THE EFFECT OF MISSING TWO-MODE TIE DATA ON PARAMETER ESTIMATION WHEN THE INFLUENCE MODEL IS USED

By

Tingqiao Chen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2020

ABSTRACT

THE EFFECT OF MISSING TWO-MODE TIE DATA ON PARAMETER ESTIMATION WHEN THE INFLUENCE MODEL IS USED

By

Tingqiao Chen

Missing data is a phenomenon that cannot be ignored in network analysis, especially due to the complex nature of network data and the plethora of models in this field. This dissertation studies the effect of missing two-mode tie data on coefficient estimates of the influence mode in two-mode network analysis. A new imputation method based on the log odds of attending events within- vs. outside- cluster is proposed. The new imputation method is compared with the multiple imputation method under the missing at random mechanism. Network data are simulated based on different parameter values, including the network density, number of actors, number of events, and the odds ratio (i.e., clustering effect). Fifty-four unique network settings are examined, and 2000 replicates are generated for each unique setting. The multiple imputation method performs the best in terms of bias, empirical standard error, and root mean square error, partly because the missing data generation mechanism favors the multiple imputation method. The proposed imputation method performs well when there are medium to strong clustering effect in the network.

To my dearest parents
Thank you for supporting me unconditionally for my ten years' Ph.D. study.

To my beloved grandma, who gave me beautiful childhood memories

ACKNOWLEDGMENTS

Many thanks to my advisor and dissertation chair, Dr. Ken Frank, who discussed with me and answered my questions while I worked on my dissertation. I would also like to thank my committee members, Dr. Frank Lawrence, Dr. Ran Xu, and Dr. Kim Kelly. I thank Dr. Lawrence for spending time to help with my R code, making my first conference presentation possible, and encouraging me to finish my dissertation. I appreciate his kindness and generosity. I thank Ran for always answering my questions promptly.

My immense gratitude goes to my parents. I am grateful for their unconditional love and consistent emotional and financial support. I thank my mom for her daily call, which made my dissertation process much easier. I thank my uncle for patiently answering my questions. I thank my cousin for discussing with me the application of Natural Language Processing and Network Analysis. I thank everyone in my extended family.

I am thankful for my friends Hope Akaeze, Ji An, Reene Beaufore, Linling Cai, Cheng Cao, Chi Chang, I-Chien Chen, Shuyi Chen, Talehsa Dokes, Yuexiao Dong, Yue Dou, Xiaofeng Fan, Jingwei Ge, Licia (Shiqing) He, Tao He, Hechuan Hou, Shihai Hu, Amal Ibourk, Dajung Jun, Nai-Cheng Kuo, Yingjie Li, Selene Li, Qingyun Lin, Yingyue Liu, Chuang Liu, Xiao Liu, Ling Liu, Xin Luo, Liyang Mao, Ya Mo, Yinghan Song, Ruoxi Sun, Weiqin Tang, Sarah (Jingyuan) Tian, Xinyi Tu, Heng Wang, Keyin Wang, Jianxun Wang, Jiwen Wu, Jiahui Zhang, Rui Zhang, Jin Zhang, Xuechun Zhou, and Qing Zhu for supporting me throughout the dissertation process. My friend Hope spent countless hours to help me and encourage me, and we studied together on Skype almost every day. My best roommate and friend, Xiaofeng, came to my home and celebrated my birthday with me every year. Yingyue and Chuang logged in study

progress with me every day, which kept me on track. Nai-Cheng, Dajung, Talehsa, and Amal frequently studied with me online. My friend Jingwei Ge, Xiao Liu, Liyang (Maomao), and Qing Zhu always called me to check on me even when I felt stressed, and even when most I could tell them was my everyday Ph.D. student life routines. You reminded me that I was not forgotten. My friend Xuechuan talked to me when I felt most stressed. My friends Ji An, Yuexiao Dong, Tao He, Hechuan Hou, Selene Li, Ling Liu, Liyang Mao, Ya Mo, Yinghan Song, Ruoxi Sun, Sarah (Jingyuan) Tian, Keyin Wang, Jiahui Zhang, Rui Zhang, and Qing Zhu landed me a hand when I needed it. My roommates Xinyi Tu and Jin Zhang helped me take care of my home when I interned in CA. Special thanks to Jin, who fixed things in my home while she was busy with job hunting. Xinyi drove me to the hospital in the middle of the night while I was sick. Xinyi and Cheng drove to the Detroit airport to send my luggage to me so that I could catch the Positive Psychology conference in Miami, FL, in 2019. A friend in need is a friend indeed.

I thank Jessica (Xinyi) Fang and Dr. David Harris, my counselors, who helped me while I worked on my dissertation.

I am thankful for everyone that I have had opportunities to get along with since Jan 2011 (when I started my Ph.D.). This experience made me become a better person.

Finally, I would like to thank myself for not giving up and trying the best to stay positive during my Ph.D. study. Please always remember to be a kind person with a brave heart and a steady soul.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Missing data in statistical analyses have long been a popular research topic. Ignoring missing data or using inappropriate methods to handle missing data usually creates biased parameter estimates and underestimated standard errors. Missing data in network analysis could make the situation more complicated. For instance, the presence of missing network ties hampers not only the ability to describe the network context of actors with missing ties but also the context of the neighboring actors (Huisman, 2014). Most research about missing data in network analyses focus on the effect of missing data on the network's structural properties, for example, indegree/outdegree, reciprocity, transitivity, geodesic distance, etc. Sometimes, besides structural properties themselves, we need to use network data to build statistical models to study how network structures and other attributes affect an actor's behavior or belief. For instance, in two-mode network analysis, the influence model examines how exposure to information on events affected actors' behavior. In the influence model, the network data of who attended which events within or outside one's cluster is used to construct an independent variable, i.e., the exposure term. When there is missing information about who attended which events, it is not a straightforward problem of how missing data on an independent variable affects parameter estimation. It is about how missing information, which, if not missing, is supposed to construct the independent variable, affects parameter estimation.

1.2 Introduction of Network Analysis

1.2.1 Two-Mode Network

Networks are categorized by the number of modes they have, one or two. When the network is one-mode, the data represents a single set of actors and relationships among them. The actors can be of different types, for example, people, organizations, etc. The relationships can be various, such as friendship, co-authorship, lending or borrowing, marriage, etc. When the network is two-mode, there are *two sets of actors* (dyadic two-mode networks) or *one set of actors and one set of events*. The '*one set of actors and one set of events'* type is of concern in this dissertation. It is also called an *affiliation network* or bi-partite network. Actors are measured for attendance at or affiliation with a set of events/activities or organizations/groups. Examples include memberships in a fitness club, people attending meetings, students taking classes, etc. (Lazega et al., 1995).

1.2.2 Clustering for Two-Mode Network Data

We do clustering in network analysis to understand the overall structure of the network, which can then be linked to the specific research question. According to Pesantez-Carera and Kalyanaraman (2016), the idea of clustering/community detection in a two-mode network is that nodes (regardless of its mode, e.g., person vs. event) are partitioned into different clusters such that nodes assigned to the same cluster have a higher density of ties among them than to the nodes from other clusters. In other words, in a network of people and events, people are more likely to attend events in the cluster they belong to than events in the rest of the network. That means people in the same cluster are exposed to similar information because they have a higher probability of attending common events than people from different clusters. Then their behaviors are likely to be affected by the information that they are exposed to. Therefore, the phenomenon of clustering is essential for studying behavior change. The graph below shows high school students' course-taking behavior in a given period. Each big circle represents a cluster; dots

represent students and squares represent courses; a line from a dot to a square indicates that the student took a particular course. From the graph, we can see that students took courses outside their clusters, but they mostly took courses in the clusters they belong to. Students' behaviors are most affected by the courses they took and other students' behavior in their clusters.

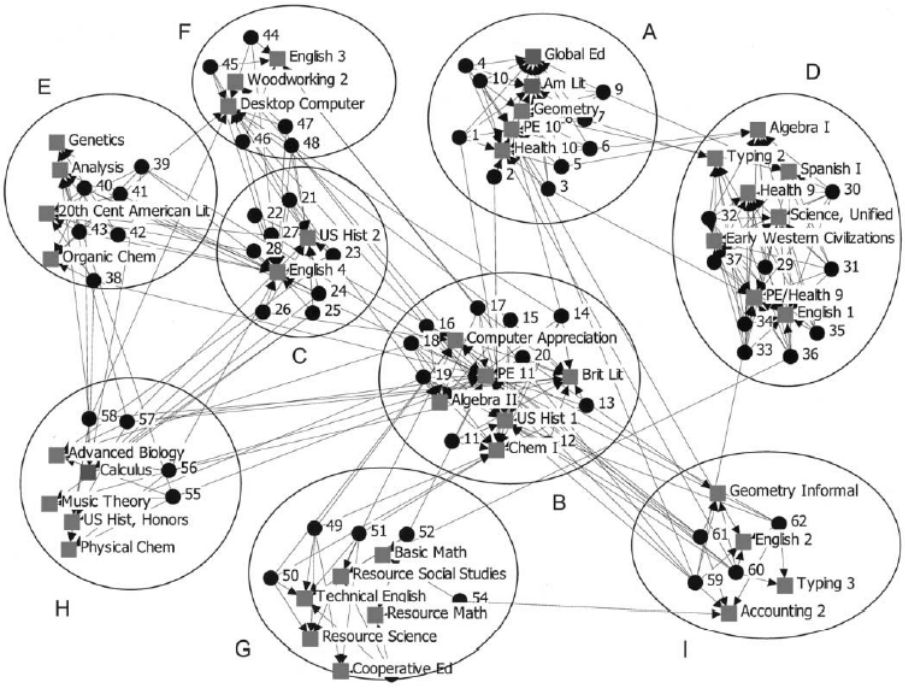

Figure 1 Course-taking pattern in Miller high school (Frank et al., 2008)

1.2.3 Influence Model

With two-mode data, we can use the influence model to explore the relationship between cluster membership and actors' behavior or attitudes. In general, the hypotheses are (1) Actors' behavior or attitudes are related to other actors' behavior or attitudes in the same cluster, and (2) Actors' behavior or attitudes are related to the information presented at events in the same cluster. Two examples of two-mode network analysis using the influence model will be given below.

One example of a two-mode network analysis is to study the dissemination of lake level knowledge in the Great Lakes region. We collected two types of two-mode data, attendance lists of related conferences/meetings and author lists of related documents (e.g., white papers, academic papers, etc.). All conferences, meetings, and documents in our data set are related to climate change. The academic papers could study how lake levels are likely to change in 50 years, etc. At a conference, an actor could be exposed to two types of information, information from the conference itself (e.g., presentation) and information or norms from other actors who attend the conference (e.g., people talking to each other). When several actors co-author a paper, an actor could be exposed to two types of information, the content of the paper itself and information directly from interacting with other co-authors in the process of discussing, editing, etc. The goal is to study how people's opinions change through attending conferences/meetings and co-authoring papers with others.

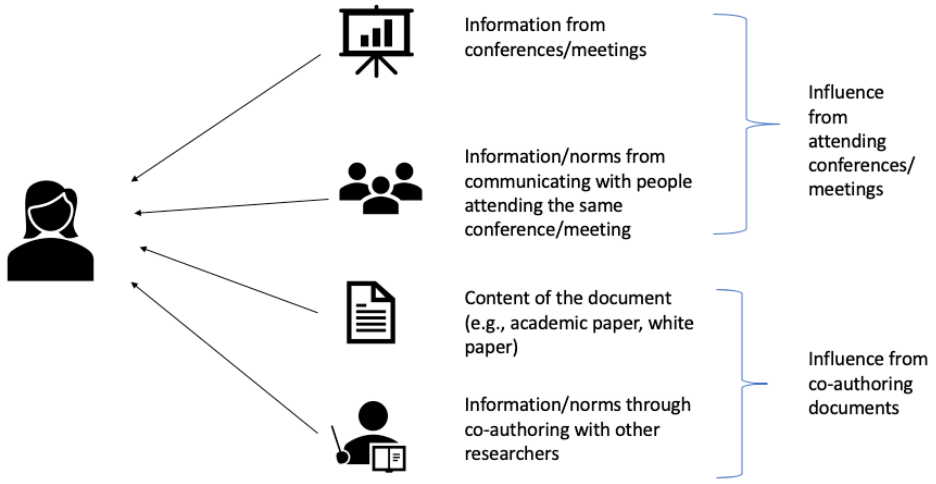

Figure 2 Example of two-mode network analysis

Another example of two-mode network analysis is to examine the course-taking behavior of high school students (Frank et al., 2008). The two-mode data used in the study are transcripts data. It was assumed that students who took the same course had a higher probability of

interacting with/being exposed to each other; for example, they had a higher probability of being classmates, etc. When a student takes a course, he or she could be exposed to two types of information, the course itself (e.g., the instructor, the common course materials used, etc.), and other students who take the same course (e.g., discussing homework together, reviewing for exams together, etc.). The research objective was to study whether other students' math level at time t-1 would influence a student's decision to advance in math or not in the same cluster.

The influence model that will be studied in this dissertation is equation (1).

$$y_{i,t} = \alpha + \rho_0 y_{i,t-1} + \beta_1 \frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1} + \epsilon_{it} \quad (1)$$

$y_{i,t}$ is the dependent variable, representing actor $i's$ behavior at time t. $y_{i,t-1}$ is the prior, representing actor $i's$ behavior at time $t-1$. The term $v_{i,q}^1$ indicates whether actor i attended event q given actor i and event q have the same cluster membership. $z_{q,t-1}$ represents information presented at event q which happened over the time interval from $t-1$ to t. h_i^1 is the number of events actor i attended given that actor i and events are in the same cluster. Therefore, $\frac{1}{h_i^1}\sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1}$ is actor $i's$ exposure to events he/she attended in the cluster he/she belongs to. For explanations of the notation, please look at the appendix A.

1.3 Introduction of Missing Data

Graham (2012) defined *missingness* as the state of being missing. A convenient representation of the state of missingness is a binary variable, *R*, which takes the value 1 if the variable of concern is observed, and 0 if it is missing. For example, $z_{q,t-1}$is the information presented on event q. If $z_{q,t-1}$ is missing, then $R_{z_{q,t-1}} = 0$; if $z_{q,t-1}$ is observed, then $R_{z_{q,t-1}} = 1$. *R* is treated as a set of random variables having a joint probability distribution. The distribution of *R* is referred to as the *missing mechanism*. It is the process by which some data are collected,

while others are missed (Rubin, 1987). Other researchers refer to the missing mechanism as the process causing the missing data (Graham, 2012). The mechanisms of missingness generally fall into three categories: *Missing Completely At Random (MCAR), Missing At Random (MAR),* and *Missing Not At Random (MNAR).*

In a statistical model, either the dependent variable(s) or the independent variable(s) could be missing. This dissertation studies the effect of missing information, which, if not missing, is supposed to construct the independent variable on parameter estimations in the influence model.

1.3.1 Missing at Random

Rubin (1976) defined missing data as MAR if the distribution of missingness (i.e., distribution of R) does not depend on Y_{mis}. In other words, under MAR, the distribution of missingness could depend on Y_{obs}, but not on Y_{mis} (Schafer 1997; Schafer and Graham, 2002). This definition is based on the assumption that there is only one variable *Y* in the data set. When there are other measured variables (*X*) in the data set and *Y* is the only variable with missing data, under MAR, the distribution of missingness could depend on any measured variable (*X* or Y_{obs}), but not on Y_{mis}. Schafer (1997) stated that despite its name, MAR does not mean that missing data values are a simple random sample of all data values. Instead, MAR implies that a systematic relationship exists between one or more measured variables and the distribution of missingness. (Enders, 2010) Practically, the notion of MAR is that we have information on an individual's records of observed responses, plus any other information that we gathered on that person, which reduces the uncertainty about what the missing value is.

1.3.2 Missing Completely at Random

MCAR indicates that the values of missing data are a simple random sample of all data values. According to Little and Rubin (2002), under MCAR, the distribution of missingness depends on neither the missing data nor the observed data. MCAR is considered a special and more restricted case of MAR. (Enders, 2010)

1.3.3 Missing Not at Random

The missing data mechanism is defined as MNAR when the missing data distribution depends on the missing data itself. MNAR is also said to be *nonignorable nonresponse* (Schafer and Graham, 2002). The association between the distribution of missingness and missing data itself could happen in two scenarios: (1) when there is a direct relationship between the distribution of missingness and the missing data itself, and (2) when the distribution of missingness and the missing data itself are mutually correlated with an unmeasured variable. (Enders, 2010)

1.3.4 Missing Two-Mode Tie Data in Network Analysis

In this dissertation, the focus is on missing person-to-event ties $v_{i,q}^{1}$. This section will discuss the definition of missing person-to-event tie and what it means under different missing mechanisms.

Assume that there are N actors and K events in the network. That is, there are N*K potential two-mode ties. Let $v_{i,q} = 1$ if actor i attended event q; $v_{i,q} = 0$ otherwise. For those actor-and-event pairs that we have information about whether there is a tie or not, they are considered *observed tie data*.

Assume that we can get the person-to-event tie from two sources: attendance lists and surveys. In a survey, two-mode network data can be collected by asking questions such as "in the past year, what related events (e.g., meeting, conference, discussion group, etc.) did you attend,

and what was the frequency of attendance if it was a re-occurring event?" Attendance lists and surveys can be complements of each other when collecting two-mode network data. Table 1 below contains examples of each missing mechanism. The content in each cell will be explained in detail. Note that table 1 only listed examples for each scenario. It is NOT a comprehensive list of all possible situations for each missing mechanism.

Table 1

Examples of missing mechanism in practical situation

	Attendance List (complete list)	**Attendance List (single entry)**	**Survey**
MCAR	• Lost • Did not keep on file.	• The event attendant forgot to sign in.	• The survey respondent did not see the item.
MAR	• Not available for a particular *reason*, AND we know what the *reason* is. (e.g., An organization did not keep attendance lists on file for a particular period.)	• The event attendant did not sign in for a particular *reason*, AND we know what the *reason* is. (e.g., Some people attended a conference from the 2nd day, but the sign-in table was there only on the 1st day.)	• The survey respondent did not answer the item for a particular *reason*, AND we know what the *reason* is. It is better to ask about the *reason* in the survey directly.
MNAR	• Confidential	• The event attendant did not sign in intentionally, AND we do not know the reason.	• The survey respondent did not answer the item for a particular reason, AND we do not know the reason.

When $v_{i,q}^{1}$ is MCAR. MCAR means that the missingness depends neither on the missing data itself nor on any other measured variables. Examples of *an attendance list* being MCAR include (1) it got lost, (2) the organizer of the event did not keep the attendance list on file initially for no particular reason. In both examples, the attendance list is missing neither because of the attendance list itself nor because of anything measured. Examples of *a single entry of an attendance list* being MCAR can be the event/meeting attendant forgot to sign in. Again, in this situation, the missingness is neither due to the missing entry itself nor due to anything else measured. If the person-to-event tie was collected from a survey, an example of MCAR could be

that the survey respondent did not see the item or forgot to answer it. It is missing neither because the respondent did not want to answer that particular item nor because of anything measured.

When $v_{i,q}^{1}$ is MAR. MAR means that the missingness does not depend on the missing data itself but *could* (*but not necessarily*) depend on other measured variables. Say that there are ten organizations in the Chicago area which manage ravines, and we want to collect two-mode data of events/meetings about managing ravines in the past year. Out of these ten organizations, organization A never kept any attendance list due to management failure. That is, whenever organization A held an event, the attendance list was missing. In this case, it is not anything related to the attendance list itself that causes it to be missing. It is missing because of the poor management in organization A. We know the reason for missingness, and we have the data; for example, there can be a binary variable management_failure = 1 when the event was held by organization A, = 0 otherwise. This is an example that the *entire attendance list* is MAR. Assume that at a conference, the registration table was there for only the first day. However, some people attended the conference from the second day or even the third day. Therefore, all the conference attendants who did not show up on the first day were not on the attendance list. We know why the data were missing in this situation, and we had a record for it. For instance, there can be a variable 1st_day_missing = 1 for all people who registered but did not show up on the 1st day, = 0 otherwise. This is an example that *a single entry of the attendance list* is MAR. If we collected the person-to-event tie from surveys, it is MAR when the survey respondent did not answer that item for a reason, and we know what that reason is and have data on it. Although it might be challenging to know why people do not want to answer a particular question, we can design the survey in a better way to make it MAR. For instance, following the network question,

we can ask a question such that 'if you did not respond to the previous network question, what is the reason for not answering it?'. Research shows that by designing a survey like this, we make the missingness MAR, which will make the parameter estimation much more manageable.

When $v_{i,q}^{1}$ is MNAR. The missing data is MNAR when the missingness depends on the missing value itself. No matter if it is an entire attendance list, a single entry of the attendance list, or a survey item collecting person-to-event tie, they are MNAR if the missingness is due to the data itself. A typical example is that the organization does not want to share an attendance list since it should be kept confidential. Another example of MNAR is that respondents did not answer a survey item because they wanted to keep the information private.

1.3.5 Diagnosis of Missing Mechanism

The diagnoses for missing mechanisms are scarce. Little (1988) developed a global test for MCAR. The test's null hypothesis is that all variables with missing values in the data set are MCAR. A statistically significant test statistic provides evidence against the MCAR mechanism. However, we will not be able to identify a specific variable that is MCAR if there is more than one variable in the dataset with missing values. Besides, the MCAR test is criticized because simulation studies show that it has low power, especially when the number of variables violating MCAR is small (Thoemmes & Enders, 2007). Therefore, this test is likely to produce type II errors (i.e., fail to reject the null hypothesis when the null hypothesis is false). It may lead to a misleading sense of security about the missing mechanism (Enders, 2010).

For the MAR mechanism, it is impossible to confirm that the missingness is exclusively a function of other measured variables. That is, the MAR mechanism is not testable (Enders, 2010). The two mainstream approaches to handle missing data (FIML and MI) are MAR-based.

Researchers believe that we should always expect departures from MAR (Schafer & Graham, 2002).

1.3.6 Missing Data Scenarios Considered in this Dissertation

Recall the influence model equation (1).

$$y_{i,t} = \alpha + \rho_0 y_{i,t-1} + \beta_1 \frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1} + \epsilon_{it} \quad (1)$$

In this dissertation, it is assumed that $y_{i,t}$, $y_{i,t-1}$, and $z_{q,t-1}$ have complete data. There will be missing values on $v_{i,q}^1$ (the two-mode network tie data), and it is assumed MAR.

1.3.7 Missing Data Estimation Methods

Complete-case analysis (CC) is also called *listwise deletion.* It deals with missing data such that cases with any missing value are removed (Little, 1992). It gives unbiased parameter estimation when the missingness is MCAR, a strict condition. One common pitfall of the CC approach is that when there are several independent variables, even a sparse pattern of missing independent variables could result in a considerable number of incomplete cases (Enders, 2010; Little, 1992). Therefore, CC is often criticized for being an inefficient way to handle missing data. We should always consider that subjects with missing value on one independent variable may also have information on other variables. Despite this, the CC approach outperforms other methods in some particular situations. In regression analysis, if a covariate is MNAR (i.e., the covariate's missingness depends on the value of the covariate itself) and conditionally independent of the dependent variable (conditioning on all independent variables in the model), CC gives consistent parameter estimates. (Glynn and Laird, 1986; Little and Zhang, 2010; White and Carlin, 2010; Bartlett et al., 2014) Bartlett et al. (2014) developed the augmented complete case analysis estimation method to address the CC approach's inefficiency.

Available-Case Analysis (AC) uses the biggest possible set of data to estimate each parameter. For instance, for regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$, when we estimate the first two moments of $(X_1, \ldots, X_p, Y)$, μ_{X_j} and $\sigma^2_{X_j}$ are estimated using $n^{(j)}$ cases, i.e., the number of cases with X_j observed. Note that j is an index for $X's$, with $j = 1, \ldots, p$. σ_{X_j, X_k} $(j \neq k)$ is estimated using $n^{(jk)}$ cases, i.e., the number of cases with X_j and X_k observed (Glasser, 1964). There are different versions of AC analysis based on different choices of parameterization. AC has the advantage of using information from incomplete cases. However, it works well only when independent variables (X's) are not highly correlated (relative to degrees of freedom). When independent variables are highly correlated, the estimated covariance matrix of X's is likely to be not positive definite, which is a major disadvantage of AC. In the context of regression analysis, in general, AC gives unbiased (or consistent in the context of asymptotic analysis) coefficient estimate when independent variables are not highly correlated and when the missingness is MCAR (Haitovsky, 1968; Kim & Curry, 1977). Recall that MCAR is a stringent condition. In general, AC does not give unbiased (or consistent) estimates for standard errors Praag et al., 1985). There are simulation studies that compare AC regression estimates with Maximum Likelihood (ML) estimates under the normality assumption. It suggests that ML estimation performs better even when normality assumptions are violated (Azen et al., 1989; Little 1988; Muthén et al., 1987).

Single imputation techniques with Ordinary Least Square (LS) or Weighted Least Square (WLS) estimation. The idea of Single Imputation is to replace the missing values with a single set of imputed values. Then OLS or WLS estimation is used on the dataset with filled-in imputed values. When WLS is used, usually, less weight is put on imputed values. (Little, 1992). There is an extensive collection of single imputation methods including Unconditional Mean Imputation

(arithmetic mean), Conditional Mean Imputation (predicted values from a regression equation), Hot Deck/Similar Response Pattern Imputation (observed scores from a subject with similar background characteristics), Person Mean Imputation (a subject's average score across a set of observed item responses), Last Observation Carried Forward (scores from a previous wave in a longitudinal study), etc. (Baraldi and Enders, 2010). In the context of regression analysis, *Unconditional Mean Imputation* replaces missing *X's* by its unconditional sample *mean.* This method produces biased (or inconsistent) estimates for X's variance-covariance matrix, even when assuming MCAR. Estimated regression coefficients based on this method are biased (or inconsistent), and standard errors are under-estimated. In general, unconditional mean imputation is not recommended (Little and Rubin, 2002). The idea of *Conditional Mean Imputation* is to regress the missing variables on other variables without missing values using complete cases. Then use the estimated regression equation to estimate missing values for incomplete cases. The primary concern about this approach is that standard errors of the regression coefficients from OLS or WLS on the filled-in data are, in general, underestimated since the uncertainty of imputed values is not considered. Formulas for standard errors are difficult to derive for general missing data patterns (Little, 1992).

Stochastic regression imputation is the only single imputation method that has some merit. It gives unbiased parameter estimates under MAR, and its estimates are very similar to those from *full information maximum likelihood (FIML)* and *multiple imputation (MI)* approaches (Enders, 2010; Gold & Bentler, 2000; Newman, 2003). The idea of this method is very similar to the *Conditional Mean Imputation.* It also uses the regression equation to predict the missing values from the complete data. However, it goes one step further. It adds a normally distributed residual term to the predicted score to restore the variability of the data. This

approach's main disadvantage is that it underestimates the standard errors because it uses only one set of imputed values for missing data. It does not take into consideration the uncertainty of imputed values. The bias of standard errors can be corrected using the bootstrap resampling approach. However, implementing bootstrap resampling typically requires more effort than the FIML and MI approaches (Enders, 2010). What is worth knowing is that stochastic regression and the *imputation stage* of the MI approach share the same imputation routine. Enders (2010) indicated that the multiple imputation approach is conceptually an iterative version of stochastic regression imputation.

Despite the advantages of stochastic regression imputation, among missing data estimation methods, single imputation techniques generally produce biased estimates under any missing data mechanisms. Their shortcomings are widely documented (Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002; Widaman, 2006).

Full Information Maximum Likelihood. The maximum likelihood-based approach to handling missing data is referred to as *full information maximum likelihood* (FIML). To identify the set of parameter estimates that most likely produce the sample data, Anderson (1957) proposed the important idea of factored likelihood methods. It obtains explicit ML estimates for *special patterns* of missing data. *Iterative* method*s* are needed to find ML estimates for *the general pattern* of missing data with few exceptions. Trawinski and Bargmann (1964) and Hartley and Hocking (1971) developed scoring algorithms for the normal model. Orchard and Woodbury (1972) proposed an alternative approach, and it was later called the EM algorithm by Dempster et al. (1977). The E stands for expectation, and the M stands for maximization. The EM procedure has E and M steps. In the context of regression analysis with missing data, the EM process starts with an initial set of estimates of parameters. The E step constructs a unique set of

regression equations that predict the incomplete variables given the observed data and current parameter estimates for each missing data pattern. Specifically, it replaces the missing parts of the sufficient statistics with conditional expectations. An algorithm named *sweep operator* can automate constructing a set of regression equations for each missing data pattern. The M step uses standard complete-data formulas to produce updated parameter estimates, which are carried forward to the next E step. The process stops when parameter estimates in two consecutive M steps do not change, i.e., when the algorithm converges to ML estimates.

Researchers compared the ML approach with listwise deletion under MCAR, MAR, and MNAR. When the data was MCAR, both ML and listwise deletion yielded unbiased estimates. The standard errors under listwise deletion are 7% to 40% larger than those under the ML approach. The ML approach maximizes the statistical power by borrowing information from the observed data (Enders, 2010). What does 'borrowing information' mean? For instance, in a multiple regression context, we have X_1, X_2 and X_3 in the data set. Person A has data on X_1, X_2, but has missing data on X_3. With ML missing data handling approach, person A's log-likelihood is calculated using A's information on X_1 and X_2. Although person A has missing data on X_3, the data entry for person A is not deleted (unlike the listwise deletion method). Instead, it is approximated in the E step. When the data is MAR, the listwise deletion method produces biased estimates, while the ML approach gives unbiased estimates. Additionally, the confidence interval coverage when using ML is close to 95%, which indicates that standard errors are almost unbiased. When the data is MNAR, both listwise deletion and ML approach yield biased estimates. However, for the ML approach, the biases are restricted to a subset of the parameter estimates. In addition to listwise deletion, traditional missing data handling approaches generally yield biased parameter estimates for all parameters under MNAR (Enders, 2010).

ML approach is regarded as one state-of-art missing data handling approach (Shafer & Graham, 2002). In general, it is considered a better approach than traditional missing data handling approaches.

Multiple imputation (MI) is another state-of-the-art missing data handling technique, in addition to FIML (Schafer & Graham, 2002). Rubin (1987) proposed this approach within the Bayesian framework. There are, in general, three phases in MI: imputation, analysis, and pooling. The imputation phase produces numerous copies of the data set, each of which consists of different estimates of the missing values. In the analysis phase, it analyses all data sets created in the imputation phase. Finally, it combines all sets of parameter estimates and standard errors from the analysis phase in the pooling phase. Different algorithms can be used in the imputation phase. *Data augmentation* is one of the most popular algorithms, which assumes multivariate normal distribution (Schafer, 1997; Tanner & Wong, 1987). It is an iterative algorithm that repeatedly performs an *imputation* step (*I-step*) and a *posterior* step (*P-step*). In the I-step, it uses the stochastic regression procedure to impute the missing values. The P-step uses imputed data from the previous I-step to construct posterior distribution for parameters of interest and then produces new parameter estimates based on the posterior distribution. The MI approach has several inviting features. First, it allows researchers to analyze the data using estimation methods for complete data. Second, the approach does not distinguish whether the missing values are on the dependent variable or independent variables. Third, it can generate unbiased estimates with correct confidence interval with a small number of imputations. (Rubin, 1987 and van Buuren, 2018). The most important advantage of the MI approach is that it gives unbiased estimates when the missingness is MAR.

Number of Imputations in Multiple Imputation. The classical advice is to set the number of imputations between 3 and 5. This suggestion originated from the relationship between T_∞ (the estimate's total variance when there is an infinite number of imputations) and T_m(the estimate's total variance when there are m imputations). m is the number of imputations. Note that 'estimate' refers to the estimate for the parameter of interest. Rubin (1987) showed that $T_m = \left(1 + \frac{\gamma_0}{m}\right) T_\infty$ where γ_0 is the true population fraction of missing information. For $\gamma_0 = .3$ (e.g., a single variable with 30% missing) and m = 5, $T_m = 1.06T_\infty$. That is, when the there are five imputations, and the fraction of missing information is 30%, the total variance of the estimate is 1.06 times the ideal variance T_∞. There is little advantage to use more imputations. (Schafer 1997, Schafer and Olsen, 1998). There are different perspectives regarding the number of imputations in the multiple imputation procedure, where researchers suggested to use a larger number of imputations (Royston, 2004; Graham, Olchowski, and Gilreath, 2007; Bonder, 2008; Von Hippel, 2018; White, Royston, and Wood, 2011). These researchers have different opinions because they based the calculation of m on different criteria.

1.3.8 Handling Missing Network Tie Data in Latent Variable Models

In this section, *first*, the latent space/factor model for network analysis will be introduced. *Second*, the idea of controlling for distance in latent space/latent positions/cluster membership in network analysis models will be reviewed. *Last but not least*, how researchers handle missing tie data in latent variable models will be described.

Latent space/factor model for network analysis. In network analysis, the probability of forming a tie is one thing that researchers are interested in modeling. In the latent space/factor model, the probability of forming a tie between two actors depends on the *distance* between them in the latent space or individuals' *latent positions* (i.e., an unobserved vector of characteristics).

In some networks, individuals who have similar characteristics have higher probabilities of forming ties between them. Then the situation that a subset of individuals with a large number of ties between them may be suggestive that these individuals have nearby positions in the *space of characteristics* or *social space* (Hoff et al., 2002). This social space is called *latent space*, and an individual's position in this latent space is called a *latent position*. When *distance* is mentioned in this context, it means *distance* between two individuals in the latent space. In addition to the individual characteristics controlled in the model, the latent space, or the unknown individual characteristics which constitute the latent space are *unobserved*. When building models based on the collected data, it is probable that not all relevant individual attributes are included in the model. Some attributes may not be measured/observed. Including *distance* or *latent positions* in the model counts for those unmeasured/unknown nodal effects, and therefore decreases the residual variance (Hoff, 2018). This is one motivation of the latent space/factor model. Figure 3 below is a circle plot of estimated two-dimensional latent factors for a trading network between different countries. Estimated directions of the sender effect (countries who export) and the receiver effect (countries who import) are shown in red and blue. The country names' sizes represent the magnitude of the latent sender vector and the latent receiver vector. Dashed lines between countries indicate higher than expected trade after controlling for the additive sender and receiver effects and other covariates. From this plot, we can identify countries similar to each other in terms of trading behaviors after controlling for latent factors and other covariates. For instance, on the lower right side of the graph, it is shown that countries on the Pacific rim, such as the USA, China, Japan, etc. have high trade volume between them.

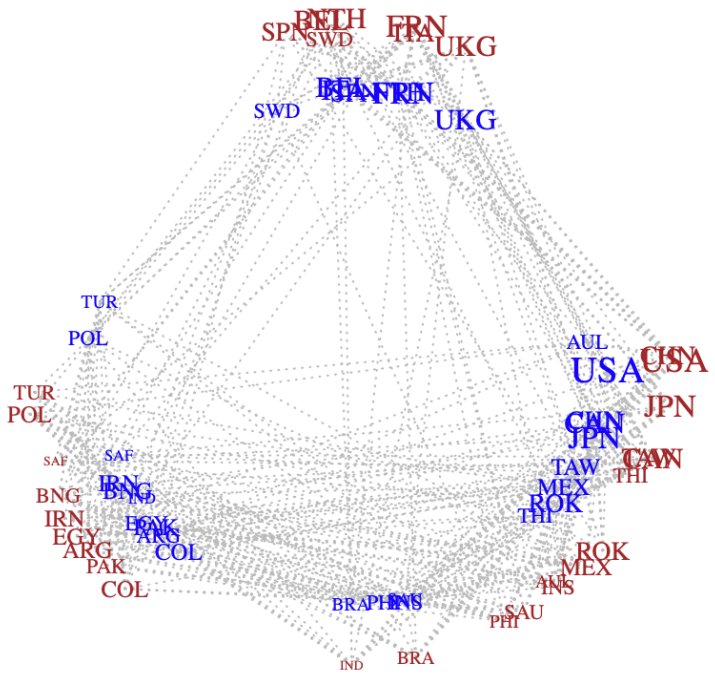

Figure 3 Circle plot of estimated latent factors for trading network (Hoff, 2015)

Another important motivation is that the distance or latent positions can account for *third-order dependencies*, which are ubiquitous phenomena in social networks. Third-order dependency is defined as dependency between triads, which can emanate from common characteristics among actors, which affects the probability of tie formation. Another example of third-order dependency is *stochastic equivalence;* it is defined such that "a pair of actors *ij* are stochastically equivalent if the probability of *i* relating to, and being related to, by every other actor is the same as the probability for *j*" (Minhas et al., 2019). The second motivation is for completeness of the literature review, but not of concern of this dissertation.

Recall that earlier in the previous paragraph, it was mentioned that in latent variable models, the probability of forming a tie depends on the *distance* between two actors or individual *latent positions*. While underlying individual characteristics constitute the latent space, and individuals with many ties between them tend to be close to each other in the latent space, it is more straightforward to understand the situation in this way: the probability of

forming a tie depends on the presence of other ties (Hoff et al., 2002). This is the intuition of controlling for latent features or network cluster memberships in network models, which will be reviewed next.

Control for latent features/cluster membership in network analysis models. The intuition of controlling for latent features/cluster membership in network models is not novel. For networks in which individuals belonging to prespecified groups, Wang and Wong (1987) introduced the *stochastic blockmodel*, where parameters representing differential probabilities of between-group and within-group ties are included. For networks in which group membership is not prespecified, Snijder and Nowicki (2001) proposed a model in which the probability of forming a tie depends on *latent class membership*. Individuals within the same latent class are treated as *stochastically equivalent.* Hoff et al. (2002) introduced the *distance model*, in which the probability of tie formation depends on the *Euclidean distance* between two actors and characteristics of dyads. Here the Euclidean space represents the latent space. Later in 2009, Hoff proposed the *latent factor model* where the probability of forming a tie depends on individuals' latent positions, i.e., individual-specific unobserved vectors of characteristics. Whether the probability of tie formation depends on latent class membership, distance in latent space, or individual latent positions, researchers are essentially controlling for dependency among a subgroup of individuals. The dependency originates from similarities among these individuals, and the similarities encourage tie formation. In other words, the formation of ties depends on the presence of other ties. Similarly, we may be able to achieve the same thing by controlling for *cluster membership*. *Clustering* is when a subset of actors have a large number of within-group ties and relatively few between-group ties (Hoff, 2009). In this dissertation, a

continuous latent factor approach will not be employed. Instead, cluster memberships will be used to represent a discrete latent space.

Handling missing tie data in the latent variable model. Hoff (2009) described how he handled missing tie data with the *multiplicative latent factor model*. In the multiplicative latent factor model, $logodds(y_{i,j} = 1) = \beta' x_{i,j} + u_i' D v_j + \epsilon_{i,j}$, where $y_{i,j} = 1$ if there is a tie from *i* to *j*. $x_{i,j}$ represents observed predictor variables. u_i is a vector of *latent* sender-specific factors, and v_j is a vector of *latent* receiver-specific factors. A fuller version of this model is $logodds(y_{i,j} = 1) = \beta' x_{i,j} + a_i + b_j + u_i' D v_j + \epsilon_{i,j}$, where a_i is the sender effect and b_j is the receiver effect. The difference between model versions does not affect the way of handling missing data. Let's focus on $logodds(y_{i,j} = 1) = \beta' x_{i,j} + u_i' D v_j + \epsilon_{i,j}$. Assume that there are *n* actors in the network. The researcher divided all *n(n-1)* directed ties into two parts, a *training set* and a *test set*. Then he assumed that the *training set* has complete data and removed some tie data from the *test set* so that the *test set* has missing tie data, $y_{i,j}$. The researcher first used data from the *training set* and estimated model parameters with the MCMC algorithm. Based on the results, the researcher estimated probability of forming a tie for each missing tie in the *test set* such that $\hat{p}_{i,j} = p(y_{i,j} = 1 | y_{i,j\ observed}) = E\left[\frac{\exp\{\theta_{i,j}\}}{1+\exp\{\theta_{i,j}\}} \middle| y_{i,j\ observed}\right]$, where $\hat{\theta}_{i,j} = \hat{\beta}' x_{i,j} + \hat{u}_i' \hat{D} \hat{v}_j$. Then, the researcher compared $\hat{p}_{i,j}$ with a threshold p. If $\hat{p}_{i,j} > p$, the estimated missing tie $y_{i,j} = 1$. The threshold p is set up by researchers based on their preference of making a balance between $P(y_{i,j} = 1 | \hat{y}_{i,j} = 1)$ and $P(\hat{y}_{i,j} = 1 | y_{i,j} = 1)$. $P(y_{i,j} = 1 | \hat{y}_{i,j} = 1)$ is the percentage of predicted ties that truly exist. $P(\hat{y}_{i,j} = 1 | y_{i,j} = 1)$ is the percentage of existing ties that are being predicted correctly. Notice how the *latent feature* plays a role in this process of handling missing data. First, complete data is used to estimate parameters and unknown

quantities in the model including u_i (the vector of latent sender-specific factors) and v_j (the vector of latent receiver-specific factors). Then, estimated parameters and quantities are used on missing entries (missing on $y_{i,j}$, but have complete information on other variables) to predict $p_{i,j}$. Then, according to the value of predicted $p_{i,j}$ (i.e., $\hat{p}_{i,j}$), the researcher decides whether the missing tie $y_{i,j}$ equals to 1 or 0. This approach assumes that missing ties are MAR.

Sewell and Chen (2015) also talked about handling missing tie data in the *dynamic latent space model* (a latent space model with time dimension), $logodds(y_{i,j,t} = 1) = \beta_{IN}\left(1 - \frac{d_{i,j,t}}{r_j}\right) + \beta_{OUT}\left(1 - \frac{d_{i,j,t}}{r_i}\right)$. In this model, $y_{i,j,t} = 1$ if there is a tie from *i* to *j* at time *t*. $d_{i,j,t}$ is the distance between *i* and *j* at time *t*. Specifically, $d_{i,j,t} = \| X_{it} - X_{jt} \|$ where X_{it} is the p dimensional vector of the i^{th} actor's latent position. r_i 's are actor-specific parameters that represent each actor's social reach. The researchers imputed missing value by drawing from a Bernoulli distribution with probability determined by parameter estimates for complete data from the model stated above. MH Gibbs sampling was used in the estimation procedure. The most updated imputed missing tie values were used in the next sampling procedure. This approach also assumes that missing ties are MAR.

CHAPTER 2: METHODOLOGY

2.1 Proposed Imputation Method for Missing Two-Mode Tie Data

In this dissertation, a new method is proposed to impute person-to-event ties' probability. If the imputed probability is greater or equal to a threshold, the imputed tie is 1. That is, the tie exists, and the person attended the event. If the imputed probability is smaller than the threshold, the imputed tie is 0. That is, the person did not attend the event. The next paragraph will discuss the model for imputing the probability, and an example will be given.

The probability of tie formation for missing actor-to-event ties will be imputed using equation (2).

$$\widehat{logit}\left[p\left(v_{i,q}=1\right)\right] = ln\left[\frac{p\left(v_{i,q}=1\right)}{1-p\left(v_{i,q}=1\right)}\right] = (1-x)\alpha_c + x\gamma_c \qquad (2)$$

$v_{i,q} = 1$ if actor i attended event q. $\ln\left[\frac{p(v_{i,q}=1)}{1-p(v_{i,q}=1)}\right]$ is the log odds of attending an event. Consequently, the two main terms on the right side of the equation are on a log-odds scale. α_c is the log odds of attending events outside one's cluster for actors in cluster c. It is also the log odds that an actor outside cluster c attending events in cluster c. γ_c is the log odds of attending events within one's cluster for actors in cluster c. α_c and γ_c are cluster specific to reflect the effect that people in the same cluster tend to have similar characteristics, and therefore tend to have similar behavior regarding events attendance. X is a dummy variable, where $X = 1$ if actor i and event q are in the same cluster, = 0 otherwise.

There are four categories for the relationship between an actor and an event: an actor attended the event given that the actor and the event are in the same cluster (A), an actor attended the event given that the actor and the event are in different clusters (B), an actor did not attend

the event given that the actor and the event are in the same cluster (C), and an actor did not attend the event given that the actor and the event are in different clusters (D). These categories will be used in the example below, which shows how to impute missing two-mode ties.

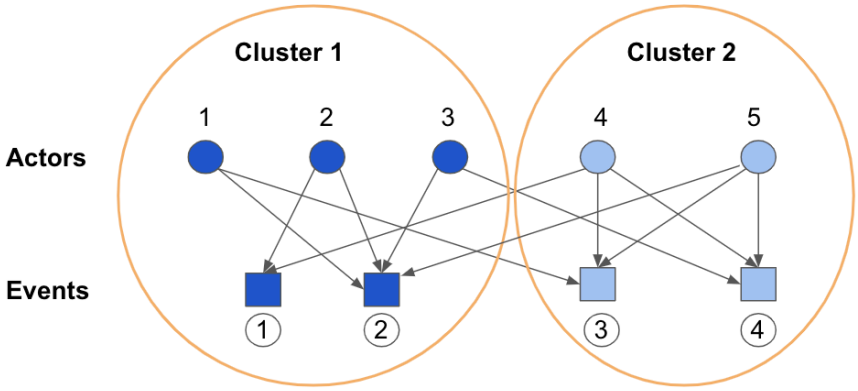

Figure 4 A simplified example of event-attendance network

Table 2

A simple example of event-attendance network

	Actors in Cluster 1											
Person	1	1	1	1	2	2	2	2	3	3	3	3
Event	①	②	③	④	①	②	③	④	①	②	③	④
Situation	.	A	C	D	A	A	D	D	C	A	D	B
	Actors in Cluster 2											
Person	4	4	4	4	5	5	5	5				
Event	①	②	③	④	①	②	③	④				
Situation	C	D	A	A	D	B	A	A				

Figure 4 and table 2 above depict a simple example of a two-mode event attendance network. There are five people and four events in the network. On the graph, round dots represent actors, and squares represent events; An arrow from a round dot to a square indicates that the actor attended the event. The dark blue color indicates cluster membership. The light blue color indicates membership in cluster 2. In table 2, A, B, C, and D correspond to the four categories of actor-and-event relationship defined in the last paragraph. There is a dot in the

'Situation' row of person 1 and event ①, which means that we do not know whether person 1 attended event ① or not. This information is missing. The information in figure 4 and table 2 are summarized in tables 3 and 4. In cluster 1, four ties belong to category A, two ties belong to category B, one tie belongs to category C, and four ties belong to category D. Table 3 and 4 can be interpreted in the same way.

Table 3

Count of event attendance for actors in cluster 1

Cluster 1		**Same Cluster**	
		Yes	**No**
Attended	**Yes**	A (4)	B (1)
	No	C (2)	D (4)
Marginal Sum		6	5

Table 4

Count of event attendance for actors in cluster 2

Cluster 2		**Same Cluster**	
		Yes	**No**
Attended	**Yes**	A (4)	B (1)
	No	C (1)	D (2)
Marginal Sum		5	3

Recall the model for imputing the probability of event attendance in equation (2).

$$\widehat{logit}\left[p\left(v_{i,q}=1\right)\right] = ln\left[\frac{p\left(v_{i,q}=1\right)}{1-p\left(v_{i,q}=1\right)}\right] = (1-x)\alpha_c + x\gamma_c \qquad (3)$$

α_c is the log odds of attending events outside one's cluster for actors in cluster c. In this example, $\alpha_1 = \ln\left(\frac{1/5}{4/5}\right)$ for people in cluster 1 and $\alpha_2 = \ln\left(\frac{1/3}{2/3}\right)$ for people in cluster 2. γ_c is the log odds of attending events within one's cluster for actors in cluster c. Here, $\gamma_1 = \ln\left(\frac{4/6}{2/6}\right)$ for people in

cluster 1 and $\gamma_2 = \ln(\frac{4/5}{1/5})$ for people in cluster 2. Back to figure 4 and table 2, from table 2 we know that the information between actor 1 and event ① is missing. Using the model proposed in this section, the log odds that actor 1 attended event ① is

$$\widehat{logit}\left[p(v_{1,1}=1)\right] = \ln\left[\frac{p(v_{i,q}=1)}{1-p(v_{i,q}=1)}\right] = (1-1)\alpha_1 + 1 \times \gamma_1 = 0 + \ln(\frac{4/6}{2/6}) \quad (4)$$

Note that X = 1 in this case because both actor 1 and event ① belong to cluster 1. Solving equation (4), we get $p(v_{1,1}=1)$ = .67, which is the imputed probability of having a tie between actor 1 and event ①. Then, we need to compare .67 to a threshold. If .67 is greater or equal to that threshold, it indicates a tie; otherwise, it indicates the absence of a tie. The next session will discuss how to decide the value of the threshold.

The threshold value is decided by considering two indices: *precision* and *recall***.** *Precision* is defined as $p(v_{i,q}=1|\hat{v}_{i,q}=1)$. It is the probability of having a tie between actor i and event q, given that we predict there is a tie between i and q, that is, the probability of correctly identifying the presence of a tie. *Recall* is defined as $p(\hat{v}_{i,q}=1|v_{i,q}=1)$; it is the probability of predicting a tie between i and q, given that there is indeed a tie between them. It tells us among all existing ties, what proportion we correctly identified as having a tie. We aim for high values of *precision* and *recall* while making predictions. *Recall* is monotonically decreasing as the classification threshold increases. In most cases, *precision* is monotonically increasing as the classification threshold increases. Therefore, when we try to decide the *threshold* and classify data into one class or the other (e.g., whether a person attended an event or not), we usually need to balance *precision* and *recall***.** For example, in the context of COVID-19, we have incomplete information about whether people attended certain events. If the *recall* index $p(\hat{v}_{i,q}=1|v_{i,q}=1)$ is low, it means that in some situations, we fail to recover the information

that people attended certain events. This is a severe mistake in the context of COVID-19 because we may not be able to track potentially infected people accurately. If the *precision* index $p(v_{i,q} = 1|\hat{v}_{i,q} = 1)$ is low, it means that in some cases, we predict that people attended certain events, but actually, they did not. In the context of COVID-19, this is not a very serious mistake compared to the situation when *recall* is low. How to choose the classification threshold depends on the specific research context. In this dissertation, a threshold which maximizes *precision* and *recall* and a threshold of .5 are used. Figure 5 shows an example of how to find the threshold which maximizes *precision* and *recall* at the same time. The horizontal axis represents the values of the threshold. The vertical axis is the probability, which represents the values of *precision* and *recall*. The orange curve is the recall curve, and the blue curve is the precision curve. The x-coordinate of these two curves' intersection is the threshold maximizing the *precision* and *recall* at the same time. A threshold of .5 is also chosen because .5 is the median of the interval [0, 1], containing all possible values of a probability.

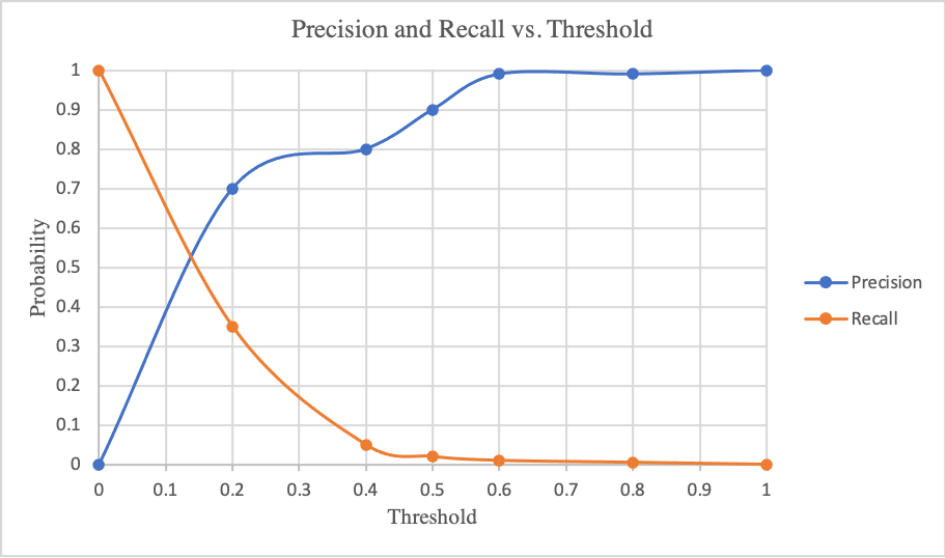

Figure 5: Precision and recall curves to find the threshold of the probability of having a tie

2.2 Use Multiple Imputation Method for Missing Two-Mode Tie Data

There are two general approaches in multiple imputation, *Joint Modeling (JM)* and *Fully Conditional Specification (FCS)*, also known as *Multivariate Imputation by Chained Equations*. The JM approach (Schafer, 1997) specifies a multivariate distribution for the missing data and draws imputation from conditional distributions by the Markov chain Monte Carlo technique. It is often used when we can use the multivariate distribution to describe the data. The FCS approach (van Buuren et al., 2006) does not assume a joint distribution for the data. Instead, it uses a separate conditional distribution for each variable with missing data and specifies the imputation model on a variable-by-variable basis. It is often used when the variable with missing data must only take specific values, for example, a binary variable for a logistic model or a count variable for a Poisson model.

The FCS is used in this dissertation to impute missing person-to-event tie data. The two-mode network data contains four columns: *person_id*, *event_id*, *attended*, and *formal*. *Attended* = 1 if the person attended the event, = 0 otherwise. *Formal* = 1 if the event was a formal event, = 0 otherwise. There are missing values in the variable *attended* and no missing information in other variables. The conditional density *P(Attended |Formal, model parameter)* is used to impute missing values in *Attended*. The variable *Formal* is in the imputation model. Recall that the missing person-to-event tie data are MAR, and the missingness depends on whether the event was formal or not. Note that the variable $Formal$ only appears in the imputation model, but not in the influence model (model for analyzing the data). Therefore, $Formal$ is an auxiliary variable. See the 'generate missing data' section for how the missing data were generated. Note that the *cluster membership* is not in the network tie data directly, but it is embedded. See the '*assign two-mode (actor-to-event) ties'* section for details. The number of imputations is set to be

five. The ordinary least square method is used on the imputed datasets to estimate coefficients in the influence model. After obtaining the coefficient estimates from the imputed datasets, Rubin's formulas are used to combine these estimates as the final coefficient estimate. In particular, the multiple imputation point estimate is $\bar{\beta}_1 = \frac{1}{m}\sum_{t=1}^{m}\hat{\beta}_{1,t}$, where $\hat{\beta}_{1,t}$ is the estimated coefficient of the exposure term for the t^{th} imputation, and m is the number of imputations. The estimated variance for $\bar{\beta}_1$ is $V_T = V_W + V_B + \frac{V_B}{m}$. V_W is the within-imputation variance. $V_W = \frac{1}{m}\sum_{t=1}^{m} SE_t^2$, where SE_t is the standard error of the estimate from the t^{th} imputation. V_B is the between-imputation variance. $V_B = \frac{1}{m}\sum_{t=1}^{m}\left(\hat{\beta}_{1,t} - \bar{\beta}_1\right)^2$, which quantifies the variability of estimates from different imputed datasets. $\frac{V_B}{m}$ is a correcting factor for using finite number of imputations.

2.3 Data Simulation

2.3.1 Simulate Complete Data

To generate complete two-mode networks, we need to consider *network density,* the *total number of actors*, the *total number of events, distribution of first-mode degree*, *number of clusters*, *number of actors per cluster*, and *number of events per cluster*.

Density. There are a few pieces of literature studying densities of real-world *two-mode* networks. Valente (2010) states that there are many redundant ties for *one-mode* networks with density values above 50%. For such networks, removing ties or even nodes will not affect overall network properties. Such networks do not contain much structural information, and often researchers are not interested in them or need to "prune" the network to find the hidden structure. Researchers think that there is a practical limit to the number of relationships that one actor can establish with other actors in a network (Valente, 2010). A reasonable range for density in typical one-mode networks is [0, 0.5]. In one-mode networks, actors only have limited time and

energy to interact with a certain number of other actors. The same reasoning applies to two-mode networks. It is assumed that participation in events is proportional to the interaction with people in the network. That is, actors have limited energy to attend events in a particular period. In this dissertation, it is assumed that the density $\Delta A \in \{0.25, 0.45\}$ representing networks with low and high densities for social networks. Say that we have a two-mode network with *N* actors and *K* events. Let *L* be the total number of ties present. Then the *density* of this two-mode network is $\frac{total\ number\ of\ ties}{possible\ number\ of\ ties} = \frac{L}{N \times K} = \frac{L/N}{K} = \frac{\lambda}{K}$, where λ is the expected outdegree (per actor). We are borrowing information about outdegree in the one-mode network and graphing it to the two-mode network.

The total number of actors in the network. Literature about *network size for two-mode networks* (including the number of actors in the network and the number of events in the network) is rare. Nevertheless, researchers have different opinions about reasonable *network size for one-mode networks*. According to Valente (2010), some researchers studied organizations with 100 to 250 employees; some evidence showed that the optimal size for a human group is 100 (Dunbar, 1993); some other researchers estimated that the average size of acquaintance network in the U.S. is about 280 (Killworth et al., 2006). Even though actors may know several thousand other people, the number of names they can give on any topic is usually much smaller (Valente, 2010). For the same reason, in two-mode networks, actors have limited energy and time to attend events and could only be exposed to a certain number of other actors. The number of actors in the network is set to be $N \in \{20, 50, 200\}$ to represent small, medium and large social networks.

Total number of events. A two-mode network's size is decided by the *number of actors* and the *number of events*. The number of events is set to be $K \in \{10,\ 50, 100\}$ to represent scenarios where there is small, medium and large number of events.

Distribution of first-mode degree. Next, the decision needs to be made regarding the distribution of *the first-mode* degree. Say that we are studying a network of people and events. Then the *first mode* is people; the *second mode* is events. The *first-mode degree* for a particular actor is the number of events that this actor attended. One assumption is that the distribution of the first-mode degree is *homogenous*, i.e., actors are similar in terms of the number of events participated, which implies "normal and expected behavior when taking an actor at random" (Fujimoto et al., 2011; Latapy et al., 2008). Another reasoning for assuming homogeneous first-mode degrees is that actors have limited energy to attend events in a given period. The distribution of first-mode degrees will be generated using Poisson distribution. Let $A = \{A_{ij}\}$ represent the two-mode social matrix, where $A_{ij} = 1$ indicates that actor *i* participated in event *j*. Let ΔA represent density of the two-mode network A. By definition, $\Delta A = \frac{total\ number\ of\ ties}{possible\ number\ of\ ties} = \frac{L}{N \times K} = \frac{L/N}{K} = \frac{\lambda}{K}$, where λ is the *expected one-mode degree*. Earlier it was decided that $\Delta A \in \{.25,\ .45\}$ and $K \in \{5, 50, 100\}$. Therefore, $\lambda = \Delta A \times K$ has $2 \times 3 = 6$ potential values. Earlier the number of events was set to be $N \in \{20, 50, 200\}$. The *first-mode degrees* will be obtained by drawing random numbers from Poisson (λ) for *N* times. The draws represent the number of events participated by each actor. Note that for now it has not been assigned exactly which events each actor attended. Before that, actors and events need to be assigned to different clusters. There are $2 \times 3 \times 3 = 18$ different networks (in terms of different density levels, number of events, and number of actors). Also note that $\Delta A \in \{.25,\ .45\}$ will not

be the exact densities for generated networks. Those are expected densities. Drawing from Poisson distribution will introduce some randomness.

Number of Clusters. Next, the *number of clusters* and corresponding *cluster size* (i.e., the *number of actors* and the *number of events* in each cluster) for each simulated network need to be decided. There are few guidelines or literature about the number of clusters or cluster sizes for two-mode networks. The number of clusters presented in a two-mode network may depend on the network density, local densities for subsets of the network, etc. A densely connected network is more likely to have clusters compared to a very sparse network. However, there is no clear rule about the relationship between the number of clusters/cluster size and network-level properties. For an event participation network, how it is clustered depends on how actors participated in different events, which means every network is unique. In this dissertation, it is assumed that there are two clusters. Defining the system allows the exploration of how the presence of clustering affects estimation. Systems with more clusters can be explored in further research.

Number of actors in each cluster. Earlier it was decided that there are N actors and 2 clusters in the network where $N \in \{20, 50, 200\}$. Random numbers will be drawn from Uniform (5, N-5), Uniform(12.5, N-12.5), and Uniform(50, N-50) to decide the number of actors in each cluster for different settings of N. The reason to set lower bounds for the range of uniform distributions to be 5, 12.5, and 50 is to make sure that clusters generated are not extreme cases. For instance, if drawn from Uniform (1, 50-1) for a network containing 50 actors, there may be two actors in one cluster and 48 actors in the other cluster, which is a very extreme case. 5, 12.5, and 50 are 25% of the total number of actors for different settings of N. Say that we get N_{C1} from

one draw. That is, one cluster has N_{C1} actors. Then the other cluster has $N - N_{C1}$ actors. Next, assign actors 1 to N_{C1} to cluster 1, and actors $N_{C1} + 1$ to N to cluster 2.

Number of events per cluster. Earlier, it was decided that there are *K* events in the network, where $K \in \{10, 50, 100\}$. Random numbers will be drawn from Uniform(2.5, K-2.5), Uniform(12.5, K-12.5), and Uniform(25, K-25) to decide the number of events in each cluster for different settings of K. The reason to set lower bounds for the range of uniform distributions to be 2.5, 12.5, and 25 is to avoid extreme cases. For example, if drawn from Uniform (1, 50-1) for a network containing 50 events, one cluster may have 1 event and the other cluster may have 49 events, which is an extreme situation. 2.5, 12.5, and 25 are 25% of the total number of events for different settings of K. Say that we get K_{C1} from one draw. That is, one cluster has K_{C1} events. Then the other cluster has $K - K_{C1}$ events. Next, assign events 1 to K_{C1} to cluster 1, and events $K_{C1} + 1$ to K to cluster 2.

Assign two-mode (actor-to-event) ties. When 'pair' is mentioned in this dissertation, it refers to the actor-to-event pair. By this stage of data simulation, we already know the number of actors, the number of events, and their corresponding cluster memberships. We know how many within-cluster pairs there are and how many outside-cluster pairs there are at both cluster and individual levels. Note that 'pair' is used here instead of 'tie,' indicating that all possible pairs are considered in the network. There is not necessarily a tie between each possible pair. We do not know how many events each actor attended within vs. outside their cluster yet. This is decided by the *odds ratio* of attending an event within vs. outside one's cluster. The odds ratio is set to be 1, 2, and 5. A value of 1 indicates no clustering; 2 indicates low clustering; 5 indicates high clustering. Next, an example of how to calculate the number of within-cluster ties and outside-cluster ties will be given. Table 5 below describes an actor's event attendance information.

Table 5

An actor's event attendance

		Same Cluster	
		Yes	**No**
Attended	**Yes**	a	b
	No	c	d

Assume that there are k events in the network, then for any actor, there are k possible actor-to-event pairs. The actor attended some events (a + b) and did not attend others (c + d). Some events are in the actor's cluster (a + c), while other events are outside the actor's cluster (b + d). For each actor a + b + c + d = k, the total number of events in the network. Note that a + b is the actor's outdegree, i.e., the number of events the actor attended. a + c is the number of events within the actor's cluster, despite whether the actor attended it or not. The odds of attending events within-cluster is a/c; the odds of attending events outside-cluster is b/d. Therefore, the odds ratio of attending events within-cluster vs. attending events outside-cluster is $\frac{ad}{bc}$. Summarizing the above information, we have the following four equations for each actor.

$$a + b = outdegree$$

$$a + c = the\ number\ of\ within - cluster\ pairs$$

$$a + b + c + d = the\ number\ of\ events\ in\ the\ network$$

$$\frac{ad}{bc} = 1, 2\ or\ 5$$

There are four equations and four unknowns (a, b, c and d). So, for each individual, we will be able to calculate a, the number of within-cluster ties and b, the number of outside-cluster ties. Once a and b are known for each actor, a simple random sample of size a is taken from all possible within-cluster pairs, and a simple random sample of size b is taken from all possible outside-cluster ties. That is how specific actor-to-event tie is assigned.

Recall the influence model in equation (1).

$$y_{i,t} = \alpha + \rho_0 y_{i,t-1} + \beta_1 \frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1} + \epsilon_{it} \quad (5)$$

The values for $v_{i,q}^1$ and h_i^1 will be from generated network data. The next session will discuss how to generate other variables.

Generate $Y_{i,t-1}, \alpha, and\ \rho_0$. $Y_{i,t}$ represents people's behavior at time t, and $Y_{i,t-1}$ represents people's behavior at time $t-1$. $Y_{i,t-1}$ will be generated from N(0, 1). People's behaviors at time t, i.e., $Y_{i,t}$, should have *a fairly strong positive correlation* with their behaviors at time $t-1$, i.e., $Y_{i,t-1}$. Therefore, ρ_0 should be between 0 and 1. ρ_0 will be generated from a Beta distribution with parameters $\alpha = 5$ and $\beta = 2$.

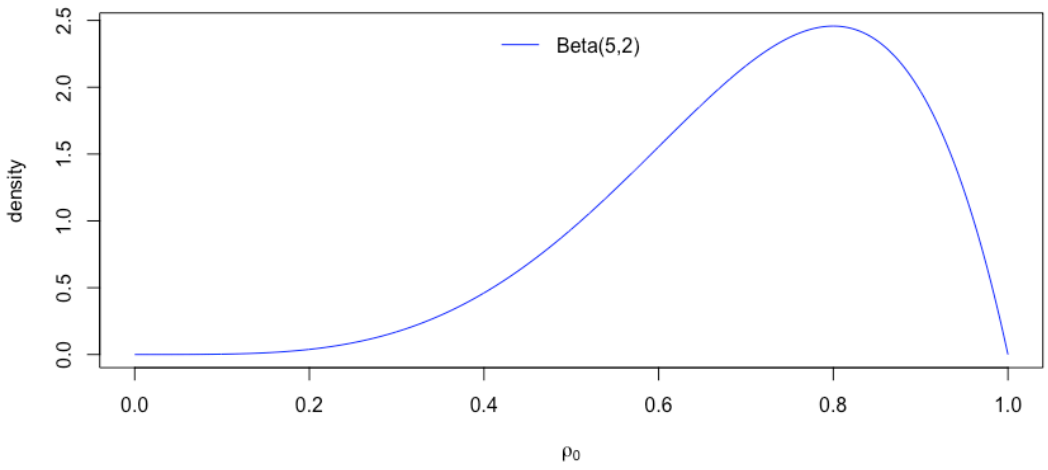

Figure 6 Beta distribution for generating coefficients for the prior

In the influence model $y_{i,t} = \alpha + \rho_0 y_{i,t-1} + \beta_1 \frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1} + \epsilon_{i,t}$, α is $Y_{i,t}$'s expected value when holding all independent variables zero. That is, when holding the exposure term $\frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1}$ zero, $y_{i,t} = \alpha$ when $y_{i,t-1} = 0$. Therefore, α is the average change in y_i

from time $t-1$ to time t when holding the exposure term zero. α is set to be 2. The value of α is not the focus of this dissertation. A small positive value is set to reflect the fact that behavior change is gradual. The key is to ensure that the conditional means of $y_{i,t}$ and $y_{i,t-1}$ are different because we expect people's behavior to change from time $t-1$ to time t. One example is a study about people's perceptions of lake levels in the Great Lakes. At time one, people believed that lake levels would decrease in 50 years; at time two, people realized that lake levels would decrease but not as much as they thought. Between time one and time two, these people attended conferences, meetings, workshops, etc. related to lake levels. All these events they attended affected their beliefs about lake levels.

Generate $Z_{q,t-1}$. Recall that $Z_{q,t-1}$ is the information presented on event q. The whole term $\frac{1}{h_i^1}\sum_{q=1}^{Q} V_{i,q}^1 Z_{q,t-1}$ describes the effects of attending events on people. $V_{i,q}^1 = 1$ if actor i participated in event q, = 0 otherwise. The term $\frac{1}{h_i^1}\sum_{q=1}^{Q} V_{i,q}^1 Z_{q,t-1}$ takes the average of all information presented on events that an individual attended within cluster. For instance, Bill participated in events 1, 4, and 5 in his cluster. The information presented on events 1, 4, and 5 is 9, 10, and 11. Therefore, for Bill, the term $\frac{1}{h_i^1}\sum_{q=1}^{Q} V_{i,q}^1 Z_{q,t-1}$ is (9+10+11)/3 = 10. Data for $V_{i,q}^1$ and h_i^1 are from generated network data. h_i^1 is the number of events the actor attended *within-cluster*. $\frac{1}{h_i^1}\sum_{q=1}^{Q} V_{i,q}^1 Z_{q,t-1}$ is Bill's exposure term because he was exposed to all the information presented on these events and his behavior might be influenced. $Z_{q,t-1}$ will be generated from $N(0,2)$. We want to make sure that the variance of $Z_{q,t-1}$ is different from the variance of $Y_{i,t-1}$ (i.e., people's behaviors/beliefs at time $t-1$).

Generate β_1. β_1 is the regression coefficient for the exposure term. The coefficient for the exposure term is often positive, falling between 0 and 1. On the other hand, the exposure term's

effect is usually not as strong as the effect of the prior ($Y_{i,t-1}$). β_1 will be generated from a Beta distribution with parameters $\alpha = 2$ and $\beta = 5$. The graph below shows the distribution for β_1.

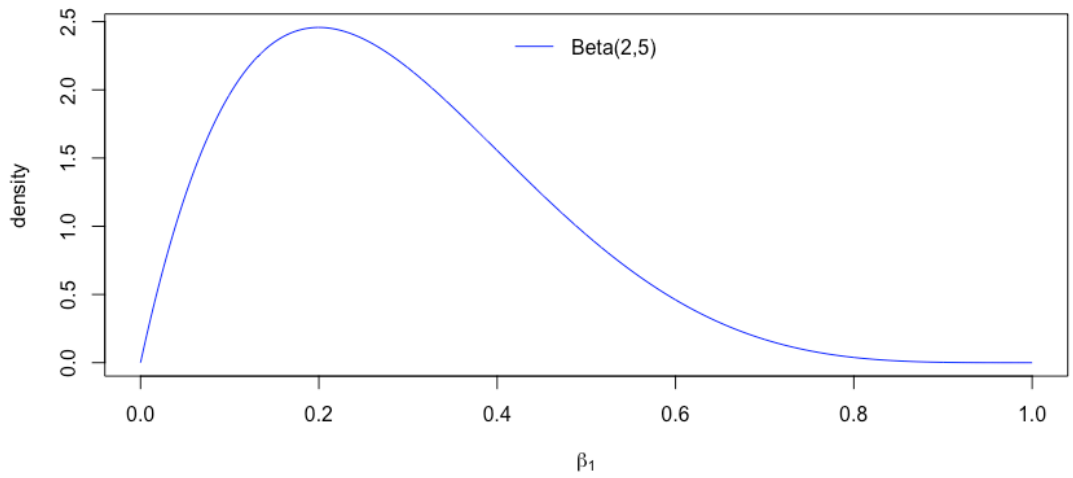

Figure 7 Beta distribution for generating coefficients for the exposure term

Generate $\epsilon_{i,t}$. $\epsilon_{i,t}$ is the random error term from the model. It will be generated from $N(0,1)$.

Compute $y_{i,t}$. Recall the influence model $y_{i,t} = \alpha + \rho_0 y_{i,t-1} + \beta_1 \frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1} + \epsilon_{i,t}$. The above paragraphs described how to generate α, ρ_0, β_1, $y_{i,t-1}$, $\epsilon_{i,t}$, $v_{i,q}^1$ and h_i^1. We have everything on the right side of the influence model. The dependent variable $y_{i,t}$ will be computed using the generated values.

Covariance structure of the influence model. The influence model is essentially a linear regression model with two independent variables: $y_{i,t-1}$ and the exposure term $\frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1}$. It is assumed that $y_{i,t}$'s are independent. Therefore, the variables do not have a special covariance structure. The covariance structure of the influence model is the variance

covariance matrix of $y_{i,t}$, $y_{i,t-1}$ and $\frac{1}{h_i^1}\sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1}$. The coefficients ρ_0 and β_1 guarantee that there is correlation between $y_{i,t}$ and $y_{i,t-1}$, as well as $y_{i,t}$ and $\frac{1}{h_i^1}\sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1}$. Recall that the coefficient of $y_{i,t-1}$, ρ_0, is generated using the Beta(5, 2) distribution to ensure that $y_{i,t}$ and $y_{i,t-1}$ are highly correlated.

2.3.2 Simulate Missing Data

Recall the definition of missing person-to-event tie in this dissertation. All possible actor-to-event pairs are considered. Those pairs that we have information about whether there is a tie or not are considered *observed data*. An observed tie can be 1 or 0. Those pairs that we do not have information about whether there is a tie or not are considered *missing data*. A missing tie can be 1 or 0.

Proportion of missing data. The *proportion of missing data* is the proportion of actor-and-event pairs that we do not have information about whether there is a tie or not. It is assumed that there is 50% missing data. It is easy to have a large proportion of missing tie data in two-mode network studies because it could be challenging to know a comprehensive list of events and get all attendance lists. 50% is a starting point for the research. Other proportions of missing tie data could be tried in the future.

Missing mechanism. Missing data are assumed to be Missing at Random (MAR). The missingness depends on whether the event is ***formal*** or ***informal***. It is assumed that 50% of events are ***formal*** events, and 50% events are ***informal*** events. Let $x_q = 1$ if event q is formal; $x_q = 0$ if event q is informal. When $x_q = 1$, the probability of missing in $v_{i,q}$ is 20%; when $x_q = 0$, the probability of missing in $v_{i,q}$ is 80%. Therefore, the proportion of missing is $20\% \times \frac{1}{2} + 80\% \times \frac{1}{2} = 50\%$.

2.3.3 Number of Replications

When simulating the network data, there are two settings for the network density, three settings for the number of actors in the network, and three settings for the number of events in the networks. Therefore, there are $2 \times 3 \times 3 = 18$ unique network settings. 2000 replicates were generated for each setting. According to Morris, White and Crowther (2018), *coverage* refers to the confidence interval coverage for the parameter of interest: $P(\hat{\theta}_{low} \leq \theta \leq \hat{\theta}_{upp})$. The estimate of *coverage* is equation (6).

$$\widehat{coverage} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{\text{low,i}} \leq \theta \leq \hat{\theta}_{upp,i}) \tag{6}$$

The Monte Carlo *standard error* of the *coverage* estimate is equation (7).

$$\sqrt{\frac{\widehat{coverage} \times (1 - \widehat{coverage})}{n_{sim}}} \tag{7}$$

n_{sim} is the number of replications. Rearranging equation (7), we have equation (8).

$$n_{sim} = \frac{E(Coverage) \times (1 - E(Coverage))}{(Monte\ Carlo\ Standard\ Error)^2} \tag{8}$$

Therefore, if we want to keep the Monte Carlo standard error below 0.5% for a coverage of 95%, we need at least $n_{sim} = \frac{95 \times 5}{0.5^2} = 1900$ replications. That's why the number of replications is set to be 2000 in this dissertation.

CHAPTER 3: RESULTS

Recall the influence model: $y_{i,t} = \alpha + \rho_0 y_{i,t-1} + \beta_1 \frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1} + \epsilon_{i,t}$. The second independent variable with coefficient β_1 is the *exposure term*. The influence model was run in the following situation: 1) when the data was complete, 2) when there were missing data, but no imputation was used, 3) when there were missing data, and the proposed imputation method was used with thresholds maximizing *Recall* and *Precision* indices, 4) when there were missing data, and the proposed imputation method was used with the threshold of .5, and 5) when there were missing data, and the *multiple imputation* method was used. The *ordinary least square* estimation method was used in all situations. Bias, empirical standard error and the Root Mean Square Error (RMSE) for the coefficient estimate of the *exposure term* are examined.

3.1 Bias

Bias for the Coefficient Estimate of the Exposure Term. The coefficient estimates of the exposure term's biases were calculated for each of the five situations described above. When 'bias' is mentioned in this dissertation, it refers to the coefficient estimate of the exposure term's bias. The formula for the bias is $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \widehat{\beta}_{1,i} - \beta_1$. Recall that there are 54 network settings in terms of density, number of actors, number of events, and the clustering effect (i.e., odds ratio of attending events within- vs. outside- cluster). For each network setting, 2000 replicates (i.e., networks) were generated. For each replicate, the influence model was run in five situations. That is, there are 2000 coefficient estimates for each situation under different network settings: (1) when the data are complete (complete), (2) when there are missing data, but no imputation used (missing), (3) when using the proposed imputation method with thresholds maximizing precision and recall (imp), (4) when using the proposed imputation method with a threshold of .5

(imp_2), and (5) when using the multiple imputation method (multiple_imp). $\hat{\beta}_{1,i}$ is the exposure term's coefficient estimate for each replicate. Initially, the exposure term's coefficient was generated from the distribution Beta(2, 5). The parameter value of β_1 used to calculate the bias is the mean of Beta(2, 5): $\frac{2}{2+5} = 0.2857$. $n_{sim} = 2000$. This explains why there are biases for complete data in the boxplot presented below.

Figure 8 depicts coefficient estimates of the exposure term's biases for the five situations.

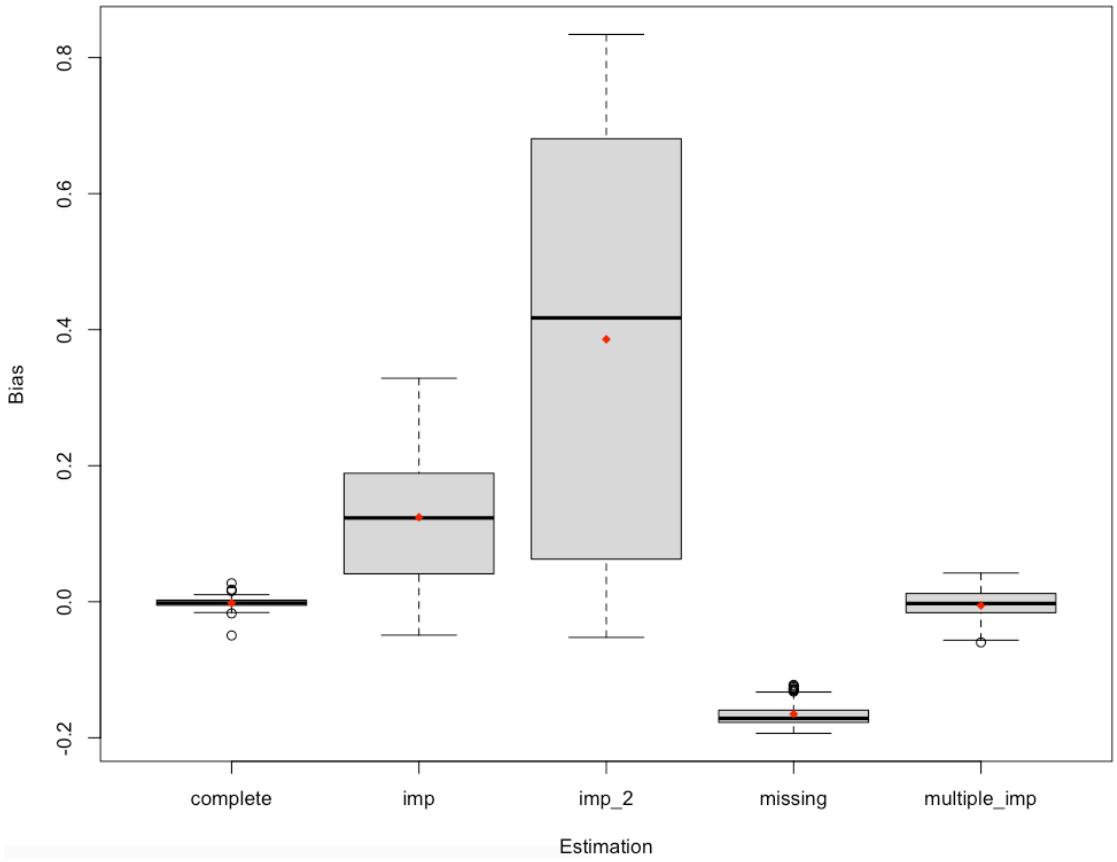

Figure 8 Bias of the coefficient estimate for the exposure term

The thick black line in the middle of each box represents the medium of the biases. The red dot represents the mean of the biases. When the data are complete, the average bias for the 2000 replicates is -0.002, with a standard deviation of 0.010. It is very close to zero, which aligns with

the OLS estimator's property: mean-unbiased estimator when regressors are exogenous, errors are homoscedastic and serially uncorrelated, and errors have finite variances. The average bias when using the multiple imputation method is -0.005, with a standard deviation of 0.024. The average bias when using the proposed imputation method with thresholds maximizing precision and recall is 0.124, with a standard deviation of 0.104. The average bias when using the proposed imputation method with a threshold of .5 is 0.386, with a standard deviation of 0.321. When there are missing data but no imputation used, the average bias is -0.165 with a standard deviation of 0.019. The Kruskal-Wallis rank test was run to examine whether there are significant differences among the estimated coefficients' biases (exposure term) in the five situations. The p-value is 0 for this test, indicating that biases are significantly different for the five situations. The Kruskal-Wallis rank test is a non-parametric version of the one-way ANOVA test. It was used here because the normality assumption of the one-way ANOVA test was violated. The pairwise Wilcoxon signed test was then used to compare biases for each pair of the five situations. The results showed no significant difference between biases when the data were complete and when using the multiple imputation method to handle missing ties, which indicates that the multiple imputation method gave coefficient estimates with very small biases. Biases are significantly different when using the multiple imputation method and using the proposed imputation method with thresholds maximizing precision and recall. The multiple imputation method performed better than the proposed imputation method in terms of bias. One possible reason for this result is that the missing two-mode ties were generated as MAR depending on whether the event is formal or not. The multiple imputation method uses the variable *Formal* in the imputation model, which leverages the missing mechanism. On the other hand, the proposed imputation method does not use the variable *Formal* to impute the missing ties. Instead, it

utilizes the clustering effect in the imputation. Note that for the Kruskal-Wallis rank test and the pairwise Wilcoxon signed test described above, the sample size is 270 (54 network settings × 5 situations). Next, the relationship between biases and network properties will be examined.

Bias vs. Clustering Effect. The clustering effect is represented by the odds ratio of attending events within- vs. outside- one's cluster. The odds ratio is set to be 1, 2, and 5 representing no clustering effect, low clustering effect, and high clustering effect. The relationship between bias and clustering effect are examined separately when using the proposed imputation method and the multiple imputation method.

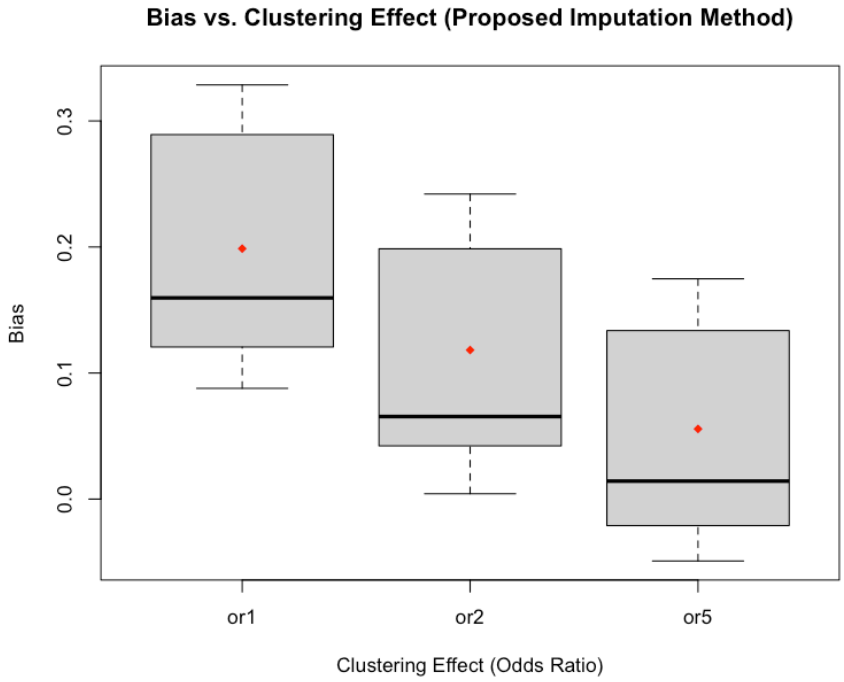

Figure 9 Relationship between bias and clustering effect (proposed imputation method)

Figure 9 shows the relationship between bias and the clustering effect when using the proposed imputation method with thresholds maximizing precision and recall. The thick black bars represent the medium of biases. The red dots represent the mean of the biases. From the figure, we can see that the larger the clustering effect, the smaller the bias. The Kruskal-Wallis rank test

results show that biases are significantly different for networks with different clustering effects (p-value = 0). The pairwise Wilcoxon signed test results show that biases are significantly different for each pair of the three situations: odds ratio = 1 (no clustering), odds ratio = 2 (low clustering), and odds ratio = 5 (high clustering). The results confirm that the proposed imputation method leverages the clustering effect in the imputation process such that it gives smaller biases when the clustering effect is higher.

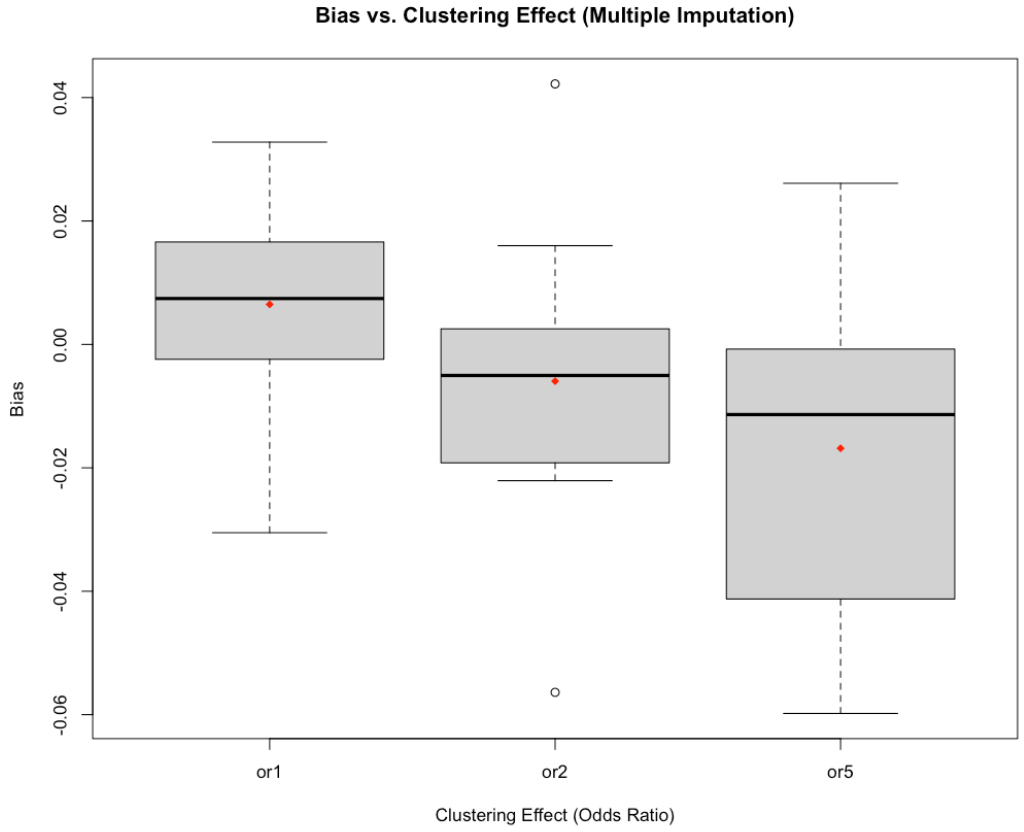

Figure 10 Relationship between bias and clustering effect (multiple imputation)

Figure 10 shows the relationship between the *bias* and the clustering effect when using the multiple imputation method. From this figure, we can see that there are positive and negative biases. Therefore, the relationship between the *absolute bias* and the clustering effect (figure 11 is also examined to make it easier to see the biases' magnitude.

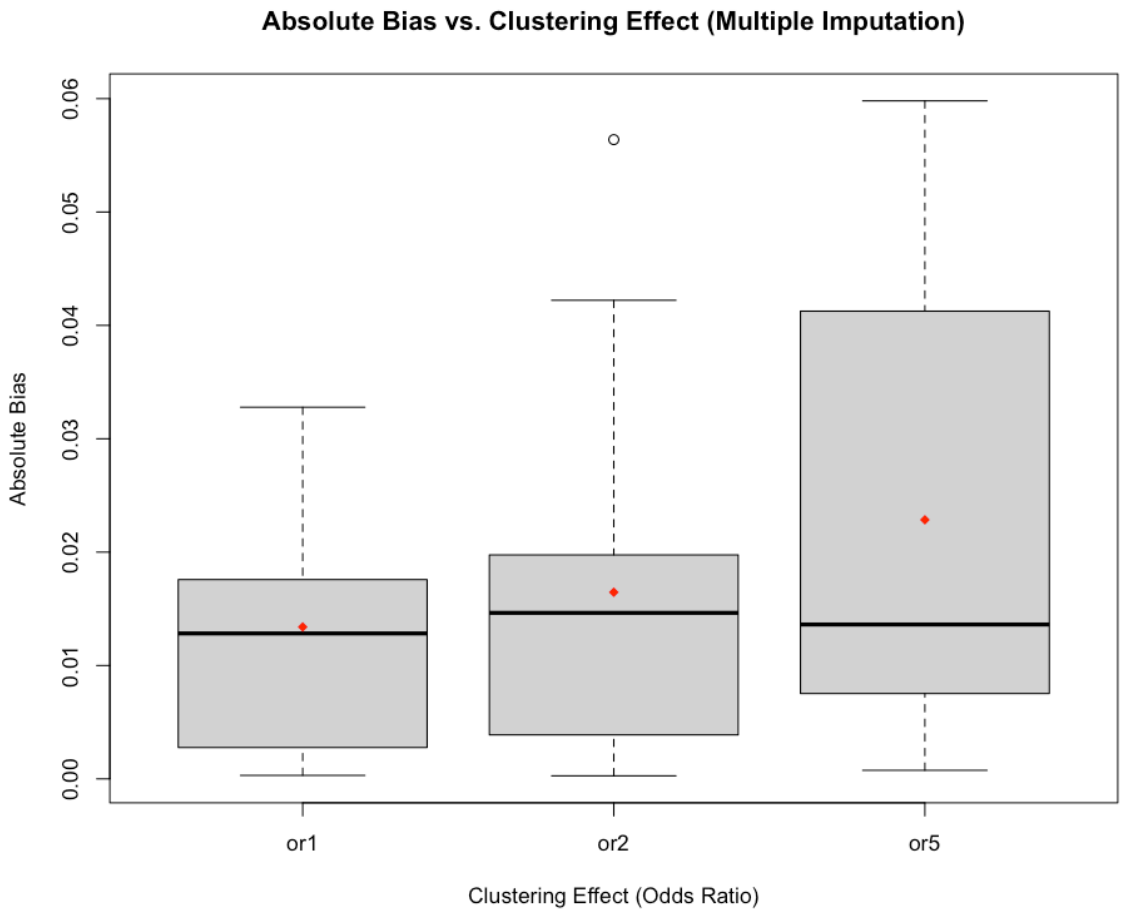

Figure 11 Relationship between absolute bias and clustering effect (multiple imputation)

Figure 11 shows that the larger the clustering effect, the larger the absolute bias when using the multiple imputation method. The results are the opposite of those when using the proposed imputation method. The Kruskal-Wallis rank test result shows that there is no significant difference between biases for networks with different clustering effects when using the multiple imputation method. One possible reason is that the imputation model used for the multiple imputation method does not include the clustering effect.

3.2 Empirical Standard Error

The estimated coefficients of the exposure term's empirical standard errors are examined for the five situations described earlier under different network settings. When 'empirical

standard error' is mentioned in this dissertation, it refers to the estimated coefficient of the exposure term's empirical standard error. Figure 12 shows the results. The formula for the estimated empirical standard error is $\sqrt{\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\left(\hat{\beta}_{1,i}-\bar{\hat{\beta}}_1\right)^2}$. For how $\hat{\beta}_{1,i}$ and n_{sim} are defined, please see the description in the '*Bias for the Coefficient Estimate of the Exposure Term*' section. $\bar{\hat{\beta}}_1$ is the average of $\hat{\beta}_{1,i}$'s for each 2000 replicates for a specific network setting.

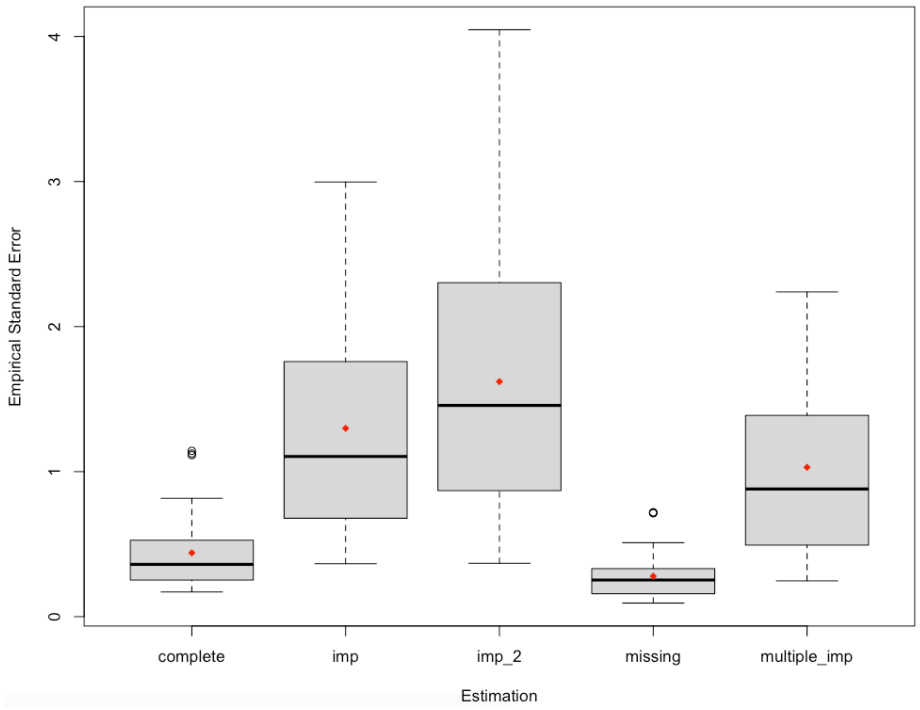

Figure 12 Empirical standard errors of the coefficient estimates for the exposure term

When the data are complete, the average empirical standard error is 0.44, with a standard deviation of 0.25. When there are missing ties but no imputation used, the average empirical standard error is 0.28 with a standard deviation of 0.16. If we do nothing about the missing tie data, the empirical standard error will be under-estimated. When using the proposed imputation method with thresholds maximizing precision and recall, the average empirical standard error is 1.30, with a standard deviation of 0.76. When using the multiple imputation method, the average empirical standard error is 1.03, with a standard deviation of 0.16. The multiple imputation

method gives a lower empirical standard error. However, these two results are not significantly different (pairwise Wilcoxon signed test, p-value = 0.19). Nevertheless, the multiple imputation method performed the best in terms of the empirical standard error.

The relationship between the empirical standard error and the network properties, including the clustering effect, network density, number of actors, and number of events, were examined. No significant relationship was found.

3.3 Root Mean Square Error

The root mean square error (RMSE) considers the bias and the empirical standard error at the same time. It was calculated for the five situations under different network settings. Figure 13 shows the results. The formula for the RMSE is $\sqrt{\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\left(\hat{\beta}_{1,i}-\beta_1\right)^2}$. For how $\hat{\beta}_{1,i}$, β_1, and n_{sim} are defined, please see the description in the '*Bias for the Coefficient Estimate of the Exposure Term*' section.

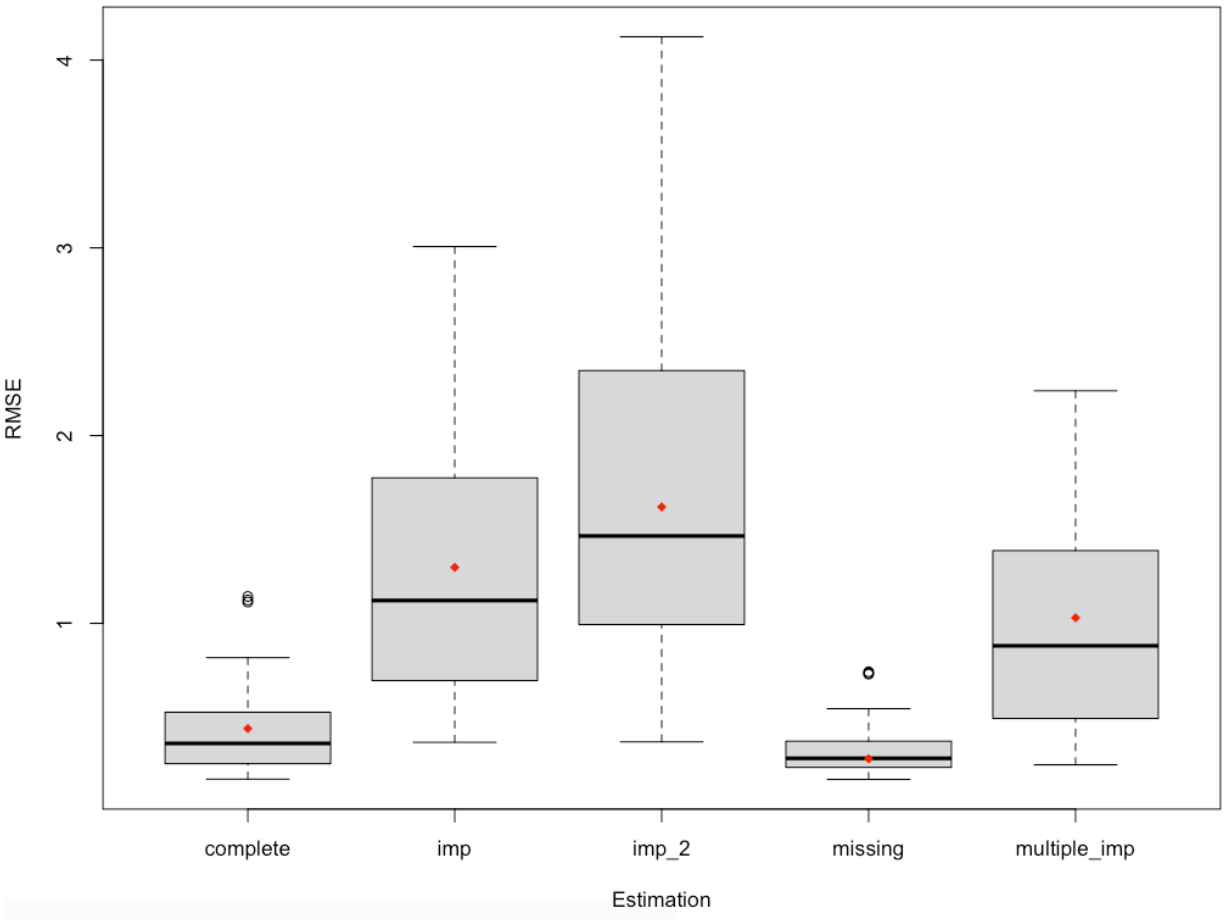

Figure 13 RMSEs of the coefficient estimates for the exposure term

In figure 13, black bars represent the median of RMSEs, and red dots represent the mean of RMSEs. When using the proposed imputation method with threshold maximizing precision and recall, the average RMSE is 1.31, with a standard deviation of 0.76. When using the multiple imputation method, the average RMSE is 1.03, with a standard deviation of 0.59. The pairwise Wilcoxon signed test shows that these two results are not significantly different (p-value = 0.19). Nevertheless, the multiple imputation method gives a smaller RMSE.

In summary, based on the *influence model*, when there are *missing-at-random* data in *two-mode ties*, the *multiple imputation* method performed the best in terms of absolute values of bias, empirical standard error, and RMSE. The multiple imputation method gives significantly

smaller biases than the proposed imputation method. This could be because the imputation model used in the multiple imputation method leverages the missing data mechanism used to generate the missing data. The proposed imputation method produced significantly smaller biases when the clustering effect is larger. This aligns with the fact that the proposed imputation method utilizes the clustering effect in the imputation process. The multiple imputation method and the proposed imputation method with thresholds maximizing precision and recall are not significantly different in terms of empirical standard error and RMSE. The proposed imputation method with a threshold of .5 performed the worst in terms of all criteria.

CHAPTER 4: LIMITATIONS AND FUTURE WORK

4.1 Limitations

In this dissertation, the two-mode network's boundary was clearly defined in terms of the number of actors and the number of events. In real life, it might be challenging to have such a clear boundary. For example, when we studied climate change knowledge dissemination in the Great Lakes region, actors were climate change researchers, and events were occasions that actors communicated with each other. The network data used were attendance lists of conferences, meetings, workshops, etc. related to climate change research. We collected the list of events by interviewing experts in this field. Nevertheless, it can't be guaranteed that all relevant events in that period were included. Climate change researchers might interact with each other in some circumstances, but we don't know. Consequently, the collected network data would not reflect all exposures that happened. Similarly, it was not easy to know every climate change researcher's name in the Great Lakes region.

Additionally, the missing two-mode ties were assumed to be MAR. That is, the missingness depends on something that we measured, but not the missing data itself. In practical situations, the missingness could be MNAR. That is, the missingness depends on the values of the missing data. For instance, some event attendance lists might be confidential. Although we know that there was such an event where actors interacted with each other, we will not be able to use that information. MNAR is a more challenging situation.

4.2 Future Work

In other contexts of missing data, the *multiple imputation* method not only performs well when the missingness is MAR, it also performs better than other methods when the missingness

is MNAR. It is worth trying the multiple imputation method and the proposed imputation method when the missingness is MNAR and comparing the two methods' results. At the same time, researchers need to be aware of the complexity of the MNAR mechanism. MNAR means that the missingness depends on the missing values themselves. To impute missing values under MNAR, one essential step is to break the dependency between the missingness and the missing values, which needs to be thoughtfully integrated into the methodology.

Besides, when generating the outdegree of actors for the two-mode network, it was assumed that the outdegree is homogenous. That is, actors are similar in terms of the number of events participated. Another plausible assumption is that the outdegree is heterogeneous. In other words, some actors are more likely to attend events, while others are less likely. Researchers found that for some complex two-mode networks in the real world, the outdegree distribution for the first mode fits the Power Law distribution very well, with the exponent of the Power Law falls (Latapy et al., 2008). We could examine the effect of missing two-mode ties on parameter estimation in this situation. The proposed imputation method and the multiple imputation method could be used, and their results compared.

Additionally, the influence mode in the context of two-mode network analysis could have different types of exposure terms: exposure to events, and exposure to other actors. See Appendix A for details. A more complicated influence model could be examined.

Another idea is regarding variables used in the imputation model when using the *multiple imputation* method. In this dissertation, the only variable used in the imputation model is *formal*, with *formal* = 1 representing the event being a formal event. Either information about actors (e.g., actors' characteristics, odds of attending events) or information about events (e.g., events' characteristics, odds of being attended) could be used in the imputation model.

Finally, the proposed imputation method and the multiple imputation method could be used on one-mode network analysis when the influence model is used.

APPENDICES

APPENDIX A: INFLUENCE MODEL IDEAS AND NOTATIONS

Person-to-Event Relationship. Let $v_{i,q}^{p}$ be an indicator of person-to-event relationship, where i is an index for actors with $i = 1, \ldots, I$, q is an index for events with $q = 1, \ldots, Q$ and p is an index for types of person-to-event relationship with $p = 1, 2, 3, 4$. I is the total number of actors in the sample. Q is the total number of events in the sample. Below is a table representing $v_{i,q}^{p}$, where $p = 1, 2, 3, 4$.

Table A1

Two-mode event attendance network: person-to-event ties

Person \ Event	**Same Cluster**	**Different Cluster**
Attended	$v_{i,q}^{1}$	$v_{i,q}^{2}$
Did not Attend	$v_{i,q}^{3}$	$v_{i,q}^{4}$ (no exposure)

The term $v_{i,q}^{1}$ is an indicator of whether actor i attended event q given actor i and event q have the same cluster membership; the term $v_{i,q}^{2}$-is an indicator of whether actor i attended event q given actor i and event q have different cluster memberships; the term $v_{i,q}^{3}$-is an indicator of whether actor i did not attend event q, given actor i and event q have the same cluster membership; the term $v_{i,q}^{4}$-is an indicator of whether actor i did not attend event q given actor i and event q have different cluster memberships (no exposure). For every unique pair of actor i and event q, $v_{i,q}^{1} + v_{i,q}^{2} + v_{i,q}^{3} + v_{i,q}^{4} = 1$. That is, each unique pair of actor and event can only be in one single situation listed in table A1. $h_{i}^{p} = \sum_{q=1}^{Q} v_{i,q}^{p}$ is the number of events of actor i in cell

p (the four cells in table A1). For each actor i, $h_i^1 + h_i^2 + h_i^3 + h_i^4 = Q$. Q is the total number of events in the network. It is the same value for everyone.

Person-to-Person Relationship. Let $u_{i,i'}^j$ $(i \neq i')$ be an indicator of a person-to-person relationship, where i and i' are indexes for actors and j is an index for types of person-to-person relationship with $j = 1, 2, 3, 4$. Below is a table representing $u_{i,i'}^j$.

Table A2

Two-mode event attendance network: person-to-person ties

Person \ Person	**Same Cluster**	**Different Clusters**
Attended Event(s) Together	$u_{i,i'}^1$	$u_{i,i'}^2$
Did not Attend Event(s) Together	$u_{i,i'}^3$	$u_{i,i'}^4$ (no exposure)

The term $u_{i,i'}^1$ is an indicator of whether actors i and i' attended common event(s) given that they have the same cluster membership; the term $u_{i,i'}^2$ is an indicator of whether actors i and i' attended common event(s) given that they have different cluster memberships; the term $u_{i,i'}^3$ is an indicator of whether actors i and i' did not attend any event together, given that they have the same cluster membership; the term $u_{i,i'}^4$ is an indicator of whether actors i and i' did not attend any event together given that they have different cluster memberships (no exposure). For every unique pair of actors i and i', $u_{i,i'}^1 + u_{i,i'}^2 + u_{i,i'}^3 + u_{i,i'}^4 = 1$. That is, each unique pair of actors can only be in one situation listed in table A2. $g_i^j = \sum_{i'=1}^{I} u_{i,i'}^j$ is number of other actors of actor

i in cell j (the four cells in table A2). For each actor i, $g_i^1 + g_i^2 + g_i^3 + g_i^4 = I$. I is the total number of actors in the network. The value is the same for everyone.

Influence Model. Given the notation above, the influence model can be expressed as equation (A1).

$$y_{i,t} = \alpha + \rho_0 y_{i,t-1} + \sum_{p=1}^{4} \beta_p \frac{1}{h_i^p} \sum_{q=1}^{Q} v_{i,q}^p z_{q,t-1} + \sum_{j=1}^{4} \rho_j \frac{1}{g_i^j} \sum_{i'=1}^{I} u_{i,i'}^j y_{i',t-1} + \epsilon_{it} \qquad \text{(A1)}$$

$$\epsilon_{it} \sim iid\ N(0, \sigma^2)$$

$y_{i,t}$ is the dependent variable, representing actor $i's$ behavior at time t. $y_{i,t-1}$ is the prior, representing actor $i's$ behavior at time $t-1$. $y_{i',t-1}$ represents actor $i-'.'s$s behavior at time $t-1$. $\frac{1}{g_i^j}\sum_{i'=1}^{I} u_{i,i'}^j y_{i',t-1}$ represents exposure to other actors. $z_{q,t-1}$ represents information presented at event q which happened over the time interval from $t-1$ to t. $\frac{1}{h_i^p}\sum_{q=1}^{Q} v_{i,q}^p z_{q,t-1}$ represents exposure to events. To make it clearer, the influence model could be re-written as equation (A2).

$$y_{i,t} = \alpha + \rho_0 y_{i,t-1} \qquad \text{(A2)}$$

$$+\beta_1 \frac{1}{h_i^1} \sum_{q=1}^{Q} v_{i,q}^1 z_{q,t-1} \quad \text{(exposure to events that the actor attended with the actor and events in the same cluster)}$$

$$+\beta_2 \frac{1}{h_i^2} \sum_{q=1}^{Q} v_{i,q}^2 z_{q,t-1} \quad \text{(exposure to events that the actor attended with the actor and events in different clusters)}$$

$$+\beta_3 \frac{1}{h_i^3} \sum_{q=1}^{Q} v_{i,q}^3 z_{q,t-1} \quad \text{(exposure to events that the actor didn't attend with the actor and events in the same cluster)}$$

$$+\beta_4 \frac{1}{h_i^4} \sum_{q=1}^{Q} v_{i,q}^4 z_{q,t-1}$$ (expected zero exposure to events that the actor didn't attend with the actor and events in different clusters)

$$+\rho_1 \frac{1}{g_i^1} \sum_{i'=1}^{I} u_{i,i'}^1 y_{i',t-1}$$ (exposure to other actors who are in the same cluster and attended events together)

$$+\rho_2 \frac{1}{g_i^2} \sum_{i'=1}^{I} u_{i,i'}^2 y_{i',t-1}$$ (exposure to other actors who are in different clusters and attended events together)

$$+\rho_3 \frac{1}{g_i^3} \sum_{i'=1}^{I} u_{i,i'}^3 y_{i',t-1}$$ (exposure to other actors who are in the same cluster but didn't attend events together)

$$+\rho_4 \frac{1}{g_i^4} \sum_{i'=1}^{I} u_{i,i'}^4 y_{i',t-1}$$ (expected zero exposure to other actors who are in different clusters and didn't attend events together)

$+\epsilon_{it} \quad \epsilon_{it} \sim iid\ N(0, \sigma^2)$ $\epsilon_{it} \sim iid\ N(0, \sigma^2)$

In this dissertation, a simpler version of the influence model was be used. Instead of having eight exposure terms as stated in equation (A2), there was one exposure term, exposure to events that the actor attended with the actor and events in the same cluster. Please see equation (1) in the main text.

APPENDIX B: DERIVATION: FORMULA FOR ASSIGNING TWO-MODE TIES

An actor's event-attendance information can be summarized in table B1.

Table B1

An actor's event attendance

		Same Cluster	
		Yes	**No**
Attended	**Yes**	a	b
	No	c	d

Assume that there are k events in the network, then for any actor, there are k possible actor-to-event pairs. The actor attended some events (a + b) and did not attend others (c + d). Some events are in the actor's cluster (a + c), while other events are outside the actor's cluster (b + d). For each actor a + b + c + d = k, the total number of events in the network. Note that a + b is the actor's outdegree, i.e., the number of events the actor attended. a + c is the number of events within the actor's cluster, despite whether the actor attended it or not. The odds of attending events within-cluster is a/c; the odds of attending events outside-cluster is b/d. Therefore, the odds ratio of attending events within-cluster vs. attending events outside-cluster is $\frac{ad}{bc}$. Summarizing the above information, we have the following four equations for each actor.

$$a + b = \alpha \ = outdegree \tag{B1}$$

$$a + c = \beta = \ the\ number\ of\ within - cluster\ pairs \tag{B2}$$

$$a + b + c + d = k = \ the\ number\ of\ events\ in\ the\ network \tag{B3}$$

$$\frac{ad}{bc} = r = \ 1, 2\ or\ 5 \tag{B4}$$

From equation (11), $b = \alpha - a$. From equation (12), $c = \beta - a$. From equation (13), $d = k - a - b - c = k - a - (\alpha - a) - (\beta - a) = k + a - \alpha - \beta$. From equation (14), $rbc - ad = 0$.

Substitute b, c, and d, $r(\alpha - a)(\beta - a) - a(k + a - \alpha - \beta) = 0$. Rearrange, we get $(r-1)a^2 - [k + (\alpha + \beta)(r-1)]a + r\alpha\beta = 0$. This is a quadratic equation with respect to a. The solutions are

$$a = \frac{k + (\alpha + \beta)(r-1) \pm \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

Then,

$$b = \frac{-k + (\alpha - \beta)(r-1) \mp \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

$$c = \frac{-k + (\beta - \alpha)(r-1) \mp \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

$$d = \frac{k(2r-1) - (\alpha + \beta)(r-1) \pm \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

For the first set of solutions (+, -, -, + before the square root), at least one of b and c is smaller than zero. In the context of this dissertation, we need positive values for a, b, c, and d. Therefore, the correct set of solutions for the current context are

$$a = \frac{k + (\alpha + \beta)(r-1) - \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

$$b = \frac{-k + (\alpha - \beta)(r-1) + \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

$$c = \frac{-k + (\beta - \alpha)(r-1) + \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

$$d = \frac{k(2r-1) - (\alpha + \beta)(r-1) - \sqrt{[k + (\alpha + \beta)(r-1)]^2 - 4r(r-1)\alpha\beta}}{2(r-1)}$$

REFERENCES

REFERENCES

Anderson, T. W. (1957). Maximum Likelihood Estimates for a Multivariate Normal Distribution when some Observations are Missing. *Journal of the American Statistical Association*. https://doi.org/10.2307/2280845

Azen, S. P., van Guilder, M., & Hill, M. A. (1989). Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomly incomplete data. Statistics in Medicine. https://doi.org/10.1002/sim.4780080208

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psychology. https://doi.org/10.1016/j.jsp.2009.10.001

Bartlett, J. W., Carpenter, J. R., Tilling, K., & Vansteelandt, S. (2014). Improving upon the efficiency of complete case analysis when covariates are MNAR. Biostatistics. https://doi.org/10.1093/biostatistics/kxu023

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*. https://doi.org/10.1080/10705510802339072

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*. https://doi.org/10.1037/1082-989X.6.4.330

Dempster, A. P. P., Laird, N. M., D.B. Rubin, & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical Society. Series B (Methodological)*. https://doi.org/10.1.1.133.4884

Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X00032325

Enders, C. K. (2010). *Applied Missing Data Anlysis.* New York, NY: The Guilford Press.

Frank, K. A., Muller, C., Schiller, K. S., Riegle-Crumb, C., Mueller, A. S., Crosnoe, R., & Pearson, J. (2008). The social dynamics of mathematics coursetaking in high school. *American Journal of Sociology*. https://doi.org/10.1086/587153

Fujimoto, K., Chou, C. P., & Valente, T. W. (2011). The network autocorrelation model using two-mode data: Affiliation exposure and potential bias in the autocorrelation parameter. *Social Networks*. https://doi.org/10.1016/j.socnet.2011.06.001

Glasser, M. (1964). Linear Regression Analysis with Missing Observations among the Independent Variables. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1964.10480730

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1993.10476366

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*. https://doi.org/10.1207/S15328007SEM0703_1

Graham, J. W. (2012). Missing data: Analysis and design. In *Missing Data: Analysis and Design*. https://doi.org/10.1007/978-1-4614-4018-5

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*. https://doi.org/10.1007/s11121-007-0070-9

Guillaume, J. L., & Latapy, M. (2004). Bipartite structure of all complex networks. *Information Processing Letters*. https://doi.org/10.1016/j.ipl.2004.03.007

Haitovsky, Y. (1968). Missing Data in Regression Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1968.tb01507.x

Hartley, H. O., & Hocking, R. R. (1971). The Analysis of Incomplete Data. *Biometrics*. https://doi.org/10.2307/2528820

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*. https://doi.org/10.1007/s10588-008-9040-4

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*. https://doi.org/10.1198/016214502388618906

Huisman, M. (2014). Imputation of Missing Network Data: Some Simple Procedures. In *Encyclopedia of Social Network Analysis and Mining*. https://doi.org/10.1007/978-1-4614-6170-8_394

Killworth, P. D., McCarty, C., Johnsen, E. C., Bernard, H. R., & Shelley, G. A. (2006). Investigating the variation of personal network size under unknown error conditions. In *Sociological Methods and Research*. https://doi.org/10.1177/0049124106289160

Kim, J. on, & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*. https://doi.org/10.1177/004912417700600206

Latapy, M., Magnien, C., & Vecchio, N. Del. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*. https://doi.org/10.1016/j.socnet.2007.04.006

Lazega, E., Wasserman, S., & Faust, K. (1995). Social Network Analysis: Methods and Applications. *Revue Française de Sociologie*. https://doi.org/10.2307/3322457

Little, R. J., & Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*. https://doi.org/10.1111/j.1467-9876.2011.00763.x

Little, R J A, & Rubin, D. B. (2002). Statistical Analysis with Missing Data: Second Edition. In *Wiley Series in Probability and Statistics*. https://doi.org/10.7710/2162-3309.1095

Little, Roderick J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1988.10478722

Little, Roderick J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1992.10476282

Little, Roderick J.A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician*. https://doi.org/10.1080/00031305.1983.10483106

Minhas, S., Hoff, P. D., & Ward, M. D. (2019). Inferential Approaches for Network Analysis: AMEN for Latent Factor Models. *Political Analysis*. https://doi.org/10.1017/pan.2018.50

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. https://doi.org/10.1002/sim.8086

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*. https://doi.org/10.1007/BF02294365

Newman, D. A. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. *Organizational Research Methods*. https://doi.org/10.1177/1094428103254673

Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*. https://doi.org/10.1198/016214501753208735

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*.

Pesantez-Cabrera, P., & Kalyanaraman, A. (2016). Detecting communities in biological bipartite networks. *ACM-BCB 2016 - 7th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. https://doi.org/10.1145/2975167.2975177

Royston, P. (2004). Multiple Imputation of Missing Values. *The Stata Journal: Promoting Communications on Statistics and Stata*. https://doi.org/10.1177/1536867x0400400301

Rubin, D. B. (1976). Inference and missing data. *Biometrika*. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys Donald B. Rubin. In *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/9780470316696

Schafer, J. L. (1997). Analysis of incomplete multivariate data. *Statistics in Medicine*. https://doi.org/10.1002/(SICI)1097-0258(20000415)19:7<1006::AID-SIM384>3.0.CO;2-T

Schafer, Josepn L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*. https://doi.org/10.1037/1082-989X.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. In *Multivariate Behavioral Research*. https://doi.org/10.1207/s15327906mbr3304_5

Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*. https://doi.org/10.1080/00949655.2018.1491577

Sewell, D. K., & Chen, Y. (2015). Latent Space Models for Dynamic Networks. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2014.988214

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1987.10478458

Trawinski, I. M., & Bargmann, R. E. (1964). Maximum Likelihood Estimation with Incomplete Multivariate Data. *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177703562

Valente, T. W. (2010). Social Networks and Health: Models, Methods, and Applications. In *Social Networks and Health: Models, Methods, and Applications*. https://doi.org/10.1093/acprof:oso/9780195301014.001.0001

Van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition. In *Flexible Imputation of Missing Data, Second Edition*. https://doi.org/10.1201/9780429492259

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*. https://doi.org/10.1080/10629360600810434

Von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*. https://doi.org/10.1111/j.1467-9531.2007.00180.x

Van Praag, B. M. S., Dijkstra, T. K., & Van Velzen, J. (1985). Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem. Psychometrika. https://doi.org/10.1007/BF02294145

Wang, Y. J., & Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the* American Statistical Association. https://doi.org/10.1080/01621459.1987.10478385

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. https://doi.org/10.1002/sim.3944

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. https://doi.org/10.1002/sim.4067

Widaman, K. F. (2006). III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, *71*(3), 42–64. https://doi.org/10.1111/j.1540-5834.2006.00404.x

Wilkinson, L. (1999). Statistical methods in psychology journals. *American Psychologist*.