Statistical Methods for Networks with Node Covariates

by

Yumu Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Ji Zhu, Chair
Assistant Professor Yang Chen
Assistant Professor Walter Dempsey
Professor Elizaveta Levina

Yumu Liu

liuyumu@umich.edu

ORCID ID: 0000-0001-5108-1654

ACKNOWLEDGEMENTS

I would like to express my greatest appreciation to my advisor Professor Ji Zhu for his kindness and providing helpful advice and discussions on the research projects, and also for his endless support and encouragement outside academic during my PhD study. I am also very grateful to Professor Elizaveta Levina, who has provided a lot of helpful advice and guidance as a close member of the research group. Further, I would like to thank Professor Yang Chen and Professor Walter Dempsey for serving as my committee member and providing insightful suggestions. Last but not least, many thanks should be attributed to my parents who provided me this opportunity to start this journey and continuously provided me with their support and encouragement both physically and mentally along these years.

Five years ago, I was entering a vast forest in the field of science. I stood in awe before the lush green tree of statistics. I was driven to explore it in hopes of making a discovery. The various forms and amazing colors of even the smallest leaves have inspired me to pursuit my study. I am fortunate to be a part of the family of the Department of Statistics at the University of Michigan, the time here will be a precious and memorable part of my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Network data, which represent relations or interactions between individual entities, together with nodal covariates information, arise in many scientific and engineering fields such as biology and social science. This dissertation focuses on developing statistical models and theory that utilize information from both the network structure and node covariates to improve statistical learning tasks, such as community detection and missing value imputation.

The first project studies the problem of community detection for degree-heterogeneous networks with covariates, where we aim to cluster the nodes into groups that share similar patterns in link connectivity and/or covariates distribution. We consider incorporating node covariates via a flexible degree-corrected block model by allowing the community memberships to depend on node covariates, while the link probabilities are determined by both node community memberships and degree parameters. We develop two algorithms, one using the variational inference and the other based on the pseudo-likelihood for estimating the proposed model. Simulation studies indicate that the proposed model can obtain better community detection results compared to methods that only utilize the network information. Further, we show that under mild conditions, the community memberships and the covariate parameters can be estimated consistently.

The second project considers the problem of missing value imputation when individuals are linked through a network. We assume the edges in the network are

related with the distances in the covariates of the individuals through a latent space network model. We propose an iterative imputation algorithm that is flexible and utilizes both the correlation among node variables and the connectivity between observations given by the network. We relate the proposed method to a Bayesian model and discuss the convergence of the imputation distribution when the specified conditional models for imputation are compatible with the true underlying model of the covariates. We also use simulation studies and a data example to illustrate empirically that the imputation accuracy can be improved by incorporating network information.

The final contribution of this dissertation is on incorporating covariates under the edge exchangeable framework. Edge exchangeable models have attractive theoretical and practical properties which make them appropriate for modeling many sparse real-world interaction networks constructed through edge sampling mechanisms. However, as far as we know, there is no edge exchangeable network model that allows for node covariates. In the third project, we propose a model that incorporates node covariates under the edge exchangeable model framework and show that it enjoys properties such as sparsity, and partial exchangeability. We further develop a maximum likelihood estimation method to estimate the model parameters and demonstrate its performance through both simulation studies and a data example.

CHAPTER I

Introduction

Network data arise naturally in many areas nowadays due to the advances in technology. In these network data, researchers use edges to represent relations or interactions between entities represented by nodes. Examples include but not limited to social networks where nodes are individual persons and edges are relations like friendships, biological networks where nodes are proteins and edges are their interactions, and etc (*Karrer and Newman*, 2011). Many works in the past decades have built up various tools for analyzing the structures or the development of the networks and have different focuses.

On analyzing the structure of a network, community detection is one of the most important questions that was widely studied. Community detection aims to cluster the nodes in the networks into communities with similar connectivity patterns (*Fortunato*, 2010). The study of community structures in network can be dated back to *Zachary* (1977) with empirical observations that real-world networks typically showed a pattern that nodes form groups with more connections within the same group than between groups. Many statistical models have been established to understand and uncover the community structure, among which the stochastic blockmodel (*Holland et al.*, 1983) and its extensions, including the mixed membership stochastic blockmod-

els (*Airoldi et al.*, 2008) and the degree corrected stochastic blockmodels (*Karrer and Newman*, 2011) are the most popular ones. Model free methods based on modularity criteria (*Newman*, 2006), spectral methods (*Rohe et al.*, 2011; *Qin and Rohe*, 2013; *White and Smyth*, 2005), and other methods (*Veldt et al.*, 2018; *Zhao et al.*, 2011; *Wang et al.*, 2011; *Amini and Levina*, 2014) are also available. Theoretical guarantees of community detection has also been established for various models based on stochastic blockmodels and various methods (*Zhao et al.*, 2012; *Bickel et al.*, 2013; *Celisse et al.*, 2012; *Lei et al.*, 2015). Another set of literature that focus on explaining the network structure assume that the nodes of the networks live in a low dimensional euclidean space. This includes the latent space models (*Hoff et al.*, 2002; *Hoff*, 2005), and the random dot product graph model (*Young and Scheinerman*, 2007). These models have the potential to explain some higher order characteristics in networks like abundance of triangles, which is not captured in stochastic blockmodels (*Hoff*, 2005). And the random dot product graph shows nice limiting properties (*Athreya et al.*, 2016). More details of the models can be seen in the survey paper by *Athreya et al.* (2017).

Other than focusing on the structure of a snapshot of the network, researchers also showed interests in understanding the mechanisms for how the networks are evolved over time, and also in explaining some commonly observed characteristics in real-world networks. Specifically, the sparse network phenomenon and the power-law degree distribution are of attention (*Barabási and Albert*, 1999; *Leskovec et al.*, 2008). The sparse network and power-law degree distributions are related to the study of preferential attachment models (*Newman*, 2001; *Wan et al.*, 2017; *Vázquez*, 2003; *Jeong et al.*, 2003), and are important factors that initiates the study of edge-exchangeable models (*Crane and Dempsey*, 2018; *Cai et al.*, 2016). On the exchangeability structure of network models, most of the literature have been focusing on node

exchangeable models including random dot product graph (*Young and Scheinerman*, 2007), stochastic blockmodels (*Holland et al.*, 1983), and graphon models (*Wolfe and Olhede*, 2013; *Choi et al.*, 2014). The edge changeable framework developed in *Crane and Dempsey* (2018); *Cai et al.* (2016) has been attracting the attention of researcher due to its nice limiting properties mentioned above and interpretation in terms of sampling. Specifically, edge exchangeable framework is natural in the case where the network data is collected by directly sampling edges instead of sampling the nodes.

Along with the network, often the traditional covariates information are also collected on each node, such as the characteristics of each person in a social network (*Leskovec and Mcauley*, 2012; *Van de Bunt et al.*, 1999). These covariates may contain information that are related to the network structures. For example, the nodes that are connected in a social network may have similar covariates, which is known as the homophily phenomenon (*McPherson et al.*, 2001; *Fujimoto and Valente*, 2012; *Christakis and Fowler*, 2007). Thus, such information can be very helpful in understanding the network structures of interest. Some models and methods that incorporate the covariates information to assist community detection have been developed in the literature, see for example (*Newman and Clauset*, 2016; *Xu et al.*, 2012; *Yang et al.*, 2013; *Binkiewicz et al.*, 2014; *Weng and Feng*, 2016). In more traditional multivariate analysis, the covariates themselves can be of interests. In that case, the network information may be helpful to improve the analysis to the covariates. Only a handful of literature considered such setting and focused on prediction (*Asur and Huberman*, 2010; *Wolf et al.*, 2009; *Li et al.*, 2019). In social science studies, methods have been proposed to make inference on causal effects with network interference (*Shalizi and Thomas*, 2011; *Manski*, 2013; *Kao*, 2017; *Basse and Airoldi*, 2015). In machine learning literature, heuristics like label propagation algorithms (*Zhur and Ghahramanirh*, 2002) have been developed for classifying nodes on a network with a part of the node

labels observed, which can be viewed as imputation for a single categorical variable utilizing network information.

This dissertation aims to develop models and statistical procedures that incorporate both network and node covariates information to improve the performance in community detection, missing value imputation, and in enriching the edge exchangeable models. For all the three targets, we view the observed network as a random object generated from some underlying distributions or mechanisms. The distribution of the covariates may or may not be considered random depending on specific applications.

The rest of the thesis is organized as follows:

Chapter II focuses on improving community detection by utilizing covariates in degree heterogeneous networks. Specifically, we proposed a model that combines the degree-corrected stochastic blockmodel and models for multivariate classification or clustering. We developed two algorithms for estimating the model based on variational inference or pseudo likelihood method. We established consistency results for the pseudo likelihood algorithm and illustrated that the community detection result was improved with covariates information incorporated.

Chapter III focuses on improving multivariate missing value imputation by incorporating network information. We assumed that the probability of two nodes connecting with each other in the network is correlated to the distance between the covariates of the two node through a latent space model. We considered combining the flexible iterative imputation with chained equation framework with the network model and developed gradient based methods for making imputations. We discussed the connection between the iterative imputation with chained equation framework and the

Gibbs sampling and Bayesian models.

Chapter IV extends the Hollywood model, a canonical edge exchangeable model to incorporate node covariates. The main difficult of incorporating covariates into edge exchangeable models is that the exchangeability structure can be broken easily. We address the question that to what extent we can preserve the edge exchangeability and what it means from a sampling perspective. We developed an estimation algorithm for the model and illustrated the model using simulation and Enron email data. We also showed that the proposed model inherits the sparsity property in limit.

CHAPTER II

Network Community Detection via the Degree-corrected Block Model with Node Covariates

2.1 Introduction

A commonly asked question when studying a network is that "can we identify groups of nodes that share similar connectivity patterns", which leads to the community detection problem, one of the fundamental problems in network analysis. Many methods have been proposed and studied, and they can be mainly divided into two categories: (1) model-free methods that do not try to fit a generative probabilistic model, and (2) model-based approaches using probabilistic network models. Notice that the two categories are not totally divided, many model-free methods also exhibit good performance under commonly used probabilistic network models.

In the model-free regime, different methods have been considered for community detection. For example, *Newman* (2006) proposed modularity as a criterion representing the "strength" of a community assignment and transformed the community detection problem to optimization of a certain criterion. Another approach to the problem is by exploring spectral properties of the adjacency matrix or the corre-

sponding Laplacian matrix. Literature including *Jin* (2015); *Qin and Rohe* (2013); *Rohe et al.* (2011) have developed various spectral clustering algorithms and analyzed their performances.

In model-based approaches, the stochastic blockmodel (*Holland et al.*, 1983) is probably the most commonly used model. For a network with n nodes, given node labels $c_i \in 1, 2, ..., K$, the probability of having an edge between node i and j is

$$P(A_{ij} = 1 | c_i, c_j) = B_{c_i c_j} \ ,$$

where $\{B_{ab}\}$ is a $K \times K$ parameter matrix. The model intuitively explains the community structure that nodes within the same group share similar link patterns. Extensions such as the mixed membership models (*Airoldi et al.*, 2008) and the degree-corrected blockmodel (*Karrer and Newman*, 2011) have been proposed to accommodate different real network properties. Specifically, the degree-corrected blockmodel assumes that there is a degree parameter θ_i associating with node i, and A_{ij} is Poisson distributed with

$$E(A_{ij} | c_i, c_j) = \theta_i \theta_j B_{c_i c_j} \ .$$

This allows for degree heterogeneity even within the same community, which makes the model much more flexible and works better in many real-world networks. Another important family of the network models is the latent space model that is studied by literature including *Hoff* (2005), *Hoff et al.* (2002). The latent space model assumes that each node has a latent position in some euclidean space and the probability of forming an edge between two nodes depends on some form of distance between their latent positions.

Fitting blockmodels is non-trivial as the problem essentially requires optimization

over all possible community label assignments. Estimation using MCMC under the Bayesian framework has been developed in the early stage (*Nowicki and Snijders*, 2001) and methods based on variational inference have been developed and studied to make the computation tractable recently (*Airoldi et al.*, 2008; *Bickel et al.*, 2013; *Celisse et al.*, 2012). Another way of fitting blockmodels is by the profile likelihood (*Bickel and Chen*, 2009; *Zhao et al.*, 2012), which establishes criteria that only depend on the label assignments by profiling out parameters for any fixed label assignments. Blockmodels are then fitted by optimizing these criteria via greedy algorithms. Recently, *Amini and Levina* (2014) proposed semi-definite relaxation that transforms the problem into an optimization where the argument is a semi-definite matrix by relaxing some constraints in the maximum likelihood problem. Last but not least, *Amini et al.* (2013) proposed a fast pseudo likelihood algorithm that scales well to large networks for both the stochastic blockmodel and the degree corrected extension by neglecting the dependence resulting from symmetry in undirected network.

For the community detection problem, the main theorectical interest lies in studying the consistency of estimated community label assignments. A commonly used notion of consistency is given in *Bickel and Chen* (2009) and *Zhao et al.* (2012):

$$\text{strong consistency: } P(\hat{\mathbf{c}} = \mathbf{c}) \to 1, \text{ as } n \to \infty$$

$$\text{weak consistency: } P(\frac{1}{n}\sum_{i=1}^{n} 1(\hat{\mathbf{c}} \neq \mathbf{c}) < \epsilon) \to 1, \text{ for any } \epsilon \text{ as } n \to \infty$$

Strong consistency of the clustering result has been established in general for profile likelihood based methods under the stochastic blockmodel family and its extension (*Bickel and Chen*, 2009; *Bickel et al.*, 2015; *Zhao et al.*, 2012) when the average degree of the graph is growing fast enough. Specifically, the average degree needs to grow faster than $\log n$. The consistency of variational inference under the stochas-

tic blockmodel has been obtained in *Mariadassou and Matias* (2015); *Celisse et al.* (2012). With mild assumptions on the initialization, pseudo likelihood method has been shown to be weakly consistent (*Amini et al.*, 2013). Spectral methods have also been shown to have similar theoretical guarantees (*Jin*, 2015; *Lei et al.*, 2015; *Qin and Rohe*, 2013; *Rohe et al.*, 2011). On top of the studies on consistency of clustering, asymptotic theories regarding the parameter estimates have also been established under the stochastic blockmodel for methods based on maximum likelihood (*Celisse et al.*, 2012) or its approximation via variational inference (*Bickel et al.*, 2013).

The work mentioned above focus on utilizing the network information alone. However, in real applications, especially in social networks, the structured network data collected using modern technologies often contain additional information on the nodes, or node covariates, about the individuals in the network. In many cases, it is natural to believe that these node covariates can be helpful in refining the communities that are given only using the network information in terms of both accuracy and interpretation. For example, in social networks, two people with similar background may have a higher probability to be connected.

Several methods have been developed to incorporate the node covariates into the community detection procedures. For examples, *Binkiewicz et al.*(2014) proposed a variant of spectral clustering by using the weighted sum of the graph laplacian and the gram matrix as input; *Yan and Sarkar*(2016) extended the semi-definite relaxation by adding in a k-means type penalty to the objective function; *Zhang et al.*(2015) proposed a joint community detection criterion representing the community strength together with node covariates similarities. *Weng and Feng*(2016) and *Newman and Clauset*(2016) have also considered to extend the stochastic blockmodel or the degree-corrected blockmodel to incorporate node covariates. However, *Weng and Feng*(2016) focuses only on the stochastic blockmodel, while *Newman and Clauset*(2016) can only

allow one categorical covariate.

In consideration of the flexibility, in this chapter, we propose a model that naturally combines the degree-corrected stochastic blockmodel and the classical logistic regression for the community detection problem. We choose the degree-corrected blockmodel to model the network as it is not only flexible theoretically, but also has been proven to perform well in fitting many real-world networks. We choose to use the multinomial logistic regression to model the relation between covariates and community labels as it does not make many assumptions on the distribution of covariates and is thus flexible. It is also possible to use mixture models, which might be helpful if we have prior knowledge about the distribution of the covariates. In section 2.3, we develop a variational EM algorithm and a pseudo likelihood algorithm for its estimation and study asymptotic properties for the pseudo likelihood algorithm in Section 2.4. We further illustrate their performances under various simulation settings and on a data example in Sections 2.5 and 2.6 respectively.

2.2 Model

Suppose we observe an undirected network of n nodes with self-loop and multi-edges allowed. We represents the network by an symmetric adjacency matrix $A = \{A_{ij}\}, i, j = 1, 2..., n$, whose diagonal element A_{ii} is equal to twice the number of edges from node i to itself. An $n \times p$ covariate matrix $X = \{X_{ij}\}, i = 1, 2..., n, j = 1, 2..., p$ is also observed. Besides, there is an unobserved community membership matrix $C = \{C_{ik}\}, i = 1, 2..., n, k = 1, 2..., K$ where K is the total number of communities, and $C_{ik} = 1$ if node i is in community k. We use c_i to denote the community label of node i, i.e. $c_i = k$ if node i is in community k.

Figure 2.1 illustrates the model we consider. Specifically the figure shows the de-

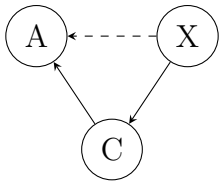


Figure 2.1: Graphical Representation of the Model

pendence structure between A, X and C. The joint probability of this model would be $P(A, X, C) = P(X)P(C|X)P(A|X, C)$. We do not model $P(X)$ as it does not involve C and instead, we work with $P(C, A|X) = P(A|X, C)P(C|X)$.

Given node covariates X and community membership C, we assume the network $A|X, C$ is generated from the Degree-Corrected Stochastic Blockmodel (DCBM)(*Karrer and Newman*, 2011). Then the network A is dependent on X implicitly through a set of latent degree correction parameters $\theta = \{\theta_i\}, i = 1, 2, ..., n$, which is represented by the dashed line in the graphical representation. Conditional on the node labels, the number of edges between a node pair (i, j) is Poisson distributed with mean $\theta_i\theta_j B_{c_i c_j}$, independent of any other node pairs, where $B = \{B_{rs}\}, r, s = 1, 2..., K$, is a $K \times K$ matrix determining the propensity of forming edges between nodes from community r and s. Since the network is undirected, B should be symmetric. Following this setting of DCBM we will have

$$\begin{aligned} P(A|X, C; \theta, B) &= \prod_{i<j} \frac{(\theta_i\theta_j B_{c_i c_j})^{A_{ij}}}{A_{ij}!} \exp(-\theta_i\theta_j B_{c_i c_j}) \\ &\times \prod_{i} \frac{(\frac{1}{2}\theta_i^2 B_{c_i c_j})^{A_{ii}/2}}{(A_{ii}/2)!} \exp(\frac{1}{2}\theta_i^2 B_{c_i c_j}) \\ &= \prod_{i<j} \prod_{k,l=1}^{K} \frac{(\theta_i\theta_j B_{kl})^{A_{ij}C_{ik}C_{jl}}}{(A_{ij}!)^{C_{ik}C_{jl}}} \exp(-\theta_i\theta_j B_{kl}C_{ik}C_{jl}) \\ &\times \prod_{i} \prod_{k=1}^{K} \frac{(\frac{1}{2}\theta_i^2 B_{kk})^{A_{ii}C_{ik}/2}}{((A_{ii}/2)!)^{C_{ik}}} \exp(-\frac{1}{2}\theta_i^2 B_{kk}C_{ik}) \ . \end{aligned} \tag{2.1}$$

We model $P(C|X)$ using a multinomial logistic regression model with parameter $\beta = \{\beta_k\}, k = 1, 2, ...K$. For identifiability, β_K is set to 0. Here we use a compact notation by implicitly including the intercept into β and let the covariates matrix X have a corresponding dummy column of 1,

$$\begin{aligned} P(C|X,\beta) &= \prod_{i=1}^{n} \frac{\exp(\beta_{c_i} x_i)}{\sum_{k=1}^{K} e^{\beta_k x_i}} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} (\frac{\exp(\beta_k x_i)}{\sum_{l=1}^{K} e^{\beta_l x_i}})^{C_{ik}} \ . \end{aligned} \tag{2.2}$$

2.3 Estimation

In this section, we develop a variational EM algorithm and a pseudo likelihood algorithm to estimate the model described in the previous section.

2.3.1 Variational EM algorithm

Since the community membership C is unknown, we can use the EM algorithm with C being the latent variable to find the maximum likelihood estimate of the parameters $\Theta = \{B, \beta, \theta\}$. As an intermediate step of EM algorithm, the estimation of the conditional distribution $P(C|A, X)$ is also calculated.

For the likelihood $P(A|X, \Theta)$ we have

$$\begin{aligned} \log P(A|X;\Theta) &\geq \log P(A|X,\Theta) - D_{KL}(R(C|A,X)||P(C|A,X,\Theta)) \\ &= \int_C R(C|A,X)[\log P(A,C|X,\Theta) - \log R(C|A,X)] \\ &:= \int_C R(C|A,X) l(\Theta|C) \\ &:= L(R,\Theta) \ , \end{aligned} \tag{2.3}$$

where $R(C|A,X)$ is any conditional distribution of C given A, X. EM algorithm tries to maximize L by alternately maximizing L over R and Θ, which gives

E-step: given $\Theta^{(t)}$, take $R^{(t+1)} = P(C|A,X,\Theta^{(t)})$ and compute

$$l(\Theta|\Theta^{(t)}) = \sum_C P(C|A,X,\Theta^{(t)}) l(\Theta|C)$$

M-step: compute

$$\Theta^{(t+1)} = \arg\max_{\Theta} l(\Theta|\Theta^{(t)}) \ .$$

In practice, however, there are K^n configurations of C and $P(C|A,X,\Theta^{(t)})$ cannot be factorized, which makes the E-step computationally intractable. To handle this difficulty, we use the variational approximation (*Jordan et al.*, 1999) which considers a restricted family of the conditional probability $P(C|A,X)$. Specifically, we consider $R(C|A,X) = \prod_{i=1}^{n} \prod_{k=1}^{K} \tau_{ik}^{C_{ik}}$, where τ is a set of variational parameters. With this choice of $R(C|A,X)$, we make the restriction that $C_i, i = 1, 2, ..., n$ are independent given A, X and follows a multinomial distribution with K-dimensional probability vector $\tau_{i\cdot}$.

The objective function to maximize now becomes

$$\begin{aligned} J(\tau,\Theta) &:= \sum_{C} R(C|A,X)[\log P(A,C|X,\Theta) - \log R(C|A,X)] \\ &= E[\log P(A,C|X,\Theta) - \log R(C|A,X)] \\ &= \sum_{i<j}\sum_{k,l=1}^{K} \tau_{ik}\tau_{jl}[A_{ij}\log(\theta_i\theta_j B_{kl}) - \log(A_{ij}!) - \theta_i\theta_j B_{kl}] \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}[\frac{A_{ii}}{2}\log(\frac{1}{2}\theta_i^2 B_{kk}) - \log((\frac{A_{ii}}{2})!) - \frac{1}{2}\theta_i^2 B_{kk}] \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}[\beta_k x_i - \log(\sum_{k=1}^{K} e^{\beta_k x_i})] - \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}\log(\tau_{ik}) \ . \end{aligned} \tag{2.4}$$

The maximization problem can then be solved using an EM-like algorithm with:

E-step: $\tau^{(t+1)} = \arg\max J(\tau, \Theta^{(t)})$

M-step: $\Theta^{(t+1)} = \arg\max J(\tau^{(t+1)}, \Theta)$.

2.3.1.1 E-step

The E-step is to maximize $J(\tau,\Theta)$ with respect to τ for given Θ; this can be solved by fixed point iteration. For $i = 1, 2, .., n$

$$\begin{aligned} \tau_{ik} \propto & (\frac{1}{2}\theta_i^2 B_{kk})^{\frac{A_{ii}}{2}} \exp(\beta_k x_i - \frac{1}{2}\theta_i^2 B_{kk}) \\ & \times \prod_{j=1, j\neq i}^{n}\prod_{l=1}^{K}[\frac{(\theta_i\theta_j B_{kl})^{A_{ij}}}{(A_{ij}!)}\exp(-\theta_i\theta_j B_{kl})]^{\tau_{jl}} \end{aligned} \tag{2.5}$$

for $k = 1, 2, ..., K$ and subject to $\sum_{k=1}^{K} \tau_{ik} = 1$.

2.3.1.2 M-step

The M-step is to maximize $J(\tau, \Theta)$ with respect to Θ for given τ; this can be divided into two sub-problems that optimize with respect to (θ, B) and β respectively. For (θ, B), the objective function is

$$\begin{aligned} g(\theta, B) = &\sum_{i<j} \sum_{k,l=1}^{K} \tau_{ik}\tau_{jl}[A_{ij} \log(\theta_i \theta_j B_{kl}) - \theta_i \theta_j B_{kl}] \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}[\frac{A_{ii}}{2} \log(\frac{1}{2}\theta_i^2 B_{kk}) - \frac{1}{2}\theta_i^2 B_{kk}] \ , \end{aligned} \tag{2.6}$$

with the constraints $B_{ij} = B_{ji}$ for undirected network, and $\theta_i > 0$ by definition. The parameters θ are only identifiable within a multiplicative constant that will be absorbed into B, so we also need the constraint $\sum_{i=1}^{n} \theta_i = 1$ for identifiability. To compute the maximizers, we set the gradient to 0 and iteratively solve the equation system. The objective function is concave with respect to each θ_i and the update for θ_i with all other variables fixed has analytical solutions as follows:

$$\begin{aligned} \hat{\theta}_i &= \frac{-b - \sqrt{b^2 - 4ac}}{2a} \\ a &= -\sum_{k=1}^{K} B_{kk}\tau_{ik} \\ b &= -\sum_{j \neq i} \sum_{k,l=1}^{K} \theta_j B_{kl} \tau_{ik} \tau_{jl} \\ c &= \sum_{j \neq i} \sum_{k,l=1}^{K} A_{ij} \tau_{ik} \tau_{jl} + \sum_{k=1}^{K} A_{ii} \tau_{ik} \ . \end{aligned} \tag{2.7}$$

θ_i are normalized to have sum equal to 1 to satisfy the identifiability constraint after each iteration of updates. In practice we run a fix number of iterations (2.7) to get a result close to convergence.

Updating B with θ fixed can be done using the following formulas as the objective

function is concave with respect to B.

$$\begin{aligned}\hat{B}_{kl} &= \frac{\sum_{i<j}[\tau_{ik}\tau_{jl}A_{ij} + \tau_{il}\tau_{jk}A_{ij}]}{\sum_{i<j}[\tau_{ik}\tau_{jl}\theta_i\theta_j + \tau_{il}\tau_{jk}\theta_i\theta_j]}, \text{ for } k \neq l \\ \hat{B}_{kk} &= \frac{\sum_{i<j}\tau_{ik}\tau_{jk}A_{ij} + \sum_i \tau_{ik}\frac{A_{ii}}{2}}{\sum_{i<j}\tau_{ik}\tau_{jk}\theta_i\theta_j + \sum_i \tau_{ik}\frac{\theta_i^2}{2}} .\end{aligned} \tag{2.8}$$

We then iterate (2.7) and (2.8) until convergence.

Maximizing $J(\tau, \Theta)$ with respect to β is a multinomial logistic regression problem. For numerical stability we include a ridge penalty and maximize

$$L(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}[\beta_k x_i - \log(\sum_{k=1}^{K} e^{\beta_k x_i})] - \lambda||\beta||_2^2 \tag{2.9}$$

where $\lambda > 0$ is a small positive number. This can be solved using existing packages, such as the glmnet package in R.

2.3.1.3 Initialization

The speed of convergence often depends on the initial values of the algorithm. We initialize the parameters by first initializing community labels using regularized spectral clustering(RSC) and initialize θ_i proportional to d_i, the degree of node i. Note this requires the network to have no isolated node. Then we set τ to be a binary matrix corresponding to the initial community labels. With τ, θ initialized, B and β can be calculated using (2.8) and (2.9).

2.3.2 Pseudo likelihood based algorithm

As mentioned in the variational EM algorithm, the main challenge in maximizing the joint likelihood using EM algorithm lies in the E-step as it is intractable. The main

idea of pseudo likelihood is to simplify the likelihood and make it tractable by ignoring some of the dependency structure. Specifically for stochastic blockmodels, a pseudo likelihood can be established by ignoring the symmetry of the adjacency matrix. In this section, we first give an brief review on the pseudo likelihood method introduced in *Amini et al.* (2013) and then make modifications to involve node covariates.

2.3.2.1 Pseudo likelihood for blockmodel

Let the true community label be denoted by $c_i, i = 1, 2, ..., K$. Given an initial labeling $\mathbf{e} = \{e_i\}, i = 1, 2, ..., n, e_i \in 1, 2, ..., K$, we will work with the following quantity,

$$b_{ik} = \sum_j A_{ij} 1(e_j = k), i = 1, ..., n, j = 1, ..., K. \tag{2.10}$$

Let $\mathbf{b}_i = (b_{i1}, b_{i2}, ..., b_{iK})$ and further let R be a $K \times K$ matrix with

$$R_{ka} = \frac{1}{n} \sum_{i=1}^{n} 1(e_i = k, c_i = a)$$

and $R_{k\cdot}$ be the k-th row. Let B be the $K \times K$ parameter matrix for the blockmodel and $B_{\cdot\ell}$ be the ℓ-th column of B. Let $\lambda_{\ell k} = nR_{k\cdot}B_{\cdot\ell}$ and $\Lambda = \{\lambda_{\ell k}\}$.

The pseudo likelihood for a stochastic blockmodel is established based on the following key observations: for each node i, given the true labels $\mathbf{c}$ with $c_i = \ell$,

(A) $\{b_{i1}, b_{i2}, ..., b_{iK}\}$ are mutually independent, and

(B) b_{ik} is approximately Poisson distributed with mean $\lambda_{\ell k}$.

With true label c_i unknown, $\mathbf{b}_i$ is a mixture of Poisson vectors. Then by ignoring the dependence between $\mathbf{b}_i, i = 1, 2, ..., n$, we have the following pseudo log-likelihood (up to a constant):

$$l_{PL}(\pi, \Lambda; \{\mathbf{b}_i\}) = \sum_{i=1}^{n} \log(\sum_{\ell=1}^{K} [\pi_\ell e^{-\lambda_\ell} \prod_{k=1}^{K} \lambda_{\ell k}^{b_{ik}}]), \tag{2.11}$$

where $\lambda_\ell = \sum_k \lambda_{\ell k}$ and π_ℓ is the probability of a node being in community ℓ.

2.3.2.2 Conditional pseudo likelihood for DCBM

Extending the pseudo likelihood to DCBM is non-trivial. The degree corrected model has n degree parameters, one for each node, which makes the pseudo likelihood and estimation much more complicated. *Amini et al.* (2013) proposed a simple alternative that considers the pseudo likelihood conditional on the observed node degrees. By conditioning on the observed node degrees, the degree parameters do not play a role in the pseudo likelihood anymore and we only need to focus on the block structures. The key observation in the conditional pseudo likelihood is that conditioning on the observed node degree $d_i = \sum_k b_{ik}$ and true community label $c_i = \ell$, the variables $(b_{i1}, b_{i2}, ..., b_{iK})$ are multinomially distributed with parameters $(d_i; \psi_{\ell 1}, \psi_{\ell 2}, ..., \psi_{\ell K})$, with $\psi_{\ell k} = \frac{\lambda_{\ell k}}{\lambda_\ell}$. Then we have the conditional log pseudo likelihood (up to a constant):

$$l_{CPL}(\pi, \{\psi_{lk}\}; \{\mathbf{b}_i\}) = \sum_{i=1}^{n} \log(\sum_{\ell=1}^{K} [\pi_\ell \prod_{k=1}^{K} \psi_{\ell k}^{b_{ik}}]). \tag{2.12}$$

2.3.2.3 Conditional pseudo likelihood for DCBM with node covariates

We now introduce the conditional pseudo likelihood with node covariates and develop a corresponding estimation algorithm.

Note that in our proposed model, the node covariates only affect the formation of the network through community probability. Thus, we replace π_ℓ by $\frac{\exp(\beta_\ell X_i)}{\sum_k \exp(\beta_k X_i)}$, which is the probability of node i being in community ℓ under the logistic regression model. This gives us the following conditional log pseudo likelihood (up to a

constant):

$$L_{CPL}(\beta, \{\psi_{\ell k}\}; \mathbf{b}_i) = \sum_{i=1}^{n} \log(\sum_{\ell=1}^{K} [\frac{\exp(\beta_\ell^T x_i)}{\sum_{k=1}^{K} \exp(\beta_k^T x_i)} \prod_{k=1}^{K} \psi_{\ell k}^{b_{ik}}]). \tag{2.13}$$

Then we can obtain the estimate of $\beta, \{\psi_{\ell k}\}$ by maximizing the conditional log pseudo likelihood via the EM algorithm for mixture models. With the parameter estimate, we update the initial label $\mathbf{e}$ and repeat the procedure for T iterations.

Let $n_k(e) = \sum_i 1(e_i = k)$, $n_{k\ell}(e) = n_k(e)n_l(e)$, $n_{kk}(e) = n_k(e)(n_k(e) - 1)$ and $O_{k\ell}(e) = \sum_{ij} A_{ij} 1(e_i = k, e_j = \ell)$. The algorithm consists of the following steps:

- Initialize label $\mathbf{e}$ using regularized spectral clustering, and initialize β correspondingly. Let $\hat{\pi}_\ell = n_\ell/n$, let $R = diag(\hat{\pi})$, $\hat{B}_{\ell k} = O_{\ell k}/n_{\ell k}$, $\hat{\lambda}_{\ell k} = n\hat{R}_{k\cdot}\hat{B}_{\cdot\ell}$ and initialize $\{\psi_{\ell k}\}$ by row normalization of Λ.
- Repeat T times:
  - (1) compute block sums $\mathbf{b}_i$ under current $\{e_i\}, i = 1, 2, ..., n$
  - (2) E-step: Given $\hat{\beta}, \{\hat{\psi}_{\ell k}\}$, compute $P_{i\ell} := P(c_i = \ell | \mathbf{b}_i, x_i)$

$$P_{i\ell} := P(c_i = \ell | \mathbf{b}_i, x_i) = \frac{\hat{\pi}_{i\ell} \prod_{m=1}^{K} \hat{\psi}_{\ell m}^{b_{im}}}{\sum_{k=1}^{K} \hat{\pi}_{ik} \prod_{m=1}^{K} \hat{\psi}_{km}^{b_{im}}} \tag{2.14}$$

where $\hat{\pi}_{i\ell} = \frac{\exp(\hat{\beta}_\ell^T x_i)}{\sum_{k=1}^{K} \exp(\hat{\beta}_k^T x_i)}$.

– (3) M-step: Given P_{il}, update β by logistic regression and update Θ

$$\hat{\psi}_{\ell k} = \frac{\sum_{i=1}^{n} P_{i\ell} b_{ik}}{\sum_{i=1}^{n} P_{i\ell} d_i} \tag{2.15}$$

– (4) repeat (2) and (3) until the parameters converge

– (5) update label $e_i = \arg\max_l P_{i\ell}$ and return to step (1) .

The algorithm typically only needs a few label updates until convergence, but the performance relies on suitable initial labels.

2.4 Theoretical properties

The main theoretical property that community detection methods pursue is consistency of the community estimates $\hat{\mathbf{c}}$. A commonly used definition of consistency is from *Bickel and Chen* (2009) and *Zhao et al.* (2012):

$$\text{strong consistency: } P(\hat{\mathbf{c}} = \mathbf{c}) \to 1, \text{ as } n \to \infty$$

$$\text{weak consistency: } P(\frac{1}{n}\sum_{i=1}^{n} 1(\hat{\mathbf{c}} \neq \mathbf{c}) < \epsilon) \to 1, \text{ for any } \epsilon \text{ as } n \to \infty .$$

Note that the consistency notion is up to label permutations. For example, switching the label of community 1 and community 2 does not change the community structure.

Although we have developed the algorithm with multi-edges allowed, we will study the theoretical property based on binary networks to simplify the problem. In most real applications, we only observe binary networks and care more about whether an edge is present or not rather the multiplicity of the edges, thus this simplification is

reasonable.

Methods based on maximum profile likelihood have been shown to be strongly consistent under both stochastic blockmodel and its degree-corrected extension (*Zhao et al.*, 2012), but consistency of community label estimates under variational inference is only established under stochastic blockmodel (*Mariadassou and Matias*, 2015; *Weng and Feng*, 2016), with indispensable dependence on the consistency result for estimating the $K \times K$ matrix parameter B that controls inter and intra community connectivity (*Bickel et al.*, 2013; *Celisse et al.*, 2012).However, extending these consistency results to variational inference under the degree-corrected blockmodel is challenging, since the consistency of the estimates for B cannot be easily obtained due to identifiability issues in the degree correction parameters. Nonetheless, the consistency of maximum profile likelihood still holds following *Zhao et al.* (2012) and the weak consistency result for pseudo likelihood algorithm similar to that in *Amini et al.* (2013) can be shown. We will focus on showing the consistency of the pseudo likelihood, based on the results from *Amini et al.* (2013).

For theoretical analysis of the pseudo likelihood algorithm, we only consider the case $K = 2$. Further, for simplicity, we assume that among the n nodes, $m = \frac{n}{2}$ nodes are in community 1 and assume the initial label $\mathbf{e}$ is also balanced, which means $\mathbf{e}$ assigns m nodes to community 1 and m nodes to community 2. We will first consider a directed graph. For the directed graph, we will use $\tilde{A}_{ij}$ to denote the adjacency matrix and $\tilde{B}$ to denote the $K \times K$ matrix parameter. The directed graph model is actually natural for pseudo likelihood approach since it is the model where the row independence assumption holds. We let the edge probability matrix of the directed graph $\tilde{B} = \frac{1}{m} \begin{pmatrix} a & b \\ b & a \end{pmatrix}$ with a and b scales with n.

The key assumption is that among the m nodes with initial labels being community 1, γm nodes are truly in community 1. It is not difficult to see that this also implies that the initial labels $\mathbf{e}$ also have γm correctly labeled nodes in community 2. We do not assume we know the value of γ or which labels are matched. But we do assume $\gamma \in (0,1)\backslash\{\frac{1}{2}\}$ and $m\gamma$ is an integer. Let $\mathcal{E}^\gamma$ denote the collection of all such initial labeling

$$\mathcal{E}^\gamma = \mathcal{E}_n^\gamma = \left\{ \mathbf{e} \in \{1,2\}^n : \sum_{i=1}^{m} 1(e_i = 1) = m\gamma = \sum_{i=m+1}^{n} 1(e_i = 2) \right\} .$$

We will focus on the E-step of the pseudo likelihood algorithm. With some initial estimates $\hat{a}, \hat{b}$ as well as $\hat{\beta}$, together with initial labeling $\mathbf{e}$, the labels are estimated by

$$\hat{c}_i(e) = \arg\max_{k\in\{1,2\}} \{\hat{\beta}_k X_i + \sum_{m=1}^{2} \tilde{b}_{im}(e) \log \hat{\gamma}_{km}(e)\} , \tag{2.16}$$

where $\hat{\gamma}_{km}$ are the elements of the row normalized matrix of $\tilde{\Lambda} = [nR(e)\tilde{B}]^T$, with $\tilde{B}$ estimated by plugging in $\hat{a}, \hat{b}$.

Let $C(\gamma) := \left[\log \frac{\hat{\psi}_{11}(e)}{\hat{\psi}_{21}(e)}\right]^{-1}$, and define the mismatch ratio

$$\tilde{M}_n(e) := \min_{\phi\in\{(1,2),(2,1)\}} \frac{1}{n} \sum_{i=1}^{n} 1[\hat{c}_i(e) \neq \phi(c_i)],$$

where ϕ is considering the fact that the labels are identified up to a permutation. We will show a consistency result based on the convergence of this mismatch ratio, which corresponds to the weak consistency definition.

Let us consider the initial estimates $(\hat{a}, \hat{b})$ that have the same ordering as true parameter (a, b) such that $(\hat{a} - \hat{b})(a - b) > 0$. We have the following result.

Theorem II.1. *Under the balanced communities assumption, let* $\gamma \in (0,1)\backslash\{\frac{1}{2}\}$.

Let the adjacency matrix $\tilde{A}$ be generated under the directed graph model with edge-probability matrix $\tilde{B}$ and assume $a \neq b$. In addition, assume $|\hat{\beta} X_i| \leq M$, where M is a constant, and

$$\begin{aligned} m &:= (1-2\gamma)(a-b) - |MC(\gamma)| \geq 0 \\ (1-2\gamma)(a-b) &+ |MC(\gamma)| \leq 3(a+b) \end{aligned} \tag{2.17}$$

Then there exists a positive sequence $\{u_n\}$ such that

$$\log u_n + \log\log u_n \geq \log(\frac{4}{e}) + \frac{m^2}{4(a+b)}$$

and the mismatch ratio has

$$P\left[\sup_{\sigma \in \mathcal{E}^\gamma} \tilde{M}_n(e) > \frac{4h(\gamma)}{\log u_n}\right] \leq \exp\{-[h(\gamma) - \kappa_\gamma(n)]n\} \tag{2.18}$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function and $\kappa_\gamma(n) := \frac{1}{n}[\log(\frac{n}{4\pi\gamma(1-\gamma)}) + \frac{1}{3n}] = o(1)$.
In particular, if $\frac{m^2}{4(a+b)} \to \infty$, we have $u_n \to \infty$ and the pseudo likelihood estimate is consistent.

Remark II.2. $\frac{m^2}{4(a+b)} \to \infty$ implies $(1-2\gamma)(a-b) \to \infty$ and the assumptions (2.17) are then satisfied for sufficiently large n, indifferent of the value of γ. The condition $\frac{m^2}{4(a+b)} \to \infty$ itself is also not strong if we assume γ fixed. For example, if we let $a = \log n$ and $b = ra, r \in (0,1)$, the condition is satisfied when n is large enough.

Remark II.3. With more assumptions on the initialization, the result of Theorem II.1 can be extended to a more general case where the communities are unbalanced, with size n_1, n_2 and edge probability matrix $\tilde{B} = \frac{1}{n}\begin{pmatrix} a_1 & b \\ b & a_2 \end{pmatrix}$. The assumption on initial labels is also relaxed to have $n_1\gamma_1$ nodes matching the true label in community 1 and $n_2\gamma_2$ in community 2, with $\gamma_1 \neq \gamma_2$. The details are discussed in the appendix.

Next we extend the result to the undirected case. Let $a_\gamma = \gamma a + (1-\gamma)b$. The undirected case is studied by introducing a coupling between the directed case and the undirected case. Specifically, the undirected adjacency matrix A is generated from the directed adjacency matrix $\tilde{A}$ by removing the edge directions, i.e.

$$A = T(\tilde{A}),\ [T(\tilde{A})]_{ij} = 1 - 1(\tilde{A}_{ij} = \tilde{A}_{ji} = 0)\ . \tag{2.19}$$

Then we have the edge probability matrix for the undirected graph:

$$B_{kl} = P(A_{ij} = 1) = 1 - P(\tilde{A}_{ij} = 0)P(\tilde{A}_{ji} = 0) = 2\tilde{B}_{kl} - \tilde{B}_{kl}^2\ .$$

Define the mismatch ratio for undirected case $M_n(e)$ similarly as in the directed case.

Theorem II.4. *Under the undirected model generated with edge-probability matrix* $\{B_{kl}\}$*, let* $\gamma \in (0,1)\backslash\{\frac{1}{2}\}$ *and assume* $a \neq b$*. In addition, we assume* $|\hat{\beta}X_i| \leq M$*, where* M *is a constant and*

$$2(1-\epsilon)a_\gamma \leq \epsilon(1-2\gamma)(a-b) \tag{2.20}$$

$$\begin{aligned} &m := (1-\epsilon)(1-2\gamma)(a-b) - |MC(\gamma)| \geq 0 \\ &(1-\epsilon)(1-2\gamma)(a-b) + |MC(\gamma)| \leq 3(a+b) \end{aligned} \tag{2.21}$$

for some $\epsilon \in (0,1)$*. Then there exist sequence* $\{u_n\}, \{v_n\}$ *such that*

$$\log u_n + \log\log u_n \geq \log(\frac{4}{e}h(\gamma)) + \frac{m^2}{4(a+b)}$$

$$\log v_n + \log\log v_n \geq \log(\frac{4}{e}h(\gamma)) + \frac{\epsilon^2}{1+\epsilon/3}a_\gamma$$

and

$$P\left[\sup_{e\in\mathcal{E}_n^\gamma} M_n(e) \geq 4h(\gamma)(\frac{1}{\log u_n}+\frac{2}{\log v_n})\right] \leq 3\exp(-n[h(\gamma)-\kappa_\gamma(n)])$$

where $h(\cdot)$ is the binary entropy function and $\kappa_\gamma(n) = o(1)$ defined as before. In particular, if $\frac{m^2}{4(a+b)} \to \infty$, $a_\gamma \to \infty$ we have $u_n \to \infty$, $v_n \to \infty$ and the pseudo likelihood estimate is consistent.

Remark II.5. Similar to the directed case, the assumptions (2.21) is satisfied for sufficiently large n if $\frac{m^2}{4(a+b)} \to \infty$. The condition (2.20) can be satisfied for fixed ϵ by letting γ small and upper bound $\frac{b}{a}$ in terms of γ. This means that there should not be too many inter-community edges comparing to within community edges in that case.

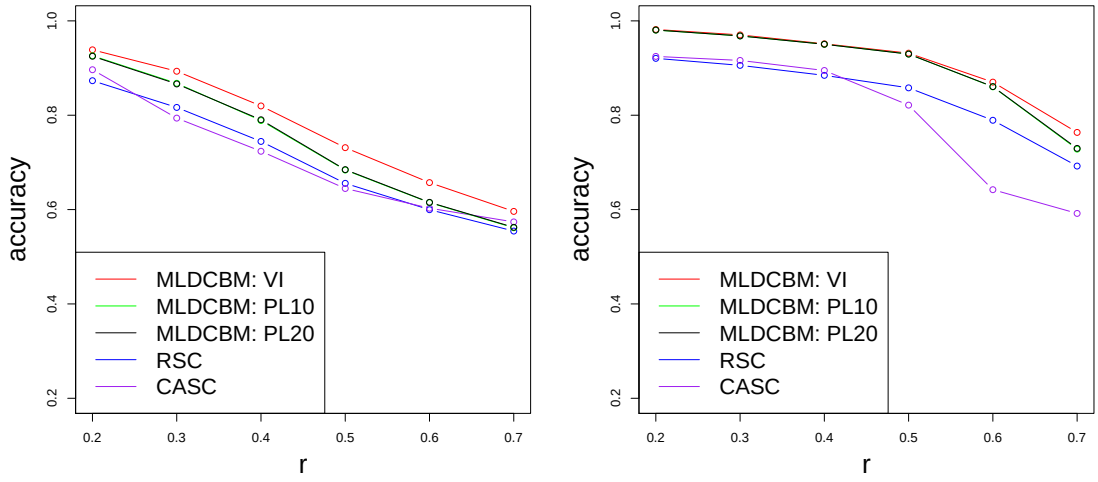
With the weak consistency in estimating community labels, we can further consider estimating the logistic regression parameter β by

$$\hat{\beta}_{PL} = \arg\max_\beta \frac{1}{n}\left[\sum_{i=1}^n 1(\hat{c}_i(e)=1)\beta X_i - \log(1+\beta X_i)\right]$$

Corollary II.6. *If the assumptions for weak consistency of pseudo likelihood hold, $\hat{\beta}_{PL}$ is consistent estimator of β.*

2.5 Simulation studies

In this section, we apply the proposed methods to simulated data generated from the model under different settings and compare the performance with regularized spectral clustering(RSC)(*Qin and Rohe*, 2013) and covariates assisted spectral clustering(CASC)(*Binkiewicz et al.*, 2014). We try $T = 10$ and $T = 20$ for the pseudo likelihood approach. We also give the computing time for the proposed algorithms

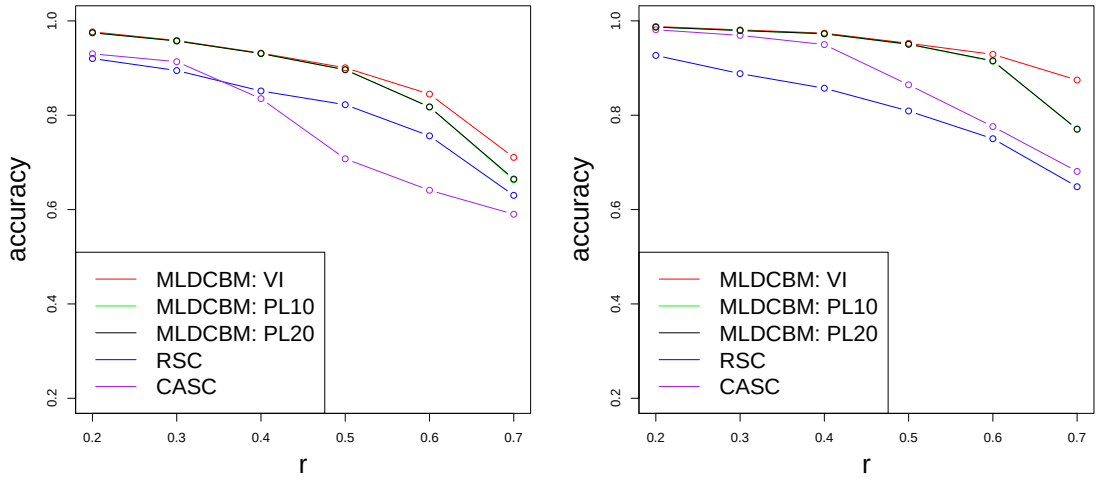


(a) p_{kk}=0.05 (b) p_{kk}=0.20

Figure 2.2: Mean accuracy vs r for $K = 2, \beta \sim U(-1, 1)$

under a specific setting.

We generate θ_i from $Beta(1, 5)$ distribution to approximate the Power-law degree distribution often found in real network data. Then we set the $B_{ab} = \omega_{within}$ for $a = b$ and $B_{ab} = r\omega_{within}$ where r is a number between 0 and 1. The ω_{within} controls p_{kk}, the mean number of edges between a pair of nodes within a group while r controls the inter-community edge density $p_{k\ell}$. We generate covariates independently from standard normal distribution and generate the logistic regression coefficients β from uniform distribution centered at 0. We change ω_{within} with β fixed to see how does edge density affect the performance and change β with ω_{within} fixed to evaluate the impact of the level of information contained in the covariates. We tested the model with $K = 2, n = 200, p = 5$, and $K = 5, n = 500, p = 10$, and take the average accuracy of clustering over 50 repetitions under each setting. In the figures our algorithms are labeled as "MLDCBM", which stands for "Multinomial Logistic Degree Corrected Blockmodel". Further, "VI" is short for "Variational Inference" and "PL" is short for "Pseudo Likelihood".

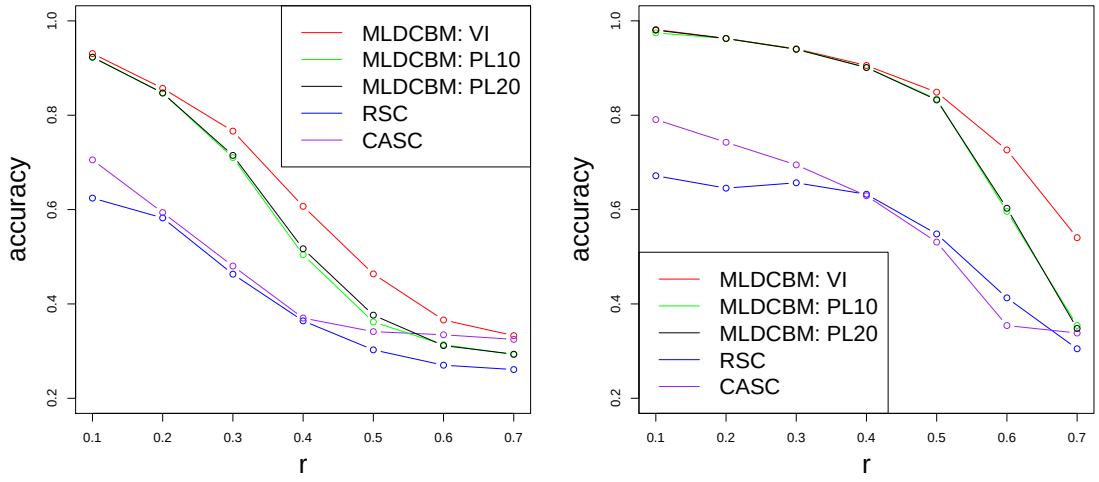


(a) p_{kk}=0.15, $\beta \sim U(-1, 1)$ (b) p_{kk}=0.15, $\beta \sim U(-5, 5)$

Figure 2.3: Mean accuracy vs r for $K = 2$

Figure 2.2 shows the result for 2 clusters and $\beta \sim U(-1, 1)$ with p_{kk} being 0.05 and 0.2 respectively. It can be seen that the proposed methods in general perform better than RSC and CASC. There is no obvious difference between $T = 10$ and $T = 20$ for the pseudo likelihood algorithm, which suggests the algorithm converges and the results suggest that pseudo likelihood algorithm works well in most cases. It should be mentioned that CASC itself is not necessarily asymptotically consistent, which may explain its inferior performance comparing to RSC in terms of clustering accuracy. As the graph becomes denser, the accuracy of clustering should become higher in general since the networks are more informative. Figure 2.3 shows the result for 2 clusters with p_{kk} fixed at 0.15 and β varies. The superior performance of our methods over RSC becomes more significant when β increases, which is natural as the covariates are more informative with larger β. Similar results are observed for 5 clusters and are shown in Figure 2.4 and Figure 2.5.

It should also be mentioned the performances of the proposed algorithms depend

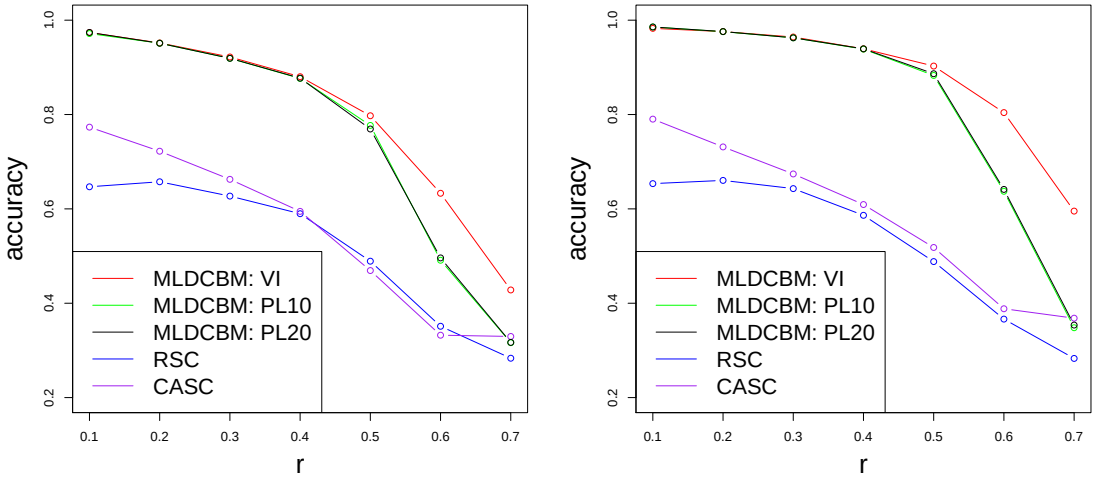


(a) p_{kk}=0.05 (b) p_{kk}=0.20

Figure 2.4: Mean accuracy vs r for $K = 5$, $\beta \sim U(-1, 1)$

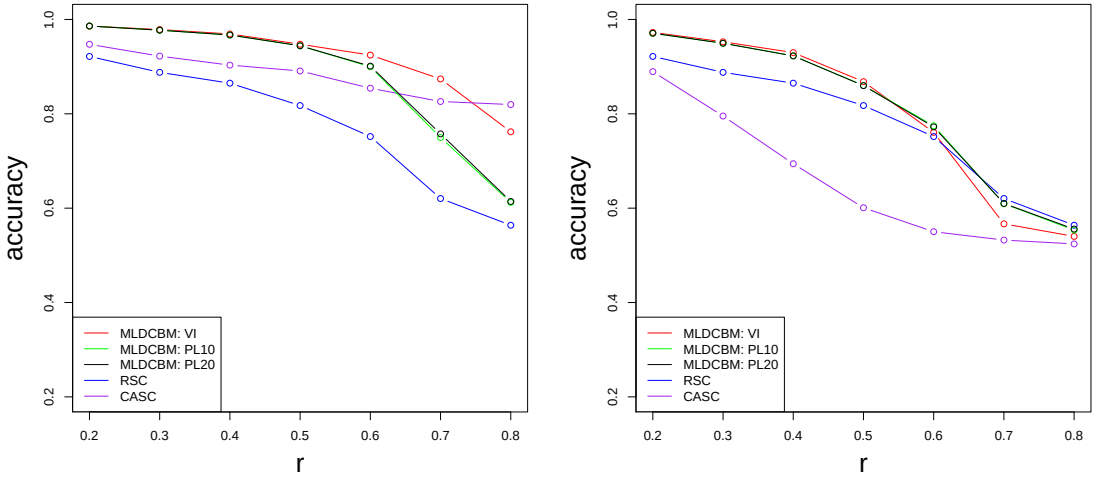
on the initialization, in the case when RSC cannot give better initial labels than pure guessing, the algorithms do not perform well and may not converge. Also, the performance of the pseudo likelihood algorithm for $K = 5$ is not as competitive as variational EM when r is big. It might be suggesting that the pseudo likelihood algorithm is more sensitive to the initialization comparing to variational EM in the case where signal is weak.

Figure 2.6 shows the case where the relation between covariates and community labels does not follow the logistic regression structure. For 2.6(a), the dependency structure between the community label and the covariates forms a mixture model where the covariates are generated from Gaussian distributions with mean 0 or 1 depending on which community the node is in. We can see that our method works better than CASC when the network is informative but the performance is not as competitive when the network becomes less informative. A possible reason is that when the network is uninformative, the initialization using RSC leads to a poor initial β estimate, which makes it difficult for the algorithm to utilize the covariates information



(a) p_{kk}=0.15, $\beta \sim U(-1, 1)$ (b) p_{kk}=0.15, $\beta \sim U(-5, 5)$

Figure 2.5: Mean accuracy of clustering vs r for $K = 5$



(a) p_{kk}=0.15, covariates from mixture (b) p_{kk}=0.15, covariates independent

Figure 2.6: Mean accuracy vs r for $K = 2$ with misspecified covariates

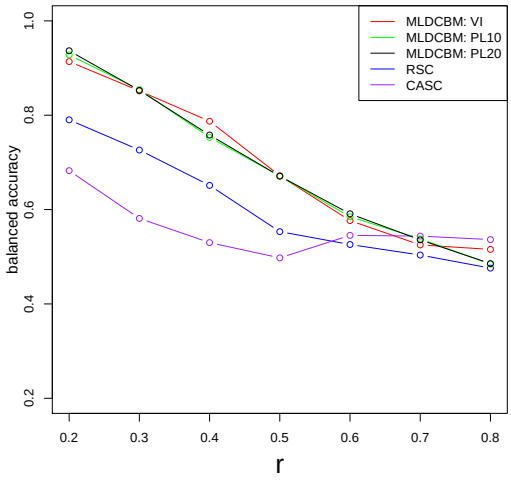


Figure 2.7: Mean balanced accuracy vs r for $K = 2$ with unbalanced communities, p_{kk}=0.15, $\beta = 1$

correctly. Similar to previous results, we observe that the accuracy of the pseudo likelihood algorithm drops more when r becomes larger. For 2.6(b), the covariates are generated independently from the community label, the result suggests that our methods still work well when the networks are informative although the covariates are irrelevant.

We further consider the case when the communities are unbalanced. We simulate data from the proposed model with β set to 1 and covariates from $N(0.5, 1)$. Under this setting, about 18.8% of nodes are in the minority community with a standard error of roughly 3%. Figure 2.7 shows the balanced accuracy, which is the average of sensitivity and specificity. The result shows that our methods have better performance when the network is informative and there is no obvious difference between the variational EM algorithm and the pseudo likelihood algorithm.

Another setting that we are interested in is when the network is not assortative, where nodes are more likely to form an edge if they are not from the same community. We

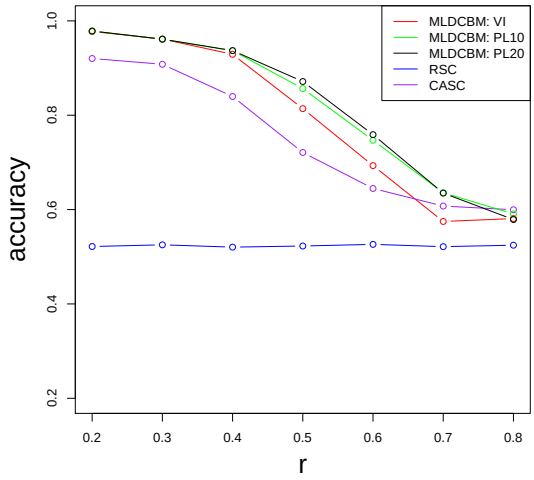


Figure 2.8: Mean accuracy vs r for $K = 2$ with non-assortative networks, initialized with CASC, $p_{k\ell}$=0.15

simulate data from the proposed model under the same setting as Figure 2.3(a) except that here we set $p_{k\ell} = 0.15$ and $p_{kk} = rp_{k\ell}$. It is known that spectral clustering does not perform well on non-assortative network while CASC still has a reasonable performance (*Binkiewicz et al.*, 2014). Since our algorithms rely on the initialization, we decide to use the CASC result as the initialization under the non-assortative setting. Figure 2.8 shows the clustering accuracy. The result suggests that our method can still improve the community detection result given by CASC even when the networks are non-assortative.

Lastly, Table 2.1 shows the computing time of the proposed algorithms under the setting $K = 5, n = 500, p = 10, p_{kk} = 0.15, r = 0.4$. It can be seen that pseudo likelihood algorithm is computationally much more efficient than the variational EM and CASC.

Algorithm	PL10	PL20	VI	CASC
Time (secs)	2.23	3.98	34.37	22.10

Table 2.1: Mean computation time of community detection algorithms under degree corrected blockmodels with covariates over 10 replications

2.6 Data example

We applied our method to a friendship network of 71 lawyers in a Northeastern US corporate law firm in New England(*Lazega*, 2001). The dataset also contains information on seven covariates including status, gender, office, year with the firm, age, practice, and law school. We notice there are only 4 people in the Providence office, which might not be informative as a source for dividing the network into clusters since the number of people in the office is too small comparing to 19 of the Hartford office and 48 for the Boston office. Thus we removed these 4 people and left with 67 people in the following analysis. The original network is directed with an edge from i to j if person i nominates person j as a friend. We converted it to an undirected network by letting person i and j be connected if either one of them nominates the other as a friend.

The covariates status, gender, office, practice, and law school are categorical and basically balanced between each category. The covariate year varies from 1 to 32 and has a median 7. The covariate age ranges from 26 to 67 with median 38. The degree distribution of the undirected network is shown in Figure 2.9.

We applied the variational EM algorithm and the pseudo likelihood algorithm with $K = 3$ and obtained same clustering result. The result is shown in Figure 2.10. Figure 2.10a shows the community membership given by the algorithm. In Figure 2.10b, the colors of the nodes represent office information and the node size is proportional

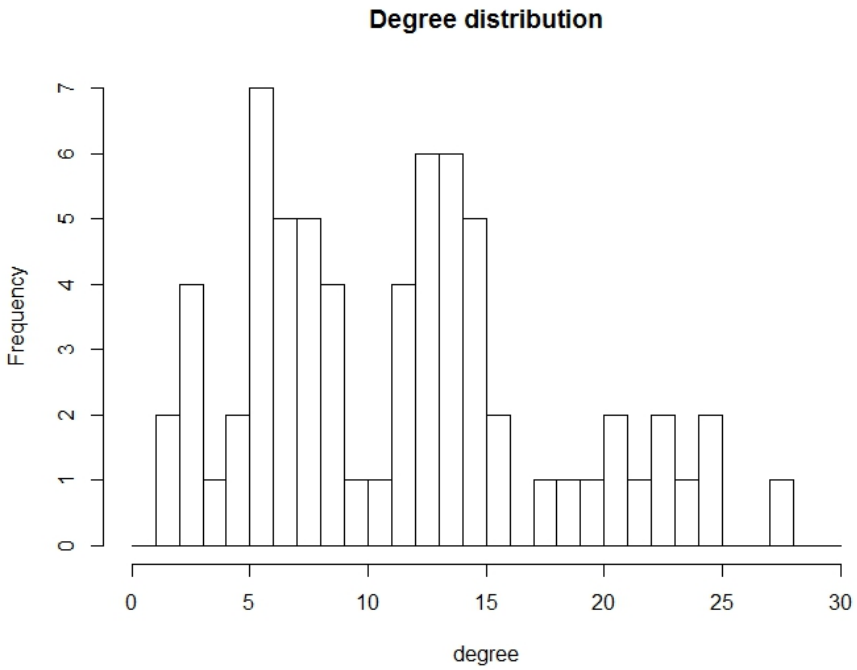
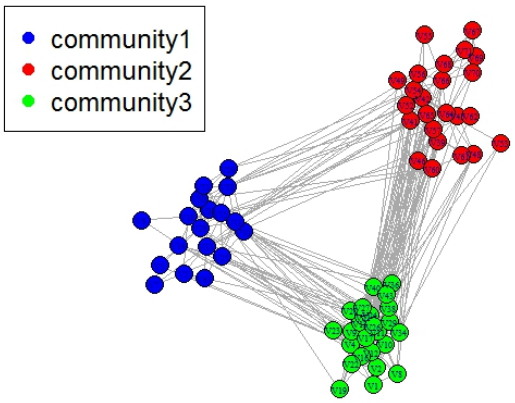
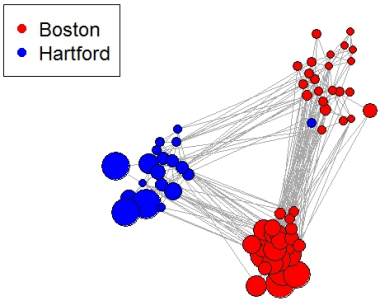


Figure 2.9: Degree distribution

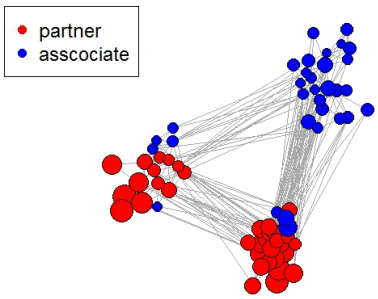
to years plus 10. It can be seen that community 1 contains people only from Hartford office, and people from Boston office are divided into 2 clusters by their years in the firm. Colors of nodes in Figure 2.10c show the status information and the node size is proportional to the age. It can be seen that nodes in community 2 are all associates and community 3 mainly consists of partners. Also, nodes in community 2 are younger comparing to community 3. We then calculated mean degree of each community. It turns out that the mean degree of community 3 is 15, which is much higher than 9.4 of community 1 and 10.3 of community 2. This is in some sense expected since the senior partners in a same office may have worked with each other for many years. To summarize, the nodes are separated into 3 communities, community of people from Hartford office, community of young associates in Boston office, and community of senior partners in Boston office.



(a) Community membership



(b) Covariate: office and years



(c) Covariate: status and age

Figure 2.10: Community detection result

2.7 Discussion

We have considered the community detection problem for networks with node covariates based on a principled statistical model combining the degree-corrected stochastic blockmodel and the logistic regression model. We have developed efficient estimation algorithms via the variational EM and the pseudo likelihood and illustrated their good performance in simulation studies and on a real-world data. We have also studied the asymptotic properties of the pseudo likelihood algorithm and obtained weak consistency result under mild assumptions.

Regarding future work, various directions might be considered. As mentioned, the model can be extended to the case where the relation between the community label C and node covariates X is specified by a mixture model $P(X|C)$. In the case where we have some prior knowledge that the covariates might be from mixtures of some distributions, the model may have better performance. Another perspective that naturally arises is the semi-supervised setting. If we already observe a part of the community labels of the nodes in the network, how do we estimate the unobserved community labels? The semi-supervised setting has many real-world applications especially in social networks where we may have a survey on a fraction of the targeting people on their community labels and need to inference for the others'. Going further from semi-supervised setting, we may consider the supervised setting where all the community labels are known and our target is to make predictions on new nodes. This can be viewed as a network information assisted classification problem which might be of interest when the response variables in classification are dependent of each other and the dependency structure is given by a network.

From the theoretical perspective, the consistency of variational inference under the degree corrected blockmodel remains unknown, although it is natural to conjecture

so as the maximum profile likelihood has shown to be consistent without covariates. Also, we observed from simulations that the covariates do help improve the clustering performance although the degree-corrected blockmodel itself without covariates provides label consistency in the asymptotic setting. It is thus interesting to study how helpful it is to incorporate covariates in non-asymptotic settings.

CHAPTER III

Missing Data Imputation with Network Information

3.1 Introduction

Missing data problem is widely encountered in real-world data analysis. One commonly used technique for handling missing data is imputation since most statistical procedures and algorithms rely on complete data. Thus it is essential to develop appropriate imputation procedures that produce imputations with high quality.

Many imputation methods have been proposed to handle univariate and multivariate missing data using only the information within the data set. However, in modern datasets, together with the traditional multivariate data, networks representing the relations between the entities are often also collected, where nodes in the network represent entities and edges between the nodes represent relations between the entities. With real-world networks, it is widely observed that there is always some kind of cohesion effect between connected entities, i.e. the connected entities have similar covariates.

In this work, we consider the imputation problem under the setting where in addition

to the multivariate data set, we also observe a network between the observations that provides information on the affinity. One example is the online social network, where we observe the friendship network between the users, but only partially observe covariates such as age, gender, income etc. for each user. The administrator of the platform may wish to impute the missing information of the users. To the best of our knowledge, this problem is not well addressed by existing imputation methods.

3.1.1 Missing data imputation methods

In the simple setting of imputing a single variable, many methods have been proposed and well studied. These methods can be conceptually divided into two groups, regression-based methods and hot-deck methods. Imputation through regression-based methods is straightforward. For example, one may perform a univariate regression, potentially with generalized linear models or nonparametric models, and also deal with post-processing such as a necessary truncation. Once the regression model is fitted, the predicted values for the missing entries may be used as imputation. See *Van Buuren* (2018) (Chapter 3) for a more detailed review. The hot-deck methods, similar to nearest-neighbor methods, typically define a distance metric between two observations using the observed covariates, and the imputation for a missing value will be borrowed from a completely observed observation that is close under the metric. *Andridge and Little* (2010) provides a review on some commonly used hot-deck methods.

Imputation for multivariate missing data can be roughly divided into two categories. A joint modeling approach would model the observed and missing variables through some joint distribution, e.g. the multivariate normal or t-distribution. See *Murray et al.* (2018) for a comprehensive review for such kind of methods. Although joint

modeling methods are easy to understand and typically have good theoretical properties, they are restrictive in many data analysis settings due to its lack of flexibility in handling complex mixed type of variables (*Van Buuren*, 2007).

Comparing to the joint modeling approach, the fully conditional specification is a more flexible framework for multivariate imputation (*Van Buuren*, 2007). Specifically, one specifies the conditional model for each variable conditioning on all other variables. For multivariate imputation, an iterative imputation procedure that starts with some simple imputation and then conducts univariate imputations using the specified conditional models sequentially is often used (*Buuren and Groothuis-Oudshoorn*, 2010). The iterative imputation procedure could be viewed as a Gibbs sampler from a Bayesian perspective: in each iteration, the sampler draws from the conditional distribution on the missing entries. The specified conditional models are also flexible. In principle, one may specify any existing univariate regression model according to their need. People have studied the performances of using Predictive-Mean Matching (*Buuren and Groothuis-Oudshoorn*, 2010), Classification and Regression Tree (*Burgette and Reiter*, 2010), Support Vector Machines (*Wang et al.*, 2006), and the Random Forest (*Stekhoven and Bühlmann*, 2011), ect.

Despite the flexibility of the fully conditional specification framework, there is limited result on the convergence property of the framework. *Liu et al.* (2013) compared the iterative imputation that uses a set of Bayesian regression models g as the conditional distributions to a proper MCMC algorithm under a joint model f. They showed that under the assumption that both Markov chains have unique stationary distributions, the iterative imputation has the same stationary distribution as the joint model provided that the conditional models g are compatible with f. *Zhu and Raghunathan* (2015) showed the convergence of the iterative imputation algorithm without the as-

sumption of requiring unique stationary distributions, but under the setting that each observation can have only one missing entry.

3.1.2 Network models

A network can be represented using an adjacency matrix A, where A_{uv} could be binary indicating whether there is an edge between nodes i and j, or weighted indicating the strength of the connection between nodes u and v. Many network models have been proposed to model a network alone, without relating to covariates, including stochastic block models (*Holland et al.*, 1983), and exponential random graph (*Robins et al.*, 2007) etc. See *Goldenberg et al.* (2010) for a detailed review of such models. The latent space model (*Hoff*, 2005; *Hoff et al.*, 2002) provides a natural way of relating the edges in a network to covariates. In one form of the latent space model, the probability of having an edge between nodes u and v depends on their latent positions Z_u, Z_v, their individual connectivity parameters b_u, b_v, the edge related covariate x_{uv}, and model parameters. Specifically, conditional on the above mentioned quantities, the probability between nodes u and v is given by

$$\text{logit}[P(A_{uv} = 1)] = \alpha x_{uv} + b_u + b_v + Z_u^T Z_v.$$

The latent space models are flexible in the sense that they can be easily modified to cover a wide range of commonly observed network properties such as degree heterogeneity, transitivity etc. Recently *Ma and Ma* (2017) developed gradient based algorithms that can fit the latent space model efficiently instead of using computationally expensive MCMC algorithms.

3.1.3 Imputation with networks

In the setting where a network is observed and each node has a categorical label with some node labels being unobserved, heuristics such as label propagation (*Zhur and Ghahramanirh*, 2002) has been proposed to infer the unobserved labels. Starting with some initialization, label propagation iteratively assigns each node with unobserved label the label that dominates in its neighbors and iterates until convergence. This could be viewed as imputing a single categorical variable with network information. *Chakrabarti et al.* (2017) proposed a model that can simultaneously infer multiple missing labels on a network by encouraging each edge in the network to be explained by at least one common label. However, such methods can only be applied to categorical variables, and do not take advantage of the correlation among the variables.

The main contribution of this work is that we propose an imputation method that can flexibly impute mixed type missing data while taking the network information into consideration. The idea of the method relies on combing the full conditional specification framework and the network model.

The rest of the chapter is organized as follows: Section 3.2 introduces the proposed model and method for missing data imputation with network information available; Section 3.3 provides theoretical results of the framework under a similar setting to *Liu et al.* (2013); Sections 3.4 and 3.5 illustrate the performance of the proposed framework using simulated studies and a real-world data example respectively; Section 3.6 discusses limitations and extensions of the framework.

3.2 Model and method

Our proposed imputation method builds on the full conditional specification framework that imputes one variable at a time, and we consider modeling the network using a latent space model conditional on the covariates.

Suppose we observe an incomplete data matrix X with n observations and p variables. Together with X, we observe a network characterized by its $n \times n$ binary adjacency matrix A that represents connectivity between the observations.

Let X_j denotes the j-th variable and X_{-j} denotes the other variables. Let X_j be the variable we are imputing in the current iteration, $M_j \subset \{1, 2, ..., n\}$ denotes the set of indices i for which X_{ij} is missing, and $O_j = \{1, 2, ..., n\} \backslash M_j$ be the index set for the observed. Our target is to impute all missing entries $\{X_{ij}, i \in M_j, j = 1, 2, ..., p\}$ using information from A and $\{X_{ij}, i \in O_j, j = 1, 2, ..., p\}$.

We assume the network is generated from a latent space model conditional on X:

$$a_{uv} := logit(P(A_{uv} = 1|X)) = \sum_{j=1}^{p} \alpha_j d(X_{uj}, X_{vj}) + b_u + b_v + Z_u^T Z_v$$

with parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)^T$, $\mathbf{b} = (b_1, ..., b_n)^T$ and latent positions $\mathbf{Z} = (Z_1, Z_2, ...Z_n)^T$.

The overall algorithm contains the following steps:

(i) Initialize the imputation using some imputation methods. Fit the latent space model using the imputed X to initialize the model parameters and the latent positions.

(ii) For each variable $j = 1, 2, ..., p$, update the missing entries $\{X_{ij}, i \in M_j\}$ with

X_{-j} fixed.

(iii) Update the model parameters and latent positions.

(iv) Iterate between steps (ii) and (iii) until convergence.

For the remaining of the chapter, we set $d(X_{uj}, X_{vj}) = (X_{uj} - X_{vj})^2$ for continuous variables and $d(X_{uj}, X_{vj}) = 1_{(X_{uj} \neq X_{vj})}$ for discrete variables. Other choices of the distance measure are also possible and will be discussed.

Without loss of generality, we consider the problem of imputing X_1 with other variables, the latent space model parameters, and the latent positions fixed. The general framework of imputation proceeds as follows. Suppose we are given the conditional distribution of the missing entries $P(\{X_{u1}, u \in M_1\}|X_{-1}, \{X_{v1}, v \in O_1\})$, possibly specified by a regression model, we may consider the following criterion:

$$\begin{aligned} &\max P(A, \{X_{u1,u\in M_1}\}|X_{-1}, \{X_{v1}, v \in O_1\}) \\ &\qquad = \max P(A|X_1, X_{-1}) P(\{X_{u1,u\in M_1}\}|X_{-1}, \{X_{v1}, v \in O_1\}) \end{aligned} \tag{3.1}$$

One interpretation of this criterion is that we have a prior on the missing entries $\{X_{u1,u\in M_1}\}$ given by $P(\{X_{u1,u\in M_1}\}|X_{-1}, \{X_{v1}, v \in O_1\})$, with likelihood $P(A|X_1, X_{-1})$, we are looking for the posterior mode for the missing entries $\{X_{u1}, u \in M_1\}$.

3.2.1 Imputation of continuous variables

Suppose X_1 is a continuous variable, with all the other variables X_{-1} fixed, the conditional log likelihood of the network is

$$\begin{aligned}
&\log P(A|X_1, X_{-1}) = \sum_{u,v} \alpha_1 (x_{u1} - x_{v1})^2 A_{uv} \\
&- \sum_{u,v} \log \left\{ 1 + \exp[\alpha_1 (x_{u1} - x_{v1})^2 + \sum_{j=2}^{p} \alpha_j d(x_{uj}, x_{vj}) + b_u + b_v + Z_u^T Z_v] \right\} \\
&+ \sum_{u,v} \sum_{j=2}^{p} \left\{ \alpha_j d(x_{uj}, x_{vj}) + b_u + b_v + Z_u^T Z_v \right\} A_{uv}.
\end{aligned}$$

Note that the terms in the last line do not depend on $\{x_{u1}, u \in M_1\}$.

Suppose the specified conditional distribution $P(\{x_{u1}, u \in M_1\}|X_{-1}, \{x_{v1}, v \in O_1\})$ has the form such that for $u \in M_1$, $x_{u1}|x_{\{u,-1\}} \sim N(\hat{x}_{u1}, \sigma^2)$, then the general criterion (3.1) has the following form

$$\begin{aligned}
&\max_{x_{u1}, u \in M_1} \left\{ \sum_{u,v} \alpha_1 (x_{u1} - x_{v1})^2 A_{uv} \right. \\
&- \sum_{u,v} \log \left[1 + \exp[\alpha_1 (x_{u1} - x_{v1})^2 + \sum_{j=2}^{p} \alpha_j \mathrm{d}(x_{uj}, x_{vj}) + b_u + b_v + z_u^T z_v] \right] \qquad (3.2) \\
&\left. - \frac{1}{2\sigma^2} \sum_{u \in M_1} (x_{u1} - \hat{x}_{u1})^2 \right\}.
\end{aligned}$$

When σ is unknown, we may replace the term $\frac{1}{2\sigma^2}$ with an estimate or a specified tuning parameter λ, which plays the role of balancing between the network and the

specified conditional distribution, and obtain the following criterion

$$\max_{x_{u1},u\in M_1}\Bigg\{\sum_{(u,v)}\alpha_1(x_{u1}-x_{v1})^2A_{uv} - \sum_{u,v}\log\left[1+\exp[\alpha_1(x_{u1}-x_{v1})^2+\sum_{j=2}^{p}\alpha_j d(x_{uj},x_{vj})+b_u+b_v+Z_u^TZ_v]\right] -\lambda\sum_{u\in M}(x_{u1}-\hat{x}_{u1})^2\Bigg\}$$

The solution to this optimization problem is then an imputation of $\{x_{u1}, u \in M_1\}$. We may use gradient based methods to compute the solution with gradient in the following form.

$$\begin{aligned}\frac{\partial}{\partial x_{u1}} &= \sum_v 2\alpha_1(x_{u1}-x_{v1})A_{uv} - \sum_v 2\alpha_1(x_{u1}-x_{v1})\sigma(a_{uv}) - 2\lambda(x_{u1}-\hat{x}_{u1}) \\ &= 2\alpha_1\langle A_u-\sigma(a_u), x_{u1}\mathbf{1}_n - x_{\cdot 1}\rangle - 2\lambda(x_{u1}-\hat{x}_{u1}),\end{aligned}$$

where A_u is the u-th row of A and $a=(a_{uv})_{u,v\le n}$ with a_u being the u-th row, $\sigma(\cdot)$ is the sigmoid function, and $\langle X,Y\rangle = Tr(X^TY)$.

3.2.2 Imputation of discrete variables

When X_1 is discrete with K categories, with all the other variables X_{-1} fixed, the conditional log-likelihood of the network becomes

$$\begin{aligned}
&\log P(A|X_1, X_{-1}) = \sum_{u,v} \alpha_1 1_{(x_{u1} \neq x_{v1})} A_{uv} \\
&- \sum_{u,v} \log \left\{ 1 + \exp[\alpha_1 1_{(x_{u1} \neq x_{v1})} + \sum_{j=2}^{p} \alpha_j d(X_{uj}, X_{vj}) + b_u + b_v + Z_u^T Z_v]) \right\} \\
&+ \sum_{u,v} \left\{ \sum_{j=2}^{p} \alpha_j d(X_{uj}, X_{vj}) + b_u + b_v + Z_u^T Z_v \right\} A_{uv}.
\end{aligned}$$

Thus maximizing $P(A|X_1, X_{-1})P(\{x_{u1}, u \in M_1\}|X_{-1}, \{x_{v1}, v \in O_1\})$ with respect to the missing entries $\{x_{u1}, u \in M_1\}$ would be an intractable combinatorial problem. To handle this problem, we relax the hard assignment of $\{x_{u1}, u \in M\}$ to probability assignment $q_u = (q_u^1, q_u^2, ..., q_u^K)$ with $q_u^T \mathbf{1} = 1$ and compute the conditional distribution $P(\{x_{u1,u\in M_1}\}|A, X_{-1}, \{X_{v1}, v \in O_1\})$. However, this conditional distribution is also intractable due to the complex dependence structure.

We propose to find a variational approximation $Q(\{x_{u1,u\in M_1}\})$ to this conditional distribution by restricting the distribution of $\{x_{u1,u\in M_1}\}$ to be in a fully factorized form, i.e. $Q(x_{u1} = k_u, x_{v1} = k_v, \ldots, x_{w1} = k_w) = q_u^{k_u} q_v^{k_v} ... q_w^{k_w}$. For a specific node u, fixing all the q_v for other nodes $v \neq u$, the assignment of q_u satisfies

$$\begin{aligned}
q_u^k &\propto P(A, x_{u1} = k|X_{-1}, \{x_{v1}, v \in O_1\}, q_v) \\
&= P(A|x_{u1} = k, X_{-1}, \{x_{v1}, v \in O_1\}, q_v) P(x_{u1} = k|X_{-1}, \{x_{v1}, v \in O_1\}, q_v).
\end{aligned}$$

We compute the q vector for every node $u, u \in M_1$ for variable 1, we then assign a hard imputation for $x_{u1}, u \in M_1$ to simplify the future computation before we proceed to the next variable.

3.2.3 Updating network model parameters

To update the parameters α, $\mathbf{b}$ and the latent positions $\mathbf{Z}$ for the latent space network model, we take the projected gradient approach described in *Ma and Ma* (2017). Ideally, the parameters should be updated after imputing each variable, namely p times in a full iteration over all the variables. But for the consideration of computational cost, we only do the update after iterations over all the p variables.

The parameters can be updated in the following way:

$$Z^{t+1} = Z^t + 2\eta_Z(A - \sigma(\mathbf{a}^t))Z^t$$

$$\alpha^{t+1} = \alpha^t + \eta_\alpha\langle A - \sigma(\mathbf{a}^t), \mathbf{d}(X_j)\rangle$$

$$b^{t+1} = b^t + 2\eta_b(A - \sigma(\mathbf{a}^t))1_n$$

with centering on Z after each iteration, where $\langle X, Y\rangle = Tr(X^TY)$ and $\mathbf{d}(X_j)$ is a matrix with the (u, v)-th entry being $d(X_{uj}, X_{vj})$. The step sizes are suggested in *Ma and Ma* (2017) to be $\eta_Z = \eta/||Z^0||^2_{op}$, $\eta_b = \eta/2n$ and $\eta_\alpha = \eta/2||X||^2_F$. For simplicity, we use the hard imputation as mentioned for the discrete missing variables in fitting the latent space model.

3.2.4 Initialization and choice of λ

Choosing the initialization of the algorithm is important. In principle, one may use any existing imputation method to provide an initialization depending on the practitioner's need. If it is known that the correlation between the covariates are strong, one may use existing imputation methods such as the misforest to initialize. If the correlation between the covariates are unclear, one may use simple imputation pro-

cedures that are less informative, such as the mean imputation.

The λ could be tuned through a cross-validation like procedure. Given a data set, we randomly sample a small set of entries to be "missing" and impute these entries together with the true missing entries, and compute the error of imputation for the missing entries we selected. We then select the λ value that gives the lowest error on the these entries.

3.3 Theoretical properties

3.3.1 Relation to Gibbs sampling

As mentioned earlier, the iterative imputation framework is essentially a Gibbs sampler type algorithm. Under the proposed model, one may view the objective function $P(\{X_{u1,u\in M_1}\}|X_{-1},\{X_{v1},v\in O_1\})P(A|X_1,X_{-1})$ as a posterior distribution for the missing entries. The proposed gradient based algorithm is looking for the mode of this posterior distribution.

Ideally, in the iterative imputation framework, one may want to sample the missing entries $\{X_{ij},i\in M_j\}$ from this posterior distribution. However, with the network information involved, the missing entries are not independent and thus this posterior distribution cannot be factorized. Directly drawing samples from this posterior would be hard to implement when the number of missing entries is large. One potential solution is to induce a second layer of Gibbs sampler that draws one missing entry at a time. But this will still be computationally inefficient and require a balancing between the inner layer of Gibbs sampling that iterates between missing entries within a variable, and the outer layer of Gibbs sampling that iterates between different variables.

Despite the above mentioned Gibbs sampling algorithm is infeasible in practice, it provides an idealized algorithm under the framework for us to consider theoretical aspects of the framework.

3.3.2 Convergence to a Bayesian model

We follows the results in *Liu et al.* (2013) and show that the iterative imputation Markov chain converges to the same stationary distribution as a Bayesian model under the assumption that the Markov chain admits a unique stationary distribution.

We first define the Bayesian model and its corresponding Gibbs sampler. Denote x_j^{mis} and x_j^{obs} the observed and missing subsets of variable j and let $x^{mis} = \{x_j^{mis}, j = 1, 2, ..., p\}$, $x^{obs} = \{x_j^{obs}, j = 1, 2, ..., p\}$. Further we use x_{-j} to denote the variables excluding the j-th variable. Let $\Theta = (\theta_R, \theta_A)$ denotes all the model parameters, with $f(x; \theta_R)$ corresponds to the joint distribution for x and θ_A corresponds to the parameters and latent positions for the latent space network model. We assume the missing mechanism is missing completely at random throughout. Under the proposed model, the likelihood could be decomposed into two parts

$$\begin{aligned} p(x^{mis}|x^{obs}, A, \Theta) &\propto p(x^{mis}, A|x^{obs}, \Theta) \\ &= f(x^{mis}|x^{obs}, \theta_R)p(A|x^{mis}, x^{obs}, \theta_A). \end{aligned} \tag{3.3}$$

With a prior $\pi(\Theta)$, the posterior predictive distribution is

$$p(x^{mis}|x^{obs}, A) = \int_{\Theta} p(x^{mis}|x^{obs}, A, \Theta)p(\Theta|x^{obs}, A)d\Theta, \tag{3.4}$$

where $p(\Theta|x^{obs}, A) \propto \pi(\Theta)p(x^{obs}, A|\Theta)$. A standard way to draw samples from the posterior predictive distribution is to use Gibbs sampler that iteratively draws Θ and X^{mis}. Under standard regularity conditions, the Markov chain is ergodic and has limiting distribution $p(x^{mis}, \Theta|x^{obs}, A)$ (*Geman and Geman*, 1984).

We modify the Gibbs sampling procedure in order to compare to the iterative imputation framework. Let $x^{(k-1)}$ be the entire dataset with both observed and imputed values and $\Theta^{(k-1)}$ be the parameter estimates at iteration $k-1$. At iteration k, the Gibbs chain evolves as follows:

- Set $x \leftarrow x^{(k-1)}$ and update the variables of x one at a time.
- For $j = 1, ..., p$, draw $\theta_R \sim p(\theta_R|x_j^{obs}, x_{-j})$ and $x_j^{mis} \sim p(x_j^{mis}|x_j^{obs}, x_{-j}, A, \theta_R, \theta_A^{(k-1)})$
- Draw $\theta_A \sim p(\theta_A|x, A)$
- Set $x^{(k)} \leftarrow x$ and $\Theta^k \leftarrow (\theta_R, \theta_A)$

Under regularity conditions (*Rosenthal*, 1995), the Markov chain converges to the posterior distribution of the corresponding model.

For iterative imputation, the user specifies p conditional regression models, denoted as $g_j(x_j|x_{-j}, \theta_j)$, with θ_j being the corresponding parameters with prior $\pi_j(\theta_j), i = 1, ..., p$. The iterative imputation scheme can be described as follows:

- Set $x \leftarrow x^{(k-1)}$ and update the variables of x one at a time.
- For $j = 1, ..., p$, draw $\theta_j \sim p_j(\theta_j|x_j^{obs}, x_{-j})$, which is the posterior distribution of θ_j with g_j and π_j, and $x_j^{mis} \sim p_j(x_j^{mis}|x_j^{obs}, x_{-j}, A, \theta_j, \theta_A^{(k-1)})$
- Draw $\theta_A \sim p(\theta_A|x, A)$
- Set $x^{(k)} \leftarrow x$ and $\Theta^k \leftarrow (\theta_R, \theta_A)$

Notice that under the proposed framework, similar to (3.3), we have

$$p_j(x_j^{mis}|x_j^{obs}, x_{-j}, A, \theta_j, \theta_A^{(k-1)}) \propto g_j(x_j^{mis}|x_j^{obs}, x_{-j}, \theta_j)p(A|x_j^{mis}, x_j^{obs}, x_{-j}, \theta_A^{(k-1)})$$

We consider when the specified conditional regression models g_j's are compatible with f. A set of condition models $g_j(x_j|x_{-j}, \theta_j), \theta_j \in \Theta_j$ is said to be compatible with $f(x|\theta), \theta \in \Theta$ if for all j, there exist a collection of surjective maps $t_j : \Theta \to \Theta_j$ such that there exists $\theta \in \Theta$ with $g_j(x_j|x_{-j}, \theta_j) = f(x_j|x_{-j}, t_j(\theta_R))$.

Let K_1 and K_2 denote the transition kernel of the Gibbs chain and the iterative imputation chain respectively, $\nu_1^{X_{obs}}$ and $\nu_2^{X_{obs}}$ be their corresponding stationary distributions. *Liu et al.* (2013) showed that K_1 and K_2 are close to each other on a large set $A_n = \{x : |\hat{\Theta}| < \gamma\}$ where $\hat{\Theta}$ is the complete data maximum likelihood estimator of the parameters and γ is a positive constant. Let $\tilde{K}_1$ and $\tilde{K}_2$ be the transition kernel corresponding to K_1 and K_2 conditioning on A_n in the sense that $\tilde{K}_j(\omega, B) = \frac{K_j(\omega, B \cap A_n)}{K_j(\omega, A_n)}$, which means we restrict the update of the missing entries to be inside A_n. Then with mild assumptions similar to those in *Liu et al.* (2013), which are provided in the supplementary material, we have the following result.

Theorem III.1. *Suppose the specified conditional models g_j's are compatible with a joint model f, the iterative imputation chain and the Gibbs chain are positive Harris recurrent and have unique stationary distribution ν_j such that $\nu_j(A_n) \to 1$ with sufficiently large γ in probability as $n \to \infty$. Further, $\tilde{K}_j$ are geometrically recurrent. Then $d_{TV}(\nu_1^{X_{obs}}, \nu_2^{X_{obs}}) \to 0$ in probability as $n \to \infty$.*

The result suggests that under the assumptions, if the true underlying joint distribution of the covariates X is in the family of joint models that are compatible with the specified conditional models, then as the sample size grows, the iterative imputation Markov Chain will have stationary distribution converging to the stationary distribu-

tion of the Gibbs chain of the Bayesian model where the underlying joint distribution is used. This means we will have an imputation that is "consistent" as if we know the underlying joint distribution.

3.4 Simulation studies

In this section we investigate the performance of the proposed method in simulation studies under several different settings and compare with the widely used iterative imputation method implemented in the MICE package (*Buuren and Groothuis-Oudshoorn*, 2010). It should be mentioned that the main purpose of the simulation study is to illustrate that through considering the network information, we may produce imputation with higher quality.

The simulation for imputing continuous variables proceeds as follows. We generate n=100, 200, and 400 and $p = 5$ dimensional continuous X from a multivariate normal distribution with the covariance matrix in the form of $(1-r)\mathbf{I} + r\mathbf{1}\mathbf{1}^T$ where r is set to 0.1, 0.3, and 0.5 to reflect different levels of correlation between the covariates. The latent positions are generated i.i.d from a 3-dimensional normal distribution with standard deviation 0.4. We fixed these generated covariates and latent positions throughout the experiments. The latent space parameter α is set to -(1,1,1,1,1) so that each variable contributes equally and if two nodes have similar covariates, they are more likely to form an edge. The parameters b are generated from a uniform distribution with the mean set to control the overall edge probability to be close to 0.3, 0.55, and 0.1. Once the parameters are generated, the network is generated according to the latent space model. We generate 20% missingness in the data matrix X completely at random. We consider two possible initializations, one using the mean imputation, the other using result from the MICE algorithm. For both MICE and the proposed method, the user specified conditional distributions are set to be linear

regression, as it is compatible to the true generating conditional distribution. For the proposed method, we also consider a variant that only uses the network information by setting $\lambda = 0$. After imputing the missing data using different methods, we compute the mean squared error for each variable on the imputed entries. Each setting is repeated 40 times and we calculate the mean over the replications. We provide results on the first variable since the variables are equivalent.

Figure 3.1 shows the simulation results for imputing continuous variables. Overall, the proposed method outperforms its MICE counterpart which does not utilize the network information. Panels (a)(c)(e) compare the cases when the correlation between the covariates is mild but the edge density of the network varies. It suggests that as the edge density increases, the improvement by including network information increases. Panels (b)(c)(d) compare the cases when the edge density is fixed but the correlation between the covariates varies. It shows that as the correlation increases, the improvement by including network information decreases.

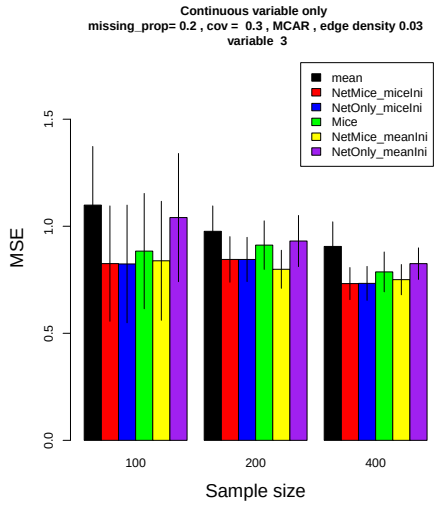
For discrete variables, we generate n=100, 200, and 400 and $p = 4$ dimensional binary variables Y following the procedure in *Cario and Nelson* (1997). Specifically, $Y_{ip}|W_{ip}, \epsilon_{ip}$ follows a logistic model with specified β that controls the marginal distribution of Y to be roughly balanced. W_{ip} are generated iid normal and ϵ_{ip} are iid normal with mean 0 and covariance $(1-r)\mathbf{I} + r\mathbf{1}\mathbf{1}^T$ where r is set to 0.3, 0.5, and 0.7 to form different levels of correlation between the generated covariates. We set α to be 0.1, 0.2, and 0.3 to reflect weak, medium, and strong relations between the generated network and the covariates. We set b to be fixed to control the overall edge density of the network and generate the latent positions similar to the continuous case. We generate 20% missingness in the data matrix Y completely at random. We initialize using mode or MICE. For both MICE and the proposed methods, the user

specified conditional distributions are set to be logistic regression. We also consider a variant that only uses the network information. After imputing the missing entries, our methods naturally provide a probability prediction for the missing entries. For MICE, we refit a logistic regression to each variable using other variables as predictors to obtain a probability prediction on the missing entries. We then compute the area under the receiver operating characteristic curve (AUC) for each variable on the missing entries. We provide results on the first variable.

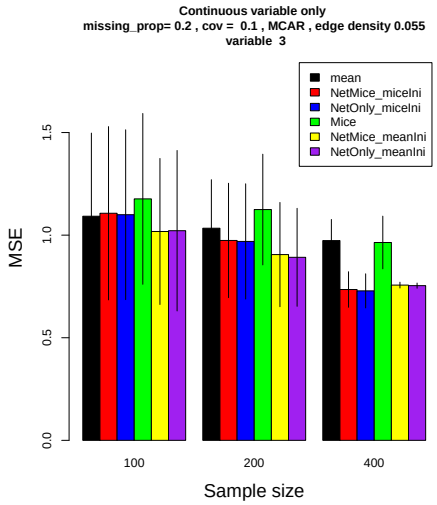
Figure 3.2 shows the results for imputing discrete variables. The proposed method performs better when the relation between the network and the covariates is strong. Panels (a)(c)(e) compare the cases when the correlation between the covariates is mild but the relation between the network and the covariates varies. It can be seen that the improvement of utilizing network information becomes larger as the relation is stronger. Panels (b)(c)(d) compare the cases when the strength of network information is fixed, but the correlation between the variables varies. When the correlation between the covariates becomes larger, the improvement of using network information becomes smaller.

Lastly, we test the case when the network is irrelevant. For this set of simulation, we use the same setting as in the simulation for continuous variables, except that the network is now generated from an Erdos-Renyi random graph model with probability 0.3, 0.5, and 0.7, which is independent of the covariates.

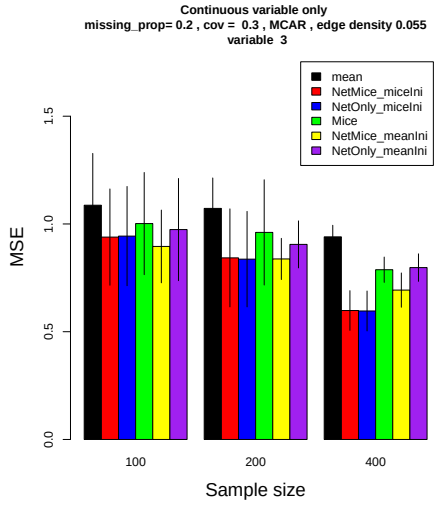
Figure 3.3 shows the results when the network is irrelevant. It can be seen that except for the method that only uses the network information and initializes with the mean imputation, the proposed methods do not suffer much in comparison to MICE even though the network is irrelevant.



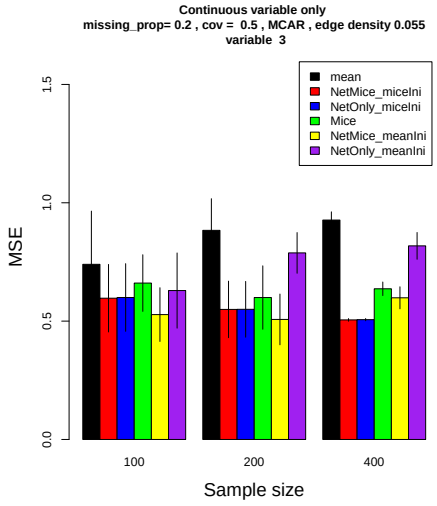
(a)



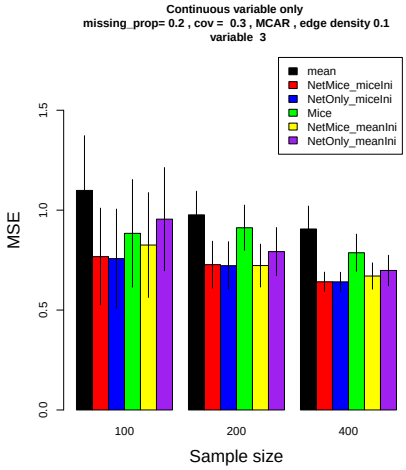
(b)



(c)

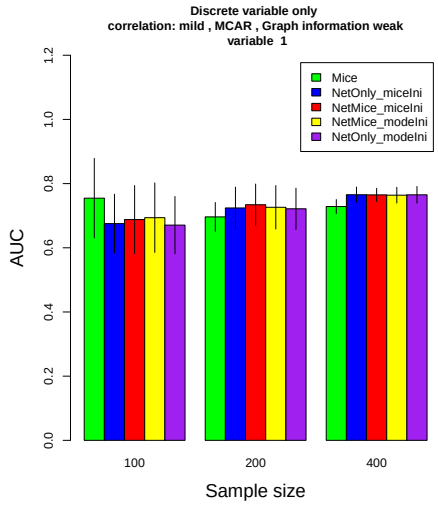


(d)

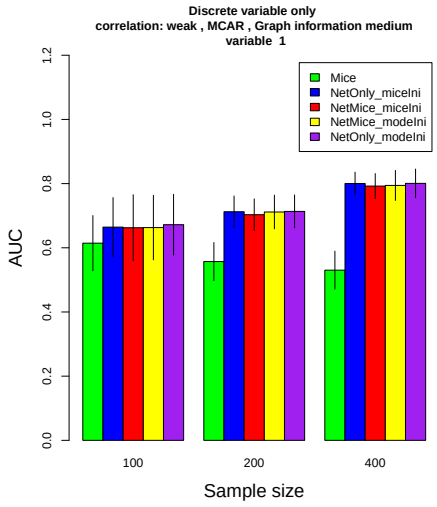


(e)

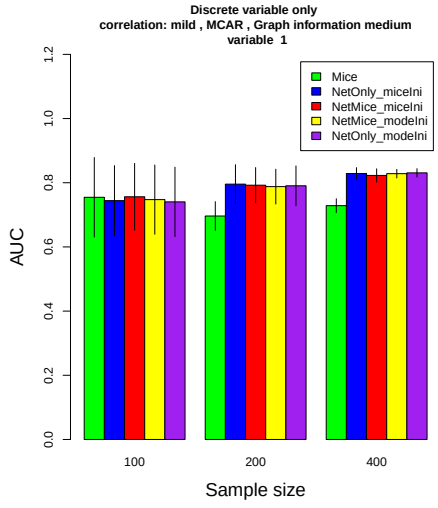
Figure 3.1: Imputation results for continuous variable



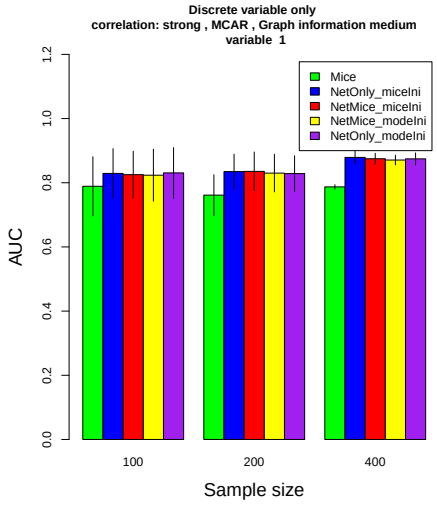
(a)



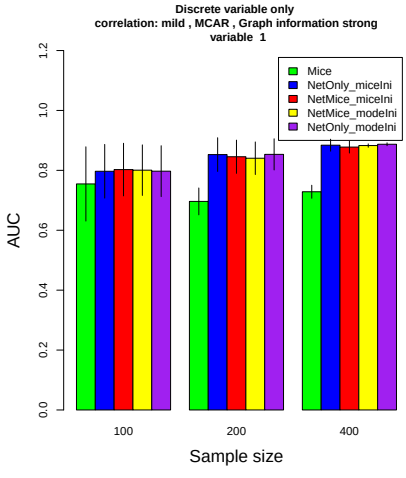
(b)



(c)

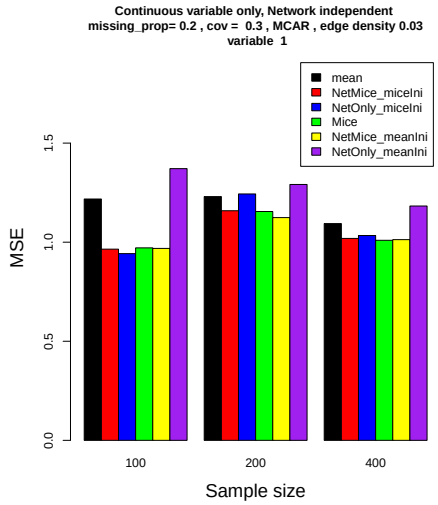


(d)

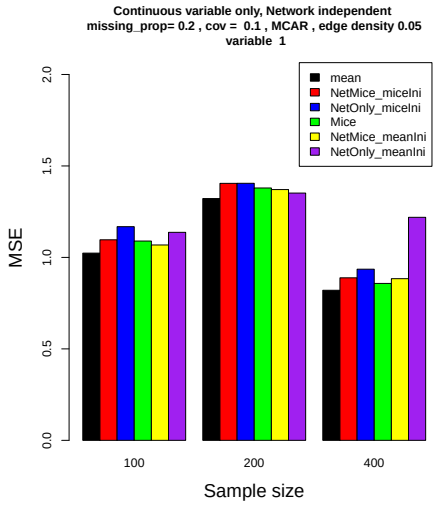


(e)

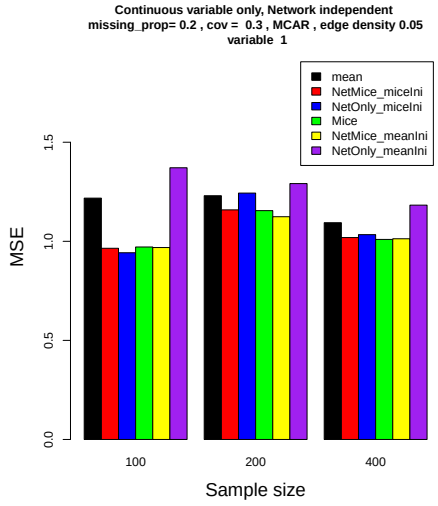
Figure 3.2: Imputation results for binary variables



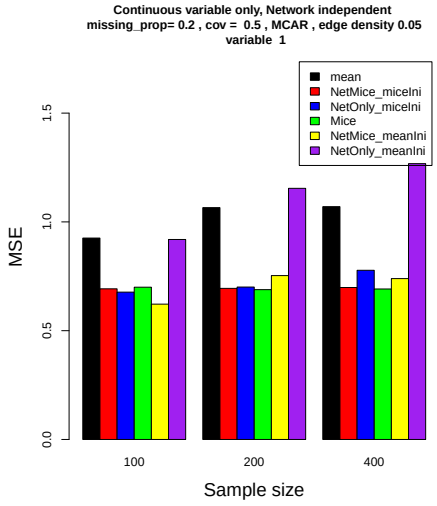
(a)



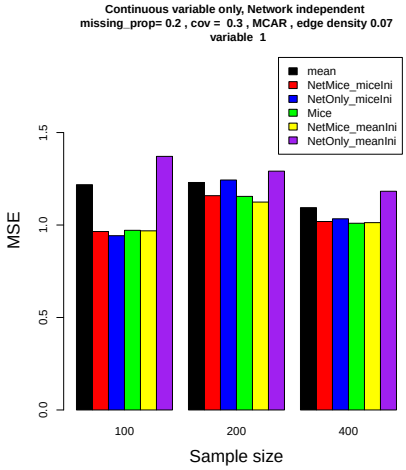
(b)



(c)



(d)



(e)

Figure 3.3: Imputation results for continuous variables when network is irrelavent

3.5 Data example

In this section, we use a data example to illustrate the potential benefit of using network information for missing value imputation. The data set consists of the social relationship network among members of the Provisional Irish Republican Army (PIRA) from 1970 to 1998 (*Gill et al.*, 2014). The whole study is divided into 6 periods, which are grouped into 5 data sets by periods 1, 2, 3, 4&5, and 6. We used all the data except for period 6 as it contains much fewer nodes comparing to previous periods. The networks are undirected with nodes representing different people and edges representing relationships between the nodes. The existence of an edge represents that the two people have at least one of the following relations: (i) involvement in a PIRA activity together, (ii) friends before joining PIRA, (iii) blood relatives, and (iv) married. The networks are sparse with average node degree between 1 and 2. The covariates we used include gender, age, marital status, attending university or not, brigade memberships, and violent characteristics. Age is continuous and contains missing entries, marital status and brigade memberships are categorical, and the other variables are binary. Several other binary covariates related to role and specific activities are not used as they are highly unbalanced and may be mutually exclusive in nature.

A basic summary of the PIRA network is shown in Table 3.1. To evaluate the performances of our imputation algorithms, we randomly generated 20% missing entries on the brigade memberships and violent characteristics, and imputed the data set using the proposed method with MICE initialization. We used 3 as the dimension of the latent space. We computed the area under the receiver operative curve for the binary and categorical variables. For categorical variables, we used the multiclass AuROC defined in *Hand and Till* (2001). We compared with the MICE imputation. We repeated the process 20 times. The average result suggests that the proposed

Period	1	2	3	4&5
Number of Nodes	334	260	526	367
Avg Degree	1.2	1.3	1.95	1.4

Table 3.1: Summary of the PIRA network

Period	1	2	3	4&5
Brigade_NetMice	0.639(0.071)	0.742(0.078)	0.649(0.065)	0.683(0.053)
Violent_NetMice	0.615(0.060)	0.639(0.070)	0.544(0.050)	0.547(0.046)
Brigade_NetOnly	0.682(0.048)	0.726(0.086)	0.680(0.043)	0.712(0.068)
Violent_NetOnly	0.556(0.060)	0.608(0.075)	0.530(0.038)	0.541(0.047)
Brigade_Mice	0.520(0.048)	0.540(0.069)	0.591(0.054)	0.522(0.055)
Violent_Mice	0.609(0.092)	0.598(0.088)	0.641(0.044)	0.549(0.042)

Table 3.2: AuROC of the imputation on the PIRA data set

method produces significantly better imputation in brigade, which is expected as the variable is strongly related to the edges in the networks by the definition of the edges. The imputation accuracy of the violent characteristic stays at a similar level as MICE imputation for periods 1, 2, 4&5 but is poorer in period 3. We computed the number of matches in violent characteristics over the connected pairs, only 63.4% of the connected pairs had same violent characteristics in period 3, which did not show strong homophily nor heterogeneity. According to *Gill et al.* (2014), period 3 had higher proportion of high-degree stars, suggesting a small group of leaders connected and coordinated the brigades, and had a high proportion of membership based cliques. These may suggest that the violent characteristics are not strongly associated with the edges, which resulted in the relative poor performances in imputing violent characteristics in period 3.

3.6 Discussion

We used the squared distance to illustrate the method, but other choices of the distance measure are also possible. An example is the Mahalanobis distance. In that case, we have

$$a_{uv} = \alpha(x_u - x_v)^T \Sigma^{-1}(x_u - x_v) + b_u + b_v + Z_u^T Z_v$$

$$\frac{\partial}{\partial x_u} = \sum_{v:(u,v)\in E} 2\alpha(x_u - x_v)^T \Sigma^{-1} - \sum_v 2\alpha\sigma(a_{uv})(x_u - x_v)^T \Sigma^{-1} - 2\lambda(x_u - \hat{x}_u)^T$$

Then updating the first variable $x_{u1}, u \in M_1$ will utilize the first element of the obtained gradient vector.

One limitation of the proposed framework is that the computational complexity will depend on the choice of the network model as we need to estimate the model parameters during the updates. For the latent space model specifically, the computational cost may grow as $O(n^2)$ where n is the number of nodes.

To summarize, we have proposed an iterative method for missing data imputation with network information available. The method combines the flexible full conditional specification in multivariate missing data imputation and the latent space network model. The method has been illustrated in numerical experiments using both simulation and real-world networks.

CHAPTER IV

A Partially Edge Exchangeable Model with Node Covariates

4.1 Introduction

Interaction networks are very common in modern data related to collaboration (*Barabási and Albert*, 1999) and social relations (*Opsahl and Panzarasa*, 2009; *Leskovec and Mcauley*, 2012). Many of the popular models are not well suited for modeling such interaction networks. One set of models focus on analyzing the networks based on the assumption that the nodes in the model are exchangeable, this includes the stochastic block model (*Holland et al.*, 1983), graphon models (*Airoldi et al.*, 2013; *Wolfe and Olhede*, 2013), etc. These models produce dense networks as the number of nodes grows to infinity (*Lloyd et al.*, 2012; *Cai et al.*, 2016), which contradicts the fact that most real-world networks are sparse. Other popular models such as exponential random graph models do not have a clear sampling mechanism for interpreting the formation of the network (*Crane and Dempsey*, 2018).

For the above mentioned consideration in modeling interaction networks, *Cai et al.* (2016) and *Crane and Dempsey* (2018) developed the edge exchangeable framework where the edges are the statistical units for modeling. On one hand, edge exchangeable

models admit an interpretation in terms of edge sampling. Under edge exchangeable models, the observed network is a result of an edge sampling process from a population. This corresponds to the data collection procedure of many interaction networks. For example, a paper citation network can be constructed by random sampling the published papers. On the other hand, edge exchangeable models can produce sparse networks as the network grows.

Crane and Dempsey (2018) proposed a simple parametric family of edge exchangeable models called the Hollywood model. The Hollywood model is a generative procedure that generates edges sequentially as follows: suppose $n-1$ edges $Y_1, ..., Y_{n-1}$ have already been generated, then

- first generate the number of nodes k_n in the n-th edge from a distribution ν,
- then given k_n, select k_n nodes $Y_{n,1}, ..., Y_{n,k_n}$ sequentially according to

$$P(Y_{n,j} = i | Y_1, ..., Y_{n-1}, Y_{n,1}, ...Y_{n,j-1}) \propto \begin{cases} D_{n,j}(i) - \alpha, i = 1, 2, ..., V_n(j) \\ \theta + \alpha V_n(j), i = V_n(j) + 1 \end{cases} \tag{4.1}$$

where $D_{n,j}(i)$ is the degree of node i, and $V_n(j)$ the number of existing nodes, computed at the time after the $(j-1)$-th node of the n-th edge was chosen, and with parameters $\alpha \in (0, 1)$ and $\theta > -\alpha$ corresponding to a growing network. The Hollywood model exhibits some desired limiting properties as a network model including sparsity and power law degree distribution with index $(1+\alpha)$ defined as follows (*Crane and Dempsey*, 2018):

Definition IV.1. (Sparse networks) A sequence of network $(\mathcal{E}_m)_{m \geq 1}$ is sparse as $m \to \infty$ if

$$\limsup_{m \to \infty} \frac{e(\mathcal{E}_m)}{v(\mathcal{E}_m)^{m_*(\mathcal{E}_m)}} = 0$$

where $e(\mathcal{E}_n)$ is the number of edges, $v(\mathcal{E}_n)$ is the number of nodes, and $m_*(\mathcal{E}_n)$is the average number of nodes in each edge in $\mathcal{E}_n$.

Definition IV.2. (Power law) Let the degree distribution of a network $\mathcal{E}$ be defined as $d(\mathcal{E}) = (N_k(\mathcal{E})/v(\mathcal{E}))_{k\geq 1}$, where $N_k(\mathcal{E})$ is the number of nodes with degree k and $v(\mathcal{E})$ is the total number of nodes in $\mathcal{E}$. A sequence of networks $(\mathcal{E}_m)_{m\geq 1}$ has power law degree distribution with index γ, if for some slowly varying function $\ell(x)$, that is, $\lim_{x\to\infty} \ell(tx)/\ell(x) = 1$ for all $t > 0$,

$$\lim_{k\to\infty} \lim_{m\to\infty} \frac{d_k(\mathcal{E}_m)}{\ell(k)k^{-\gamma}} = 1$$

for some slowly varying function $\ell(k)$

A specific example of sparse network sequence as in Definition IV.1 is when $m_*(\mathcal{E}_m) = 2$, which is the case when each edge contains 2 nodes. Then the network sequence is sparse if the number of edges in the network grows slower than the quadratic of number of nodes. In more general cases, the denominator $v(\mathcal{E}_m)^{m_*(\mathcal{E}_m)}$ is the rate at which the total number of all possible edges grows as the number of nodes increases, and the network sequence is sparse if the number of edges grows at a slower rate.

Despite the nice properties that the Hollywood model has, the process does not consider additional information that may exist together with the network. Specifically, node covariates are often collected together with the network in real-world data. Although incorporating covariates into network analysis is hard in general, some successful attempts existed, see for example *Binkiewicz et al.* (2017), *Mariadassou et al.* (2010), *Zhang et al.* (2016), and *Sweet* (2015); *Hoff et al.* (2002). However, to the best of our knowledge, there is no existing literature that attempts to consider node covariates information under edge exchangeable models. In practice, it is natural to

believe that a node's covariates may be related with the interactions that the node makes. For example, in a social network, a person may be more likely to chat with a person who is humorous. Such kind of relation is not well captured in the Hollywood model, where whether a node is likely to make new interactions or not depends only on its degree.

A main difficulty of considering the covariates under edge exchangeable models is that the edge exchangeability can be broken due to the involvement of covariates. One may ask the question: to what extent can we preserve the edge exchangeability structure with covariates taking into consideration and what is the interpretation in terms of sampling. In this paper, we attempt to address the question by developing a model that incorporates covariates and partially preserves the edge exchangeability. We describe the model and its properties in Section 4.2, establish an estimation algorithm in Section 4.3. Simulation studies and a data example are provided to demonstrate the proposed model in Sections 4.4 and 4.5 respectively.

4.2 Model setup

We introduce a new model that incorporates node covariates based on the canonical Hollywood model. Let ν be a distribution on positive integers with probabilities $\nu_1, \nu_2, ...$ We generate edges $Y_1, Y_2, ...$ as follows. Suppose $n-1$ edges have already been generated and we are generating the n-th edge. First we draw the number of nodes k_n in the next edge from ν, independently. Then we draw the k_n nodes denoted $Y_{n,1}, ..., Y_{n,k_n}$ one at a time according to the following probability:

$$P(Y_{n,j} = i | Y_1, ..., Y_{n-1}, Y_{n,1}, ...Y_{n,j-1}) \propto \begin{cases} D_{n,j}(i) - \alpha + e^{\beta X_i}, i = 1, 2, ..., V_n(j) \\ \theta + \alpha V_n(j) + \frac{\sum_{k=1}^{V_n(j)} e^{\beta X_k}}{V_n(j)}, i = V_n(j) + 1 \end{cases} \tag{4.2}$$

where $D_{n,j}(i)$ is the degree of node i, and $V_n(j)$ is the number of existing node, computed at the time after the $(j-1)$-th node of the n-th edge was chosen. $\alpha \in (0,1)$ and $\theta > -\alpha$ are parameters. One notable difference from the canonical Hollywood model is that now each node i has an "attractive" score related to its covariates given by the term $e^{\beta X_i}$, and the node i is more likely to be chosen when edges are generated if it has a higher score. The term $\frac{\sum_{k=1}^{V_n(j)} e^{\beta X_k}}{V_n(j)}$ is an average score of all the nodes in the network, and the probability that new nodes would join the network will depend on this average score. It is also noticeable that the effect of the covariates diminishes as the degree of a node grows and the total number of nodes in the network grows. The interpretation of the parameters β largely relies on the sign. A positive β suggests that the corresponding covariate is positively correlated with whether a person is likely to make a new connection.

A real-world example that may exhibit the effect described above is a social network where a node represents a person and edges represent interactions between people. When a person first joins the social network, whether he or she is likely to make new interactions highly depends on his or her own characteristics, e.g. whether he or she plays sports. If a person has already made many interactions, suggesting he or she is highly active and well-known in the network, the effect of his or her characteristics may not have much impact on whether he or she will make new interactions in comparison to someone who is new to the network. Similarly, when the social network is at its initial stage, how likely a person would like to join the network may depend on who are already in the network and how attractive they are, while when there are already many people in the network, a person may no longer consider who are in the network.

Because the model depends on the covariates of the nodes that are currently involved in the network, the edge exchangeability is broken whenever a new node joins

the network. However, it can be shown that the edges generated between the time after one new node joins, to the time when next new node joins, are exchangeable. Formally, suppose at each time point when a new node is introduced into the network, we observe a snapshot. Denote the snapshots $\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_m$ at the time when a new node joins such that $\mathcal{E}_s$ has $s-1$ nodes for $s < m$, with $\mathcal{E}_1$ being an empty network. We do not model the distribution of the covariates X. Then we have the following conditional probability for $s < m$:

$$P(\mathcal{E}_s|\mathcal{E}_{s-1}, X_1, ..., X_{s-1}) = \prod_{k\geq 1} \nu_k^{M_k(\mathcal{E}_s \backslash \mathcal{E}_{s-1})} \times \frac{\prod_{i=1}^{s-1} \prod_{j=D_{\mathcal{E}_{s-1}}(i)}^{D_{\mathcal{E}_s}(i)-1} (e^{\beta X_i} - \alpha + j)}{\prod_{i=m(\mathcal{E}_{s-1})}^{m(\mathcal{E}_s)-1} (\sum_{k=1}^{s-1} e^{\beta X_k} \frac{s}{s-1} + \theta + i)} \times \frac{\theta + (s-1)\alpha + \frac{\sum_{k=1}^{s-1} e^{\beta X_k}}{s-1}}{\sum_{i=1}^{s-1} e^{\beta X_i} \frac{s}{s-1} + m(\mathcal{E}_s) + \theta} \tag{4.3}$$

where $D_{\mathcal{E}_s}(i)$ counts the degree of node i in snapshot $\mathcal{E}_s$, $m(\mathcal{E}_s)$ counts the total degree of the snapshot $\mathcal{E}_s$, and $M_k(\mathcal{E}_s \backslash \mathcal{E}_{s-1})$ counts the number of k-node edges in snapshot $\mathcal{E}_s$ that is not in $\mathcal{E}_{s-1}$. It could be seen that this conditional probability depends on the network only through $D_{\mathcal{E}_s}(i)$, $D_{\mathcal{E}_{s-1}}(i)$, $m(\mathcal{E}_{s-1})$, $m(\mathcal{E}_s)$, and $M_k(\mathcal{E}_s \backslash \mathcal{E}_{s-1})$. These are quantities that do not depend on the labeling of the edges between the two snapshots. Thus we have the following:

Proposition IV.3. *The edges generated between any two consecutive snapshots $\mathcal{E}_{s-1}$ and $\mathcal{E}_s$ are exchangeable.*

An interpretation of this exchangeability from a sampling perspective is as follows: the population of edges changes slightly when a new node joins the network, and the edges we observe between two consecutive new nodes are representative of the population of edges during the period.

It should be mentioned that although theoretically there is no restriction to the parameters β and covariates X as long as they are bounded, in practice we may need

to standardize X and restrict the range of β so that the term $\exp(\beta X)$ does not dominate the probability. Also, a special case of the model is when $\beta = -\infty$, where the Hollywood model is recovered.

The proposed model shows sparsity as the network grows, which is a desired property for network models as real-world networks are often sparse. This properties is similar to what the Hollywood model exhibits. Specifically, we can show the following result:

Proposition IV.4. *Let $\mathcal{E}_1, ..., \mathcal{E}_n, ...$ be a sequence of networks generated following the proposed model as in (4.2) labeled by the number of nodes, with $\nu = 1$, and n is number of nodes in the network. Assuming the covariate X_i and the parameter β are bounded, we have that the expected total degree $E(m(\mathcal{E}_n))$ grows at a rate at least $n^{\frac{1}{\alpha}}$, but satisfies $E(m(\mathcal{E}_n)) = O(n^{\frac{1}{\alpha}})$ as $n \to \infty$.*

This result suggests that the proposed model generates networks that are at least as dense as the Hollywood model with parameter α, but has the same sparsity level as the Hollywood model with parameter α in the limit. Then following the definition of sparse networks in definition IV.1, we have

Corollary IV.5. *The model described in 4.2 generates sparse networks if $E(\nu)\alpha > 1$.*

4.3 Estimation

Based on the conditional likelihood (4.3), the joint likelihood is

$$P(\mathcal{E}_1, \mathcal{E}_2, ...\mathcal{E}_s) = P(\mathcal{E}_1) \prod_{i=2}^{s} P(\mathcal{E}_i | \mathcal{E}_{i-1})$$

with $P(\mathcal{E}_1) = 1$. Thus, the log-likelihood is given by

$$\ell(\nu, \alpha, \beta, \theta; X, \mathcal{E}_1, ...\mathcal{E}_s) = \sum_{k\geq 1} M_k(\mathcal{E}_s) \log \nu_k$$
$$+ \sum_{i=1}^{s-1} \sum_{j=1}^{D_{\mathcal{E}_s}(i)-1} \log(e^{\beta X_i} - \alpha + j) - \sum_{m=2}^{s} \sum_{j=m(\mathcal{E}_{m-1})+1}^{m(\mathcal{E}_m)-1} \log(\sum_{k=1}^{m-1} e^{\beta X_k} \frac{m}{m-1} + \theta + j)$$
$$+ \sum_{m=2}^{s-1} \log\left[\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k} \frac{1}{m-1})\right] - \sum_{m=2}^{s-1} \log\left[\sum_{k=1}^{m-1} e^{\beta X_k} \frac{m}{m-1} + \theta + m(\mathcal{E}_s)\right]$$

The maximum likelihood estimator of ν_k is straightforward, given by computing the frequency of k-node edges in the network. To estimate parameters α, β, and θ, we use a block-coordinate descent approach. The gradients for β, α, and θ are as follows:

$$\frac{\partial \ell}{\beta} = \sum_{i=1}^{s-1} \sum_{j=1}^{D_{\mathcal{E}_s}(i)-1} \frac{X_i e^{\beta X_i}}{e^{\beta X_i} - \alpha + j} - \sum_{m=2}^{s} \sum_{j=m(\mathcal{E}_{m-1})+1}^{m(\mathcal{E}_m)-1} \frac{\sum_{k=1}^{m-1} X_k e^{\beta X_k \frac{m}{m-1}}}{\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + j}$$
$$+ \sum_{m=2}^{s-1} \frac{\sum_{k=1}^{m-1} X_k e^{\beta X_k \frac{1}{m-1}}}{\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}}} - \sum_{m=2}^{s-1} \frac{\sum_{k=1}^{m-1} X_k e^{\beta X_k \frac{m}{m-1}}}{\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + m(\mathcal{E}_s)}$$

$$\frac{\partial \ell}{\alpha} = \sum_{i=1}^{s-1} \sum_{j=1}^{D_{\mathcal{E}_s}(i)-1} \frac{-1}{e^{\beta X_i} - \alpha + j} + \sum_{m=2}^{s-1} \frac{m-1}{\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}}}$$

$$\frac{\partial \ell}{\theta} = -\sum_{m=2}^{s} \sum_{j=m(\mathcal{E}_{m-1})+1}^{m(\mathcal{E}_m)-1} \frac{1}{\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + j}$$
$$+ \sum_{m=2}^{s-1} \frac{1}{\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}}} - \sum_{m=2}^{s-1} \frac{1}{\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + m(\mathcal{E}_s)}$$

Further, the Hessian can be derived and the elements are listed as follows. In practice, the Hessian can be used to construct to estimate the standard error of the maximum likelihood estimates by taking the diagonal of the inverse negative Hessian, which is

the observed Fisher information matrix.

$$\frac{\partial^2 \ell}{\partial \beta^2} = \sum_{i=1}^{s-1} \sum_{j=1}^{D_{\mathcal{E}_s}(i)-1} \frac{X_i X_i^T e^{\beta X_i}(-\alpha + j)}{(e^{\beta X_i} - \alpha + j)^2} - \sum_{m=2}^{s} \sum_{j=m(\mathcal{E}_{m-1})+1}^{m(\mathcal{E}_m)-1} \frac{\sum_{k=1}^{m-1} X_k X_k^T e^{\beta X_k \frac{m}{m-1}}(\theta + j)}{(\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + j)^2}$$
$$+ \sum_{m=2}^{s-1} \frac{\sum_{k=1}^{m-1} X_k X_k^T e^{\beta X_k \frac{1}{m-1}}(\theta + (m-1)\alpha)}{(\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}})^2} - \sum_{m=2}^{s-1} \frac{\sum_{k=1}^{m-1} X_k X_k^T e^{\beta X_k \frac{m}{m-1}}(\theta + m(\mathcal{E}_s))}{(\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + m(\mathcal{E}_s))^2}$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \alpha} = \sum_{i=1}^{s-1} \sum_{j=1}^{D_{\mathcal{E}_s}(i)-1} \frac{X_i e^{\beta X_i}}{(e^{\beta X_i} - \alpha + j)^2} - \sum_{m=2}^{s-1} \frac{\sum_{k=1}^{m-1} X_k e^{\beta X_k}}{(\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}})^2}$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \theta} = \sum_{m=2}^{s} \sum_{j=m(\mathcal{E}_{m-1})+1}^{m(\mathcal{E}_m)-1} \frac{\sum_{k=1}^{m-1} X_k e^{\beta X_k \frac{m}{m-1}}}{(\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + j)^2}$$
$$- \sum_{m=2}^{s-1} \frac{\sum_{k=1}^{m-1} X_k e^{\beta X_k \frac{1}{m-1}}}{(\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}})^2} + \sum_{m=2}^{s-1} \frac{\sum_{k=1}^{m-1} X_k e^{\beta X_k \frac{m}{m-1}}}{(\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + m(\mathcal{E}_s))^2}$$

$$\frac{\partial^2 \ell}{\partial \alpha^2} = \sum_{i=1}^{s-1} \sum_{j=1}^{D_{\mathcal{E}_s}(i)-1} \frac{-1}{(e^{\beta X_i} - \alpha + j)^2} - \sum_{m=2}^{s-1} \frac{(m-1)^2}{(\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}})^2}$$

$$\frac{\partial^2 \ell}{\partial \alpha \partial \theta} = - \sum_{m=2}^{s-1} \frac{(m-1)}{(\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}})^2}$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = \sum_{m=2}^{s} \sum_{j=m(\mathcal{E}_{m-1})+1}^{m(\mathcal{E}_m)-1} \frac{1}{(\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + j)^2}$$
$$- \sum_{m=2}^{s-1} \frac{1}{(\theta + (m-1)\alpha + \sum_{k=1}^{m-1} e^{\beta X_k \frac{1}{m-1}})^2} + \sum_{m=2}^{s-1} \frac{1}{(\sum_{k=1}^{m-1} e^{\beta X_k \frac{m}{m-1}} + \theta + m(\mathcal{E}_s))^2}$$

However, since the likelihood is non-convex, the resulting Hessian may not be invertible, or the observed Fisher information matrix may have negative diagonal elements.

In that case, we may perform a parametric bootstrap procedure to obtain an estimate of the variance in the maximum likelihood estimates.

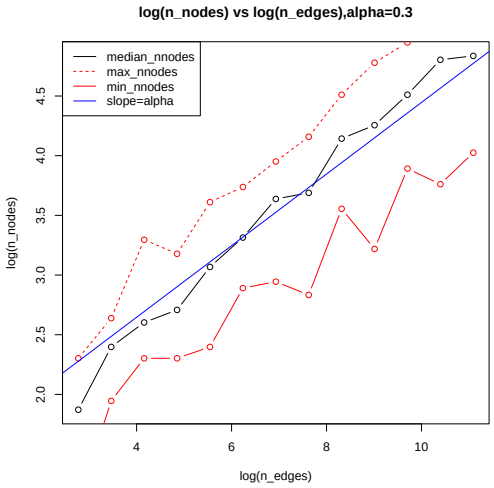
For parametric bootstrap, we generate B^* new networks based on the model with the estimated parameters, and when a new node is introduced into the network, we randomly sample its covariates by sampling one node's covariates from the observed nodes. We then estimate the parameters for all the B^* generated networks, and construct bootstrap confidence intervals for parameters by the quantiles of the parameter estimates in the B^* replications.

4.4 Simulation studies

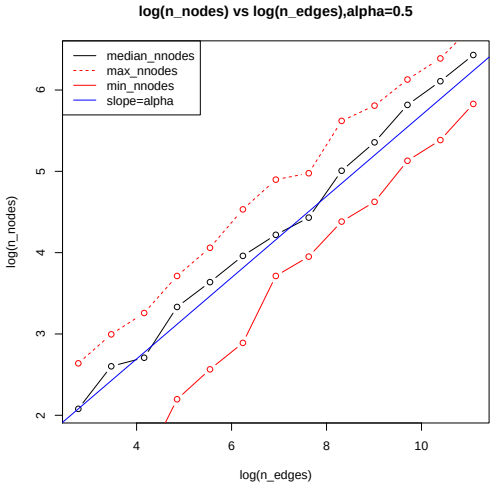
The simulation studies are divided into two parts. The first part illustrates the theoretical properties of the proposed model as the network grows, and the second part shows the performance of parameter estimation.

Figure 4.1 shows the relationship between the number of nodes and the total degree of the network on log-log scale for networks generated from the proposed model. For this set of simulations, we set θ=1, α= 0.3, 0.5, and 0.7. We generated 3-dimensional covariates, each from $U(0,1)$ distribution independently, and β was generated from the standard normal distribution. We generated 40 different networks under each setting with 2^16 edges, with $\nu \equiv 2$. We make the plot using the maximum, minimum, and median number of nodes of the generated networks when the number of edges in the networks are $2^k, k = 3, 4, ..., 16$. We also plot a line with slope equal to α.

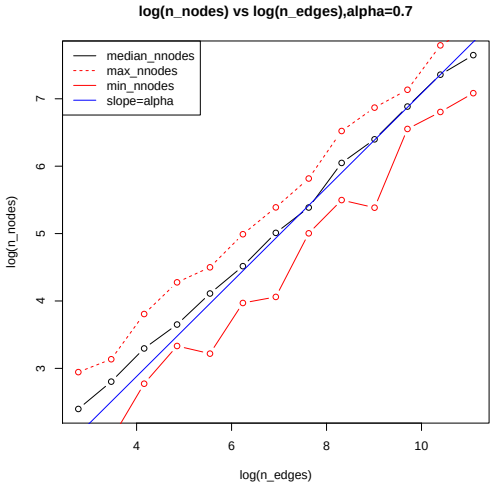
The result suggests that as the network grows, the log number of nodes and the log number of edges follows a linear relationship with the slope close to α, which confirms our result in Proposition IV.4.



(a) $\alpha = 0.3$



(b) $\alpha = 0.5$

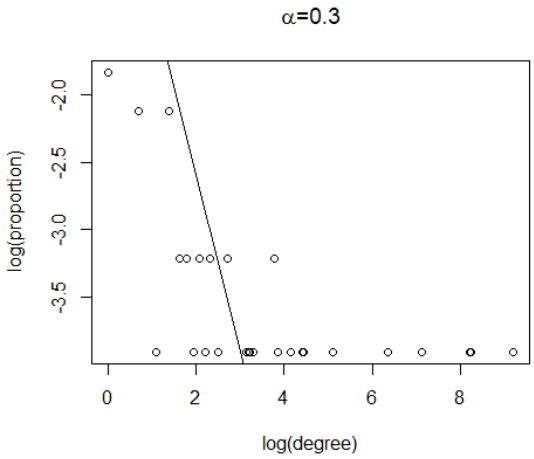


(c) $\alpha = 0.7$

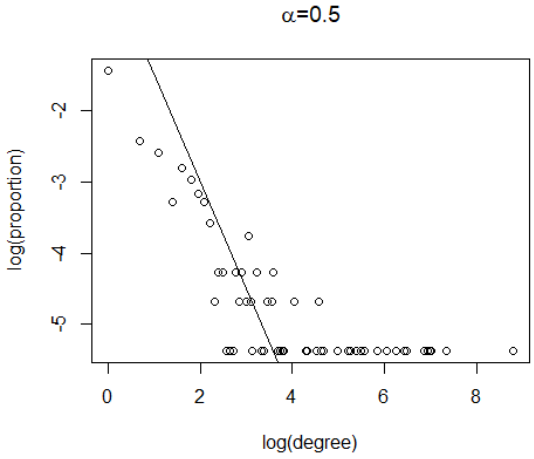
Figure 4.1: $\log(n_{\text{nodes}})$ vs $\log(n_{\text{edges}})$

Figure 4.2 plots the log proportion of nodes with degree k on $\log(k)$, with a solid line with slope $-(1+\alpha)$. The result suggests that the network generated from the proposed model may obey a power law with index $\alpha+1$ according to the definition IV.2.

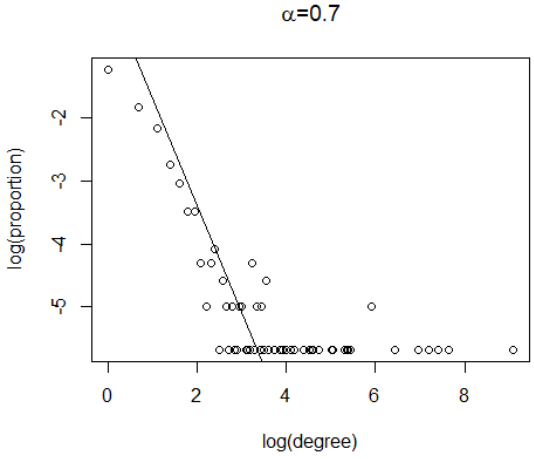
Next we show results of simulation studies for parameter estimation. We set α=0.3, 0.5, and 0.7, θ=0 and 10, and we also set $\nu \equiv 2$, $\beta = (1, 1, -1)$. Again, 3-dimensional covariates were generated from $U(0, 1)$ independently. The number of edges were set to be 2500, 5000, and 10000. For each of the settings, we generated 40 networks and estimated the parameters. We computed mean of the estimates, mean of the standard errors obtained from the diagonal of the inverse negative Hessian at the estimates, and the proportion that the interval constructed with the estimate $\pm$ 2 standard error covered the true parameter. We reported the proportion of the times when the standard error estimate was invalid due to that the Hessian was not invertible or produced negative diagonal elements in the observed Fisher information matrix. Table 4.1 shows the estimation results for α, β_1, β_3, and θ when $\alpha = 0.5$ and $\theta = 10$. The result for β_2 is not shown as it is similar to β_1. The parameter estimates for this parameter setting were accurate, and only a few cases had invalid standard error estimates from the observed Fisher information matrix. As the number of edges increases, the standard error estimate decreases. Table 4.2 show the results when $\alpha = 0.3$ and $\theta = 10$. In comparison to the results with true $\alpha = 0.5$, there were more invalid standard error estimates for β_1 and β_3, suggesting that we may need the bootstrap approach for estimation of the standard error of the estimates when α is small. Results of the other parameter settings are given in the appendix.



(a) $\alpha = 0.3$



(b) $\alpha = 0.5$



(c) $\alpha = 0.7$

Figure 4.2: $\log(proportion)$ vs $\log(degree)$

	mean(n_{nodes})	mean($\hat{\alpha}$)	mean($\hat{\sigma}_{\hat{\alpha}}$)	coverage	invalid
2500	221.0	0.489	0.066	0.925	0.025
5000	314.5	0.494	0.050	0.975	0.025
10000	445.5	0.494	0.038	0.975	0.025
	mean(n_{nodes})	mean($\hat{\beta}_1$)	mean($\hat{\sigma}_{\hat{\beta}_1}$)	coverage	invalid
2500	221.0	1.004	0.244	0.925	0.025
5000	314.5	1.003	0.189	0.975	0.000
10000	445.5	1.024	0.147	0.925	0.025
	mean(n_{nodes})	mean($\hat{\beta}_3$)	mean($\hat{\sigma}_{\hat{\beta}_3}$)	coverage	invalid
2500	221.0	-1.011	0.335	0.925	0.075
5000	314.5	-0.972	0.254	0.950	0.025
10000	445.5	-0.988	0.210	0.975	0.025
	mean(n_{nodes})	mean($\hat{\theta}$)	mean($\hat{\sigma}_{\hat{\theta}}$)	coverage	invalid
2500	221.0	11.034	4.942	1.000	0.000
5000	314.5	10.784	4.455	1.000	0.000
10000	445.5	10.790	4.130	1.000	0.000

Table 4.1: Estimate when true $\alpha = 0.5$, $\theta = 10$

	mean(n_{nodes})	mean($\hat{\alpha}$)	mean($\hat{\alpha}$)	coverage	invalid
2500	127.7	0.283	0.088	0.875	0.025
5000	164.2	0.284	0.062	0.900	0.000
10000	210.5	0.292	0.044	0.900	0.025
	mean(n_{nodes})	mean($\hat{\beta}_1$)	mean($\hat{\sigma}_{\hat{\beta}_1}$)	coverage	invalid
2500	127.7	0.966	0.385	0.725	0.225
5000	164.2	0.964	0.308	0.850	0.150
10000	210.5	0.975	0.208	0.725	0.250
	mean(n_{nodes})	mean($\hat{\beta}_3$)	mean($\hat{\sigma}_{\hat{\beta}_3}$)	coverage	invalid
2500	127.7	-0.949	0.644	0.650	0.275
5000	164.2	-0.971	0.415	0.750	0.200
10000	210.5	-0.973	0.292	0.650	0.300
	mean(n_{nodes})	mean($\hat{\theta}$)	mean($\hat{\sigma}_{\hat{\theta}}$)	coverage	invalid
2500	127.7	11.097	4.735	0.925	0.00000
5000	164.2	11.131	4.204	0.925	0.00000
10000	210.5	10.740	3.601	0.925	0.00000

Table 4.2: Estimate when true $\alpha = 0.3$, $\theta = 10$

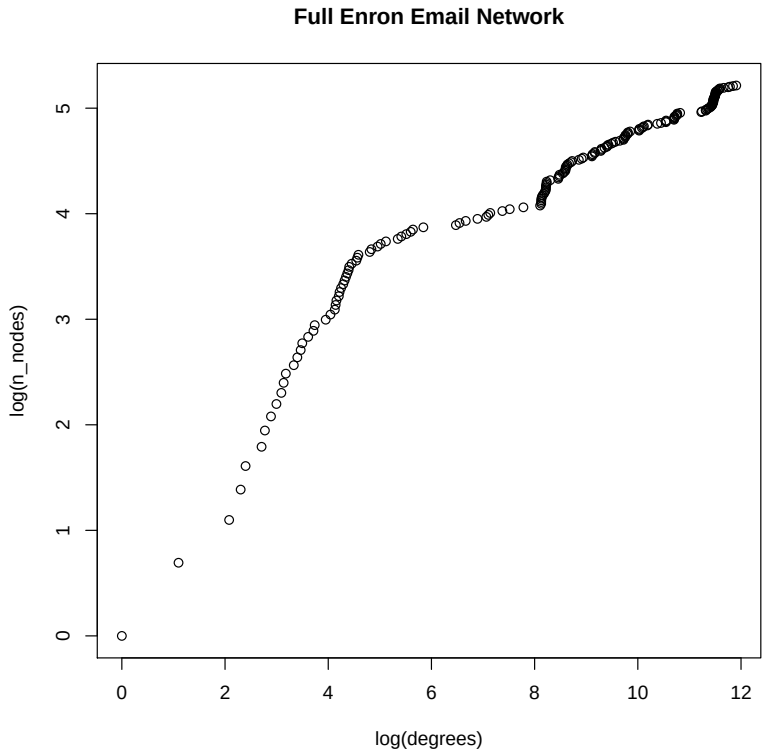


Figure 4.3: Full Enron email network, $\log(n_{\text{nodes}})$ vs $\log(degree_{\text{total}})$

4.5 Data example

In this section, we illustrate the use of the propose model on the Enron email network. The Enron email network collected the information about email interactions between 184 people who were affiliated with Enron. Each email could have one sender and multiple receivers. We assumed that one sender can only send one email at a specific time point, and a person joins the email network at the time they first appeared as a sender or receiver. Then we treated emails as edges and people as nodes. Each person was categorized into one of the following roles: Employee, Trader, Manager, Director, Vice President, President, and CEO. We used Employee as the baseline and construct 6 binary variables to indicate the role of a person. These variables were used as the covariates.

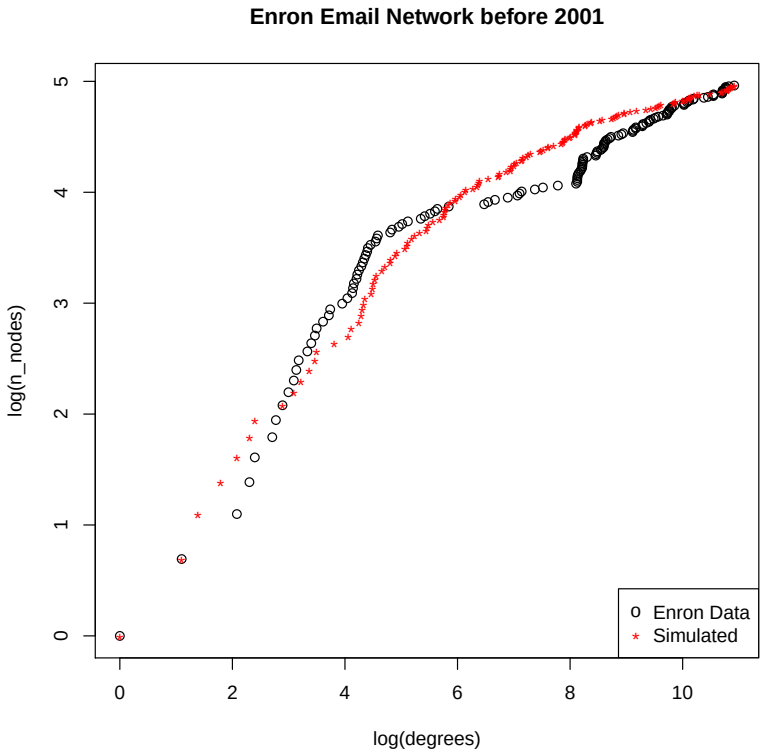


Figure 4.4: Enron email network before 2001, $\log(n_{\text{nodes}})$ vs $\log(degree_{\text{total}})$

Figure 4.3 plots the $\log(n_{\text{nodes}})$ against $\log(degree_{\text{total}})$ for the full Enron email network. We notice that there is a gap towards the right end of the plot, which corresponds to 2001 when the scandal of Enron broke out and may indicate a structural change in the network. Thus we removed that part of data and only focus on emails before 2001.

Figure 4.4 plots the $\log(n_{\text{nodes}})$ against $\log(degree_{total})$ for the Enron email network before 2001, and a simulated network using the parameter estimates from fitting the proposed model to the network. Table 4.3 shows the result of the estimated parameters and bootstrap confidence intervals from 40 bootstrap replications for the Enron email network. The result suggests that while CEO is not, people with role being Trader, Manager, Director, Vice President and President are more likely to be involved in email interactions comparing to Employee. The large θ value and small α

value suggest that the number of nodes of the network grows fast at the beginning of the network generation, and grows slowly when the total degrees in the network is large, which can also be seen from Figure 4.4.

	T	M	D	VP	P	CEO	α	θ
est	2.830	2.010	1.761	1.720	1.149	1.200	0.001	18.976
bootstrap 5%	2.632	1.767	1.596	1.583	0.441	-0.059	0.001	13.888
bootstrap 95%	3.039	2.165	1.928	1.854	1.658	1.895	0.072	21.070

Table 4.3: Estimates for Enron network, T:Trader, M: Manager, D: Director, VP/P: Vice President/President

For comparison, we also fitted the Hollywood model to the data, which gives estimated $\hat{\alpha}_{Hollywood} = 0.001$ and $\hat{\theta}_{Hollywood} = 18.17$. We computed a BIC type criterion using $-2\hat{l} + \log(n_{edges})k$ where $\hat{l}$ is the fitted log-likelihood and k is the number of parameters. The result is $BIC_{Hollywood} = 436065.2$ while the criterion for the proposed model is $BIC_{proposed} = 435578.4$, suggesting that the proposed model might be a better fit to the data.

4.6 Discussion

In summary, we have proposed a network model that can incorporate covariate information in the edge exchangeable framework. We discussed the exchangeability of the model and its interpretation. We showed the sparsity of the model and illustrated a power law behavior using simulation studies. We have developed an accurate estimation algorithm for the model and demonstrated its performance through simulation studies and a data example.

APPENDICES

APPENDIX A

Appendix for Chapter II

A.1 Proof of Theorem II.1

Without loss of generality we assume $\gamma \in (0, \frac{1}{2})$ and $a > b$. It can be checked that the same argument holds for $\gamma \in (\frac{1}{2}, 1)$ with $b > a$, while only switching the estimated label is needed for the other two cases. Thus a valid estimate of $(\hat{a}, \hat{b})$ should satisfy $\hat{a} > \hat{b}$.

With the $R(e), \tilde{B}$, we have

$$\hat{\psi}_{11}(e) = \hat{\psi}_{22}(e) = \gamma \frac{\hat{a}}{\hat{a} + \hat{b}} + (1 - \gamma) \frac{\hat{b}}{\hat{a} + \hat{b}}$$

$$\hat{\psi}_{12}(e) = \hat{\psi}_{21}(e) = \gamma \frac{\hat{b}}{\hat{a} + \hat{b}} + (1 - \gamma) \frac{\hat{a}}{\hat{a} + \hat{b}}$$

and since $\gamma \in (0, \frac{1}{2})$, $\hat{a} > \hat{b}$, we have $\hat{\psi}_{11}(e) < \hat{\psi}_{12}$.

Consider a node in community 1, in the directed case, we have $\hat{c}_i(e) = 1$ if

$$\tilde{b}_{i1}(e) \log \frac{\hat{\psi}_{11}(e)}{\hat{\psi}_{21}(e)} + \tilde{b}_{i2}(e) \log \frac{\hat{\psi}_{12}(e)}{\hat{\psi}_{22}(e)} > \hat{\beta} X_i$$

Rearranging we get $\hat{c}_i(e) \neq 1$ implies

$$\tilde{b}_{i1}(e) - \tilde{b}_{i2}(e) \geq \hat{\beta} X_i \left[\log \frac{\hat{\psi}_{11}(e)}{\hat{\psi}_{21}(e)} \right]^{-1}$$

Define $\tilde{\xi}_i(\sigma(e)) = \sum_{j=1}^{n} \tilde{A}_{ij}\sigma_j(e)$ where $\sigma_j(e) = \begin{cases} 1, & e_j = 1 \\ -1, & e_j = 2 \end{cases}$.

Then $\hat{c}_i(e) \neq 1$ if $\tilde{\xi}_i(\sigma(e)) \geq \hat{\beta} X_i \left[\log \frac{\hat{\psi}_{11}(e)}{\hat{\psi}_{21}(e)} \right]^{-1} = \hat{\beta} X_i C(\gamma)$. Further, we have

$$\begin{aligned} E((\tilde{\xi}_i(\sigma)) &= \sum_{j \in S_{11}} \frac{a}{m}(1) + \sum_{j \in S_{22}} \frac{b}{m}(-1) + \sum_{j \in S_{21}} \frac{a}{m}(-1) + \sum_{j \in S_{12}} \frac{b}{m}(1) \\ &= (a-b)\gamma + (b-a)(1-\gamma) = -(1-2\gamma)(a-b) \end{aligned} \tag{A.1}$$

and

$$v := var[\tilde{\xi}_i(\sigma(e))] = \sum_{j=1}^{n} var(\tilde{A}_{ij}\sigma_j) \leq \sum_{j=1}^{n} E(\tilde{A}_{ij}) = a + b$$

Then by Bernstein inequality, if $t/3 \leq a + b$

$$P\left[\tilde{\xi}_i(\sigma) \geq E(\tilde{\xi}_i(\sigma)) + t \right] \leq \exp(-\frac{t^2}{2(v + t/3)}) \leq \exp(-\frac{t^2}{4(a+b)})$$

And plugging in the expectation given in (A.1) we get

$$P\left[\tilde{\xi}_i(\sigma) \geq -(1-2\gamma)(a-b) + t \right] \leq \exp(-\frac{t^2}{2(v + t/3)}) \leq \exp(-\frac{t^2}{4(a+b)}) \tag{A.2}$$

Let $\tilde{M}_{n,1}(e) := \frac{1}{m} \sum_{i=1}^{m} 1(\hat{c}_i(e) \neq 1)$ be the mismatch ratio for community 1. Also, define $\tilde{N}_{n,1}(\sigma; r) = \sum_{i=1}^{m} 1(\tilde{\xi}_i(\sigma) \geq r_i)$ where r is an vector of size m with r_i being its

i-th element, we have

$$\tilde{M}_{n,1}(e) = \frac{1}{m}\sum_{i=1}^{m} 1(\hat{c}_i(e) \neq 1) \leq \frac{1}{m}\sum_{i=1}^{m} 1[\tilde{\xi}_i(\sigma(e)) \geq \hat{\beta}X_iC(\gamma)] = \frac{1}{m}\tilde{N}_{n,1}(\sigma; \hat{\beta}XC(\gamma)) \tag{A.3}$$

Here we use a the compact notation $\hat{\beta}XC(\gamma)$ for the vector with the i-th element $\hat{\beta}X_iC(\gamma)$. And the inequality is due to treating the boundary case $\tilde{\xi}_i(\sigma(e)) = \hat{\beta}X_iC(\gamma)$ as error.

We now bound $\tilde{N}_{n,1}(\sigma; \hat{\beta}XC(\gamma))$.

Define

$$p_i(k) = P[\tilde{\xi}_i(\sigma) \geq k] \text{ , and } \bar{p}_1(r) = \frac{1}{m}\sum_{i=1}^{m} p_i(r_i)$$

Then we have the result that

$$P\left[\frac{1}{m}\tilde{N}_{n,1}(\sigma; r) \geq eu\bar{p}_1(r)\right] \leq \exp(-em\bar{p}_1(r)u\log u), u > 1/e \tag{A.4}$$

by Lemma 5 of (*Amini et al.*, 2013).

Recall $|\hat{\beta}X_i| \leq M$, and we assumed that

$$m := (1-2\gamma)(a-b) - |MC(\gamma)| \geq 0$$

Then apply (A.2) by taking $t_i = (1-2\gamma)(a-b) - \hat{\beta}X_iC(\gamma)$, notice that we have $t_i \geq (1-2\gamma)(a-b) - |MC(\gamma)| \geq 0$ and by assumption $t_i \leq (1-2\gamma)(a-b) + |MC(\gamma)| \leq 3(a+b)$, so t_i is valid and we shall have

$$\bar{p}_1(\hat{\beta}XC(\gamma)) \leq \exp(-\frac{[\min_i(t_i)]^2}{4(a+b)}) \leq \exp(-\frac{m^2}{4(a+b)})$$

The cardinality of the set $\mathcal{E}^\gamma$ is $\binom{m}{m\gamma}^2 \leq \exp(2m[h(\gamma)+\kappa_\gamma(2m)])$ where $h(p) = -p\log p-(1-p)\log(1-p), p\in[0,1]$ is the binary entropy function and $\kappa_\gamma(2m) = \kappa_\gamma(n) = o(1)$. Then by union bound and (A.4) we have

$$\begin{aligned} &P\left[\sup_{\sigma\in\mathcal{E}^\gamma}\frac{1}{m}\tilde{N}_{n,1}(\sigma;\hat{\beta}XC(\gamma)) > eu_n\bar{p}_1(\hat{\beta}XC(\gamma))\right] \\ &\leq \exp\{m[2h(\gamma)+2\kappa_\gamma(n)-e\bar{p}_1(\hat{\beta}XC(\gamma))u_n\log u_n]\} \end{aligned} \tag{A.5}$$

Then take u_n such that $u_n\log u_n = \frac{4h(\gamma)}{e\bar{p}_1(\hat{\beta}XC(\gamma))}$ and use $n=2m$ we have

$$P\left[\sup_{\sigma\in\mathcal{E}^\gamma}\frac{1}{m}\tilde{N}_{n,1}(\sigma;\hat{\beta}XC(\gamma)) > \frac{4h(\gamma)}{\log u_n}\right] \leq \exp\{-[h(\gamma)-\kappa_\gamma(n)]n\}$$

The same bound holds for community 2 by symmetry and it follows that the mismatch ratio $\tilde{M}_n(e) = \frac{1}{2}\tilde{M}_{n,1}(e)+\frac{1}{2}\tilde{M}_{n,2}(e)$ has the same bound. This completes the proof.

A.2 Proof of Theorem A.1

Recall $|\hat{\beta}X_i|\leq M$, $a_\gamma = \gamma a+(1-\gamma)b$, and we assumed that

$$2(1-\epsilon)a_\gamma \leq \epsilon(1-2\gamma)(a-b), \text{ and } m := (1-\epsilon)(1-2\gamma)(a-b)+MC(\gamma)\geq 0$$

for some $\epsilon\in(0,1)$. Then apply (A.2) by taking $t_i = (1-2\gamma)(a-b)-2(1+\epsilon)a_\gamma - \hat{\beta}X_iC(\gamma)$, notice that $t_i\geq(1-\epsilon)(1-2\gamma)(a-b)+MC(\gamma)\geq 0$ and by assumption $t_i\leq(1-2\gamma)(a-b)-MC(\gamma)\leq 3(a+b)$, so t_i is valid and we shall have

$$\bar{p}_1(\hat{\beta}XC(\gamma)-2(1-\epsilon)a_\gamma) \leq \exp(-\frac{[\min_i(t_i)]^2}{4(a+b)}) \leq \exp(-\frac{m^2}{4(a+b)})$$

Define $\xi_i(\sigma(e))$ for the undirected case similarly as in the directed case with only $\tilde{A}_{ij}$ replaced by A_{ij}, we will upper bound $\xi_i(\sigma)$ in terms of $\tilde{\xi}_i(\sigma)$. Let $D_{ij} = A_{ij}-\tilde{A}_{ij}\geq 0$,

then

$$\xi_i(\sigma) - \tilde{\xi}_i(\sigma) = \sum_{j \in S_1} D_{ij} - \sum_{j \in S_2} D_{ij} \leq \sum_{j \in S_1} D_{ij} \tag{A.6}$$

Further, $D_{ij} \leq \tilde{A}_{ij} + \tilde{A}_{ji}$. Define

$$\tilde{A}_{i*}(\sigma) = \sum_{j \in S_1} \tilde{A}_{ij}, \;\; \tilde{A}_{*i}(\sigma) = \sum_{j \in S_1} \tilde{A}_{ji}$$

we have

$$\xi_i(\sigma) \leq \tilde{\xi}_i(\sigma) + \tilde{A}_{i*}(\sigma) + \;\; \tilde{A}_{*i}(\sigma) \tag{A.7}$$

We then apply Bernstein inequality to $\tilde{A}_{i*}(\sigma)$, the same result holds for $\tilde{A}_{*i}(\sigma)$ by symmetry.

$$\mu = E\left[\sum_{j \in S_1} \tilde{A}_{ij}\right] = \sum_{j \in S_{11}} \frac{a}{m} + \sum_{j \in S_{12}} \frac{b}{m} = a\gamma + b(1-\gamma) = a_\gamma$$

and $\sum_{j \in S_1} var(\tilde{A}_{ij}) \leq \mu$, thus we get

$$P\left[\tilde{A}_{i*}(\sigma) > \mu + t\right] \leq \exp(\frac{t^2}{2(\mu + t/3)})$$

Take $t = \epsilon a_\gamma$ we have

$$P\left[\tilde{A}_{i*}(\sigma) > (1+\epsilon)a_\gamma\right] \leq \exp(-\frac{\epsilon^2}{1+\epsilon/3} a_\gamma)$$

By (A.7) we have

$$\xi_i(\sigma) \geq \hat{\beta} X_i C(\gamma) \implies \left(\tilde{\xi}_i(\sigma) \geq \hat{\beta} X_i C(\gamma) - r\right) \vee (\tilde{A}_{i*}(\sigma) \geq \frac{r}{2}) \vee (\tilde{A}_{*i}(\sigma) \geq \frac{r}{2})$$

where $\vee$ is logic OR operator. Then, using indicator representation,

$$1\left[\xi_i(\sigma) \geq \hat{\beta} X_i C(\gamma)\right] \leq 1\left[(\tilde{\xi}_i(\sigma) \geq \hat{\beta} X_i C(\gamma) - r\right] + 1\left[\tilde{A}_{i*}(\sigma) \geq \frac{r}{2}\right] + 1\left[\tilde{A}_{*i}(\sigma) \geq \frac{r}{2}\right]$$

Averaging over $i \in C_1$ we have

$$\frac{1}{m} N_{n,1}(\sigma; \hat{\beta} XC(\gamma)) \leq \frac{1}{m} \tilde{N}_{n,1}(\sigma; \hat{\beta} XC(\gamma) - r) + \frac{1}{m} \tilde{Q}_{n,1*}(\sigma; \frac{r}{2}) + \frac{1}{m} \tilde{Q}_{n,*1}(\sigma; \frac{r}{2}) \quad \text{(A.8)}$$

where $N_{n,1}(\sigma; r)$ is defined similar to $\tilde{N}_{n,1}(\sigma; r)$ but with $\tilde{\xi}_i(\sigma)$ replaced by $\xi_i(\sigma)$. And $\tilde{Q}_{n,1*}(\sigma; t) = \sum_{i=1}^{m} 1\left[\tilde{A}_{i*}(\sigma) \geq t\right]$, similarly for $\tilde{Q}_{n,*1}(\sigma; t)$.
Let $q_i(r) = P(\tilde{A}_{i*}(\sigma) \geq \frac{r}{2}), \bar{q}_1(r) = \frac{1}{m} \sum_{i=1}^{m} q_i(r)$.
Then similar to (A.4) we have

$$P\left[\frac{1}{m} \tilde{Q}_{n,1*}(\sigma; r/2) \geq eu\bar{q}_1(r)\right] \leq \exp(-em\bar{q}_1(r) u \log u), \text{ for } u > 1/e \quad \text{(A.9)}$$

The same bound holds for $\frac{1}{m} \tilde{Q}_{n,*1}(\sigma; r/2)$. By combining the bounds on $\frac{1}{m} \tilde{N}_{n,1}(\sigma; \hat{\beta} XC(\gamma) - r)$, $\frac{1}{m} \tilde{Q}_{n,1*}(\sigma; r/2)$, and $\frac{1}{m} \tilde{Q}_{n,*1}(\sigma; r/2)$, we obtain

$$\begin{aligned}
&P\left[\sup_{\sigma \in \Sigma^\gamma} \frac{1}{m} N_{n,1}(\sigma; \hat{\beta} XC(\gamma)) \geq e[u_n \bar{p}_1(\hat{\beta} XC(\gamma) - r) + 2v_n \bar{q}_1(r)]\right] \\
&\leq P\left[\sup_{\sigma \in \Sigma^\gamma} \frac{1}{m} \tilde{N}_{n,1}(\sigma; \hat{\beta} XC(\gamma) - r) \geq eu_n \bar{p}_1(\hat{\beta} XC(\gamma) - r)\right] \\
&+ 2P\left[\sup_{\sigma \in \Sigma^\gamma} \frac{1}{m} \tilde{Q}_{n,*1}(\sigma; r/2) \geq ev_n \bar{q}_1(r)\right] \\
&\leq \exp\{m[2h(\gamma) - e\bar{p}_1(\hat{\beta} XC(\gamma) - r) u_n \log u_n + 2\kappa_\gamma(n)]\} \\
&+ 2\exp\{m[2h(\gamma) - e\bar{q}_1(r) v_n \log v_n + 2\kappa_\gamma(n)]\}
\end{aligned} \quad \text{(A.10)}$$

where $u_n, v_n > 1/e$. Now take $r = 2(1+\epsilon)a_\gamma$, we have $\bar{q}_1(r) \leq \exp(-\frac{\epsilon^2}{1+\epsilon/3} a_\gamma)$, and $\bar{p}_1(\hat{\beta} XC(\gamma) - r) \leq \exp(-\frac{m^2}{4(a+b)})$. Then we get the bound for mismatch ratio of

community 1 $M_{n,1}(e)$ by picking u_n, v_n such that

$$u_n \log u_n = \frac{4h(\gamma)}{e\bar{p}_1(\hat{\beta}XC(\gamma) - r)},\ v_n \log v_n = \frac{4h(\gamma)}{e\bar{q}_1(r)}$$

The same bound holds for $M_{n,2}(e)$ using the same argument. And The proof is finished by noting that $M_n(e) = \frac{1}{2}(M_{n,1}(e) + M_{n,2}(e))$.

A.3 Proof of Corollary A.2

$$\hat{\beta}_{PL} = \arg\max_{\beta} \frac{1}{n}\left[\sum_{i=1}^{n} 1(\hat{c}_i(e) = 1)\beta X_i - \log(1 + \beta X_i)\right]$$

Notice that

$$\begin{aligned} &\frac{1}{n}\left[\sum_{i=1}^{n} 1(\hat{c}_i(e) = 1)\beta X_i - \log(1 + \beta X_i)\right] = \\ &\frac{1}{n}\left[\sum_{i=1}^{n} 1(c_i(e) = 1)\beta X_i - \log(1 + \beta X_i)\right] + \frac{1}{n}\sum_{i=1}^{n} 1[\hat{c}_i(e) - c_i(e)]\beta X_i\ . \end{aligned} \tag{A.11}$$

By assumption $\frac{1}{n}\sum_{i=1}^{n} 1[\hat{c}_i(e) \neq c_i(e)] \xrightarrow{p} 0$, X_i bounded as $|\hat{\beta}X_i| < M$ where M is a constant. Further, the MLE estimator of logistic regression

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \frac{1}{n}\left[\sum_{i=1}^{n} 1(c_i(e) = 1)\beta X_i - \log(1 + \beta X_i)\right]$$

is consistent under regularity conditions, we have $\hat{\beta}_{PL}$ is consistent.

A.4 General directed case

We consider the case where $K = 2$ and $\tilde{B} = \frac{1}{n}\begin{pmatrix} a_1 & b \\ b & a_2 \end{pmatrix} = \frac{b}{n}\begin{pmatrix} \rho_1 & 1 \\ 1 & \rho_2 \end{pmatrix}$, with $\rho_1, \rho_2 > 1$ and b scales with n.

Suppose we have initial estimates $\hat{\rho}_1, \hat{\rho}_2, \hat{b}, \hat{\beta}$, and initial labeling $\mathbf{e}$, the labels are estimated by

$$\hat{c}_i(e) = \arg\max_{k\in\{1,2\}} \{\hat{\beta}_k X_i + \sum_{m=1}^{2} \tilde{b}_{im}(e) \log \hat{\psi}_{km}(e)\} \tag{A.12}$$

where $\hat{\psi}_{km}$ are the elements of the row normalized matrix of $\tilde{\Lambda} = [nR(e)\tilde{B}]^T$ and $\tilde{B}$ is estimated by plugging in $\hat{\rho}_1, \hat{\rho}_2, \hat{b}$. We further assume that the initial estimates have $\underline{\rho}_k \le \hat{\rho}_k \le \bar{\rho}_k$ where $\{\underline{\rho}_k\}$ and $\{\bar{\rho}_k\}$ are bounds on the estimates $\hat{\rho}_k, k = 1, 2$.

Consider an initial labeling $\mathbf{e} = \{e_i\} \in \{1, 2\}^n$ that matches $n_1\gamma_1$ labels in community 1 and $n_2\gamma_2$ labels in community 2, n_1, n_2 are the number of nodes are truly in community 1 and community 2 respectively. We use $\mathcal{E}^{\gamma_1,\gamma_2}$ do denote the collection of all such labeling Another assumption is that the covariates X and initial estimate $\hat{\beta}$ should be bounded such that $|\hat{\beta}X_i| \le M$ where M is a constant satisfying some technical conditions we will state.

Let $\tilde{M}_n(e) := \min_{\phi\in\{(1,2),(2,1)\}} \frac{1}{n}\sum_{i=1}^{n} 1[\hat{c}_i(e) \neq \phi(c_i)]$ be the mismatch ratio for the directed case, where ϕ is considering the fact that the labels are identified up to a permutation. We will show a consistency result based on the convergence of this mismatch ratio, which corresponds to the weak consistency definition. Under these assumptions, the confusion matrix $R = \begin{pmatrix} \gamma_1\pi_1 & (1-\gamma_2)\pi_2 \\ (1-\gamma_1)\pi_1 & \gamma_2\pi_2 \end{pmatrix}$.

Define $u(x) = \frac{(1-\gamma_1)x+\gamma_2\tau}{\gamma_1 x+(1-\gamma_2)\tau}$, $v(x) = u(\frac{1}{x})$, $F_1(x, y) = \log\frac{1+u(x)}{1+v(y)}$ and $F_2(x, y) = \log\frac{1+[u(x)]^{-1}}{1+[v(y)]^{-1}}$ where $\tau = \pi_2/\pi_1$

Let $\alpha_1 := F_1(\bar{\rho}_1, \bar{\rho}_2), \beta_1 := F_1(\underline{\rho}_1, \underline{\rho}_2), \alpha_2 := F_2(\underline{\rho}_1, \underline{\rho}_2), \beta_2 := F_2(\bar{\rho}_1, \bar{\rho}_2)$.

Set $\alpha = (\alpha_1, \alpha_2), |\alpha| = (|\alpha_1|, |\alpha_2|)$ and similarly for β.

Define $z_{1,n,i} := -[\tilde{\Lambda}\alpha]_1 + \hat{\beta}X_i$ and $z_{2,n,i} = [\tilde{\Lambda}\beta]_2 - \hat{\beta}X_i$, recall that we assume

$|\hat{\beta}X_i| \leq M$.

Theorem A.1. *Assume* $\gamma_1, \gamma_2 \in (0, \frac{1}{2})$,

$$\begin{aligned} m_1 &:= -[\tilde{\Lambda}\alpha]_1 - M \geq 0 \text{ and } -[\tilde{\Lambda}\alpha]_1 + M \leq 3[\tilde{\Lambda}|\alpha|]_1 \\ m_2 &:= [\tilde{\Lambda}\beta]_2 - M \geq 0 \text{ and } [\tilde{\Lambda}\beta]_2 + M \leq 3[\tilde{\Lambda}|\beta|]_2 \end{aligned} \tag{A.13}$$

Let $C := \sum_{i=1}^{2} \pi_i h(\gamma_i)$ *and* $r_n := \sum_{i=1}^{2} \pi_i \kappa_{\gamma_i}(2\pi_i n)$ *where* $h(p) = -p\log p - (1-p)\log(1-p), p \in [0,1]$ *is the binary entropy function, and* $\kappa_\gamma(n) := \frac{1}{n}[\log \frac{n}{4\pi\gamma(1-\gamma)} + \frac{1}{3n}] = o(1)$.

Then

$$P\left[\sup_{e \in \mathcal{E}^{\gamma_1,\gamma_2}} \tilde{M}_n(e) > 2C \sum_{k=1}^{2} \frac{1}{\log u_{n,k}}\right] \leq 2\exp[-n(C - r_n)]$$

where

$$\log(u_{n,1}\log u_{n,1}) \geq \log\frac{2C}{e\pi_1} + \frac{m_1^2}{4||\alpha||_\infty[\tilde{\Lambda}|\alpha|]_1}$$

$$\log(u_{n,2}\log u_{n,2}) \geq \log\frac{2C}{e\pi_2} + \frac{m_2^2}{4||\beta||_\infty[\tilde{\Lambda}|\beta|]_2}$$

A simpler result can be obtained following a same argument as Theorem 3 in (*Amini et al.*, 2013). If ρ_1, ρ_2 are large enough and γ_1, γ_2 satisfy some assumptions, we can estimate ρ_1, ρ_2 by be infinity and still get consistency. Define KL-divergence $D(p||q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. and $D(\gamma_1, \gamma_2) = \frac{D(\gamma_1||(1-\gamma_2))}{D((1-\gamma_2)||\gamma_1)}$.

Corollary A.2. *Assume that we start with* $\hat{\rho}_1 = \hat{\rho}_2 = \infty$ *and arbitrary* $\hat{b}$. *Assume* γ_1, γ_2 *satisfy*

$$\frac{\tau}{\rho_1}(1+\epsilon) \leq D(\gamma_1, \gamma_2) \leq (1-\epsilon)\rho_2\tau \tag{A.14}$$

for some $\epsilon \in (0,1)$. *Then for* $b \geq 2\epsilon^{-1}\frac{M}{\pi_1\tau(\rho_2 \wedge 1)D((1-\gamma_2)||\gamma_1)}$, *we have*

$$P\left[\sup_{e \in \mathcal{E}^{\gamma_1,\gamma_2}} \tilde{M}_n(e) > 2C \sum_{k=1}^{2} \frac{1}{\log u_{n,k}}\right] \leq 2\exp[-n(C - r_n)]$$

where

$$\log(u_{n,k}\log u_{n,k}) \geq \log\frac{2C}{e\pi_k} + b\pi_1\frac{\epsilon^2 C_k}{16C_0}D((1-\gamma_2)||\gamma_1),\ k=1,2$$

where $C_1 = \frac{\tau^2}{\rho_1 D(\gamma_1,\gamma_2)+\tau}$, $C_2 = \frac{\rho_2^2\tau^2}{\rho_1 D(\gamma_1,\gamma_2)+\tau}$,
and $C_0 = \max|\log((1-\gamma_2)/\gamma_1)|, |\log(\gamma_2/(1-\gamma_1))|$.

Remark A.3. The corollary suggest that if the parameters ρ_i, π_i, γ_i, $i = 1, 2$ are constant that does not scale with n while $b \to \infty$, the estimate start with $\hat{\rho}_1 = \hat{\rho}_2 = \infty$ is consistent as long as the assumption (A.14) is satisfied. Regarding the assumption, if we fix γ_1, γ_2 and τ, then the inequality can be satisfied by increasing ρ_1 and ρ_2.

A.4.1 Proof of Theorem A.1

The proof is mainly based on (*Amini et al.*, 2013) Proposition 1.
Let $C_l = i : c_i = l$ and $S_l = i : e_i = l$, $S_{kl} = S_k \cap C_l$. $n_1 = |C_1|$, $n_2 = |C_2|$.
Initial label $\mathbf{e}$ satisfies that $\mathbf{e}$ matches $\gamma_1 n_1$ labels in community 1 and $\gamma_2 n_2$ labels in community 2. Let

$$\mathcal{E}^{\gamma_1,\gamma_2} = \left\{\mathbf{e}\in\{1,2\}^n : \sum_{i\in C_1} 1(e_i=1) = \gamma_1 n_1, \sum_{i\in C_2} 1(e_i=2) = \gamma_2 n_2\right\}$$

be the collection of all such initial labeling
$R = (\frac{1}{n}|S_{kl}|) = \begin{pmatrix} \gamma_1\pi_1 & (1-\gamma_2)\pi_2 \\ (1-\gamma_1)\pi_1 & \gamma_2\pi_2 \end{pmatrix}$ where $\pi_i = |C_i|/n$.
$\tilde{B} = \frac{1}{n}\begin{pmatrix} a_1 & b \\ b & a_2 \end{pmatrix} = \frac{b}{n}\begin{pmatrix} \rho_1 & 1 \\ 1 & \rho_2 \end{pmatrix}$, with $\rho_1, \rho_2 > 1$ and b scales with n. Then $\tilde{\Lambda} = [nR\tilde{B}]^T$. We have initial estimate $\hat{B} = \frac{\hat{b}}{n}\begin{pmatrix} \hat{\rho}_1 & 1 \\ 1 & \hat{\rho}_2 \end{pmatrix}$ and by assumption, $\underline{\rho}_k \leq \hat{\rho}_k \leq \bar{\rho}_k$

where $\{\underline{\rho}_k\}$ and $\{\bar{\rho}_k\}$ are bounds on the estimate $\hat{\rho}_k$. Then

$$\hat{\Lambda} = \begin{pmatrix} \hat{a}_1 & \hat{b} \\ \hat{b} & \hat{a}_2 \end{pmatrix} \begin{pmatrix} \gamma_1\pi_1 & (1-\gamma_2)\pi_2 \\ (1-\gamma_1)\pi_1 & \gamma_2\pi_2 \end{pmatrix} = \hat{b}\pi_1 \begin{pmatrix} \hat{\rho}_1\gamma_1 + \tau(1-\gamma_2) & \hat{\rho}_1(1-\gamma_1) + \tau\gamma_2 \\ \gamma_1 + \tau\hat{\rho}_2(1-\gamma_2) & (1-\gamma_1) + \tau\hat{\rho}_2\gamma_2 \end{pmatrix}$$

where $\tau = \frac{\pi_2}{\pi_1}$, then $\hat{\psi} = \begin{pmatrix} \frac{\hat{\lambda}_{11}}{\hat{\lambda}_{11}+\hat{\lambda}_{12}} & \frac{\hat{\lambda}_{12}}{\hat{\lambda}_{11}+\hat{\lambda}_{12}} \\ \frac{\hat{\lambda}_{21}}{\hat{\lambda}_{21}+\hat{\lambda}_{22}} & \frac{\hat{\lambda}_{22}}{\hat{\lambda}_{21}+\hat{\lambda}_{22}} \end{pmatrix}$.

$u(x) = \frac{(1-\gamma_1)x+\gamma_2\tau}{\gamma_1 x+(1-\gamma_2)\tau}$, $v(x) = u(\frac{1}{x})$, $F_1(x,y) = \log\frac{1+u(x)}{1+v(y)}$ and $F_2(x,y) = \log\frac{1+[u(x)]^{-1}}{1+[v(y)]^{-1}}$. Then we can check

$$\frac{\hat{\psi}_{21}}{\hat{\psi}_{11}} = \frac{1+u(\hat{\rho}_1)}{1+v(\hat{\rho}_2)} = F_1(\hat{\rho}_1, \hat{\rho}_2) \text{ and } \frac{\hat{\psi}_{22}}{\hat{\psi}_{12}} = F_2(\hat{\rho}_1, \hat{\rho}_2)$$

Assume $\gamma_1, \gamma_2 \in (0, \frac{1}{2})$, $(1-\gamma_1)(1-\gamma_2) > \gamma_1\gamma_2$. We will have $u(x)$ is increasing on $(0,\infty)$ and $v(x)$ decreasing, then we will get

$$\begin{aligned} \beta_1 := F_1(\underline{\rho}_1, \underline{\rho}_2) \le \log\frac{\hat{\psi}_{21}}{\hat{\psi}_{11}} \le F_1(\bar{\rho}_1, \bar{\rho}_2) =: \alpha_1 \\ \beta_2 := F_2(\bar{\rho}_1, \bar{\rho}_2) \le \log\frac{\hat{\psi}_{22}}{\hat{\psi}_{12}} \le F_2(\underline{\rho}_1, \underline{\rho}_2) =: \alpha_2 \end{aligned} \tag{A.15}$$

Further, let $\hat{\beta}$ be the estimate of coefficient for logistic regression and assume that for any initial labeling $\mathbf{e} \in \mathcal{E}^{\gamma_1,\gamma_2}$, the corresponding $\hat{\beta}$ has $|\hat{\beta}X_i| \le M$ where M is a constant.

The conditional pseudo likelihood estimate of community label is defined as

$$\hat{c}_i(e) = \arg\max_{k\in\{1,2\}} \{\hat{\beta}_k X_i + \sum_{m=1}^{2} \tilde{b}_{im}(e)\log\hat{\psi}_{km}(e)\} \tag{A.16}$$

WLOG, we assume $\hat{\beta}_2 = 0$ and $\hat{\beta}_1 = \hat{\beta}$. Consider a node $i \in C_1$, then $\hat{c}_i(e) = 1$ if

$$\tilde{b}_{i1}(e) \log \frac{\hat{\psi}_{21}(e)}{\hat{\psi}_{11}(e)} + \tilde{b}_{i2}(e) \log \frac{\hat{\psi}_{22}(e)}{\hat{\psi}_{12}(e)} < \hat{\beta} X_i \tag{A.17}$$

If $\hat{c}_i(e) \neq 1$, then LHS of (A.17) is $\geq \hat{\beta} X_i$ and thus implies that $\alpha_1 \tilde{b}_{i1}(e) + \alpha_2 \tilde{b}_{i2}(e) \geq \hat{\beta} X_i$.

Define

$$\sigma_j(e) := \begin{cases} \alpha_1, \ e_j = 1 \\ \alpha_2, \ e_j = 2 \end{cases}$$

Then $\alpha_1 \tilde{b}_{i1}(e) + \alpha_2 \tilde{b}_{i2}(e) = \sum_j \tilde{A}_{ij} \sigma_j(e) =: \tilde{\xi}_i(\sigma(e))$. So we have that $\hat{c}_i(e) \neq 1$ implies $\tilde{\xi}_i(\sigma(e)) \geq \hat{\beta} X_i$.

Then the mismatch ratio over community 1 is

$$\tilde{M}_{n,1}(e) := \frac{1}{n_1} \sum_{i \in C_1} 1(\hat{c}_i(e) \neq 1) \leq \frac{1}{n_1} \sum_{i \in C_1} 1(\tilde{\xi}_i(\sigma(e)) \geq \hat{\beta} X_i) =: \frac{1}{n_1} \tilde{N}_{n,1}(\sigma; \hat{\beta} X) \tag{A.18}$$

By Bernstein inequality, we have

$$P\left[\tilde{\xi}_i(\sigma) \geq E[\tilde{\xi}_i(\sigma)] + t\right] \leq \exp(-\frac{t^2/2}{\sum_j var(\tilde{A}_{ij}\sigma_j) + ||\alpha||_\infty t/3}) \tag{A.19}$$

where $||\alpha||_\infty := max\{|\alpha_1|, |\alpha_2|\}$ and $|\tilde{A}_{ij}\sigma_j| \leq ||\alpha||_\infty$.

Since $i \in C_1$,

$$\begin{aligned} E[\tilde{\xi}_i(\sigma)] = \sum_j \sigma_j E(\tilde{A}_{ij}) &= \sum_{k=1}^{2} \sum_{\ell=1}^{2} \sum_j \sigma_j E(\tilde{A}_{ij}) 1(j \in S_{k\ell}) \\ &= \sum_{k=1}^{2} \sum_{\ell=1}^{2} \sum_j \alpha_k \tilde{B}_{1\ell} 1(j \in S_{k\ell}) \\ &= n \sum_{k=1}^{2} \sum_{\ell=1}^{2} \alpha_k \tilde{B}_{\ell 1} \frac{|S_{k\ell}|}{n} = n[\alpha^T R \tilde{B}]_1 \end{aligned} \tag{A.20}$$

where $[\alpha^T R\tilde{B}]_1$ is the first element of $\alpha^T R\tilde{B}$ and $\alpha = (\alpha_1, \alpha_2)$. Recall $\tilde{\Lambda} = [nR\tilde{B}]^T$ so $n[\alpha^T R\tilde{B}] = (\tilde{\Lambda}\alpha)^T$ and we have $E[\tilde{\xi}_i(\sigma)] = [\tilde{\Lambda}\alpha]_1$.

Similarly,

$$\begin{aligned}\sum_j var(\tilde{A}_{ij}\sigma_j) &= \sum_j \sigma_j^2 var(\tilde{A}_{ij}) \\ &\leq \sum_j \sigma_j E(\tilde{A}_{ij}) \leq ||\alpha||_\infty \sum_j |\sigma_j| E(\tilde{A}_{ij}) = ||\alpha||_\infty [\tilde{\Lambda}|\alpha|]_1\end{aligned} \tag{A.21}$$

where $|\alpha| = (|\alpha_1|, |\alpha_2|)$. Plug in (A.20)(A.21) to (A.19) we get

$$P\left[\tilde{\xi}_i(\sigma) \geq [\tilde{\Lambda}\alpha]_1 + t\right] \leq \exp(-\frac{t^2}{2||\alpha||_\infty([\tilde{\Lambda}|\alpha|]_1 + t/3)})$$

Take $t_i = z_{1,n,i} := -[\tilde{\Lambda}\alpha]_1 + \hat{\beta}X_i$, which is valid by assumption $z_{1,n,i} \geq -[\tilde{\Lambda}\alpha]_1 - M := m_1 \geq 0$ and $z_{1,n,i}/3 \leq [\tilde{\Lambda}|\alpha|]_1$, we obtain

$$P\left[\tilde{\xi}_i(\sigma) \geq \hat{\beta}X_i\right] \leq \exp(-\frac{z_{1,n,i}^2}{4||\alpha||_\infty[\tilde{\Lambda}|\alpha|]_1}) \leq \exp(-\frac{m_1^2}{4||\alpha||_\infty[\tilde{\Lambda}|\alpha|]_1})$$

Let $p_i(r) := P\left[\tilde{\xi}_i(\sigma) \geq r\right]$ and $\bar{p}_1(\hat{\beta}X) = \frac{1}{n_1}\sum_{i\in C_1} p_i(\hat{\beta}X_i)$ Then by (*Amini et al.*, 2013) lemma 2 we have

$$P\left[\frac{1}{n_1}\tilde{N}_{n,1}(\sigma; \hat{\beta}X) \geq eu\bar{p}_1(\hat{\beta}X)\right] \leq \exp(-en_1\bar{p}_1(\hat{\beta}X)u\log u), u > 1/e$$

And we have just obtained

$$\bar{p}_1(\hat{\beta}X) \leq \exp(-\frac{m_1^2}{4||\alpha||_\infty[\Lambda|\alpha|]_1})$$

Then we want to take a supremum over the set $\mathcal{E}^{\gamma_1,\gamma_2}$. By (*Amini et al.*, 2013) lemma 6, the cardinality of the set $\Sigma^{\gamma_1,\gamma_2} := \{\sigma(e) : e \in \mathcal{E}^{\gamma_1,\gamma_2}\}$ is

$$\binom{n_1}{\gamma_1 n_1}\binom{n_2}{\gamma_2 n_2} \leq \exp(\sum_{i=1}^{2} n_i[h(\gamma_i) + \kappa_{\gamma_i}(2n_i)]) = \exp(n(C + r_n))$$

where $C = \sum_{i=1}^{2} \pi_i h(\gamma_i), r_n = \sum_{i=1}^{2} \pi_i \kappa_{\gamma_i}(2n_i)$, $h(p) = -p\log p - (1-p)\log(1-p), p \in [0,1]$ is the binary entropy function, and $\kappa_\gamma(n) = \frac{1}{n}(\log\frac{n}{4\pi\gamma(1-\gamma)} + \frac{1}{3n})$.

We then obtain

$$P\left[\sup_{\sigma\in\Sigma^{\gamma_1,\gamma_2}} \frac{1}{n_1}\tilde{N}_{n,1}(\sigma;\hat{\beta}X) > eu_n\bar{p}_1(\hat{\beta}X)\right] \leq \exp\{n[(C + r_n) - e\pi_1\bar{p}_1(\hat{\beta}X)]u_n\log u_n\} \tag{A.22}$$

Pick u_n that $u_n \log u_n = \frac{2C}{e\pi_1\bar{p}_1(\hat{\beta}X)}$ then

$$P\left[\sup_{\sigma\in\Sigma^{\gamma_1,\gamma_2}} \frac{1}{n_1}\tilde{N}_{n,1}(\sigma;\hat{\beta}X) > \frac{1}{\pi_1}\frac{2C}{\log u_n}\right] \leq \exp\{-n(C - r_n)\}$$

Next consider a node in community 2, $i \in C_2$, similarly, $\hat{c}_i(e) \neq 2$ implies

$$\tilde{b}_{i1}(e)\log\frac{\hat{\psi}_{21}(e)}{\hat{\psi}_{11}(e)} + \tilde{b}_{i2}(e)\log\frac{\hat{\psi}_{22}(e)}{\hat{\psi}_{12}(e)} \leq \hat{\beta}X_i \tag{A.23}$$

and thus implies $\beta_1\tilde{b}_{i1}(e) + \beta_2\tilde{b}_{i2}(e) \leq \hat{\beta}X_i$. Define

$$\sigma_j(e) := \begin{cases} \beta_1, & e_j = 1 \\ \beta_2, & e_j = 2 \end{cases}$$

Then $\beta_1\tilde{b}_{i1}(e) + \beta_2\tilde{b}_{i2}(e) = \sum_j \tilde{A}_{ij}\sigma_j(e) =: \tilde{\xi}_i(\sigma(e))$.

Then the miss match ratio over community 2 is

$$\tilde{M}_{n,2}(e) := \frac{1}{n_2}\sum_{i\in C_2} 1(\hat{c}_i(e) \neq 2) \leq \frac{1}{n_2}\sum_{i\in C_2} 1(\tilde{\xi}_i(\sigma(e)) \leq \hat{\beta}X_i) =: \frac{1}{n_2}\tilde{N}_{n,2}(\sigma;\hat{\beta}X) \tag{A.24}$$

Then by Bernstein inequality and similar argument to that for community 1, we have

$$P\left[\tilde{\xi}_i(\sigma) \leq [\tilde{\Lambda}\beta]_2 - t\right] \leq \exp(-\frac{t^2}{2||\beta||_\infty([\tilde{\Lambda}|\beta|]_2 + t/3)})$$

Taking $t_i = z_{2,n,i} := [\tilde{\Lambda}\beta]_2 - \hat{\beta}X_i$, which is valid by assumption $z_{2,n,i} \geq [\tilde{\Lambda}\beta]_2 - M := m_2 \geq 0$, we get

$$P\left[\tilde{\xi}_i(\sigma) \leq \hat{\beta}X_i\right] \leq \exp(-\frac{z_{2,n,i}^2}{4||\beta||_\infty[\tilde{\Lambda}|\beta|]_2}) \leq \exp(-\frac{m_2^2}{4||\beta||_\infty[\tilde{\Lambda}|\beta|]_2})$$

Similarly define $p_i(r)$ and $\bar{p}_2(\hat{\beta}X)$ we have $\bar{p}_2(\hat{\beta}) \leq \exp(-\frac{m_2^2}{4||\beta||_\infty[\Lambda|\beta|]_2})$ and choose $u_n \log u_n = \frac{2C}{e\pi_2\bar{p}_2(\hat{\beta}X)}$ and we get a similar bound for community 2

$$P\left[\sup_{\sigma\in\Sigma^{\gamma_1,\gamma_2}} \frac{1}{n_2}\tilde{N}_{n,2}(\sigma;\hat{\beta}X) > \frac{1}{\pi_2}\frac{2C}{\log u_n}\right] \leq \exp\{-n(C - r_n)\}$$

Last, putting the two classes together, the total mismatch is $\tilde{M}_n(e) := \pi_1\tilde{M}_{1,n}(e) + \pi_2\tilde{M}_{2,n}(e)$ which completes the proof.

A.5 Proof of Corollary A.2

The proof is using the same argument as the proof for Theorem 3 in (*Amini et al.*, 2013), with minor modifications to the assumptions. Nonetheless, we repeat the proof here for completeness.

Define $D(p||q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. and $D(\gamma_1,\gamma_2) = \frac{D(\gamma_1||(1-\gamma_2))}{D((1-\gamma_2)||\gamma_1)}$

Assume $\frac{\tau}{\rho_1}(1-\epsilon) \le D(\gamma_1,\gamma_2) \le (1-\epsilon)\rho_2\tau$, for some $\epsilon \in (0,1)$.

Apply $\bar{\rho}_i = \underline{\rho}_i = \infty$, then $u(\infty) = (1-\gamma_1)/\gamma_1$, $v(\infty) = \gamma_2/(1-\gamma_2)$, and $\alpha_1 = \beta_1 = \log((1-\gamma_2)/\gamma_1)$, $\alpha_2 = \beta_2 = \log(\gamma_2/(1-\gamma_1))$. Then

$$\begin{aligned}
\tilde{\Lambda}\alpha = \tilde{\Lambda}\beta = n\tilde{B}R^T\beta &= b\pi_1 \begin{pmatrix} \rho_1 & 1 \\ 1 & \rho_2 \end{pmatrix} \begin{pmatrix} \gamma_1 & (1-\gamma_1) \\ (1-\gamma_2)\tau & \gamma_2\tau \end{pmatrix} \begin{pmatrix} \log((1-\gamma_2)/\gamma_1) \\ \log(\gamma_2/(1-\gamma_1)) \end{pmatrix} \\
&= b\pi_1 \begin{pmatrix} \rho_1 & 1 \\ 1 & \rho_2\tau \end{pmatrix} \begin{pmatrix} \gamma_1 & (1-\gamma_1) \\ (1-\gamma_2) & \gamma_2 \end{pmatrix} \begin{pmatrix} \log((1-\gamma_2)/\gamma_1) \\ \log(\gamma_2/(1-\gamma_1)) \end{pmatrix} \\
&= b\pi_1 \begin{pmatrix} \rho_1 & 1 \\ 1 & \rho_2\tau \end{pmatrix} \begin{pmatrix} -D(\gamma_1||(1-\gamma_2)) \\ D((1-\gamma_2)||\gamma_1)) \end{pmatrix}
\end{aligned} \tag{A.25}$$

Let $L_1 = D(\gamma_1||(1-\gamma_2))$, $L_2 = D((1-\gamma_2)||\gamma_1)$. Then $L_1, L_2 \ge 0$ by non-negativity of KL-divergence and

$$[\tilde{\Lambda}\alpha]_1 = b\pi_1(-\rho_1 L_1 + \tau L_2),\ [\tilde{\Lambda}\beta]_2 = b\pi_1(-L_1 + \rho_2\tau L_2)$$

By assumption $\frac{M}{\pi_1 b} \le \frac{1}{2}\epsilon\tau L_2$, then

$$\begin{aligned}
\frac{m_1}{\pi_1 b} = \frac{-[\tilde{\Lambda}\alpha]_1}{\pi_1 b} - \frac{M}{\pi_1 b} &\ge \rho_1 L_1 - \tau L_2 - \frac{\epsilon}{2}\tau L_2 \\
&= L_2[\rho_1\frac{L_1}{L_2} - \tau(1+\epsilon)] + \frac{\epsilon}{2}\tau L_2 \\
&\ge \frac{\epsilon}{2}\tau L_2 \ge 0
\end{aligned} \tag{A.26}$$

On the other hand, since $\epsilon \in (0,1)$

$$\frac{-[\tilde{\Lambda}\alpha]_1}{\pi_1 b} + \frac{M}{\pi_1 b} \leq \rho_1 L_1 - \tau L_2 + \frac{\epsilon}{2}\tau L_2 \leq 3(\rho_1 L_1 + \tau L_2) = 3\frac{[\tilde{\Lambda}|\alpha|]_1}{\pi_1 b}$$

Thus the condition that $0 \leq z_{1,n,i} \leq 3\frac{[\tilde{\Lambda}|\alpha|]_1}{\pi_1 b}$ is satisfied. And we have

$$\frac{m_1^2}{[\tilde{\Lambda}|\alpha|]_1} = b\pi_1 \frac{[m_1/(\pi_1 b)]^2}{[\tilde{\Lambda}|\alpha|]_1/(\pi_1 b)} \geq b\pi_1 \frac{(\frac{\epsilon}{2}\tau L_2)^2}{\rho_1 L_1 + \tau L_2}$$

For community 2, a similar argument shows that the conditions are also satisfied and since we assumed $\frac{M}{\pi_1 b} \leq \frac{1}{2}\epsilon\rho_2\tau L_2$. Then

$$\begin{aligned} \frac{m_2}{\pi_1 b} = \frac{[\tilde{\Lambda}\beta]_2}{\pi_1 b} - \frac{M}{\pi_1 b} &\geq -L_1 + \rho_2\tau L_2 - \frac{\epsilon}{2}\rho_2\tau L_2 \\ &= L_2(-\frac{L_1}{L_2} + \rho_2\tau(1-\epsilon)) + \frac{\epsilon}{2}\rho_2\tau L_2 \geq 0 \end{aligned} \tag{A.27}$$

And similarly,

$$\frac{m_2^2}{[\tilde{\Lambda}|\beta|]_2} = b\pi_1 \frac{[m_2/(\pi_1 b)]^2}{[\tilde{\Lambda}|\beta|]_2/(\pi_1 b)} \geq b\pi_1 \frac{(\frac{\epsilon}{2}\rho_2\tau L_2)^2}{L_1 + \tau\rho_2 L_2}$$

Putting the pieces together and noting $||\alpha||_\infty = ||\beta||_\infty = C_0$ finishes the proof.

APPENDIX B

Appendix for Chapter III

B.1 Proof of Proposition III.1

We follow the proof of (*Liu et al.*, 2013) and show a similar result under similar assumptions.

Suppose we observed a network A with n nodes, and there are p covariates for each node. The covariates are collected into $n \times p$ matrix X with variables x_i. Denote x_i^{mis} and x_i^{obs} the observed and missing subsets of variable i and let $x^{mis} = \{x_i^{mis}, i = 1, 2, ..., p\}$, $x^{obs} = \{x_i^{obs}, i = 1, 2, ..., p\}$. Further we use x_{-i} to denote the variables excluding the i-th variable. Let $\Theta = (\theta_R, \theta_A)$ denotes all the model parameters, with θ_R corresponds to the regression model and θ_A corresponds to the latent space network model. We assume the missing data is at random throughout. Under the proposed model, the likelihood could be decomposed into two parts

$$\begin{aligned} p(x^{mix}|x^{obs}, A, \Theta) &\propto p(x^{mis}, A|x^{obs}, \Theta) \\ &= f(x^{mix}|x^{obs}, \theta_R)p(A|x^{mis}, x^{obs}, \theta_A) \end{aligned} \tag{B.1}$$

where f is the joint distribution of x. With a prior $\pi(\Theta)$, the posterior predictive distribution is

$$p(x^{mis}|x^{obs}, A) = \int_{\Theta} p(x^{mix}|x^{obs}, A, \Theta)p(\Theta|x^{obs}, A)d\Theta \tag{B.2}$$

where $p(\Theta|x^{obs}, A) \propto \pi(\Theta)p(x^{obs}, A|\Theta)$. A standard way to draw samples from the posterior predictive distribution is to use Gibbs sampler with data augmentation strategy that iteratively draw Θ and X^{mis}. Under standard regularity conditions, the Markov chain is ergodic and has limiting distribution $p(x^{mis}, \Theta|x^{obs}, A)$ (**).

We modify the Gibbs sampling procedure in order to compare to the iterative imputation framework. Let $x^{(k-1)}$ and be the entire dataset with both observed and imputed values, and $\Theta^{(k-1)}$ be the parameter estimates, at iteration $k-1$. At iteration k, the Gibbs chain evolves as follows

- Set $x \leftarrow x^{(k-1)}$ and update the variables of x one at a time.
- For $i = 1, ..., p$, draw $\theta_R \sim p(\theta_R|x_i^{obs}, x_{-i})$ and $x_i^{mis} \sim p(x_i^{mis}|x_i^{obs}, x_{-i}, A, \theta_R, \theta_A^{(k-1)})$
- Draw $\theta_A \sim p(\theta_A|x, A)$
- Set $x^{(k)} \leftarrow x$ and $\Theta^k \leftarrow (\theta_R, \theta_A)$

Under regularity conditions (***), the Markov chain converges to the posterior distribution of the corresponding model.

For iterative imputation, the user specifies p conditional regression models, denoted as $g_i(x_i|x_{-i}, \theta_i)$, with θ_i being the corresponding parameters with prior $\pi_i(\theta_i), i = 1, ..., p$. The iterative imputation scheme can be described as follows.

- Set $x \leftarrow x^{(k-1)}$ and update the variables of x one at a time.

- For $i = 1, ..., p$, draw $\theta_i \sim p_i(\theta_i|x_i^{obs}, x_{-i})$, which is the posterior distribution of θ_i with g_i and π_i and $x_i^{mis} \sim p_i(x_i^{mis}|x_i^{obs}, x_{-i}, A, \theta_i, \theta_A^{(k-1)})$
- Draw $\theta_A \sim p(\theta_A|x, A)$
- Set $x^{(k)} \leftarrow x$ and $\Theta^k \leftarrow (\theta_R, \theta_A)$

Notice that under the proposed framework, similar to (B.1),

$$p_i(x_i^{mis}|x_i^{obs}, x_{-i}, A, \theta_i, \theta_A^{(k-1)}) \propto g_i(x_i^{mis}|x_i^{obs}, x_{-i}, \theta_i)p(A|x_i^{mis}, x_i^{obs}, x_{-i}, \theta_A^{(k-1)})$$

We first consider when the specified conditional regression models g_i are compatible with f. A set of condition models $g_i(x_i|x_{-i}, \theta_i), \theta_i \in \Theta_i$ is said to be compatible with $f(x|\theta), \theta \in \Theta$ if for all i, there exist a collection of surjective maps $t_i : \Theta \to \Theta_i$ such that there exists $\theta \in \Theta$ with $g_i(x_i|x_{-i}, \theta_i) = f(x_i|x_{-i}, t_i(\theta_R))$.

When g_i and f is compatible, the difference of the Gibbs sampling scheme and the iterative imputation scheme lies in the step of drawing parameters $\theta_R \sim p(\theta_R|x_i^{obs}, x_{-i})$ and $\theta_i \sim p_i(\theta_i|x_i^{obs}, x_{-i})$ as the distribution of the missing data given the parameters are the same under the joint model f and iterative imputation models g_i. The following results follow the same line of work by (*Liu et al.*, 2013). We here restate the important steps to accommodate the results to our setting for completeness.

As we are assuming compatibility, we may drop the notation g_i and use the unified notation f. Specifically, we denote $f(x_i|x_{-i}, \theta_i) = f(x_i|x_{-i}, \theta_R)$ for $t_i(\theta_R) = \theta_i$.

To compare the posterior distribution of θ_R and θ_i, the first difference we should notice is that the dimension of θ_R is higher. θ_R contains parameters describing both the conditional distribution $x_i|x_{-i}$ and the marginal distribution of x_{-i}. Thus we augment the parameter space of iterative distribution to (θ_i, θ_i^*) with $\theta_i^* = t_i^*(\theta_R)$, and

$T_i(\theta_R) = \{\theta_i, \theta_i^*\}$ is an invertible map. With this augmentation, the prior distribution π on θ_R for the Bayesian model is equivalent to a prior on (θ_i, θ_i^*) with the following form.

$$\pi_i^*(\theta_i, \theta_i^*) = \det(\partial T_i/\partial_{\theta_R})^{-1}\pi(T_i^{-1}(\theta_i, \theta_i^*))$$

The posterior distribution for θ_i under the Bayesian model is

$$p(\theta_i|x_i^{obs}, x_{-i}) = \int p(\theta_i, \theta_i^*|x_i^{obs}, x_{-i})d\theta_i^* \propto \int f(x_i^{obs}, x_{-i}|\theta_i, \theta_i^*)\pi_i^*(\theta_i, \theta_i^*)d\theta_i^*.$$

Since $f(x_i^{obs}|x_{-i}, \theta_i, \theta_i^*) = f(x_i^{obs}|x_{-i}, \theta_i)$, we can further reduce the posterior distribution to the following

$$p(\theta_i|x_i^{obs}, x_{-i}) \propto f(x_i^{obs}|x_{-i}, \theta_i) \int f(x_{-i}|\theta_i, \theta_i^*)\pi_i(\theta_i, \theta_i^*)d\theta_i^*.$$

Denote the integral in the above formula as $\pi_{i,x_{-i}}(\theta_i)$, we have

$$p(\theta_i|x_i^{obs}, x_{-i}) \propto f(x_i^{obs}|x_{-i}, \theta_i)\pi_{i,x_{-i}}(\theta_i).$$

Recall for the iterative imputation, we have

$$p_i(\theta_i|x_i^{obs}, x_{-i}) \propto g_i(x_i^{obs}|x_{-i}, \theta_i)\pi_i(\theta_i) = f(x_i^{obs}|x_{-i}, \theta_i)\pi_i(\theta_i).$$

The difference of the posterior depends only on the difference between the prior distributions $\pi_{i,x_{-i}}$ and π_i.

Lemma B.1. *Let n be the sample size, let $f_X(\theta)$ and $g_X(\theta)$ that shares the same likelihood but with different prior π_g and π_f. Let $L(\theta) = \pi_g/\pi_f$, $r(\theta) = \frac{g_X(\theta)}{f_X(\theta)} = \frac{L(\theta)}{\int L(\theta')f_X(\theta')d\theta'}$.*

Let $\partial L(\theta)$ be the partial derivative with respect to θ and let ξ be a random variable such that

$$L(\theta) = L(\mu_\theta) + \partial L(\xi)^T(\theta - \mu_\theta)$$

where $\mu_\theta = \int \theta f_X(\theta) d\theta$. If there exists a random variable $Z(\theta)$ with finite variance under f_X such that

$$\left|n^{1/2}\partial L(\xi)^T(\theta - \mu_\theta)\right| \leq |\partial L(\mu_\theta)| Z(\theta)$$

then there exists a constant $\kappa > 0$ such that for n sufficiently large,

$$\left\|\tilde{f}_X - \tilde{g}_X\right\|_1 \leq \frac{\kappa |\partial \log L(\mu_\theta)|^{1/2}}{n^{1/4}}$$

The lemma states that when the ratio between the priors satisfies the condition, the difference between the corresponding posterior predictive distribution vanishes as n grows. The condition is satisfied for most parametric models on the following set B_n. Let $\hat{\theta}(x)$ the complete-data maximum likelihood estimator and let $B_n = \{x : |\hat{\theta}(x)| \leq \gamma\}$. Specifically, it states that we are only interested in the area where the observed data and the imputation is "valid" such that the MLE exists.

Thus we have shown that the transition kernels Gibbs chain and the iterative imputation chain are close on the region B_n. The subsequent step is to show that conditioning on the set B_n, the stationary distributions $\tilde{\nu}_i^{X_{obs}}$ for the conditional processes are close to that of the original processes $\nu_i^{X_{obs}}$. Also, $\tilde{\nu}_1^{X_{obs}}$ and $\tilde{\nu}_2^{X_{obs}}$ are close in total variation and thus are $\nu_1^{X_{obs}}$ and $\nu_2^{X_{obs}}$.

We consider the chains conditional on the set B_n where the two transition kernels are

close to each other. In particular, for any set C, let

$$\tilde{K}_i(w, C) = \frac{K_i(w, C \cap B_n)}{K_i(w, B_n)}$$

By this, we restrict the update of the missing data to B_n. Next Lemma B.2 shows that the stationary distribution of the original chain and this corresponding conditional chain are close.

Lemma B.2. *Let B_n be in the form $B_n = \{x : |\hat{\theta}(x)| \leq \gamma\}$ and we pick γ sufficiently large such that $\nu_i^{X_{obs}}(B_n) \to 1$ in probability as $n \to \infty$. Let the defined conditional chains following $\tilde{K}_i$ has invariant distribution $\tilde{\nu}_i$, then*

$$\lim_{n\to\infty} d_{TV}\left(\nu_i^{X^{obs}}, \tilde{\nu}_i^{Xobs}\right) = 0.$$

Then $\|K_1(w, \cdot) - K_2(w, \cdot)\|_1$vanishes uniformly for $w \in B_n$, this implies

$$\lim_{n\to\infty} \left\| \tilde{K}_1(w, \cdot), \tilde{K}_2(w, \cdot) \right\|_1 = 0 \text{ uniformly for } w \in B_n$$

. Then we need to show that $d_{\mathrm{TV}}\left(\tilde{\nu}_1^{X^{\mathrm{obs}}}, \tilde{\nu}_2^{X^{\mathrm{obs}}}\right) \to 0$.

Lemma B.3. *With*

$$\lim_{n\to\infty} \left\| \tilde{K}_1(w, \cdot), \tilde{K}_2(w, \cdot) \right\|_1 = 0 \text{ uniformly for } w \in B_n$$

holds and suppose there exists a monotone decreasing sequence $r_t \to 0$ and a data dependent starting distribution ν such that

$$\mathrm{pr}\left\{ \left\| \tilde{K}_i^{(t)}(\nu, \cdot) - \tilde{\nu}_i^{X^{obs}}(\cdot) \right\|_1 \leq r_t, \text{ for all } t > 0 \right\} \to 1, \quad n \to \infty$$

Then

$$\left\|\tilde{\nu}_1^{X^{obs}} - \tilde{\nu}_2^{X^{obs}}\right\|_1 \to 0, \ \textit{in probability as } n \to \infty$$

The required condition can be established with a set of sufficient conditions according to (*Rosenthal*, 1995): $\tilde{K}_1$ and $\tilde{K}_2$ admits a common small set C and each admits their own drift function on C. A Gibbs chain admits a small set C and a drift function V means that $\tilde{K}_1(\omega, A) \geq q_1\mu_1(A)$ for some positive μ_1 with $\omega \in C$, $q_1 \in (0,1)$; and for some $\lambda_1 \in (0,1)$ an for all $\omega \notin C$,

$$\lambda_1 V(w) \geq \int V(w')\, \tilde{K}_1(w, \mathrm{d}w')$$

. Then Given $\tilde{K}_1$ and $\tilde{K}_2$ are close, the set C is also a small set for $\tilde{K}_2$. Proposition 2 in (*Liu et al.*, 2013) showed a weak conditions under which V is also a drift function for $\tilde{K}_2$ and thus the condition for lemma B.3 can be established.

Last, we summarize the results as follows: assuming the compatibility of the models g_i with f, the Gibbs chain and the iterative imputation chain are constructed. Lemma 1 provide conditions under which the distance between their posterior predictive distributions vanish. On the set B_n where the MLE is bounded, the distance between the condition kernels $\tilde{K}_1$ and $\tilde{K}_2$ vanishes with the condition in lemma B.1. Then with conditions in lemma B.3, we can show that the stationary distribution of the conditional chains are close. Last, by lemma B.2, the stationary distribution is close to the stationary distribution of the original chain. Combing the components, we can conclude that the stationary distribution of the Gibbs chain and the iterative imputation chain are close.

APPENDIX C

Appendix for Chapter IV

C.1 Proof of Proposition IV.4 and corollary IV.5

We want to show that the network generated by the proposed model is asymptotically sparse.

$$P(Y_{n,j} = i|Y_1, Y_2, ..., Y_{n-1}, Y_{n,1}, ...Y_{n,j-1}) \propto \begin{cases} D_{n,j}(i) - \alpha + e^{\beta X_i}, i = 1, 2, ..., V_n(j) \\ \theta + \alpha V_n(j) + \frac{\sum_{k=1}^{V_n(j)} e^{\beta X_k}}{V_n(j)}, i = V_n(j) + 1 \end{cases}.$$

Let T_i be the degrees needed to the generation of next node, starting with 1 node in the network, so that T_1 is the degree needed to the second node. Let $r_i = exp(\beta X_i)$, and we assume $r_i \in (\delta, C)$ with $\delta > 0$ and $C \leq \gamma\delta$ for constant δ, γ, C. Then we can write

$$P(T_1 = 1) = \frac{\theta + \alpha + r_1}{1 + \theta + 2r_1}$$

$$P(T_1 = k) = \frac{\theta + \alpha + r_1}{k + \theta + 2r_1}(1 - \frac{\theta + \alpha + r_1}{k - 1 + \theta + 2r_1})...(1 - \frac{\theta + \alpha + r_1}{1 + \theta + 2r_1})$$

Similarly we will have

$$P(T_2 = 1|T_1) = \frac{\theta + 2\alpha + \frac{r_1+r_2}{2}}{T_1 + 1 + \theta + \frac{3(r_1+r_2)}{2}}$$

and etc..

Let $S_n = T_1 + T_2 + ... + T_n + 1$ we can write

$$P(T_n = 1|S_{n-1}) = \frac{\theta + n\alpha + \sum_{i=1}^{n} r_i/n}{S_{n-1} + \theta + \frac{n+1}{n}\sum_{i=1}^{n} r_i}$$

When $n > \gamma$, we have

$$P(T_n = 1|S_{n-1}) \leq \frac{(\theta + C) + n\alpha}{S_{n-1} + (\theta + C)} =: P(T_n^{CRP,\alpha} = 1|S_{n-1})$$

, and similarly for $P(T_n = k|S_{n-1})$, where $T_n^{CRP,\alpha}$ can be viewed as the degrees needed to the next node for a standard Chinese Restaurant Process (CRP) with parameter α and $\theta + C$. By standard results of CRP, $S_n^{CRP,\alpha} \sim (\frac{n}{S_\alpha})^{\frac{1}{\alpha}}$ almost surely, where S_α is a positive and finite random variable(*Pitman*, 2006). Then we have that $T_n|S_{n-1} \succ T_n^{CRP,\alpha}|S_{n-1}$ for any n, where $\succ$ denotes stochastic dominance. Thus we will have $S_n \succ S_n^{CRP,\alpha}$ as $n \to \infty$.

Next,

$$P(T_n = 1|S_{n-1}) \geq \frac{\theta + n\alpha}{S_{n-1} + \theta + C + nC} =: P(T_n^{mCRP,\alpha} = 1|S_{n-1})$$

, and similarly for $P(T_n = k|S_{n-1})$ where $T_n^{mCRP,\alpha}$ is defined for a modified Chinese Restaurant process as follows, when there are $(n-1)$ existing nodes, select the next

node with the following probability

$$\propto \begin{cases} D(i) - \alpha + C, i = 1, 2, ..., n-1 \\ \theta + \alpha(n-1), i = n \end{cases} . \tag{C.1}$$

We can then define $S_n^{mCRP,\alpha}$ and we will have $S_n \prec S_n^{mCRP,\alpha}$. Then we look at the upper bound of $S_n^{mCRP,\alpha}$ for the modified CRP. Notice for a standard CRP with parameter α, θ, $P(T_n^{CRP,\alpha} = 1|S_{n-1}) = \frac{\theta+n\alpha}{S_{n-1}+\theta}$, it could be seen that

$$E(S_1^{mCRP,\alpha}) - E(S_1^{CRP,\alpha}) = E(T_1^{mCRP,\alpha}) - E(T_1^{CRP,\alpha}) = \frac{2C}{\alpha+\theta}$$

.

Then since

$$E(S_2^{mCRP,\alpha}) = E(E(S_2^{mCRP,\alpha}|S_1^{mCRP,\alpha})$$

and

$$E(S_2^{CRP,\alpha}) = E(E(S_2^{CRP,\alpha}|S_1^{CRP,\alpha})$$

, we have

$$E(S_2^{mCRP,\alpha}) - E(S_2^{CRP,\alpha}) = \frac{2C}{(\alpha+\theta)(2\alpha+\theta)} + \frac{3C}{(2\alpha+\theta)}$$

.

Then by deduction, we will have

$$E(S_n^{mCRP,\alpha}) - E(S_n^{CRP,\alpha}) = \sum_{i=1}^{n} \frac{(i+1)C}{\prod_{j=i}^{n}(\theta+j\alpha)}$$

.

Now, let $N^*, M^* > 2N^*$ be positive integer constant such that $\theta + N^*\alpha > 1$ and $(\theta+\alpha)(\theta+M^*\alpha) > 1$.

Then we shall have when $n > N^*$,

$$(E(S_n^{mCRP,\alpha}) - E(S_n^{CRP,\alpha})) - (E(S_{n-1}^{mCRP,\alpha}) - E(S_{n-1}^{CRP,\alpha})) \leq \frac{(N^*+1)C}{N^*\alpha}$$

and thus when $n > M*$,

$$(E(S_n^{mCRP,\alpha}) - E(S_n^{CRP,\alpha})) \leq \frac{2Cn}{\alpha}$$

Then since we know $S_n^{CRP,\alpha} \sim (\frac{n}{S_\alpha})^{\frac{1}{\alpha}}$, when $n > M^*$ by Markov inequality

$$P(S_n^{CRP,\alpha} > (\frac{n}{S_{\alpha-\epsilon}})^{\frac{1}{\alpha-\epsilon}}) \leq \frac{E(S_n^{CRP,\alpha}) + 2Cn/\alpha}{(\frac{n}{S_{\alpha-\epsilon}})^{\frac{1}{\alpha-\epsilon}}} \to 0 \tag{C.2}$$

Where ϵ is arbitrary small positive constant. Now for a CRP with parameter $\alpha - \epsilon$, we will have

$$P(T_n^{CRP,\alpha-\epsilon} = 1 | S_{n-1}^{CRP,\alpha-\epsilon}) = \frac{\theta + (n-1)(\alpha-\epsilon)}{S_{n-1}^{CRP,\alpha-\epsilon} + \theta}$$

and since (C.2), we have as $n \to \infty$ $S_{N-1}^{CRP,\alpha-\epsilon} \geq S^{mCRP,\alpha}$ in probability. Then we have

$$P(T_n^{CRP,\alpha-\epsilon} = 1 | S_{n-1}^{CRP,\alpha-\epsilon}) \leq \frac{\theta + (n-1)(\alpha-\epsilon)}{S_{n-1} + nC} =: P(T_n = 1 | S_{n-1})$$

And we can then establish $T_n | S_{n-1} \succ T_n^{CRP,\alpha-\epsilon} | S_{n-1}$ for n large enough and thus $S_n \prec S_n^{CRP,\alpha-\epsilon}$. Then by standard result of CRP, we have $(\frac{n}{S_\alpha})^{\frac{1}{\alpha}} \prec S_n \prec (\frac{n}{S_{\alpha-\epsilon}})^{\frac{1}{\alpha-\epsilon}}$. Then we have $S_n = O(\frac{n}{S_\alpha})^{\frac{1}{\alpha}}$, as $n \to \infty$.

Using the following definition of sparsity from *Crane and Dempsey* (2018): Let $(\mathcal{E}_n)_{n\geq 1}$ be a sequence of edge-labeled networks for which $e(\mathcal{E}_n) \to \infty$ as $n \to \infty$. The sequence $(\mathcal{E}_n)_{n\geq 1}$ is sparse if

$$\limsup_{n\to\infty} \frac{e(\mathcal{E}_n)}{v(\mathcal{E}_n)^{m_*(\mathcal{E}_n)}} = 0$$

where $m_*(\mathcal{E}_n) = e(\mathcal{E}_n)^{-1} \sum_{k\geq 1} kM_k(\mathcal{E}_n)$ is the average arity of the edges in $\mathcal{E}_n$. Following a similar argument to Theorem 4.3 in (*Crane and Dempsey*, 2018), the network is sparse whenever $\mu\alpha > 1$. Where μ is expected number of nodes in an edge. For completeness, we restate the argument in the following.

We already established that $(\frac{n}{S_\alpha})^{\frac{1}{\alpha}} \prec S_n \prec (\frac{n}{S_{\alpha-\epsilon}})^{\frac{1}{\alpha-\epsilon}}$. Reversing the relationship of number of nodes n and total degree S_n, let m be the total degree and N_m be the number of nodes until we generated m unary edges, we shall obtain the result that $m^{\alpha-\epsilon} \prec N_m \prec m^\alpha$ as $n \to \infty$. Then for arbitrary distribution ν with mean μ on the number of nodes in each edge, the number of nodes $v(\mathcal{E}_n)$ is a random subsequence of (N_m) with indices $k_1, k_2, ...$ where $k_s = \sum_{i=1}^s \kappa_i$ and κ_i i.i.d from ν. Then $v(\mathcal{E}_n) = N_{k_s}$ is bounded by $k_s^\alpha S_\alpha$ and $k_s^{\alpha-\epsilon} S_{\alpha-\epsilon}$ as $n \to \infty$. Then as $\frac{k_s}{s} \to \mu$ almost surely by law of large numbers, we shall have $(\mu n)^{\alpha-\epsilon} \prec N_{k_s} \prec (\mu n)^\alpha$. Then in order to have sparsity, we need $\liminf_{n\to\infty} \frac{1}{n} v(\mathcal{E}_n)^{m_*(\mathcal{E}_n)} = \infty$, then since $m_*(\mathcal{E}_n) \to \mu$ we will have $\frac{1}{n}(\mu n)^{\mu(\alpha-\epsilon)} S_{\alpha-\epsilon} \prec \frac{1}{n} v(\mathcal{E}_n)^{m_*(\mathcal{E}_n)} \prec \frac{1}{n}(\mu n)^{\mu\alpha} S_\alpha$. Thus $\mathcal{E}_n$ is sparse when $\mu\alpha > 1$ as ϵ is arbitrary small.

C.2 Additional simulation results for the estimation in Chapter IV

By comparing the simulation results for $\alpha = 0.3, 0.5, 0.7$ and $\theta = 0, 10$, we can conclude that it is more difficult to obtain accurate estimates and valid standard error estimates for β from the observed fisher information when α is small and β is small. One explanation for this phenomena is that when α is small and β is small, there number of new nodes joining the network is small, and is thus harder to estimate the covariates effects.

	mean(n_{nodes})	mean($\hat{\alpha}$)	mean($\hat{\sigma}_{\hat{\alpha}}$)	coverage	invalid
2500	375.6	0.685	0.061	0.950	0.000
5000	585.5	0.688	0.042	1.000	0.000
10000	925.1	0.696	0.030	1.000	0.000
	mean(n_{nodes})	mean($\hat{\beta}_1$)	mean($\hat{\sigma}_{\hat{\beta}_1}$)	coverage	invalid
2500	375.6	1.013	0.178	0.975	0.000
5000	585.5	1.007	0.132	0.950	0.000
10000	925.1	1.008	0.099	0.975	0.000
	mean(n_{nodes})	mean($\hat{\beta}_3$)	mean($\hat{\sigma}_{\hat{\beta}_3}$)	coverage	invalid
2500	375.6	-1.014	0.276	1.000	0.000
5000	585.5	-1.017	0.194	1.000	0.000
10000	925.1	-1.014	0.143	1.000	0.000
	mean(n_{nodes})	mean($\hat{\theta}$)	mean($\hat{\sigma}_{\hat{\theta}}$)	coverage	invalid
2500	375.6	12.208	5.965	1.000	0.000
5000	585.5	12.188	5.369	1.000	0.000
10000	925.1	11.800	4.944	1.000	0.000

Table C.1: Estimation when true $\alpha = 0.7, \theta = 10$

	mean(n_{nodes})	mean($\hat{\alpha}$)	mean($\hat{\sigma}_{\hat{\alpha}}$)	coverage	invalid
2500	85.5	0.446	0.088	0.900	0.050
5000	121.3	0.465	0.068	0.900	0.025
10000	174.6	0.484	0.051	0.925	0.000
	mean(n_{nodes})	mean($\hat{\beta}_1$)	mean($\hat{\sigma}_{\hat{\beta}_1}$)	coverage	invalid
2500	85.5	1.019	0.473	0.725	0.225
5000	121.3	0.990	0.372	0.825	0.150
10000	174.6	1.005	0.262	0.900	0.075
	mean(n_{nodes})	mean($\hat{\beta}_3$)	mean($\hat{\sigma}_{\hat{\beta}_3}$)	coverage	invalid
2500	85.5	-1.048	0.706	0.775	0.225
5000	121.3	-1.014	0.585	0.800	0.175
10000	174.6	-0.998	0.368	0.925	0.025
	mean(n_{nodes})	mean($\hat{\theta}$)	mean($\hat{\sigma}_{\hat{\theta}}$)	coverage	invalid
2500	85.5	1.251	1.957	1.000	0.000
5000	121.3	1.006	1.752	1.000	0.000
10000	174.6	0.750	1.578	1.000	0.000

Table C.2: Estimation when true $\alpha = 0.5, \theta = 0$

	mean(n_{nodes})	mean($\hat{\alpha}$)	mean($\hat{\sigma}_{\hat{\alpha}}$)	coverage	invalid
2500	38.2	0.231	0.109	0.925	0.025
5000	47.8	0.243	0.085	0.950	0.000
10000	60.9	0.267	0.071	0.950	0.000
	mean(n_{nodes})	mean($\hat{\beta}_1$)	mean($\hat{\sigma}_{\hat{\beta}_1}$)	coverage	invalid
2500	38.2	0.970	1.453	0.500	0.450
5000	47.8	1.041	0.502	0.525	0.425
10000	60.9	1.004	0.727	0.625	0.350
	mean(n_{nodes})	mean($\hat{\beta}_3$)	mean($\hat{\sigma}_{\hat{\beta}_3}$)	coverage	invalid
2500	38.2	-0.930	1.676	0.450	0.450
5000	47.8	-0.987	0.715	0.350	0.550
10000	60.9	-0.942	0.854	0.525	0.450
	mean(n_{nodes})	mean($\hat{\theta}$)	mean($\hat{\sigma}_{\hat{\theta}}$)	coverage	invalid
2500	38.2	1.030	1.663	0.975	0.025
5000	47.8	0.938	1.499	1.000	0.000
10000	60.9	0.710	1.379	1.000	0.000

Table C.3: Estimation when true $\alpha = 0.3, \theta = 0$

	mean(n_{nodes})	mean($\hat{\alpha}$)	mean($\hat{\sigma}_{\hat{\alpha}}$)	coverage	invalid
2500	193.4	0.665	0.072	0.900	0.050
5000	305.8	0.676	0.050	0.900	0.025
10000	485.9	0.685	0.037	0.875	0.000
	mean(n_{nodes})	mean($\hat{\beta}_1$)	mean($\hat{\sigma}_{\hat{\beta}_1}$)	coverage	invalid
2500	193.4	1.018	0.268	0.925	0.075
5000	305.8	1.007	0.199	0.950	0.025
10000	485.9	1.016	0.143	1.000	0.000
	mean(n_{nodes})	mean($\hat{\beta}_3$)	mean($\hat{\sigma}_{\hat{\beta}_3}$)	coverage	invalid
2500	193.4	-1.021	0.382	0.900	0.075
5000	305.8	-1.006	0.265	0.975	0.025
10000	485.9	-1.007	0.188	1.000	0.000
	mean(n_{nodes})	mean($\hat{\theta}$)	mean($\hat{\sigma}_{\hat{\theta}}$)	coverage	invalid
2500	193.4	1.521	2.290	0.950	0.050
5000	305.8	1.372	2.057	1.000	0.000
10000	485.9	1.256	1.942	1.000	0.000

Table C.4: Estimation when true $\alpha = 0.7, \theta = 0$

BIBLIOGRAPHY

BIBLIOGRAPHY

Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing (2008), Mixed membership stochastic blockmodels, *Journal of Machine Learning Research*, *9*(Sep), 1981–2014.

Airoldi, E. M., T. B. Costa, and S. H. Chan (2013), Stochastic blockmodel approximation of a graphon: Theory and consistent estimation, in *Advances in Neural Information Processing Systems*, pp. 692–700.

Amini, A. A., and E. Levina (2014), On semidefinite relaxations for the block model, *arXiv preprint arXiv:1406.5647*.

Amini, A. A., A. Chen, P. J. Bickel, and E. Levina (2013), Pseudo-likelihood methods for community detection in large sparse networks, *The Annals of Statistics*, *41*(4), 2097–2122.

Andridge, R. R., and R. J. Little (2010), A review of hot deck imputation for survey non-response, *International statistical review*, *78*(1), 40–64.

Asur, S., and B. A. Huberman (2010), Predicting the future with social media, in *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, vol. 1, pp. 492–499, IEEE.

Athreya, A., C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman (2016), A limit theorem for scaled eigenvectors of random dot product graphs, *Sankhya A*, *78*(1), 1–18.

Athreya, A., D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin (2017), Statistical inference on random dot product graphs: a survey, *The Journal of Machine Learning Research*, *18*(1), 8393–8484.

Barabási, A.-L., and R. Albert (1999), Emergence of scaling in random networks, *science*, *286*(5439), 509–512.

Basse, G. W., and E. M. Airoldi (2015), Optimal design of experiments in the presence of network-correlated outcomes, *ArXiv e-prints*.

Bickel, P., D. Choi, X. Chang, and H. Zhang (2013), Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels, *The Annals of Statistics*, *41*(4), 1922–1943.

Bickel, P. J., and A. Chen (2009), A nonparametric view of network models and newman–girvan and other modularities, *Proceedings of the National Academy of Sciences*, *106*(50), 21,068–21,073.

Bickel, P. J., A. Chen, Y. Zhao, E. Levina, and J. Zhu (2015), Correction to the proof of consistency of community detection, *The Annals of Statistics*, *43*(1), 462–466.

Binkiewicz, N., J. T. Vogelstein, and K. Rohe (2014), Covariate-assisted spectral clustering, *arXiv preprint arXiv:1411.2158*.

Binkiewicz, N., J. T. Vogelstein, and K. Rohe (2017), Covariate-assisted spectral clustering, *Biometrika*, *104*(2), 361–377.

Burgette, L. F., and J. P. Reiter (2010), Multiple imputation for missing data via sequential regression trees, *American journal of epidemiology*, *172*(9), 1070–1076.

Buuren, S. v., and K. Groothuis-Oudshoorn (2010), mice: Multivariate imputation by chained equations in r, *Journal of statistical software*, pp. 1–68.

Cai, D., T. Campbell, and T. Broderick (2016), Edge-exchangeable graphs and sparsity, in *Advances in Neural Information Processing Systems*, pp. 4249–4257.

Cario, M. C., and B. L. Nelson (1997), Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix, *Tech. rep.*, Citeseer.

Celisse, A., J.-J. Daudin, and L. Pierre (2012), Consistency of maximum-likelihood and variational estimators in the stochastic block model, *Electronic Journal of Statistics*, *6*, 1847–1899.

Chakrabarti, D., S. Funiak, J. Chang, and S. A. Macskassy (2017), Joint label inference in networks, *The Journal of Machine Learning Research*, *18*(1), 1941–1979.

Choi, D., P. J. Wolfe, et al. (2014), Co-clustering separately exchangeable network data, *The Annals of Statistics*, *42*(1), 29–63.

Christakis, N. A., and J. H. Fowler (2007), The spread of obesity in a large social network over 32 years, *New England journal of medicine*, *357*(4), 370–379.

Crane, H., and W. Dempsey (2018), Edge exchangeable models for interaction networks, *Journal of the American Statistical Association*, *113*(523), 1311–1326.

Fortunato, S. (2010), Community detection in graphs, *Physics reports*, *486*(3-5), 75–174.

Fujimoto, K., and T. W. Valente (2012), Social network influences on adolescent substance use: disentangling structural equivalence from cohesion, *Social Science & Medicine*, *74*(12), 1952–1960.

Geman, S., and D. Geman (1984), Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721–741.

Gill, P., J. Lee, K. R. Rethemeyer, J. Horgan, and V. Asal (2014), Lethal connections: The determinants of network connections in the provisional irish republican army, 1970–1998, *International Interactions*, *40*(1), 52–78.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, E. M. Airoldi, et al. (2010), A survey of statistical network models, *Foundations and Trends® in Machine Learning*, *2*(2), 129–233.

Hand, D. J., and R. J. Till (2001), A simple generalisation of the area under the roc curve for multiple class classification problems, *Machine learning*, *45*(2), 171–186.

Hoff, P. D. (2005), Bilinear mixed-effects models for dyadic data, *Journal of the american Statistical association*, *100*(469), 286–295.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002), Latent space approaches to social network analysis, *Journal of the american Statistical association*, *97*(460), 1090–1098.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983), Stochastic blockmodels: First steps, *Social networks*, *5*(2), 109–137.

Jeong, H., Z. Néda, and A.-L. Barabási (2003), Measuring preferential attachment in evolving networks, *EPL (Europhysics Letters)*, *61*(4), 567.

Jin, J. (2015), Fast community detection by score, *The Annals of Statistics*, *43*(1), 57–89.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999), An introduction to variational methods for graphical models, *Machine learning*, *37*(2), 183–233.

Kao, E. K. (2017), Causal inference under network interference: A framework for experiments on social networks, *arXiv preprint arXiv:1708.08522*.

Karrer, B., and M. E. J. Newman (2011), Stochastic blockmodels and community structure in networks, *Phys. Rev. E*, *83*, 016,107, doi:10.1103/PhysRevE.83.016107.

Lazega, E. (2001), *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*, Oxford University Press on Demand.

Lei, J., A. Rinaldo, et al. (2015), Consistency of spectral clustering in stochastic block models, *The Annals of Statistics*, *43*(1), 215–237.

Leskovec, J., and J. J. Mcauley (2012), Learning to discover social circles in ego networks, in *Advances in neural information processing systems*, pp. 539–547.

Leskovec, J., K. J. Lang, A. Dasgupta, and M. W. Mahoney (2008), Statistical properties of community structure in large social and information networks, in *Proceedings of the 17th international conference on World Wide Web*, pp. 695–704.

Li, T., E. Levina, J. Zhu, et al. (2019), Prediction models for network-linked data, *The Annals of Applied Statistics*, *13*(1), 132–164.

Liu, J., A. Gelman, J. Hill, Y.-S. Su, and J. Kropko (2013), On the stationary distribution of iterative imputations, *Biometrika*, *101*(1), 155–173.

Lloyd, J., P. Orbanz, Z. Ghahramani, and D. M. Roy (2012), Random function priors for exchangeable arrays with applications to graphs and relational data, in *Advances in Neural Information Processing Systems*, pp. 998–1006.

Ma, Z., and Z. Ma (2017), Exploration of large networks with covariates via fast and universal latent space model fitting, *arXiv preprint arXiv:1705.02372*.

Manski, C. F. (2013), Identification of treatment response with social interactions, *The Econometrics Journal*, *16*(1), S1–S23.

Mariadassou, M., and C. Matias (2015), Convergence of the groups posterior distribution in latent or stochastic block models, *Bernoulli*, *21*(1), 537–573.

Mariadassou, M., S. Robin, C. Vacher, et al. (2010), Uncovering latent structure in valued graphs: a variational approach, *The Annals of Applied Statistics*, *4*(2), 715–742.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001), Birds of a feather: Homophily in social networks, *Annual review of sociology*, *27*(1), 415–444.

Murray, J. S., et al. (2018), Multiple imputation: a review of practical and theoretical findings, *Statistical Science*, *33*(2), 142–159.

Newman, M. E. (2001), Clustering and preferential attachment in growing networks, *Physical review E*, *64*(2), 025,102.

Newman, M. E. (2006), Modularity and community structure in networks, *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Newman, M. E., and A. Clauset (2016), Structure and inference in annotated networks, *Nature Communications*, *7*.

Nowicki, K., and T. A. B. Snijders (2001), Estimation and prediction for stochastic blockstructures, *Journal of the American Statistical Association*, *96*(455), 1077–1087.

Opsahl, T., and P. Panzarasa (2009), Clustering in weighted networks, *Social networks*, *31*(2), 155–163.

Pitman, J. (2006), *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*, Springer.

Qin, T., and K. Rohe (2013), Regularized spectral clustering under the degree-corrected stochastic blockmodel, in *Advances in Neural Information Processing Systems*, pp. 3120–3128.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher (2007), An introduction to exponential random graph (p*) models for social networks, *Social networks*, *29*(2), 173–191.

Rohe, K., S. Chatterjee, and B. Yu (2011), Spectral clustering and the high-dimensional stochastic blockmodel, *The Annals of Statistics*, pp. 1878–1915.

Rosenthal, J. S. (1995), Minorization conditions and convergence rates for markov chain monte carlo, *Journal of the American Statistical Association*, *90*(430), 558–566.

Shalizi, C. R., and A. C. Thomas (2011), Homophily and contagion are generically confounded in observational social network studies, *Sociological methods & research*, *40*(2), 211–239.

Stekhoven, D. J., and P. Bühlmann (2011), Missforest—non-parametric missing value imputation for mixed-type data, *Bioinformatics*, *28*(1), 112–118.

Sweet, T. M. (2015), Incorporating covariates into stochastic blockmodels, *Journal of Educational and Behavioral Statistics*, *40*(6), 635–664.

Van Buuren, S. (2007), Multiple imputation of discrete and continuous data by fully conditional specification, *Statistical methods in medical research*, *16*(3), 219–242.

Van Buuren, S. (2018), *Flexible imputation of missing data*, Chapman and Hall/CRC.

Van de Bunt, G. G., M. A. Van Duijn, and T. A. Snijders (1999), Friendship networks through time: An actor-oriented dynamic statistical network model, *Computational & Mathematical Organization Theory*, *5*(2), 167–192.

Vázquez, A. (2003), Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations, *Physical Review E*, *67*(5), 056,104.

Veldt, N., D. F. Gleich, and A. Wirth (2018), A correlation clustering framework for community detection, in *Proceedings of the 2018 World Wide Web Conference*, pp. 439–448.

Wan, P., T. Wang, R. A. Davis, S. I. Resnick, et al. (2017), Fitting the linear preferential attachment model, *Electronic Journal of Statistics*, *11*(2), 3738–3780.

Wang, F., T. Li, X. Wang, S. Zhu, and C. Ding (2011), Community discovery using nonnegative matrix factorization, *Data Mining and Knowledge Discovery*, *22*(3), 493–521.

Wang, X., A. Li, Z. Jiang, and H. Feng (2006), Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme, *BMC bioinformatics*, *7*(1), 32.

Weng, H., and Y. Feng (2016), Community detection with nodal information, *arXiv preprint arXiv:1610.09735*.

White, S., and P. Smyth (2005), A spectral clustering approach to finding communities in graphs, in *Proceedings of the 2005 SIAM international conference on data mining*, pp. 274–285, SIAM.

Wolf, T., A. Schroter, D. Damian, and T. Nguyen (2009), Predicting build failures using social network analysis on developer communication, in *2009 IEEE 31st International Conference on Software Engineering*, pp. 1–11, IEEE.

Wolfe, P. J., and S. C. Olhede (2013), Nonparametric graphon estimation, *arXiv preprint arXiv:1309.5936*.

Xu, Z., Y. Ke, Y. Wang, H. Cheng, and J. Cheng (2012), A model-based approach to attributed graph clustering, in *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pp. 505–516.

Yan, B., and P. Sarkar (2016), Convex relaxation for community detection with covariates, *arXiv preprint arXiv:1607.02675*.

Yang, J., J. McAuley, and J. Leskovec (2013), Community detection in networks with node attributes, in *2013 IEEE 13th International Conference on Data Mining*, pp. 1151–1156, IEEE.

Young, S. J., and E. R. Scheinerman (2007), Random dot product graph models for social networks, in *International Workshop on Algorithms and Models for the Web-Graph*, pp. 138–149, Springer.

Zachary, W. W. (1977), An information flow model for conflict and fission in small groups, *Journal of anthropological research*, *33*(4), 452–473.

Zhang, Y., E. Levina, and J. Zhu (2015), Community detection in networks with node features, *arXiv preprint arXiv:1509.01173*.

Zhang, Y., E. Levina, J. Zhu, et al. (2016), Community detection in networks with node features, *Electronic Journal of Statistics*, *10*(2), 3153–3178.

Zhao, Y., E. Levina, and J. Zhu (2011), Community extraction for social networks, *Proceedings of the National Academy of Sciences*, *108*(18), 7321–7326.

Zhao, Y., E. Levina, and J. Zhu (2012), Consistency of community detection in networks under degree-corrected stochastic block models, *The Annals of Statistics*, *40*(4), 2266–2292.

Zhu, J., and T. E. Raghunathan (2015), Convergence properties of a sequential regression multiple imputation algorithm, *Journal of the American Statistical Association*, *110*(511), 1112–1124.

Zhur, X., and Z. Ghahramanirh (2002), Learning from labeled and unlabeled data with label propagation.