Link Prediction and Missing Data in Social Network Surveys: An Initial Exploration

Taniecea A. Arceneaux[1], Burton H. Singer[2]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 22033
[2]University of Florida, 2055 Mowry Rd, Gainesville, FL 32610

Abstract

An important problem in the analysis of social networks, particularly those modeled using survey response data, is that of non-response. Recent studies have shown the negative effects of missing actors and missing ties on the structural properties of networks [4]. To overcome these problems, we adapt link prediction methodology [6] to the problem of single imputation of item non-response in social network survey data. Empirically, we examine the accuracy of link prediction methods on relational data for social groups in Zablocki's Urban Communes Data Set [13]. Furthermore, we extend the methodology to include the use of supplementary information for link prediction in the case of networks with multiple types of relations on the same individuals.

Key Words: social network analysis, missing data, imputation, link prediction

1. Introduction

Item non-response is an important problem in social network surveys, one that is usually dealt with by context-independent imputation methods [4, 11]. As an alternative to conventional strategies, we use link prediction methodology [6] and context-dependent information to investigate the influence of even very limited behavioral information on the accuracy of imputation of network links. We study an illustrative social network, which is part of Zablocki's Urban Communes Data Set [13]. In this setting, we consider the quality of prediction of missing links when the empirical data is based on only a single type of tie, but where a defensible behavioral assumption is an essential element in the analysis. We then examine the situation in which data on multiple types of tie are available for the same set of individuals. This presents an opportunity to base imputation on linkages for the same individual across two or more types of tie simultaneously. Algebraic semigroup representations of linkages among individuals across two or more kinds of tie are utilized in a link prediction framework to carry out more informed imputation than either the single tie data or the use of conventional context-independent methods will allow.

[1] *This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.*

2. What is Link Prediction?

We begin with the premise that individuals, labeled x and y, are more likely to be linked in a true (directed) network G if there is a large overlap between their sets of neighbors (paths of length 1) in a given observed network $G_0 \subset G$, where G_0 is presumed to have missing links. To formalize an imputation method that is designed to recover G from G_0, we introduce a proximity measure *score*(x, y), where we interpret the highest values to be associated with individuals who are "closer" in the network and therefore have the highest likelihood of being linked. Several different specifications of *score*(x, y) will be considered here. These are as follows,

a. Graph distance: *score*(x, y) = the length of the shortest path between x and y in G_0.

b. Common neighbors: *score*(x, y) = $\left|R(x) \cap R(y)\right|$, the number of neighbors that x and y have in common, where $R(x)$ = the set of neighbors of x in G_0 = {z: <x, z> is a single link connection between x and z in G_0.

c. Jaccard's similarity coefficient: *score*(x, y) = $\dfrac{\left|R(x) \cap R(y)\right|}{\left|R(x) \cup R(y)\right|}$, which is the probability that both x and y have a neighbor in common, for some randomly selected neighbor of either x or y.

d. Preferential attachment: *score*(x, y) = $\left|R(x)\right| \cdot \left|R(y)\right|$. This specification is motivated by the empirical evidence that the probability of a link between x and y is correlated with the product of the number of neighbors of x and y [1, 8].

With any specification, we rank scores in decreasing order of proximity for all pairs (x, y) of vertices that are missing in G_0. Then, we require a threshold s^* such that we impute a link between x and y for the s^* pairs with the highest ranked *score*(x, y) values. Thresholds are best determined from training graphs with the same number of vertices and comparable patterns of connectivity to the graph at hand. In this paper, we set the threshold s^* to be the number of links that exist in the true complete graph G for the set of missing dyads in G_0. Additionally, there are some instances when multiple vertex pairs share the same *score* value. When this occurs, we randomly select the necessary number of vertex pairs to impute.

The above discussion concerns imputation when the available data pertains to only a single type of tie between individuals. When there is relational data on two (or more) types of tie for the same individuals, we need an alternative specification of *score*(x, y)—one that uses the multi-relational information—to impute data for the primary relation of interest. To this end, let X denote the adjacency matrix for the primary relation, and let Y denote the adjacency matrix for the secondary relation. We assume that the existence of a link between two individuals in the secondary relation implies a higher propensity for linkage in the primary relation. Consequently, pairs of individuals are more likely to be linked in the primary relation if they share larger numbers of common neighbors in both relations.

To formalize the notion of common neighbors in this context, we begin with two relations X and Y, represented by graphs $G_0(X)$ and $G_0(Y)$. Then, the adjacency matrices

for the two relations are multiplied to generate compound relations of length 2, namely X^2, XY, YX, and Y^2. The common neighbors of x and y can then be traced through a compound relationship by identifying a vertex z such that $<x, z>$ is a link in the first component, and $<z, y>$ is a link in the second component. Two examples of this kind of linkage are as follows:

1. $<x, z>$ is a link in X and $<z, y>$ is a link in Y, with no $<z, y>$ link in X. Then, z is a common neighbor of x and y in XY.

2. Both $<x, z>$ and $<z, y>$ are links in Y, with neither of these links present in X. Then, z is a common neighbor of x and y in Y^2.

Counting neighbors of x and y in this way allows us to compute $score(x, y)$ by summing across all compound relations of length 2. It is worth noting that in the case of multiple relations, there may be missing linkage data in both the X and Y relations. Our objective here is to use the observed data in Y to reliably impute missing links in X.

3. Examples

3.1 Data Source

We utilize data from the Urban Communes Project, which is a longitudinal, multi-wave panel study of participants from 60 social groups (hereafter referred to as *communes*) first surveyed in 1974 and studied extensively by Zablocki [13] and Bradley [2, 3]. In the Urban Communes Project, a commune is operationally defined as a minimum of three families or five non-blood-related adults who shared, to some degree, a common geographical location, voluntary membership, economic interdependence, and some program of common enterprise (i.e., spiritual, social-psychological, political, cultural, or some combination of these). Six Standard Metropolitan Statistical Areas (Atlanta, Boston, Houston, Los Angeles, Minneapolis-St. Paul, and New York) were selected, which were broadly reflective of the differences in geographic location and urban context in the United States. In each city, a commune census was undertaken by fieldworkers to establish contact with as many communes as possible. From this census, ten communes were chosen from each city that both represented the overall characteristics of the communes enumerated and also reflected the communal differences among the six cities [13].

After the communes were selected, three of the 60 communes were removed from analysis because of their rehabilitative nature—that is, in three communes, membership was not totally voluntary and involved an element of legal coercion [2]. Additionally, a subsequent check on the exact membership disposition of every adult in all communes reduced the actual number of permanent residents to four adults in the case of two communes. Therefore, a total of 57 communes are suitable for analysis, ranging in size from $n = 4$ to $n = 26$ members, with an average of 10 members in each commune. A total of 566 adults, aged 15 years and older, comprised the 57 communes at the time of the first wave of data collection in 1974. The sample is similar to the national population (in 1974) on employment status and class affiliation. However, the smaller family of origin, the greater probability of a middle class background, higher education, and non-blue collar occupation distinguishes the commune sample from the American population as a whole [2].

A sociometric questionnaire served to gather information about the nature and structure of informal relations among the members in each commune. In this instrument, every member was asked to answer a set of questions about his or her relation with each other resident who was 15 years of age or older. The aim was to map all possible dyadic relations among group members across a variety of different types of social bonds. Therefore, each person had to complete a set of standardized questions for $n - 1$ members (where n = the number of permanent adults in the commune). This procedure generated an exhaustive mapping of the $n(n - 1)$ dyadic relations in a commune, for each of 13 types of relational contents. In this study, we consider the *loving* and *improving* relational contents. The questionnaire item pertaining to these relations is as follows,

...indicate if [loving/improving is] involved in your relationship with [alter] by circling the appropriate answer.

Possible responses to these relational questions are "Yes," "No," or "Sometimes." In this paper, we consider unweighted networks, with both "Yes" and "Sometimes" responses indicating the existence of a relation. Since commune members were not required to answer any or all of the relational questions in the sociometric questionnaire, it is possible that a given commune exhibits some degree of non-response. However, it is observed that most communes had a reasonably high response rate. In fact, 46 of the 57 eligible communes exhibited a 25% or lower item non-response rate, and 40 of these exhibited a 15% or lower item non-response rate.

3.2 Imputation Using Data from a Single Type of Tie

Consider a commune with 7 members and complete data (i.e., full response) on the loving relationship L. The network diagram and adjacency matrix are shown in Figure 1.

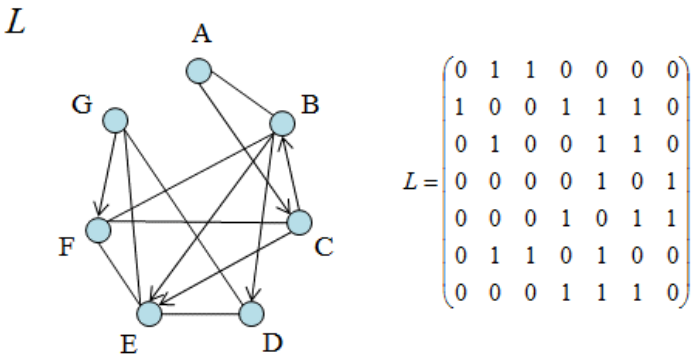
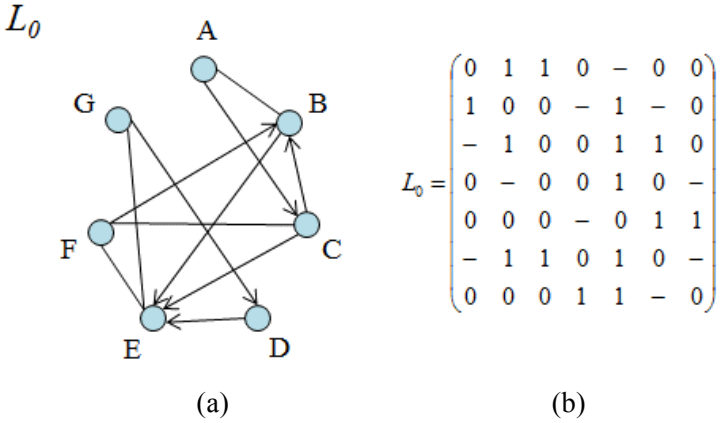


Figure 1: (a) True network diagram for loving relation;
(b) True adjacency matrix

We induce missing data scenarios by assuming that random patterns of 10, 20, and 30 out of the 42 possible directed dyads are unobserved (i.e., missing). In Figure 2 below, we illustrate the "observed" network diagram and adjacency matrix for a random removal of 10 dyads.



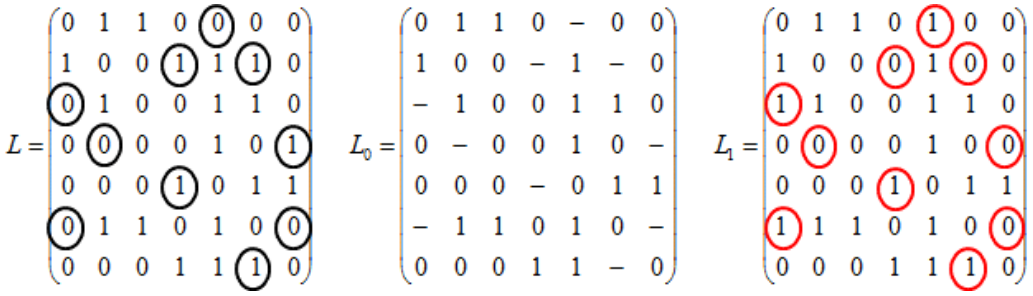
(a) (b)

Figure 2: (a) "Observed" network diagram for loving relation; (b) "Observed" adjacency matrix

1. Start with observed data	2. Set missing values = 0 to obtain L_0	3. Compute $score(x, y)$ = # common neighbors = L_0^2	4. Rank vertex pairs by $score(x, y)$
$L_0 = \begin{pmatrix} 0 & 1 & 1 & 0 & - & 0 & 0 \\ 1 & 0 & 0 & - & 1 & - & 0 \\ - & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & - & 0 & 0 & 1 & 0 & - \\ 0 & 0 & 0 & - & 0 & 1 & 1 \\ - & 1 & 1 & 0 & 1 & 0 & - \\ 0 & 0 & 0 & 1 & 1 & - & 0 \end{pmatrix}$	$L_0 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$	$L^2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$	1. $score(1,5) = 2$ 2. $score(2,6) = 1$ 3. $score(3,1) = 1$ 4. $score(4,7) = 1$ 5. $score(5,4) = 1$ 6. $score(6,1) = 1$ 7. $score(6,7) = 1$ 8. $score(7,6) = 1$ 9. $score(2,4) = 0$ 10. $score(4,2) = 0$

Figure 3: Illustration of common neighbors link prediction methodology on example commune, with 10 dyads missing

In Figure 3, we illustrate link prediction using the common neighbors method. With a threshold $s^* = 5$, we wish to impute the 5 vertex pairs with the highest ranked *score* values. It is clear that a link is imputed for the vertex pair (1, 5) with the highest score value. Then, since there are 7 vertex pairs with $score(x, y) = 1$, we randomly select 4 pairs from among these for which to impute a link in L_0. Figure 4 below illustrates one possible outcome of the link prediction procedure.



(a) (b) (c)

Figure 4: Illustration of common neighbors link prediction methodology on example commune; (a) true adjacency matrix, (b) imposed 10 dyads missing, (c) link prediction result

Since there are multiple ways to randomly select 4 pairs of vertices to impute from the seven pairs with the same *score* value, we consider all possible selections (i.e., 35 possible imputations). To evaluate the efficiency of a given prediction method, we compute a prediction accuracy rate, which is defined to be the percentage of link predictions that are correct, compared to the true network. That is, we compare the predicted values (circled in red) in Figure 4(c) with the true values (circled in black) in Figure 4(a). The average number of correct predictions and prediction accuracy rate from all 35 possible resulting imputations are reported in Table 1 below.

Table 1: Link prediction results for imputation of missing data in loving relation, single tie case

Link Prediction Method	10 dyads missing ($s^* = 5$)		20 dyads missing ($s^* = 6$)		30 dyads missing ($s^* = 8$)	
	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate
RANDOM	5.0000	0.5000	11.6000	0.5800	18.2667	0.6089
Graph Distance	5.0000	0.5000	12.3636	0.6182	22.0000	0.7333
Common Neighbors	4.5714	0.4571	12.3636	0.6182	22.0000	0.7333
Jaccard's Coefficient	5.6000	0.5600	15.0000	0.7500	23.0000	0.7667
Preferential Attachment	4.0000	0.4000	11.2000	0.5600	18.0000	0.6000

3.3 Imputation Using Data from Two Types of Tie

Consider now that we have information about an additional type of tie connecting the same 7 individuals. We wish to use this information in order to better inform our link prediction procedure. As before, we induce missing data scenarios by assuming that random patterns of 10, 20, and 30 out of the 42 possible directed dyads in the loving relation L are unobserved (i.e., missing). Additionally, we introduce the improving relation I, where we assume that the existence of an improving relation implies a propensity toward the existence of a loving relation. In Figure 5 below, we illustrate the complete network diagram for the improving relation I_0, together with the "partially observed" network diagram for the loving relation L_0, which has the same (randomly selected) 10 dyads missing as before.

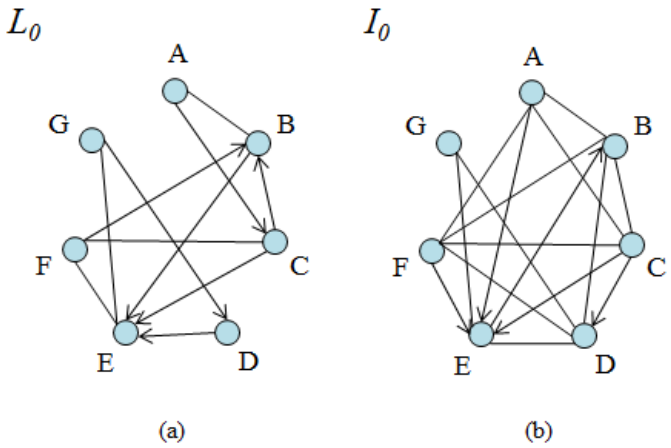


Figure 5: (a) (Partially observed) network diagram for loving relation;
(b) (Fully observed) network diagram for improving relation

In Figure 6 below, we illustrate link prediction using the common neighbors method. With a threshold of $s^* = 5$, we impute the five vertex pairs with the highest ranked *score* values. It is clear that a link is imputed for the vertex pairs (1, 5), (2, 6), and (5, 4) with the highest *score* values. Then, since there are 3 vertex pairs with $score(x, y) = 6$, we randomly select 2 pairs from among these for which to impute a link in L_0.

Again, since there are multiple ways to randomly select 2 pairs of vertices to impute from the three pairs with the same *score* value, we consider all possible selections (i.e., 3 possible imputations) in evaluating the efficiency of link prediction methodology. We then compute the average number of correct predictions and prediction accuracy rate from all 3 possible resulting imputations, which are reported in Table 2 below. Bold values in Table 2 indicate that imputation using information from multiple relations performs at least as well as or better than imputation using data from a single relational tie.

1. Start with observed data

$$L_0 = \begin{pmatrix} 0 & 1 & 1 & 0 & - & 0 & 0 \\ 1 & 0 & 0 & - & 1 & - & 0 \\ - & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & - & 0 & 0 & 1 & 0 & - \\ 0 & 0 & 0 & - & 0 & 1 & 1 \\ - & 1 & 1 & 0 & 1 & 0 & - \\ 0 & 0 & 0 & 1 & 1 & - & 0 \end{pmatrix}, \quad I_0 = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

2. Set missing values = 0 to obtain L_0 & I_0

$$L_0 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}, \quad I_0 = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

3. Compute $score(x, y)$ = # common neighbors = $L_0^2 + L_0 I_0 + I_0 L_0 + I_0^2$

$$L^2 = \begin{pmatrix} 1 & 1 & 0 & 0 & \textcircled{2} & 1 & 0 \\ 0 & 1 & 1 & \textcircled{0} & 0 & \textcircled{1} & 1 \\ \textcircled{1} & 1 & 1 & 0 & 2 & 1 & 1 \\ 0 & \textcircled{0} & 0 & 0 & 0 & 1 & \textcircled{1} \\ 0 & 1 & 1 & \textcircled{1} & 2 & 0 & 0 \\ \textcircled{1} & 1 & 0 & 0 & 2 & 2 & \textcircled{1} \\ 0 & 0 & 0 & 0 & 1 & \textcircled{1} & 1 \end{pmatrix} + LI = \begin{pmatrix} 2 & 1 & 1 & 2 & \textcircled{1} & 2 & 0 \\ 1 & 4 & 2 & \textcircled{2} & 3 & \textcircled{2} & 2 \\ \textcircled{2} & 2 & 2 & 3 & 1 & 1 & 1 \\ 0 & \textcircled{1} & 0 & 2 & 1 & 0 & \textcircled{1} \\ 1 & 2 & 1 & \textcircled{2} & 3 & 1 & 1 \\ \textcircled{2} & 2 & 1 & 3 & 1 & 2 & \textcircled{1} \\ 1 & 3 & 1 & 2 & 2 & \textcircled{1} & 2 \end{pmatrix} + IL = \begin{pmatrix} 1 & 2 & 1 & 2 & \textcircled{3} & 3 & 1 \\ 0 & 3 & 2 & \textcircled{0} & 3 & \textcircled{1} & 1 \\ \textcircled{1} & 2 & 2 & 2 & 3 & 2 & 2 \\ 1 & \textcircled{1} & 1 & 3 & 3 & 3 & \textcircled{1} \\ 1 & 0 & 0 & \textcircled{2} & 3 & 2 & 1 \\ \textcircled{1} & 2 & 1 & 2 & 3 & 3 & \textcircled{2} \\ 0 & 0 & 0 & 1 & 1 & \textcircled{1} & 2 \end{pmatrix} + I^2 = \begin{pmatrix} 3 & 3 & 2 & 4 & \textcircled{2} & 2 & 1 \\ 2 & 4 & 2 & \textcircled{2} & 4 & \textcircled{3} & 1 \\ \textcircled{2} & 4 & 3 & 3 & 3 & 3 & 2 \\ 2 & \textcircled{2} & 2 & 4 & 2 & 1 & \textcircled{1} \\ 1 & 1 & 1 & \textcircled{2} & 2 & 2 & 1 \\ \textcircled{2} & 4 & 2 & 3 & 3 & 4 & \textcircled{2} \\ 0 & 2 & 0 & 1 & 1 & \textcircled{1} & 2 \end{pmatrix}$$

3. (Continued)

$$score = \begin{pmatrix} 7 & 7 & 4 & 8 & 8 & 8 & 2 \\ 3 & 12 & 7 & 4 & 10 & 7 & 5 \\ 6 & 9 & 8 & 8 & 9 & 7 & 6 \\ 3 & 4 & 3 & 9 & 6 & 5 & 4 \\ 3 & 4 & 3 & 7 & 10 & 5 & 3 \\ 6 & 9 & 4 & 8 & 9 & 11 & 6 \\ 1 & 5 & 1 & 4 & 5 & 4 & 7 \end{pmatrix}$$

4. Rank vertex pairs by $score(x, y)$

1. $score(1,5) = 8$
2. $score(2,6) = 7$
3. $score(5,4) = 7$
4. $score(3,1) = 6$
5. $score(6,1) = 6$
6. $score(6,7) = 6$
7. $score(2,4) = 4$
8. $score(4,2) = 4$
9. $score(4,7) = 4$
10. $score(7,6) = 4$

Figure 6: Illustration of common neighbors link prediction methodology on example commune, with 10 dyads missing

Table 2: Link prediction results for imputation of missing data in loving relation, multiple tie case (with zero dyads missing in improving relation)

Link Prediction Method	10 dyads missing ($s^* = 5$)		20 dyads missing ($s^* = 6$)		30 dyads missing ($s^* = 8$)	
	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate
RANDOM	5.0000	0.5000	11.6000	0.5800	18.2667	0.6089
Graph Distance	**5.0000**	**0.5000**	**13.0000**	**0.6500**	20.0000	0.6667
Common Neighbors	4.0000	0.4000	**13.4286**	**0.6714**	**22.2857**	**0.7429**
Jaccard's Coefficient	5.2000	0.5200	13.3333	0.6667	22.0000	0.7333
Preferential Attachment	**4.0000**	**0.4000**	**12.0000**	**0.6000**	**22.6667**	**0.7556**

When multiple relations are present, it is possible that there is missing linkage data in the secondary relation. We consider three missing data scenarios, where we randomly impose patterns of 10, 20, and 30 of the 42 possible directed dyads of the improving relation *I* are missing. Again, the objective is to reliably impute missing data in the loving relation *L*. In each of these cases, we utilize the observed data in *I* to inform our prediction of links in *L*. Tables 3 to 5 outline the link prediction results for each of these missing data scenarios. Again, bold values in Tables 3 to 5 indicate that imputation using information from multiple relations performs at least as well as or better than imputation using data from a single relational tie.

Table 3: Link prediction results for imputation of missing data in loving relation, multiple tie case (with ten dyads missing in improving relation)

Link Prediction Method	10 dyads missing ($s^* = 5$)		20 dyads missing ($s^* = 6$)		30 dyads missing ($s^* = 8$)	
	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate
RANDOM	5.0000	0.5000	11.6000	0.5800	18.2667	0.6089
Graph Distance	4.2857	0.4286	**12.3636**	**0.6182**	19.3333	0.6444
Common Neighbors	**6.0000**	**0.6000**	12.0000	0.6000	**22.3333**	**0.7444**
Jaccard's Coefficient	**6.0000**	**0.6000**	14.0000	0.7000	**24.0000**	**0.8000**
Preferential Attachment	**4.0000**	**0.4000**	10.0000	0.5000	**18.8571**	**0.6286**

Table 4: Link prediction results for imputation of missing data in loving relation, multiple tie case (with twenty dyads missing in improving relation)

Link Prediction Method	10 dyads missing ($s^* = 5$)		20 dyads missing ($s^* = 6$)		30 dyads missing ($s^* = 8$)	
	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate
RANDOM	5.0000	0.5000	11.6000	0.5800	18.2667	0.6089
Graph Distance	4.2857	0.4286	**14.0000**	**0.7000**	20.0000	0.6667
Common Neighbors	4.0000	0.4000	**14.0000**	**0.7000**	**22.6667**	**0.7556**
Jaccard's Coefficient	**4.0000**	**0.4000**	14.0000	0.7000	**24.0000**	**0.8000**
Preferential Attachment	2.0000	0.2000	**12.0000**	**0.6000**	**20.0000**	**0.6667**

Table 5: Link prediction results for imputation of missing data in loving relation, multiple tie case (with thirty dyads missing in improving relation)

Link Prediction Method	10 dyads missing ($s^* = 5$)		20 dyads missing ($s^* = 6$)		30 dyads missing ($s^* = 8$)	
	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate	Average # Correct	Average Prediction Accuracy Rate
RANDOM	5.0000	0.5000	11.6000	0.5800	18.2667	0.6089
Graph Distance	**5.0000**	**0.5000**	**13.3333**	**0.6667**	20.2500	0.6750
Common Neighbors	4.0000	0.4000	11.5000	0.5750	19.2000	0.6400
Jaccard's Coefficient	4.0000	0.4000	12.0000	0.6000	19.0000	0.6333
Preferential Attachment	2.0000	0.2000	8.5000	0.4250	16.0000	0.5333

4. Discussion

It is worth noting that many real world interactions occur (or fail to occur) for reasons independent of network structure. Therefore, the raw performance of some link predictors is understandably low, with the lowest prediction accuracy rate of 0.2000. It is for this reason that we base the quality of our predictions on comparisons with the performance of random imputation—imputing a random subset of vertex pairs from among all missing pairs.

4.1 Imputation Using Data from a Single Type of Tie

By inspection of Table 1 above, we make several key observations about the accuracy of link prediction in our example network, with a single type of tie connecting individuals. In the case of the smallest degree of missingness, most of these link prediction methods perform at or below the level of random imputation, with the exception of Jaccard's coefficient. However, as the degree of missingness increases, each predictor performs increasingly well and increasingly better than random imputation. This suggests that utilizing the network topology in making link predictions is particularly important as the level of missingness in the network increases.

4.2 Imputation Using Data from Two Types of Tie

In some instances, we have supplemental information in the form of additional relations connecting individuals in a network. In Tables 2 through 5 above, we present the results of link prediction for various levels of missingness in the secondary relation.

We make several key observations about the accuracy of link prediction in our example network, now with two types of tie connecting individuals. First, we examine link prediction when we have full information on the secondary relation (Table 2). With the smallest degree of missingness, we observe that most link prediction methods do not perform better than random imputation, except Jaccard's coefficient. However, as the degree of missingness increases, all methods outperform random imputation. This is analogous to the single tie case, reaffirming the importance of utilizing network topology to make link predictions as the level of missingness increases.

We then consider various levels of missingness on the secondary relation *I*. Tables 3 to 5 outline the link prediction results when there are 10, 20, and 30 dyads missing from the *I* relation. In each instance, there is no prior imputation on *I*. Instead, all entries missing in *I* are assumed to be nonexistent for the purpose of this study. Through inspection of Tables 3 to 5, we make several key observations. With 10 dyads missing on *I*, most predictors continue to outperform random imputation at predicting linkages in *L*. In fact, this is also true for the scenario in which there are 20 dyads missing on *I*. However, as the level of missingness in *I* increases to 30 dyads, the efficiency of link prediction methods tends to decrease, with greater numbers of predictors performing worse than random prediction. We anticipate that there exists a threshold of missingness on the secondary relation, at which point the use of additional information is no longer a benefit to link prediction on the primary relation. Investigating this threshold is a topic of further research. However, the improvement upon random imputation in cases where up to half of the data in the secondary relation are missing suggest that the use of even incomplete supplementary information leads to better link prediction outcomes.

4.3 Future Directions

A caveat regarding our exploratory study concerns the fact that only a single random draw of missing links was investigated for counts of 10, 20, and 30 dyads in our example network. A much more extensive analysis should be carried out as a next step, in which many possible missing patterns of sizes 10, 20, and 30 dyads are investigated. We would then ask, "Which pattern(s) of missing data are associated with the best imputation result?" We anticipate that there is a connection between the structure of the original network and the pattern of missing data that is associated with best performance of the link prediction methodology for local imputation.

A second feature of future studies should include imputation on the secondary relation (*I*) prior to imputation on the primary relation (*L*). Here, we investigate whether missingness in the secondary relation is ignorable or non-ignorable. That is, we ask whether there is a benefit to an initial imputation of missing elements in the secondary relation over the use of raw data. To fully explore this question would require that we consider many pairs (*L*, *I*) with diverse patterns of missingness on each relation. Then, we anticipate that there will be special (*L*, *I*) pairs for which link prediction methods are particularly efficient. Understanding of these structural details lies in the future.

5. Conclusions

In this paper, we have presented link prediction methodology as an alternative to conventional and random imputation strategies when faced with the problem of item non-response in social network surveys. We have explored the use of social network topology in making informed prediction of network linkages, and we have examined the

accuracy of local predictions across various imputation methods. Additionally, we have extended these methods to investigate the influence of additional relational information on the accuracy of imputation of network links. An illustrative social network from Zablocki's Urban Communes Data Set [13] serves as an investigative example, where we have considered the quality of prediction of missing links when utilizing both single and multiple relational data. In both instances, the benefits of link prediction methodology for imputation are most apparent in the most severe levels of missingness, suggesting that the use of network topology in making predictions is particularly important when very limited network information is available. An example of this type of application is covert networks such as gangs, criminal networks, and terrorist groups. Because of the sensitive nature of the goals and functionality of covert networks, much information about network linkages is often missing. However, it is particularly important in these instances to make accurate predictions of potential linkages among network members in order to intercept and prevent any attempts at criminal and malicious behavior.

Acknowledgments

TAA thanks the Center for Statistical Research and Methodology at the U.S. Census Bureau for postdoctoral research fellowship support. Thanks also go to Ben Klemens, Joshua Tokle, and Darcy Steeg Morris of the U.S. Census Bureau for helpful comments. BHS thanks the Army Research Office Multidisciplinary Initiative for financial support under MURI Grant #5815-MA-MUR, Prime Award #W91 INF-11-1-0036. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

References

[1] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaboration. *Physica A*, 311(3-4):590-614, 2002.

[2] R. T. Bradley. *Charisma and Social Structure: A Study of Love and Power, Wholeness and Transformation*. Paragon House, New York, NY, 1987.

[3] R. T. Bradley and N.C. Roberts. Network structure from relational data: Measurement and inference in four operational models. *Social Networks*, 11:89-134, 1989.

[4] M. Huisman and C. E. G. Steglich. Treatment of non-response in longitudinal network studies. *Social Networks*, 30:297-308, 2008.

[5] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39-43, March 1953.

[6] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019-1031, 2007.

[7] M. Mitzenmacher. A brief history of lognormal and power law distributions. *Proceedings of the Allerton Conference on Communication, Control, and Computing*, 182-191, 2001.

[8] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review Letters E*, 64(025102), 2001.

[9] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167-256, 2003.

[10] J. L. Schafer and J. W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7:147-177, 2002.

[11] D. Stork and W. D. Richards. Nonrespondents in communication network studies. *Group & Organization Management*, 17:193-209, 1992.

[12] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, New York, NY, 1994.

[13] B. Zablocki. *Alienation and Charisma: A Study of Contemporary American Communes*. The Free Press, New York, NY, 1980.