00**	0.75±0.00	0.65±0.04	**0.57±0.02**
outdegree	20	0.50±0.00	0.50±0.00	0.85±0.01	**0.88±0.01**	0.82±0.01	**0.16±0.00**	0.17±0.00	0.17±0.00
	40	0.49±0.00	0.49±0.00	**0.85±0.01**	0.76±0.00	0.77±0.01	0.43±0.00	0.47±0.00	**0.41±0.01**
	60	0.49±0.00	0.51±0.00	**0.75±0.01**	0.70±0.00	0.71±0.00	0.70±0.00	0.74±0.00	**0.53±0.01**

less overlaps of the confidence intervals. The CRDPG and "Reconstruction" link imputation algorithms also 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).

VI. 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 Figure 2 we present the result of this experiment. Here, we are 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 actors 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 Figure 3. 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 a number of actors in the social networks).

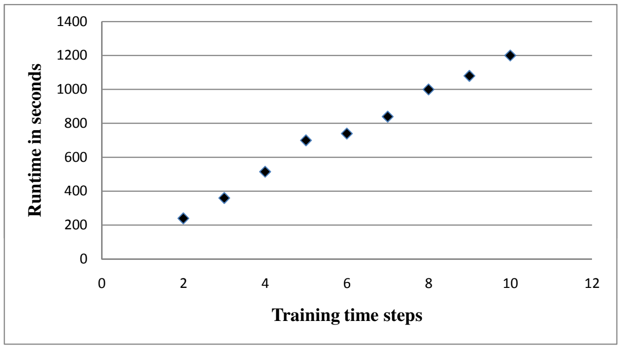

Figure 2. ITERGM runtime in seconds vs. number of surveys

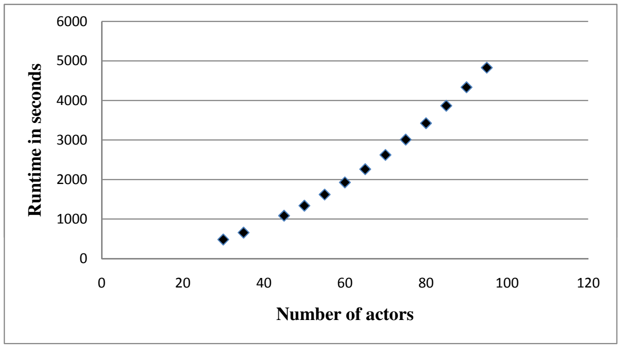

Figure 3. ITERGM runtime in seconds vs. number of actors

VII. Discussion And Future Work

We demonstrated through the empirical results that our approach can be used as a viable imputation tool for the 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 [19]), which we are planning to investigate in our future work.

Acknowledgment

This project is funded in part under a grant to Z. Obradovic with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

This research was supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854.

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

References

[1] S. Borgatti and J. Molina, "Ethical and strategic issues in organizational social network analysis," *Journal of Applied Behavioral Science*, vol. 39, no. 3, pp. 337–349, 2003.

[2] R. Burt, "A note on missing network data in the general social survey," *Social Networks*, 1987.

[3] M. Huisman and C. Steglich, "Treatment of non-response in longitudinal network studies," *Social Networks*, vol. 30, no. 4, pp. 297 – 308, 2008.

[4] O. Frank and D. Strauss, "Markov graphs," *Journal of the American Statistical Association*, vol. 81, 1986.

[5] M. Huisman, "Imputation of missing network data: Some simple procedures," *Journal of Social Structure*, vol. 10, 2009.

[6] T. Snijders, "Models for longitudinal network data," in *Models and Methods in SNA*, 2005.

[7] V. Ouzienko, Y. Guo, and Z. Obradovic, "Prediction of attributes and links in temporal social networks," in *Proc. Euro. Conf. Artificial Intelligence*, 2010, pp. 1121–1122.

[8] D. Stork and W. Richards, "Nonrespondents in communication network studies," *Group and Organization Management*, 1992.

[9] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, 1999.

[10] D. Marchette and C. Priebe, "Predicting unobserved links in incompletely observed networks," *Computational Statistics And Data Analysis*, vol. 52, 2008.

[11] L. Li, J. McCann, N. Pollard, and C. Faloutsos, "Dynammo: mining and summarization of coevolving sequences with missing values," in *Proc. 15th ACM SIGKDD*, 2009.

[12] S. Hanneke and E. Xing, "Discrete temporal models of social networks," in *Proceedings of the International Conference on Machine Learning Workshop on Statistical Network Analysis*. New York, NY, USA: Springer-Verlag, 2006.

[13] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Temporal link prediction using matrix and tensor factorizations," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, pp. 1–10, February 2011.

[14] Z. Huang and D. Lin, "The time-series link prediction problem with applications in communication surveillance," *Institute for Operations Research and the Management Sciences Journal on Computing*, vol. 21, 2009.

[15] T. Snijders, "Markov chain monte carlo estimation of exponential random graph models," *Journal of Social Structure*, vol. 3, 2002.

[16] J. L. Schafer, "Multiple imputation: a primer," *Statistical Methods in Medical Research*, vol. 8, no. 1, p. 3, 1999.

[17] T. Snijders, C. Steglich, and G. van de Bunt, "Introduction to stochastic actor-based models for network dynamics," *Social Networks*, vol. 32, 2009.

[18] L. Michell and A. Amos, "Girls, pecking order and smoking," *Social Science and Medicine*, vol. 44, 1997.

[19] G. Robins, T. Snijders, P. Wang, and M. Handcock, "Recent developments in exponential random graph (p*) models for social networks," *Social Networks*, vol. 29, 2006.