Statistical Tools for Directed and Bipartite Networks

by

Hyesun Yoo

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
Associate Professor Jian Kang
Professor Elizabeta Levina
Assistant Professor Gongjun Xu

Hyesun Yoo

yoohs@umich.edu

ORCID iD: 0000-0002-4811-0490

ACKNOWLEDGEMENTS

Working towards PhD has been a challenging and rewarding journey. I would like to give thanks to those who were supportive for my accomplishment over the past five years.

I would like to express my sincere gratitude to my advisor Prof. Ji Zhu for unwavering support and invaluable guidance. He has taught me not only statistics but also aptitude toward research and life. His understanding and encouragement made me overcome difficult moments in PhD.

I thank Prof. Liza Levina for giving constructive advice and sharing her experience. I am also grateful to the rest of my thesis committee members, Prof. Jian Kang and Prof. Gonjun Xu, for their time and thoughtful feedback.

I would like to thank many friends whom I met during PhD study. They made my PhD life much happier and easier than the one without them.

Finally but not least, I would like to thank my family members who always supported me no matter what. I am extremely grateful to my husband, PhD. Kanghwan Kim, who always helps me and shares his thoughts.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Directed networks and bipartite networks, which exhibit unique asymmetric connectivity structures, are commonly observed in a variety of scientific and engineering fields. Despite their abundance and utility, most network analysis methods only consider symmetric networks. In this thesis, we develop statistical methods and theory for directed and bipartite networks.

The first chapter focuses on matched community detection in a bipartite network. The detection of matched communities, i.e. communities that consist of nodes of two types that are closely connected with one another, is a fundamental and challenging problem. Most widely used approaches for matched community detection are either computationally inefficient or prone to non-ideal performance. We propose a new two-stage algorithm that uses fast spectral methods to recover matched communities. We show that, for bipartite networks, it is critical to adjust for the community size in matched community detection, which had not been considered before. We also provide theoretical error bounds for the proposed algorithm on the number of mis-clustered nodes under a variant of the stochastic block model. Numerical studies indicate that the proposed method outperforms existing spectral algorithms, especially when the sizes of the matched communities are proportionally different between the two types.

The second chapter of the thesis introduces a new preference-based block model for community detection in a directed network. Unlike existing models, the proposed

model allows different sender nodes to have different preferences to communities in the network. We argue that the right singular vectors of a graph Laplacian matrix contain community structures under the model. Further, we propose a spectral clustering algorithm to detect communities and estimate parameters of the model. Theoretical results show insights on how the heterogeneity of preferences and out-degrees contribute to an upper bound of the number of mis-clustered nodes. Numerical studies support the theoretical results and illustrate the outstanding performance of the proposed method. The model can also be naturally extended to bipartite networks.

In the third chapter, we propose a dyadic latent space model which accommodates the reciprocity between a pair of nodes in directed networks. Nodes in a pair in directed networks often exhibit strong dependencies with each other, though most widely used approaches usually account for this phenomenon with limited flexibility. We propose a new latent space model for directed networks that incorporates the reciprocity in a flexible way, allowing for important characteristics such as homophily and heterogeneity of the nodes. A fast and scalable algorithm based on projected gradient descent has been developed to fit the model by maximizing the likelihood. Both simulation studies and real-world data examples illustrate that the proposed model is effective in various network analysis tasks including link prediction and community detection.

CHAPTER I

Introduction

Networks represent interacting relationships among the components inside complex systems such as social networks, technology networks, citation networks, and biological networks, to name a few (Girvan and Newman, 2002; Goldenberg et al., 2010). In recent years, the advances in technology have provided more network data with increasing size and complexity (Fortunato, 2010). Because of the plethora of network data and valuable insights about the patterns of the connections inside a complex system that network analysis provides, network data have drawn attention from many scientific fields. Over the past few decades, extensive network analysis tools have been developed in a broad range of fields to understand the structures and features of complex network systems (Newman, 2018).

Among several forms of networks that exist in nature, directed networks and bipartite networks are two commonly observed networks. These two types of networks are distinguished by the asymmetric connectivity structures among the nodes inside the networks. A bipartite network is a network which has nodes of two types and whose connections exist only between nodes of different types (Newman, 2018). Nodes of the same type are not connected. Bipartite networks can represent many systems consisting of objects of two types, such as a network of actors and movies,

a network of papers and authors, and a network of users and items. Let n_1 denote the number of type-1 nodes and n_2 be the number of type-2 nodes, then the (binary) adjacency matrix A corresponding to a bipartite network is of $n_1 \times n_2$ dimension with elements A_{ij} such that $A_{ij} = 1$ if type-1 node i connects to type-2 node j, and 0 otherwise.

A directed network, also called digraph, has directionality on each edge, pointing from one node to another. Many real-world networks are directed such as the World Wide Web, email networks and social networks (e.g. Twitter, instagram). In a directed network, every node has two degrees, the out-degree and the in-degree, where the in-degree of a node is the number of incoming edges for that node and the out-degree is the number of outgoing edges. The in-degree and out-degree indicate the node's tendency to receive and send edges. In directed networks, a commonly observed phenomenon for a pair of nodes is that the incoming edge and the outgoing edge are often dependent, the so-called reciprocal relationship. Let n denote the number of nodes, then the (binary) adjacency matrix A is of $n \times n$ dimension with elements A_{ij} such that $A_{ij} = 1$ if node i sends an edge to node j, and 0 otherwise.

Community detection is a fundamental problem in network analysis, either as a goal or as an stepping stone for other learning tasks (Abbe, 2018). The goal of community detection is to partition the nodes in a network into clusters whose components are densely connected with each other. In general, a community refers to a groups of nodes whose connectivity behavior to other groups of nodes are similar. There are several approaches for community detection. One approach is to optimize some global criteria over possible partitions, such as graph cuts, spectral clustering and modularity. There are also model-based methods, i.e. fitting a probabilistic model with community memberships (Zhao et al., 2012). Popular models include the

stochastic block model and its variants, the latent position cluster model (Handcock et al., 2007) and the random dot product graph (Young and Scheinerman, 2007).

Perhaps the best studied models in community detection are the stochastic block model (SBM) (Holland et al., 1983) and its variants, such as the degree-corrected stochastic block model (Karrer and Newman, 2011; Zhao et al., 2012), and the mixed stochastic block model (Airoldi et al., 2008). The stochastic block model assumes that edge probabilities only depend on community memberships. Numerous studies have been conducted on several types of spectral clustering methods under the stochastic block model or the degree-corrected stochastic block model (Rohe et al., 2011; Sussman et al., 2012; Jin et al., 2015; Lei et al., 2015) (See review for SBM in (Abbe, 2018).

Though the stochastic block model and its variants enjoy the simple structure and ability to summarize a network, they sometimes can be too restrictive. Latent space models, first proposed by (Hoff et al., 2002), have been popular network models due to its flexibility and interpretability. In latent space models, each node is represented by a vector z_i in a low-dimensional Euclidean space. Given latent positions of two nodes, the edge probability is modeled as a function of their positions. For instance, (Hoff et al., 2002) used $-\|z_i - z_j\|^2$ as the distance model, i.e. two nodes are more likely to be connected if z_i and z_j are close to each other. This concept has been extended in several follow-up works. For example, the multiplicative effect and the random effect modeling were introduced to capture second or third order dependency in networks. Random effects for degree heterogeneity were also introduced. Markov Chain Monte Carlo is often used for model fitting and inference for these models. However, using Markov Chain Monte Carlo makes it difficult to apply these models to large networks. To overcome this challenge and build more general models, Ma

et al. (2020) and Wu et al. (2017) introduced variants of the latent space model and proposed to fit models using projected gradient descent algorithms. There have been other approaches based on matrix decomposition. For example, graph embedding methods based on matrix decomposition (Belkin and Niyogi, 2003; Kunegis and Lommatzsch, 2009; Athreya et al., 2017) have been popular. In these methods, the leading eigenvectors or singular vectors of the graph laplacian matrix or the adjacency matrix are used for estimating latent positions. Several graph embedding algorithms for large networks with stochastic gradient descent (Grover and Leskovec, 2016; Perozzi et al., 2014) have also been developed so the computational time scales linearly with respect to the size of the network.

Despite their abundance and utility, directed networks and bipartite networks were given less attention compared to symmetric (undirected) networks. For example, in practice, the most common approach to analyze directed networks is to transform them into undirected ones first and then apply techniques that have been developed for undirected networks. Specifically, bipartite adjacency matrix can be considered as a special case of symmetric adjacency matrix by embedding bipartite adjacency matrix into a larger block matrix. Directed networks are commonly transformed into undirected networks by removing directionality. However, these approaches might result in a considerable loss of information. In particular, we may need different definitions for communities in bipartite networks when there are two types of nodes. The directionality may carry useful information about the community structure and link probability. Therefore, statistical methods designed for directed networks and bipartite networks are demanded.

In this thesis, we develop statistical methods and theory for bipartite networks and directed networks for three problems: (1) detection of matched communities

between two types of nodes, (2) detection of communities under a preference-based block model, and (3) accommodation of the reciprocal property in modeling. The rest of the thesis is organized as follows:

Chapter II focuses on matched community detection for bipartite networks. Most existing work (Dhillon, 2001; Rohe et al., 2016; Razaee et al., 2019) are computationally inefficient, or prone to non-ideal results. Thus, we propose an efficient and robust new two-stage algorithm based on spectral clustering to identify matched communities. We show that it is crucial to adjust community sizes in matched community detection. Theoretical results on the upper bound on the number of mis-clustered nodes are provided. Simulation studies and data analysis are also shown to support the performance of the proposed algorithm.

Chapter III presents a method that considers individual differences in directed networks for community detection. In directed networks, each node can have different preferences to communities (Cantwell and Newman, 2019; Altenburger and Ugander, 2018). We introduce a preference-based block model that takes into account these individual differences. We propose a spectral clustering algorithm for community detection under the model and estimate the community memberships and model parameters. Theoretical results for the number of mis-clustered nodes are provided in terms of the preference heterogeneity and degree heterogeneity.

In Chapter IV, we propose a novel dyadic latent space model that considers reciprocal relationships in directed networks. Existing models (Hoff, 2015; Holland et al., 1983) have considered the reciprocity with strong assumptions such as a constant tendency of reciprocation in a network. We provide an efficient projected gradient descent algorithm for estimation. Theoretical results on the error bounds for parameters and the probability matrix are also developed. Simulation studies and real-world

data application demonstrate the outstanding performance of the proposed method.

CHAPTER II

A Two-stage Spectral Co-clustering Algorithm for Matched Communities

2.1 Introduction

Networks have been an important representation of relationships between entities or objects. They are commonly observed in many scientific and engineering fields, such as social networks, biological networks, telecommunication networks etc. A bipartite network is a network which has nodes of two different types and the connections exist only between nodes of different types (Newman, 2018). Many real-world systems can be represented as bipartite networks, such as actors and movies network, papers and authors network. In addition, directed networks have also often been treated as bipartite networks (Malliaros and Vazirgiannis, 2013) and can be analyzed using techniques for bipartite networks.

In study of networks, it is often of interest to detect communities that consist of nodes that are closely linked to one another. One common approach to community detection for bipartite networks is to first use one-mode projection to transform the bipartite network into two regular networks and then apply algorithms for regular networks to the projected network. However, this transformation may involve information loss. Moreover, bipartite networks show relationships between nodes of two types more effectively than two transformed one-mode projections.

The stochastic block model (SBM), proposed by Holland et al. (1983) provides a simple way to incorporate community structures. In SBM, connectivity between nodes are determined based on latent membership variables. For detailed survey on SBM, see Abbe (2018). Most community detection methods under SBM have been developed for symmetric networks. These methods can be applied to bipartite networks as well since a bipartite network can be considered as a symmetric matrix by embedding the bipartite network into the symmetric block matrix. To be specific, a bipartite network with n_1 type-1 nodes and n_2 type-2 nodes can be expressed as a symmetric network with n_1+n_2 nodes. However, if we transform a bipartite network into a symmetric network and apply the community detection methods developed for the symmetric networks, this process is most likely to have type-1 nodes and type-2 nodes as in different communities since type-1 and type-2 nodes behave differently, i.e., there is no edge between type 1 node and type-2 node. Thus, this procedure provides community memberships for type-1 nodes and for type-2 nodes separately. To discover community structure between two types of nodes, the transformation may not be sufficient for bipartite networks.

Co-clustering, introduced by Hartigan (1972), simultaneously clusters the rows and columns of a data matrix, each representing information of different types. Dhillon (2001) studied co-clustering documents and words by concatenating the left and the right singular vectors of the data matrix followed by a k-means algorithm. Rohe et al. (2016) and Razaee et al. (2019) applied co-clustering to network analysis, yet concatenation of the left and right singular vectors was not explicitly justified. Matched co-clustering tries to obtain one-to-one matched clusters of two different types, which is especially useful and interpretable for assortative networks.

In this chapter, we focus on matched co-clustering. We first observe that imbal-

anced community sizes between two types of nodes can affect the performance of existing algorithms. Then we propose a two-stage spectral co-clustering algorithm, in which we adjust for the effect of imbalanced community sizes in singular value decomposition. We further provide an upper bound on the proportion of the mis-clustered nodes under a special case of the stochastic block model.

The chapter is organized as follows. Section 2.2 defines a matched stochastic block model for bipartite networks. A two-stage algorithm for estimating the model is also proposed. Section 2.3 studies the upper bound of the error rate to theoretically validate the performance of the algorithm. Section 2.5 uses simulation studies to investigate the performance of the algorithm. Applications to real-world data sets are presented in Section 2.6.

2.2 Methodology

In this section, we first review related methods then introduce our proposed method for matched co-clustering.

2.2.1 A brief review of existing spectral co-clustering algorithms

We first give an overview of previous spectral matched co-clustering algorithms for bipartite networks. Dhillon (2001) proposed a spectral co-clustering algorithm to cluster documents and words simultaneously. Specifically, they posed co-clustering as a bipartite graph partitioning problem and solved the problem using singular value decomposition (SVD) of the bipartite Laplacian matrix. Unlike regular spectral clustering, concatenation of the right and the left singular vectors is the key step for simultaneous co-clustering. The concatenation was motivated by the observation that a partitioning vector whose elements' values are the same if the corresponding nodes are in the same community, minimize the normalized-cut objective function

regardless of the node's type. However, the observation does not exactly hold true when relaxation is allowed to apply SVD.

Rohe et al. (2016) and Razaee et al. (2019) proposed variants of the algorithm for networks along with regularization for sparsity and row normalization for degree heterogeneity, which are frequently used in network analysis. They both introduced stochastic block models for bipartite networks. Both algorithms concatenate the left and the right singular vectors to cluster two different types of nodes simultaneously. Specifically, Rohe et al. (2016) focused on discovery of asymmetric nodes and directional communities considering that the sending behavior and the receiving behavior of nodes can be different. The resulting directed network can be naturally perceived as a special case of bipartite network. They also proved that their algorithm can estimate the clusters of each type consistently under certain conditions. Razaee et al. (2019) attempted to match communities of different types. They focused on incorporation of node covariates, which might contain additional information, based on a variational inference method. They also proposed a variant of the spectral co-clustering algorithm of Dhillon (2001) as the initialization step for the variational inference.

Both algorithms (Rohe et al., 2016; Razaee et al., 2019) used the concatenation of the left and the right singular vectors and applied spectral clustering on the concatenated singular vectors. This procedure clusters relatively close points of two types together, using similar techniques as in correspondence analysis. In other words, similar points from the concatenated singular vectors are clustered together even if the population version of the points from the concatenated singular vectors do not match. The right and the left singular vectors can be considered as some representation in a low-dimensional latent space and similar positions imply their similarity in

stochastic behavior in the network. The process is akin to correspondence analysis, which is further studied in Zha et al. (2001). None of the above algorithms take into account the fact that the equality of the population centroids of the left and the right singular vectors is not guaranteed. In order to address this issue, we propose an algorithm that adjusts the population centroids of two types to be the same under a model for matched communities. The algorithm not only enables more robust clustering results but also is theoretically valid.

2.2.2 Model setup

We propose our model based on the Stochastic Block Model (SBM), a popular and well-studied model for community detection (Rohe et al., 2011; Amini et al., 2013). In the context of bipartite networks, consider a network with n_1 nodes of type-1 and n_2 nodes of type-2 represented by an adjacency matrix $A \in \{0,1\}^{n_1 \times n_2}$. Let $\mathbb{M}_{n,K} \in \{0,1\}^{n \times K}$ be the set of all $n \times K$ matrices where each row has exactly one 1 and $(K-1)$ 0's. For any $Z_1 \in \mathbb{M}_{n_1,K}$ ($Z_2 \in \mathbb{M}_{n_2,K}$), we call Z_1 (Z_2) a membership matrix of type-1 (type-2). Denote the community label of a type-1 node i by $z_{1i} \in \{1, \cdots, K\}$, thus the ith row of Z_1 is 1 in column z_{1i} and 0 elsewhere. Similarly, denote the community label of a type-2 node j by $z_{2j} \in \{1, \cdots, K\}$, and the jth row of Z_2 is 1 in column z_{2j} and 0 elsewhere. Let $G_{1,k} = G_{1,k}(Z_1) = \{1 \leq i \leq n_1 : z_{1i} = k\}$, $G_{2,k} = G_{2,k}(Z_2) = \{1 \leq j \leq n_2 : z_{2j} = k\}$, $n_{1,k} = |G_{1,k}|$, and $n_{2,k} = |G_{2,k}|$ for $1 \leq k \leq K$. Let $n_{1,min} = \min_{1 \leq k \leq K} n_{1,k}$, $n_{2,min} = \min_{1 \leq k \leq K} n_{2,k}$, $n_{1,max} = \max_{1 \leq k \leq K} n_{1,k}$, and $n_{2,max} = \max_{1 \leq k \leq K} n_{2,k}$.

Definition II.1. Matched stochastic block model (MSBM) for bipartite network. Let $Z_1 \in \mathbb{M}_{n_1,K}$ and $Z_2 \in \mathbb{M}_{n_2,K}$ be membership matrices. Assume the block probability matrix $B \in [0,1]^{K \times K}$ is positive definite. Given Z_1, Z_2 and B, the edge

variables A_{ij}'s are independent Bernoulli random variables with

$$\mathbb{E}[A_{ij}|Z_1, Z_2] = B_{z_{1i}z_{2j}}. \tag{2.1}$$

Further, the node i of type-1 and the node j of type-2 are in the same matched community if $z_{1i} = z_{2j}$.

Note the expression (2.1) can be written in a matrix form as $P = \mathbb{E}[A|Z] = Z_1 B Z_2^\intercal$. The MSBM is parametrized by matrices (Z_1, Z_2, B). In MSBM, nodes of different types are in the same matched community if they are labeled the same.

We wish to match two communities of different types if they have assortative relationship. Assortative community structure in a bipartite network is slightly different from that of a SBM. By its nature, a bipartite network has strong disassortative structures among nodes of type-1 (or of type-2) as no edge exists between nodes of the same type. This requires the assortative relationship in a bipartite network be defined on the edges between type-1 and type-2 nodes, where edges exist. Although there are multiple definitions for assortativeness in SBM in the literature (Binkiewicz et al., 2017; Amini et al., 2018), all definitions refer to the rough concept that there are more edges within a community than between communities. Here, we adopt positive definiteness to define assortativeness for bipartite networks (Binkiewicz et al., 2017).

It is also important that B is a symmetric matrix. A SBM of undirected network can naturally be extended to bipartite networks by restricting the SBM to have no edge between nodes of the same type (Larremore et al., 2014; Razaee et al., 2019). Such a restriction on SBM results in the symmetry of B in MSBM. Therefore, for type-1 nodes $i \neq i'$ and type-2 nodes $j \neq j'$, if i and j are in the same matched community and i' and j' are in the same matched community, then $B_{g_i g_{j'}} = B_{g_{i'} g_j}$

by Definition II.1.

Note the MSBM can be thought of as a special case of SBM with $2K$ communities. Let $n = n_1 + n_2$ denote the total number of nodes of both types. For any adjacency matrix $A \in \{0,1\}^{n_1 \times n_2}$ generated from MSBM, we have

$$\mathbb{E}\begin{bmatrix} 0 & A \\ A^\intercal & 0 \end{bmatrix} = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} 0 & B \\ B & 0 \end{bmatrix} \begin{bmatrix} Z_1^\intercal & 0 \\ 0 & Z_2^\intercal \end{bmatrix} = ZB'Z^\intercal \tag{2.2}$$
$$\text{where } Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \in \mathbb{R}^{n \times 2K} \text{ and } B' = \begin{bmatrix} 0 & B \\ B & 0 \end{bmatrix} \in \mathbb{R}^{K \times K}.$$

Note $Z \in \mathbb{M}_{n,2K}$. Thus, SBM by itself perceives that there are $2K$ communities in MSBM. To obtain MSBM, however, we need to know the one-to-one matching between the communities in both types, i.e. which label in type-1 is matched to a label in type-2. Since communities are only detectable up to permutation, applying community detection methods for SBM directly to MSBM may not be able to match communities. This necessitates the definition of MSBM for the purpose of co-clustering.

2.2.3 Spectral clustering

Investigating the SVD structure of the mean matrix P provides heuristics for spectral clustering because A can be treated as a noisy version of P. The following lemma explains the structure of the singular vectors of A under MSBM.

Lemma II.2. *Basic SVD-structure of the mean matrix $P = \mathbb{E}[A]$. Let (Z_1, Z_2, B) parametrize a MSBM with K communities, where B is a positive definite matrix with full rank K. Let $\mathcal{U}D\mathcal{V}^\intercal$ be the singular value decomposition of P. Then, $\mathcal{U} = Z_1 C_\mathcal{U}$ where $C_\mathcal{U} \in \mathbb{R}^{K \times K}$ and $\mathcal{V} = Z_2 C_\mathcal{V}$ where $C_\mathcal{V} \in \mathbb{R}^{K \times K}$. In addition, the directions of the row vectors in $C_\mathcal{U}$ and $C_\mathcal{V}$ depend on the size of the communities.*

Proof Let $\Delta_1 = diag(n_{1,1}, \cdots, n_{1,K})$, $\Delta_2 = diag(n_{2,1}, \cdots, n_{2,K})$,

$$P = Z_1 B Z_2^\intercal = Z_1 \Delta_1^{-1/2} \left(\Delta_1^{1/2} B \Delta_2^{1/2} \right) \Delta_2^{-1/2} Z_2^\intercal.$$

Let singular value decomposition of $\Delta_1^{1/2} B \Delta_2^{1/2}$ be $\mathcal{X} D \mathcal{Y}^\intercal$. Then, the SVD of P is $(Z_1 \Delta_1^{-1/2} \mathcal{X}) D (Z_2 \Delta_2^{-1/2} \mathcal{Y})^\intercal$, which leads to $\mathcal{U} = Z_1 \Delta_1^{-1/2} \mathcal{X}$ and $\mathcal{V} = Z_2 \Delta_2^{-1/2} \mathcal{Y}$. Thus, we have $C_\mathcal{U} = \Delta_1^{-1/2} \mathcal{X}$ and $C_\mathcal{V} = \Delta_2^{-1/2} \mathcal{Y}$. □

Define the Laplacian matrix $L \in \mathbb{R}^{n_1 \times n_2}$ as $L = D_1^{-1/2} A D_2^{-1/2}$, where the diagonal matrices $D_1 \in \mathbb{R}^{n_1 \times n_1}$ and $D_2 \in \mathbb{R}^{n_2 \times n_2}$ are defined as $D_1 = diag(\sum_{j=1}^{n_2} A_{ij},\ i = 1, \cdots, n_1)$ and $D_2 = diag(\sum_{i=1}^{n_1} A_{ij},\ j = 1, \cdots, n_2)$ respectively. Let the population version Laplacian matrix be $\mathcal{L} = \mathcal{D}_1^{-1/2} P \mathcal{D}_2^{-1/2}$, where $\mathcal{D}_1 = diag(\sum_{j=1}^{n_2} P_{ij},\ i = 1, \cdots, n_1)$ and $\mathcal{D}_2 = diag(\sum_{i=1}^{n_1} P_{ij},\ j = 1, \cdots, n_2)$ respectively. Then an analogous lemma for the Laplacian matrix is as follows.

Lemma II.3. *Basic SVD-structure of the mean matrix $\mathcal{L} = \mathcal{D}_1^{-1/2} P \mathcal{D}_2^{-1/2}$. Let (Z_1, Z_2, B) parametrize a MSBM with K communities, where B is positive definite with full rank K. Let $\mathcal{U} D \mathcal{V}^\intercal$ be the singular value decomposition of $\mathcal{L}$. Then, $\mathcal{U} = Z_1 C_\mathcal{U}$ where $C_\mathcal{U} \in \mathbb{R}^{K \times K}$ and $\mathcal{V} = Z_2 C_\mathcal{V}$ where $C_\mathcal{V} \in \mathbb{R}^{K \times K}$. In addition, the directions of the row vectors in $C_\mathcal{U}$ and $C_\mathcal{V}$ depend on the size of the communities.*

Note that the left and right singular vectors of the mean matrix P (or $\mathcal{L}$) depend on the community sizes in each type, thus the left and right singular vectors that belong to the same matched community may not be the same. This implies that there are more than K distinct rows which would make clustering challenging unless the ratio of community sizes are all the same so that $\Delta_1^{1/2} B \Delta_2^{1/2}$ is symmetric. Therefore, the performances of previously existing methods depend on the closeness between the centroids of type-1 and the corresponding centroids of type-2.

In other words, matched cluster memberships can only be found if the variation of singular vectors of the observed matrix A or L is not large and the matched community centroids of both types are close enough. This sheds light on when previously existing algorithms could cluster two types simultaneously and when they will fail. Also note that if we adjust for the effect of the community sizes, we could make the centroids of each matched community to be the same. Specifically, let the adjusted mean matrix $\tilde{P} = W_1^{-1/2} P W_2^{-1/2}$, where W_1=diag$(n_{1,z_{1i}},\ i = 1, \cdots, n_1)$ and W_2=diag$(n_{2,z_{2j}},\ j = 1, \cdots, n_2)$. Then we have the following lemmas.

Lemma II.4. *Basic SVD-structure of $\tilde{P}$. Let (Z_1, Z_2, B) parametrize a MSBM with K communities, where B is a positive definite matrix with full rank K. Let $\tilde{\mathcal{U}}\tilde{D}\tilde{\mathcal{V}}^\intercal$ be the singular value decomposition of $\tilde{P}$. Then, we have $\tilde{\mathcal{U}} = Z_1 C_{\tilde{\mathcal{U}}}$ where $C_{\tilde{\mathcal{U}}} \in \mathbb{R}^{K\times K}$, and $\tilde{\mathcal{V}} = Z_2 C_{\tilde{\mathcal{V}}}$ where $C_{\tilde{\mathcal{V}}} \in \mathbb{R}^{K\times K}$. In addition, we have $\Delta_1^{1/2} C_{\tilde{\mathcal{U}}} = \Delta_2^{1/2} C_{\tilde{\mathcal{V}}}$, where $\Delta_1 = diag(n_{1,1}, \cdots, n_{1,K})$, $\Delta_2 = diag(n_{2,1}, \cdots, n_{2,K})$.*

Proof Note

$$\tilde{P} = W_1^{-1/2} P W_2^{-1/2} = W_1^{-1/2} Z_1 B Z_2^\intercal W_2^{-1/2} = Z_1 \Delta_1^{-1/2} B \Delta_2^{-1/2} Z_2^\intercal.$$

Let the singular value decomposition of B be $\mathcal{X}\tilde{D}\mathcal{X}^\intercal$. Then, the SVD of $\tilde{P}$ is $(Z_1\Delta_1^{-1/2}\mathcal{X})\ \tilde{D}(Z_2\Delta_2^{-1/2}\mathcal{X})^\intercal$, which leads to $\tilde{\mathcal{U}} = Z_1\Delta_1^{-1/2}\mathcal{X}$ and $\tilde{\mathcal{V}} = Z_2\Delta_2^{-1/2}\mathcal{X}$. Then, we have $C_{\tilde{\mathcal{U}}} = \Delta_1^{-1/2}\mathcal{X}$ and $C_{\tilde{\mathcal{V}}} = \Delta_2^{-1/2}\mathcal{X}$. □

Lemma II.5. *Basic SVD-structure of $\tilde{\mathcal{L}}$. Let (Z_1, Z_2, B) parametrize a MSBM with K communities, where B is a positive definite matrix with full rank K. Let $\tilde{\mathcal{U}}\tilde{D}\tilde{\mathcal{V}}^\intercal$ be the singular value decomposition of $\tilde{\mathcal{L}}$. Then, we have $\tilde{\mathcal{U}} = Z_1 C_{\tilde{\mathcal{U}}}$ where $C_{\tilde{\mathcal{U}}} \in \mathbb{R}^{K\times K}$, and $\tilde{\mathcal{V}} = Z_2 C_{\tilde{\mathcal{V}}}$ where $C_{\tilde{\mathcal{V}}} \in \mathbb{R}^{K\times K}$. In addition, we have $\Delta_1^{1/2} C_{\tilde{\mathcal{U}}} = \Delta_2^{1/2} C_{\tilde{\mathcal{V}}}$, where $\Delta_1 = diag(n_{1,1}, \cdots, n_{1,K})$, $\Delta_2 = diag(n_{2,1}, \cdots, n_{2,K})$.*

Proofs for Lemma II.3 and Lemma II.5 are provided in the appendix. Note the above observation suggests how we can adjust the adjacency matrix or the Laplacian matrix to match the centroids of two types.

If the true community size of each type corresponding to each node is known, such adjustment will result in K distinct directions in the row vectors, and also the population centroids for the two types being exactly the same. However, in practice, the true community sizes are unknown and therefore should be estimated.

2.2.4 Proposed algorithms

In this subsection, we propose two spectral co-clustering algorithms using the adjacency matrix or the Laplacian matrix based on Lemma II.4 and Lemma II.5 respectively. Algorithm II.1 uses the adjacency matrix, and Algorithm II.2 uses the Laplacian matrix. Both algorithms consist of two stages, where the result from the first stage is committed to estimating corresponding community sizes. In both algorithms, Steps 1-2 are standard whereas the rest steps are committed for making adjustments. Denote $\hat{n}_{1,k}$ ($\hat{n}_{2,k}$) as the estimated community size of type-1 (type-2). Similarly, $\hat{W}_1$ denotes $\text{diag}(\hat{n}_{1,z_{1i}},\ i = 1, \cdots, n_1)$ and $\hat{W}_2$ denotes $\text{diag}(\hat{n}_{2,z_{2j}},\ j = 1, \cdots, n_2)$.

Algorithm II.1 Using the adjacency matrix

1: Input: bipartite adjacency matrix $A \in \{0,1\}^{n_1 \times n_2}$ and number of communities K
2: Compute K left and K right singular vectors $U \in \mathbb{R}^{n_1 \times K}$ and $V \in \mathbb{R}^{n_2 \times K}$ corresponding to the K largest singular values of A. Run k-means separately on rows of U and rows of V.
3: Based on the result from 2, construct diagonal matrices $\hat{W}_1$ and $\hat{W}_2$, where each diagonal element is the estimated size of the community that the corresponding node belongs to.
4: Let $\hat{A} = \hat{W}_1^{-1/2} A \hat{W}_2^{-1/2}$. Compute K left and K right singular vectors $\hat{U} \in \mathbb{R}^{n_1 \times K}$ and $\hat{V} \in \mathbb{R}^{n_2 \times K}$ corresponding to the K largest singular values of $\hat{A}$.
5: Concatenate $\hat{W}_1^{1/2}\hat{U}$ and $\hat{W}_2^{1/2}\hat{V}$ and run k-means on the concatenated matrix to obtain K clusters.

Algorithm II.2 Using the Laplacian matrix

1: Input: bipartite adjacency matrix $A \in \{0,1\}^{n_1 \times n_2}$ and number of communities K
2: Form $L = D_1^{-1/2} A D_2^{-1/2}$. Compute K left and K right singular vectors $U \in \mathbb{R}^{n_1 \times K}$ and $V \in \mathbb{R}^{n_2 \times K}$ corresponding to the K largest singular values of L. Run k-means separately on rows of U and rows of V.
3: Based on the result from 2, construct diagonal matrices $\hat{W}_1$ and $\hat{W}_2$, where each diagonal element is the estimated size of the community that the corresponding node belongs to.
4: Let $\hat{A} = \hat{W}_1^{-1/2} A \hat{W}_2^{-1/2}$, $\hat{D}_1$=diag$(\sum_j A_{ij}/\hat{n}_{2,\hat{z}_{2j}}, i = 1, \cdots, n_1)$ and $\hat{D}_2$=diag$(\sum_i A_{ij}/\hat{n}_{1,\hat{z}_{1i}}, j = 1, \cdots, n_2)$.
5: Let $\hat{L} = \hat{D}_1^{-1/2} \hat{A} \hat{D}_2^{-1/2}$. Compute K left and K right singular vectors $\hat{U} \in \mathbb{R}^{n_1 \times K}$ and $\hat{V} \in \mathbb{R}^{n_2 \times K}$ corresponding to the K largest singular values of $\hat{L}$.
6: Concatenate $\hat{W}_1^{1/2}\hat{U}$ and $\hat{W}_2^{1/2}\hat{V}$ and run k-means on the concatenated matrix to obtain K clusters.

Understandably, the performance of the second stage depends on the performance of the first stage, and when the result from the first stage is moderately good, we expect improved performance in the second stage. Further, note the singular vectors multiplied by square root of the corresponding estimated community size are used as the input of the k-means algorithm in the last step of the above algorithms. One can also use regularization techniques for better concentration of singular vectors when the network is sparse (Chaudhuri et al., 2012; Amini et al., 2013; Joseph et al., 2016). For example, L and $\hat{L}$ in Algorithm II.2 can be replaced by regularized versions.

The proposed algorithms are expected to set the centroids of the matched communities to be the same so that co-clustering would be more effective than previously existing methods. However, it should be noted that is an interplay between the variance of singular vectors and the distance between the centroids of the two types. The proposed algorithm reduces the distance between matching centroids, but it may increase the variance of the singular vectors, which is also an important factor for clustering. This phenomenon might be explained by using similar techniques presented in Sarkar et al. (2015), where they compared the performance of the Laplacian matrix with that of the adjacency matrix. The performance of a particular spectral

clustering algorithm depends on parameter regimes, while obtaining exact parameter regimes could be complicated due to the two stages in the algorithms. In simulation studies, we show the proposed algorithms work well especially when the ratio of community sizes between two types are imbalanced.

2.3 Theoretical results

In this section, we investigate theoretical properties of the proposed algorithms under the MSBM model. For simplicity, we use the adjacency matrix to illustrate the theoretical results, rather than the Laplacian matrix. It has been shown that concentration of eigenvectors from adjacency matrix and laplacian matrix, which is closely related to the performance of spectral clustering, has identical rate of convergence (Sarkar et al., 2015). Throughout the analysis, we assume that the numbers of nodes of the two types do not differ much (i.e. $O(n_1/n_2) = 1$).

Our analysis consists of two parts. In the first part, bounds on the mis-clustering error of each type are obtained separately. This step is akin to the results of SBM with $2K$ communities. In the second part, matched mis-clustering rate is derived based on the mis-clustering rate of the first stage. The main component of the proof is to bound the difference between singular vectors of the adjusted adjacency matrix $\hat{A}$ and those of the correctly estimated population version matrix $\tilde{P}$.

Applying the k-means algorithm to singular vectors is a key step of the spectral clustering algorithm. The k-means algorithm minimizes $\|ZC - U\|_F^2$ over all $Z \in \mathbb{M}_{n\times K}$ and $C \in \mathbb{R}^{K\times K}$. Since solving the k-means problem is NP-hard, we consider the efficient approximate k-means algorithm (Kumar et al., 2004), which provides a solution $(\hat{Z}, \hat{C}) \in \mathbb{M}_{n\times K} \times \mathbb{R}^{K\times K}$ such that

$$\|\hat{Z}\hat{C} - U\|_F^2 \leq (1+\varepsilon) \min_{Z,C} \|ZC - U\|_F^2.$$

2.3.1 The first stage algorithm

The goal of clustering before adjustment is to estimate Z_1 and Z_2 respectively up to permutation. Since MSBM is a special case of SBM, the bound on the mis-clustering rate for SBM can be applied in the first stage. The bound on the mis-clustering rate has been extensively studied (Lei et al., 2015; Rohe et al., 2011) and we build our result on the work by Lei et al. (2015). We define the mis-clustered nodes similar to those in Rohe et al. (2011) and Lei et al. (2015) using the distance between the centroids obtained from the k-means algorithm and the centroids from the population matrix. This definition can ultimately bound the error rate

$$Error(\mathbf{z}, \hat{\mathbf{z}}) = n^{-1} \min_{\sigma} \sum_{i=1}^{n} I(z_i \neq \sigma(\hat{z}_i)) \tag{2.3}$$

where z_i is true label for node i, $\hat{z}_i$ is the estimated label for node i and σ is a permutation function. Let $\mathcal{U}\mathcal{D}\mathcal{V}^\intercal$ be the singular value decomposition of P and $UDV^\intercal$ be the singular value decomposition of A, where $U, \mathcal{U} \in \mathbb{R}^{n_1 \times K}$ and $V, \mathcal{V} \in \mathbb{R}^{n_2 \times K}$. As in Lemma II.2, $\mathcal{U} = Z_1 C_{\mathcal{U}}$ with $Z_1 \in \mathbb{M}_{n_1,K}$, $C_{\mathcal{U}} \in \mathbb{R}^{K \times K}$ and $\mathcal{V} = Z_2 C_{\mathcal{V}}$ with $Z_2 \in \mathbb{M}_{n_2,K}$, $C_{\mathcal{V}} \in \mathbb{R}^{K \times K}$. Let $(\hat{Z}_1, \hat{C}_U)$ be a $(1+\varepsilon)$-approximate solution to the k-means problem and $\bar{U} = \hat{Z}_1 \hat{C}_U$. Similarly, let $(\hat{Z}_2, \hat{C}_V)$ be a $(1+\varepsilon)$-approximate solution to the k-means problem and $\bar{V} = \hat{Z}_2 \hat{C}_V$. We denote M_{i*} as ith row of a matrix M.

Definition II.6. Let $\delta_{1,k} = \min_{l \neq k} \|C_{\mathcal{U},l*} - C_{\mathcal{U},k*}\|$ and $\delta_{2,k} = \min_{l \neq k} \|C_{\mathcal{V},l*} - C_{\mathcal{V},k*}\|$. Define $\mathcal{S}_{1,k} = \{i \in G_{1,k}(Z_1) : \|\bar{U}_{i*} - \mathcal{U}_{i*}\mathcal{Q}\| \geq \delta_{1,k}/2\|\}$ and $\mathcal{S}_{2,k} = \{j \in G_{2,k}(Z_2) : \|\bar{V}_{j*} - \mathcal{V}_{j*}\mathcal{Q}\| \geq \delta_{2,k}/2\|\}$ for an orthonormal matrix $\mathcal{Q} \in \mathbb{R}^{K \times K}$. Define the set of mis-clustered nodes of type-1 as $\mathcal{S}_1 = \cup_k \mathcal{S}_{1,k}$ and the set of mis-clustered nodes of type-2 as $\mathcal{S}_2 = \cup_k \mathcal{S}_{2,k}$.

The bounds on the number of mis-clustered nodes for each type can be obtained

by modifying the Corollary 3.2 in Lei et al. (2015). Specifically, it can be done by first creating a symmetrized square matrix for a bipartite network and then applying the results in Lei et al. (2015) accordingly.

Proposition II.7. *Let $A \in \{0,1\}^{n_1 \times n_2}$ be a bipartite adjacency matrix generated from a MSBM (Z_1, Z_2, B). Assume that $P = Z_1 B Z_2^\intercal$ is of rank K and $B = \alpha_n B_0$ for some $\alpha_n \geq \log n/n$ with B_0 having minimum singular value $\geq \lambda_K > 0$ and $\max_{s,t} B_{0,st} = 1$. Let $\hat{Z}_1$ and $\hat{Z}_2$ be the output of spectral clustering using the $(1+\varepsilon)$-approximate k-means algorithm. For an absolute constant $c > 0$, with probability at least $1 - n^{-1}$, we have*

$$|\mathcal{S}_1| \leq c^{-1}(2+\varepsilon)\frac{n_{1,max}Kn}{n_{1,min}n_{2,min}\lambda_K^2\alpha_n}, \quad |\mathcal{S}_2| \leq c^{-1}(2+\varepsilon)\frac{n_{2,max}Kn}{n_{1,min}n_{2,min}\lambda_K^2\alpha_n}, \tag{2.4}$$

where $n_{1,min}$ and $n_{2,min}$ represent the smallest community size of each type and $n_{1,max}$ and $n_{2,max}$ represent the largest community size of each type.

As the derivation is similar to the proof in Lei et al. (2015), only the different part of the proof is included in the appendix. Proposition II.7 shows how the number of mis-clustered nodes of each type is bounded by other parameters, and the result can be easily extended to bipartite networks with different numbers of communities between the two types. However, this result does not show how the communities between the two types can be one-to-one matched.

2.3.2 The second stage algorithm

Because the result in the second stage depends on that of the first stage, reasonable assumptions on the performance of the first stage are necessary. Intuitively, if the performance of the first stage is not satisfactory, good performance of the second stage cannot be expected. Here, we make an assumption that the total number of mis-clustered nodes for each type is no larger than the minimum community size of

the corresponding type. This prevents two extreme scenarios: (1) all mis-clustered nodes are from the community of the smallest size so that the estimated community size becomes zero; (2) all mis-clustered nodes are assigned to the smallest community so that the estimated smallest community size becomes much larger.

Assumption II.8. *Assume* $\eta < C_1$ *where* $\eta = \max(\frac{|\mathcal{S}_1|}{n_{1,min}}, \frac{|\mathcal{S}_2|}{n_{2,min}})$ *for some constant* $0 \leq C_1 < 1$.

Under this assumption, we are able to obtain a simple bound on the ratio of estimated community size and the true community size with respect to η. Let $r = \max(\frac{n_{1,max}}{n_{1,min}}, \frac{n_{2,max}}{n_{2,min}})$ be the maximum ratio of the largest and the smallest community sizes of each type. Combined with Proposition II.7, we also have $\eta \leq c^{-1}(2+\varepsilon)r\frac{Kn}{n_{1,min}n_{2,min}\lambda_K^2\alpha_n}$. Since the community sizes are estimated from the first stage to adjust the adjacency (or Laplacian) matrix in the second stage, the error bound on the estimated community size is needed as a result of the first stage algorithm. The following lemma provides the bound on the ratio of estimated community size and the true community size under the Assumption II.8.

In order to state the lemma, we define the estimated kth community $\hat{G}_{1,k}$ corresponding to true community $G_{1,k}$ as the aligned estimated community using the same permutation function that minimizes the error rate (2.3), which is $\hat{G}_{1,k} = \{1 \leq i \leq n_1 : \hat{z}_{1i} = \sigma(k)\}$. $\hat{G}_{2,k}$ can be defined similarly. The estimated community size $\hat{n}_{1,k}$ and $\hat{n}_{2,k}$ are then defined based on $\hat{G}_{1,k}$ and $\hat{G}_{2,k}$ respectively.

Lemma II.9. *Error bounds on the estimated community size. By using the bound on mis-clustering error from the first stage algorithm and assumption II.8, we have*

$$\begin{aligned} \left|\frac{\hat{n}_{1,k}}{n_{1,k}} - 1\right| \leq \eta \qquad \left|\frac{n_{1,k}}{\hat{n}_{1,k}} - 1\right| \leq \frac{\eta}{1-\eta} \\ \left|\frac{\hat{n}_{2,k}}{n_{2,k}} - 1\right| \leq \eta \qquad \left|\frac{n_{2,k}}{\hat{n}_{2,k}} - 1\right| \leq \frac{\eta}{1-\eta}. \end{aligned} \tag{2.5}$$

After computing the adjusted adjacency matrix $\hat{A} = \hat{W}_1^{-1/2} A \hat{W}_2^{-1/2}$, we perform singular value decomposition on $\hat{A}$ and obtain the singular vectors $\hat{U}$ and $\hat{V}$. In the second stage, we apply the k-means algorithm to the rows of the concatenated matrix $T \in \mathbb{R}^{n \times K}$, with

$$T = \begin{bmatrix} \hat{W}_1^{1/2} \hat{U} \\ \hat{W}_2^{1/2} \hat{V} \end{bmatrix}. \tag{2.6}$$

We define the population version $\mathcal{T} \in \mathbb{R}^{n \times K}$, with

$$\mathcal{T} = \begin{bmatrix} W_1^{1/2} \tilde{\mathcal{U}} \\ W_2^{1/2} \tilde{\mathcal{V}} \end{bmatrix} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \mathcal{X} = Z \mathcal{X}, \tag{2.7}$$

where $\tilde{\mathcal{U}} \in \mathbb{R}^{n_1 \times K}$ and $\tilde{\mathcal{V}} \in \mathbb{R}^{n_2 \times K}$ are the left and right singular vectors of $\tilde{P} \in \mathbb{R}^{n_1 \times n_2}$ respectively, $\mathcal{X} \in \mathbb{R}^{K \times K}$ is the singular vector matrix of B, and $Z = [Z_1^\intercal Z_2^\intercal]^\intercal \in \mathbb{R}^{n \times K}$. Let $(\hat{Z}, \hat{X})$ be a $(1+\varepsilon)$-approximate solution to the k-means problem and $\bar{T} = \hat{Z}\hat{X}$. Then, we can define the mis-clustered nodes similarly as in the first stage. If the observed centroid $\hat{X}_{i*}$ corresponding to node i is closer to the population centroid $\mathcal{X}_{i*}$ than any other observed centroids $\hat{X}_{j*}$ for $j \neq i$, then node i is correctly clustered.

Definition II.10. Define the set of matched mis-clustered nodes of both types as $\mathcal{S} = \{i : \|\bar{T}_{i*} - \mathcal{T}_{i*}\mathcal{Q}_2\| \geq 1/\sqrt{2}\|\}$ for an orthonormal matrix $\mathcal{Q}_2 \in \mathbb{R}^{K \times K}$.

Then our main result provides an upper bound on the matched mis-clustering error rate for MSBM with (Z_1, Z_2, B) in terms of model parameters.

Theorem II.11. *Let $A \in \{0,1\}^{n_1 \times n_2}$ be a bipartite adjacency matrix generated from a MSBM (Z_1, Z_2, B). Assume that $P = Z_1 B Z_2^\intercal$ is of rank K and $B = \alpha_n B_0$ for some $\alpha_n \geq \log n / n$ with B_0 having minimum absolute singular value $\lambda_K > 0$ and $\max_{s,t} B_{0,st} = 1$. Let $\hat{Z}$ be the output of spectral clustering using the $(1+\varepsilon)$-approximate k-means algorithm. Assume the clustering error from the first stage*

satisfies Assumption II.8. For an absolute constant $c > 0$, with probability at least $1 - n^{-1}$, we have

$$|\mathcal{S}|/n \leq c_1 \frac{n_{max} K}{n} K\beta \left(1 + \sqrt{c_2 \frac{1}{\lambda_K^2}(K\beta + 3)}\right)^2 + c_3(\beta + 3)\beta, \tag{2.8}$$

where $\beta = c^{-1}(2+\varepsilon) r \frac{Kn}{n_{1,min} n_{2,min} \lambda_K^2 \alpha_n}$, $n_{max} = \max(n_{1,max}, n_{2,max})$, $n_{min} = \min(n_{1,min}, n_{2,min})$, $c_1 = 2^2/(1-C_1)^2$, $c_2 = 2^6(2+\varepsilon)$, $c_3 = 2^2(2+\varepsilon)$, *and* $r = \max(\frac{n_{1,max}}{n_{1,min}}, \frac{n_{2,max}}{n_{2,min}})$.

The proof of Theorem II.11 is given in the appendix. Note β is an upper bound for η and a function of other parameters. The matched misclutering rate (2.8) may seem complicated but it converges to zero as long as $\frac{n_{max} K}{n} K\beta$ converges to zero. For simpler presentation, if we assume fixed $\lambda_K > 0$, we have

$$|\mathcal{S}|/n = O_p\left(\frac{K^3 n_{max}^2}{\alpha_n n_{min}^3}\right).$$

Consider the special case where $r = O(1)$ and the constant $\lambda_K > 0$. If $\alpha_n = \Omega(\log n/n)$, then $|\mathcal{S}|/n = o_p(1)$ as long as $K = o((\log n)^{1/4})$. Thus the proposed algorithm recovers communities as K is growing moderately. Note that λ_K does not change with growing K in planted partition model (i.e., simple SBM with only two parameters). Another example is when $\alpha_n = \Omega(1)$ and $K = O(1)$. In this case, $|\mathcal{S}|/n = o_p(1)$ as long as $n_{min} = o(n^{2/3})$.

The result seems to require more stringent conditions than the rate result for the SBM. For example, in Lei et al. (2015), the mis-clustering rate is $o_p(1)$ as long as K=$o((\log n)^{1/2})$ with constant λ_K and $\alpha_n = \Omega(\log n/n)$. In addition, the mis-clustering rate is $o_p(1)$ as long as $n_{min} = o(n^{1/2})$ with $\alpha_n = \Omega(1)$ and $K = O(1)$. Both examples show that stronger conditions are required for the proposed method. It is because the proposed algorithm consists of two stages where the final result depends on the accuracy of the first stage itself. If we look at where the additional

terms come from, r term is added to the first-stage rate by making Assumption II.8 to accommodate the extreme cases. The K^2 term is also added when we bound $\|\hat{P} - \tilde{P}\|_F^2$. Overall, it is understandable that the rate of a two-stage algorithm is not as sharp as that of a one-stage algorithm because of the interplay between the first-stage and the second-stage results.

2.4 Extension to the degree-corrected stochastic block model

The degree-corrected SBM (DC-SBM) (Karrer and Newman, 2011) extends the standard SBM by permitting different expected node degrees within the same community. It effectively models networks that contain "hub" nodes and other degree variations. The degree-corrected matched stochastic block model (DC-MSBM) can be naturally obtained by replacing (2.1) with the following,

$$\mathbb{E}[A_{ij}|Z] = \theta_{1,i}\theta_{2,j}B_{z_{1i}z_{2j}}, \tag{2.9}$$

where $\theta_{1,i}$ and $\theta_{2,j}$ are node degree parameters. Denote $\Theta_1 = diag(\theta_{1,1}, \cdots, \theta_{1,n_1})$ and $\Theta_2 = diag(\theta_{2,1}, \cdots, \theta_{2,n_2})$. Equation (2.9) can be re-expressed in matrix form as $P = \mathbb{E}[A|Z] = \Theta_1 Z_1 B Z_2^T \Theta_2^T$. If a bipartite network is generated from DC-MSBM, we can replace singular vectors in the first stage and the second stage with row normalized singular vectors to remove the effect of node degree heterogeneity. The specific algorithms incorporating this normalization technique are provided in the appendix.

2.5 Simulation studies

In this section, we assess the performance of the proposed algorithm by varying (1) the sparsity of the network, (2) the ratio of community sizes between the two types, and (3) the spectral gap. It is known than these parameters heavily affect the

clustering performance.

To compare performances, we use the normalized mutual information (NMI). NMI ranges from 0 to 1 where 1 means perfect match. Let $\mathbf{e}$ be the estimated membership and R be the confusion matrix where $R_{st} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} I(z_i = s, e_j = t)$. NMI is defined by Strehl and Ghosh (2002) as $NMI(z,e) = \sum_{s,t} R_{st} \log\left(\frac{R_{st}}{R_{s+}R_{+t}}\right) / \left(\sum_s R_{s+}\log R_{s+} \sum_t R_{+t}\log R_{+t}\right)$. We will use matched NMI to measure the simultaneous clustering. Results using the error rate as in (2.3) are given in the appendix.

In addition to the adjacency matrix and the Laplacian matrix, we also consider the regularized Laplacian, with $L_\tau = (D_1 + \tau I)^{-1/2} A (D_2 + \tau I)^{-1/2}$ and $\hat{L}_\tau = (\hat{D}_1 + \tau I)^{-1/2} \hat{A} (\hat{D}_2 + \tau I)^{-1/2}$. Specifically, we compare performances of the following methods: spectral clustering with adjacency matrix, two-stage spectral clustering with adjacency matrix, spectral clustering with Laplacian matrix, two-stage spectral clustering with Laplacian matrix, regularized spectral clustering, two-stage regularized spectral clustering. In regularized spectral clustering, we set τ as the average degree of the nodes in L_τ and as adjusted average degree of the nodes in $\hat{L}_\tau$. Each simulation was repeated 100 times and the average results were reported.

Simulation 1. In this simulation, we fix

$$K = 6, \quad B_0 = \frac{1}{5}\mathbf{1}\mathbf{1}^T + \frac{4}{5}\mathbf{I}_6,$$

$$\begin{bmatrix} n_{1,1} & n_{1,2} & n_{1,3} & n_{1,4} & n_{1,5} & n_{1,6} \\ n_{2,1} & n_{2,2} & n_{2,3} & n_{2,4} & n_{2,5} & n_{2,6} \end{bmatrix} = \begin{bmatrix} 100 & 100 & 100 & 500 & 500 & 500 \\ 500 & 500 & 500 & 100 & 100 & 100 \end{bmatrix}.$$

We change α_n to see how sparsity affects the performance. As can be seen from Figure 2.1, in the very sparse regime, none of the algorithms works well. However, as the network becomes denser, performances of all the algorithms improve, and the two-stage algorithms outperform the one-stage algorithms.

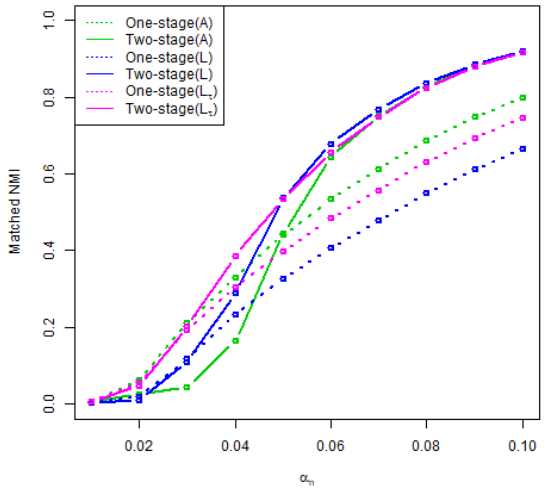


Figure 2.1: Matched NMI between true and estimated memberships as α_n varies

Simulation 2. In this simulation, we fix

$$K = 4, \quad B = \frac{1}{5}\mathbf{1}\mathbf{1}^T + \frac{4}{5}\mathbf{I}_6, \quad n_{min} + n_{max} = 1000$$

$$\begin{bmatrix} n_{1,1} & n_{1,2} & n_{1,3} & n_{1,4} \\ n_{2,1} & n_{2,2} & n_{2,3} & n_{2,4} \end{bmatrix} = \begin{bmatrix} n_{min} & n_{min} & n_{max} & n_{max} \\ n_{max} & n_{max} & n_{min} & n_{min} \end{bmatrix}.$$

We change the minimum community size, thus changing the imbalance between the two types. Identifying imbalanced communities is known to be difficult compared to balanced ones. Therefore, as n_{min} increases and the communities become more balanced, the performances of all the algorithms improve as shown in Figure 2.2. At the same time, the two-stage algorithms perform consistently better than the corresponding one-stage algorithms.

Simulation 3. In this simulation, we set

$$K = 6, \quad \alpha = 0.05, \quad B = \frac{1}{\gamma}\mathbf{1}\mathbf{1}^T + \frac{\gamma - 1}{\gamma}\mathbf{I}_6,$$

$$\begin{bmatrix} n_{1,1} & n_{1,2} & n_{1,3} & n_{1,4} & n_{1,5} & n_{1,6} \\ n_{2,1} & n_{2,2} & n_{2,3} & n_{2,4} & n_{2,5} & n_{2,6} \end{bmatrix} = \begin{bmatrix} 100 & 100 & 100 & 500 & 500 & 500 \\ 500 & 500 & 500 & 100 & 100 & 100 \end{bmatrix}.$$

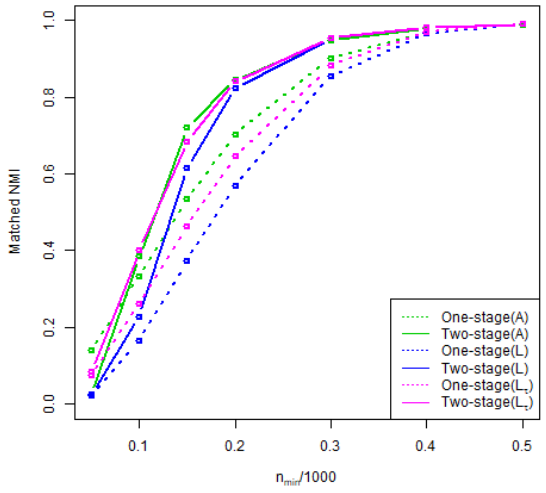


Figure 2.2: Matched NMI between true and estimated memberships as $n_{min}/1000$ varies

We investigate how the performance changes as we vary γ. Changing γ changes the out-in ratio and thus the spectral gap. In Figure 2.3, the performances of all the algorithms improve as the spectral gap increases. As the performances of one-stage algorithms improve, all the two-stage methods outperform the one-stage algorithms.

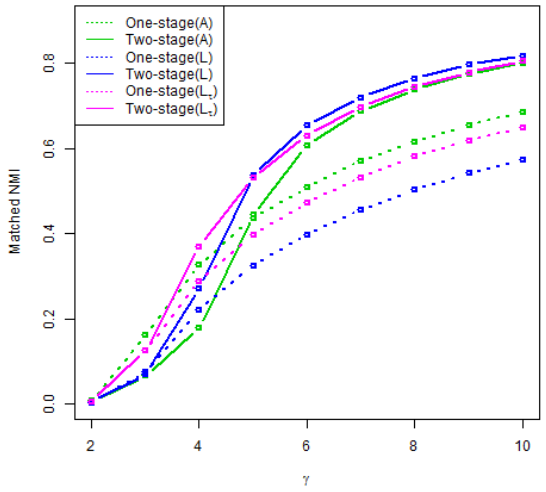


Figure 2.3: Matched NMI between true and estimated memberships as γ varies

It is observed that in all three settings, the performance of a two-stage algorithm

would improve when the performance of the corresponding first-stage algorithm was moderately good. In addition, regularization helps in challenging regimes, i.e. when the network is sparse, when the community sizes are unbalanced, or when the spectral gap is small.

2.6 Data example

In this section, we examine the performance of the proposed algorithms on a real-world data example. We consider the British Members of Parliament (MPs) twitter network curated by Greene and Cunningham (2013). The data set contains twitter interactions among 419 British Members of Parliament (MPs). Each MP belongs to one of the political parties, which include Labour, Conservative, Liberal democratic, SNP and Other. Since the sizes of SNP and Other are small (with 5 and 11 members respectively), we only focus on the three major parties. The original network is directed, with an edge from node i to node j implying that MP i follows MP j. A bipartite network is then created by considering all the MP followers in the network as one type and all popular MPs who have many followers as another type. We set a MP as popular if they are followed by at least 40 other MPs in the twitter network. The resulting bipartite network consists of 386 followers and 290 followees. Table 2.1 summarizes the number of MPs in each party for each of the two types. Note that the party sizes between the two types of nodes are proportionately different, which indicates that matched co-clustering can benefit from the proposed algorithm. We apply the two-stage spectral clustering with regularized Laplacian matrix. Because of node degree heterogeneity, we also apply the normalization technique. If we treat the parties as the ground truth community memberships, we obtained 2.7% error rate after the first stage and 1.6% after the second stage adjustment. The result

Political party	Conservative	Labour	Liberal Democratic
Community size of follower	162	183	41
Community size of followee	120	156	14
Ratio of community sizes	1.35	1.17	2.93

Table 2.1: Community size of the political party

	Error rate (%)	
Cutoff value	1-step	2-step
30	0.97	0.56
35	1.43	0.86
40	2.66	1.63

Table 2.2: Misclustering error rate for British Members of Parliament with cutoff values for popular MPs

agrees with the findings in simulation studies, i.e. the two-stage algorithm recovers communities more accurately when the community sizes are not proportional between the two types. Table 2.2 also summarizes results when different cutoff values were used for determining popular MPs.

2.7 Discussion

We have proposed a two-stage spectral clustering method for matched community detection in bipartite networks. A matched stochastic block model is proposed to define one-to-one matched communities between two types of nodes. The key component of the proposed algorithm is that it adjusts the effect of community sizes in the second adjustment stage. When the sizes of matched communities are imbalanced between the two types, the second adjustment stage can improve the performance of the first step algorithm. We derived an upper bound on the number of mis-clustered nodes for the proposed algorithm. Both simulation studies and a data example indicate that the proposed method outperforms existing methods.

The proposed method may be further extended to be applied to more general cases. In our work, we assumed that B is symmetric when we defined matched

communities. The assumption can be relaxed by modifying the definition of the matched community. It should be taken into consideration, however, that the use of algorithms based on spectral clustering becomes more challenging without the assumption. The method can also be extended so that it can be utilized in networks in which there are nodes that do not belong to any matched community. In this case, one can consider extracting matched communities from a network using a concept similar to the one introduced in Zhao et al. (2011). Finally, the work can be extended for application to networks with multiple-types of nodes, which are also known as multipartite networks.

CHAPTER III

Community Detection in Directed Networks with Individual Preferences

3.1 Introduction

Networks represent relationships among the components inside complex systems such as social networks, brain networks, and biological networks. Because the analysis of a network can provide a great deal of insight about the connections that exist inside the complex system, network analysis has been utilized in many disciplines for a long time. One of the useful network analyses is community detection, which is used to identify groups among the nodes inside a network based on the nodes' structural connectivity. Among many methods that can be used for community detection, clustering-based methods are widely used because they provide a simple data structure as the result of analysis.

Perhaps the best studied models in community detection are the stochastic block model (SBM) (Holland et al., 1983) and the variants of the SBM, such as the degree-corrected stochastic block model (DC-SBM) (Karrer and Newman, 2011) and the mixed stochastic block model (Airoldi et al., 2008). The SBM assumes a community structure, where the connectivity among nodes depends only on community memberships. See Abbe (2018) for a review on the SBM. Community detection in networks has mainly been considered and studied for undirected networks. For directed net-

works, a common approach for community detection is to first transform the network into an undirected one and then apply methods for undirected networks (Malliaros and Vazirgiannis, 2013). The use of symmetrization to remove the directionality, however, is not ideal because important information that the directionality carries is completely ignored in the process.

Both the SBM and the DC-SBM capture the average connectivity within a community and between communities, while the DC-SBM additionally considers the degree heterogeneity. Thus, these models assume that a node's preferences to communities are the same as other nodes that are in the same community. This is considered as a limitation, as there can be a considerable variation among each node's preferences to communities, even among those in the same community. In particular, for two nodes i and j in the same community, i's preference to a certain community may be stronger than j's preference. Such strong individual preferences carry important information, which may be essential for the identification of groups in networks. For example, it was observed that political blogs could link to blogs of the opposite party. One possible explanation for this is that they send links to blogs they dislike to criticize them (Rohe et al., 2016). Cantwell and Newman (2019) proposed a preference model that allows for nodes' distinct preferences and developed a Bayesian method to fit the model. They demonstrated the proposed method in real networks by showing the existance of individual nodes' different preferences given true known labels. Additionally, a few other works also proposed similar concepts. For example, Altenburger and Ugander (2018) introduced monophily to explain the overdispersion of preferences, which is similar to the concept of individual's distinct preferences, and argued that the use of the information would improve tasks such as semi-supervised learning. Peel et al. (2018) introduced a localized assortativity mea-

sure at the node level as opposed to the global assortativity measure. Sengupta and Chen (2018) proposed a popularity-adjusted block model to accommodate different individual's popularity in the community for undirected networks.

Based on the preference-based network model in Cantwell and Newman (2019), we propose a preference-based block model and develop a spectral clustering algorithm for fitting the model with a focus on directed and bipartite networks. The details are given in Section 3.2. Given the estimated communities, parameters can be estimated by maximum likelihood for inference. In Section 3.3, we show that how degree heterogeneity and preference heterogeneity affect the performance of community detection. In Section 3.4 and Section 3.5, numerical results are shown to demonstrate the performance of the proposed method and exhibit the presence of the preferences in real directed networks and bipartite networks.

3.2 Model

In this section, we introduce a preference-based block model for directed networks. A similar model has been considered by Cantwell and Newman (2019). But unlike their model, we do not assume any prior distribution for the preferences. In the proposed model, each individual can have different preferences to the communities. These distinct preferences of the nodes contain information that characterize communities.

Throughout the chapter, we use $\|\cdot\|$ to denote 2-norm of a vector and the spectral norm of a matrix. For a matrix M, $\|M\|_F$ denotes the Frobenius norm.

3.2.1 Preference-based block model

Let n denote the number of nodes and K denote the number of communities. To model a node's different preferences to communities, we introduce a parameter w_{ik},

which denotes the probability of how node i prefers group k. To deal with degree heterogeneity, degree-correction parameters θ_i and ϕ_j are introduced as in the DC-SBM (Karrer and Newman, 2011). Denote the community membership of a node i as $g_i \in \{1, \cdots, K\}$. Given nodes' community (group) labels $\mathbf{g} = (g_1, \cdots, g_n)$, the edge variables A_{ij}'s are independent Bernoulli random variables with

$$\mathbb{E}[A_{ij}] = \theta_i w_{ig_j} \phi_j. \tag{3.1}$$

Note the model is not identifiable without constraints. To ensure identifiability, we impose certain constraints; specifically, they are

$$\sum_k w_{ik} = 1 \text{ and } \sum_{j:g_j=k} \phi_j = 1 \text{ for all } k = 1, \cdots, K. \tag{3.2}$$

Thus, w_{ik} can be interpreted as node i's probability of the preference to group k. With the constraints, we have

$$\begin{aligned} \sum_{j:g_j=k} \theta_i w_{ig_j} \phi_j &= \theta_i w_{ik} \\ \sum_{j=1}^{n} \theta_i w_{ig_j} \phi_j &= \sum_{k=1}^{K} \sum_{j:g_j=k} \theta_i w_{ig_j} \phi_j = \theta_i. \end{aligned} \tag{3.3}$$

The summation over j clears away the effect of ϕ_j. The θ_i controls average outgoing degree of node i. Since w_{ik} is the probability for i's preference, $\theta_i w_{ik}$ implies how many edges from node i will go out to nodes in group k on average.

The probability matrix $P = \mathbb{E}[A]$ can be expressed as

$$P = \Theta W Z^T \Phi, \tag{3.4}$$

where $\Theta \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal elements $\theta = (\theta_1, \cdots, \theta_n)$, $\Phi \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal elements $\phi = (\phi_1, \cdots, \phi_n)$, $W \in \mathbb{R}^{n \times K}$ is a preference matrix where ith row is $w_i = [w_{i1}, \cdots, w_{ik}] \in \mathbb{R}^K$, and $Z \in \mathbb{R}^{n \times K}$ is a

membership matrix, with each row containing one 1 and $(K-1)$ zeros. If we denote the jth row of Z as z_j, then $z_{j,g_j} = 1$ and zeros for other elements. In Section 3.2.2, we will show how we recover the Z matrix.

The model we introduce is similar to the model considered by Cantwell and Newman (2019) while we impose different constraints for the parameters θ_i and ϕ_j. In addition, we treat each node's preferences w_i as parameters whereas Cantwell and Newman (2019) treated w_i as random variables generated from a Dirichlet prior distribution that depends on node i's community. Consequently, they used K^2 parameters for K Dirichlet distributions of communities, while we use $n(K-1)$ parameters for the preference parameters. However, if we are concerned about community detection, we do not need to estimate the w_i parameters. The influence of w_i parameters will be seen in the overall block quantity as explained in Section 3.3. After we identify community memberships, we may estimate the preference parameters, but it should be noted that for nodes with high degrees, such estimates may have low estimation errors, while for sparse degree nodes, the estimate of w_i tends to have a relatively large estimation error.

A natural extension of the preference-based block model is to switch the direction of the preference. For certain mechanisms in directed networks, if an individual receiver node attracts the edges from the communities, that attractiveness to the sending group might be considerably different for each receiver node. In other words, sender nodes have the community structure and each receiving node has different attractiveness to the communities. This is just reversing the sender nodes and receiver nodes in the preference-based block model. In reality, there might be both preference and attractiveness. However, we feel that the node's preference effect is probably stronger than attractiveness. In addition, it is reasonable to assume

that nodes taking an active role in making edges have relatively higher preference heterogeneity.

Another extension is to bipartite networks. Directed networks can be considered as a special case of bipartite networks if we think of sender nodes as type-1 nodes and receiver nodes as type-2 nodes. In a preference-based block model for bipartite networks, one type of nodes is considered to have preferences to the group of another type of nodes. This concept is closely related to recommender systems even though the purpose of recommender system is to give suggestions to users (Ricci et al., 2011). A common approach in bipartite networks when we are interested in the community structure within a type of nodes is to apply standard community detection algorithms to an one-mode projected network. However, projection loses the information about the preferences and can be less informative depending on their weighting scheme (Zhou et al., 2007). The preference-based model on the other hand provides more meaningful and interpretable results for bipartite networks. A data example in Section 3.5.2 illustrates the use of the model in bipartite networks.

3.2.2 Estimation

In many real-world networks, community labels are often unknown. Finding hidden community structures in networks has been an important problem. Once we have community labels, the estimation of Θ, W and Φ under the preference-based block model is straightforward via maximum likelihood. However, if community labels are unknown, we need to first identify communities, before the estimation of parameters. To estimate unknown community labels, we apply spectral clustering to the regularized Laplacian matrix as suggested in Rohe et al. (2016).

First, we make several assumptions for the parameters.

Assumption III.1. $\theta_i > 0$ *and* $\phi_j > 0$ *for all* $i = 1, \cdots, n$ *and* $j = 1, \cdots, n$

In Assumption III.1, $\phi_j > 0$ is necessary to ensure node j can be chosen by other nodes so that we can infer node j's community. $\phi_j > 0$ is assumed that every node has some preferences. Assumption III.1 assures to exclude the nodes that are not probabilistically connected to other nodes.

Assumption III.2. *There is at least one node in each community. In addition, at least* K *rows of* X *are distinct.*

Assumption III.2 is needed to ensure the rank of W is K. This is a sufficient condition that rules out the scenario in which we cannot identify communities from the $WZ^\intercal$ matrix. In certain cases, W does not need to be rank K so to identify communities. For example, if w_{ik} is different for all k for some i, we can identify communities from the probability matrix. However, we will opt out very special cases in this section for simplicity.

Both the Laplacian matrix L and the adjacency matrix A have been commonly used for community detection in networks. It has been shown that the concentration of an adjacency matrix's eigenvectors and of a Laplacian matrix's eigenvectors show the same rate of convergence. However, using the Laplacian matrix may outperform using the adjacency matrix over broader regimes (Sarkar et al., 2015). We choose to use the regularized Laplacian matrix. Using regularized version of the Laplacian matrix or adjacency matrix helps with sparse networks and degree heterogeneous networks (Qin and Rohe, 2013; Amini et al., 2013; Le et al., 2017). The regularized graph Laplacian $L_\tau \in \mathbb{R}^{n \times n}$ (Chaudhuri et al., 2012) can be defined as

$$L_\tau = (D_l + \tau I)^{-1/2} A (D_r + \tau I)^{-1/2},$$

where $D_l \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal elements $D_{l,ii} = d_i = \sum_j A_{ij}$

and $D_r \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal elements $D_{r,jj} = \sum_i A_{ij}$.

We can also define the population version of the regularized Laplacian matrix $\mathcal{L}_\tau$

$$\mathcal{L}_\tau = (\mathcal{D}_l + \tau I)^{-1/2} \mathcal{P} (\mathcal{D}_r + \tau I)^{-1/2},$$

where $\mathcal{D}_l \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal elements $D_{l,ii} = \sum_j P_{ij}$ and $\mathcal{D}_r \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal elements $D_{r,jj} = \sum_i P_{ij}$.

Note the observed L_τ matrix is a perturbed version of $\mathcal{L}_\tau$ and the singular vectors of L_τ converge to those of $\mathcal{L}_\tau$ under certain conditions, and since the singular vectors of $\mathcal{L}_\tau$ often contain community structures of the network, the singular value decomposition of L_τ also reveals the community structure. The following lemma explains the structure of the singular vectors of $\mathcal{L}_\tau$ under the preference model, and we also introduce a matrix $\mathcal{H}$ that has the same singular values as $\mathcal{L}_\tau$.

Lemma III.3. *SVD-structure of the population matrix $\mathcal{L}_\tau$. Consider a preference model (Θ, W, Z, Φ) with K communities. Let $\mathcal{U}D\mathcal{V}^T$ be the singular value decomposition of $\mathcal{L}_\tau \in \mathbb{R}^{n\times n}$. Then, we have*

1. *$\mathcal{V} = \tilde{\Phi} Z \tilde{\Psi}^{-1} \mathcal{C}$ for some $\mathcal{C} \in \mathbb{R}^{K\times K}$ where $\tilde{\Phi} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal element $\tilde{\Phi}_{jj} = \phi_j / \sqrt{\phi_j \sum_i \theta_i w_{ig_j} + \tau}$ and $\tilde{\Psi} = (Z^\intercal \tilde{\Phi}^2 Z)^{1/2} \in \mathbb{R}^{K\times K}$.*
2. *The Laplacian matrix $\mathcal{L}$ and the matrix $\mathcal{H} = \tilde{\Theta} W \tilde{\Psi} \in \mathbb{R}^{n\times K}$ have the same singular values where $\tilde{\Theta}$ is a diagonal matrix with diagonal element $\tilde{\Theta}_{ii} = \theta_i / \sqrt{\theta_i + \tau}$.*

Lemma III.3 shows that the right singular vectors of $\mathcal{L}_\tau$ under the model (3.1) contain the community information. The diagonal matrices $\tilde{\Phi}$ and $\tilde{\Psi}$ do not play an important role, since row normalization for $\mathcal{V}$ can filter out the effects of $\tilde{\Phi}$ and $\tilde{\Psi}$. Row normalization of singular vectors is a common technique for degree

heterogeneous networks, and its application to $\mathcal{V}$ results in $\mathcal{V}^* = Z\mathcal{C}$. For nodes in the same community, their corresponding rows in $\mathcal{V}^*$ will be the same, and for nodes in different communities, their corresponding rows in $\mathcal{V}^*$ will be different. Based on this observation, we propose to estimate community labels using the right singular vectors with spectral clustering for the preference model.

Estimation of model parameters is based on the recovered community labels. For the likelihood, we use the Poisson distribution rather than Bernoulli for simpler technical derivations as in Cantwell and Newman (2019) and Zhao et al. (2012). It has been discussed in the literature that replacing the Bernoulli distribution with Poisson yields almost no difference in sparse networks while enjoying simpler form for the estimator. Given community labels, the log-likelihood function can be written as

$$\sum_{i=1}^{n}\sum_{j=1}^{n} -\theta_i w_{ig_j}\phi_j + A_{ij}\log(\theta_i w_{ig_j}\phi_j) - \log(A_{ij}!). \tag{3.5}$$

Maximizing the log-likelihood using the approximated Poisson distribution gives us $\hat{\theta}_i = d_i$ and $\hat{w}_{ik} = d_{ik}/d_i$, where $d_i = \sum_j A_{ij}$ denotes the out-degree of node i and $d_{ik} = \sum_{j\in g_k} A_{ij}$ denotes the number of node i's outgoing edges to the community k. In addition, $\hat{\phi}_j = \sum_i A_{ij}/(\sum_{j':g_{j'}=s}\sum_i A_{ij'})$ when $g_j = s$, which is the proportion of in-degree of node j in the total in-degrees of the community where the node belongs to.

Overall, the algorithm proceeds as the following.

Algorithm III.1 Community Detection and Estimation of W and Θ

1: Input: directed adjacency matrix $A \in \{0,1\}^{n\times n}$
2: Form (regularized) graph Laplacian $L_\tau = (D_l + \tau I)^{-1/2} A (D_r + \tau I)^{-1/2}$. Let $L_\tau = U\Lambda V^T$ be the SVD of L_τ.
3: Use K right singular vectors corresponding to the K largest singular values. Normalize each row of V to have unit length. Denote the normalized version of V as V^*. Run k-means on rows of V^* and assign each node to one of K communities.
4: Given estimated community memberships, estimate θ_i and w_{is} as d_i and d_{is}/d_i.

3.2.3 Connections to other models

The preference-based block model is closely related to a degree-corrected stochastic block model (DC-SBM), proposed by Karrer and Newman (2011). The DC-SBM provides a simple way to incorporate community structures by modeling connectivity between nodes, which only depends on community labels of the nodes and nodes' degree heterogeneity. The original DC-SBM, which was for undirected networks, can be easily extended to directed and bipartite networks as in Larremore et al. (2014). The DC-SBM for directed networks or bipartite networks aims to summarize the network structure into a $K_1 \times K_2$ low-dimensional block structure $B \in \mathbb{R}^{K_1 \times K_2}$ for connectivity, where K_1 and K_2 are the numbers of communities for sending nodes (type-1 nodes) and receiving nodes (type-2 nodes) respectively. Analogous to (3.1), the model for DC-SBM can be written as

$$\mathbb{E}[A_{ij}] = \theta_i B_{g_{l,i} g_{r,j}} \phi_j, \tag{3.6}$$

where $g_{1,i}$ is the community label of the sender node i and $g_{2,j}$ is the community label of the receiver node j. The main difference between DC-SBM and the preference-based block model lies in $B_{g_{l,i} g_{r,j}}$ and $w_{i g_{r,j}}$. In the preference model, the case when $w_{i g_{r,j}} = B_{g_{l,i} g_{r,j}}$ implies that the DC-SBM is a special case of the preference model in which the preference of the nodes in the same community are the same. Another way to think about it is that the proposed preference model is an extreme case of the DC-SBM where each sender node can be considered as a community, i.e. $K_1 = n$. Therefore, they are similar and one can be a special case of the other with certain constraints or more flexibility. In general, the preference model is much more flexible than the DC-SBM as in DC-SBM it is often the case that $K_1 \ll n$. The number of parameters of $w_{i g_{r,j}}$ for the preference model is $n(K_1 - 1)$, while the

number of parameters of $w_{g_{l,i}g_{r,j}}$ for DC-SBM with the same constraint $\sum_{t=1}^{K} w_{g_{l,i}t} = 1$ is $K_1(K_2 - 1)$. As we will see in the next section, heterogeneous preferences of the nodes actually help us in identifying the community structure given the same average connectivity. Unlike the DC-SBM, in the preference-based block model, the community structure of the sender (type-1) nodes is not of the interest.

Overlapping SBMs, which is an another variant of the SBM, also have some similarity with the preference-based block model. In the overlapping SBM, each node can have several community labels or a continuous label, which might be more sensible for real-world networks. There have been several studies for overlapping SBMs (Zhang et al., 2014; Airoldi et al., 2008; Ball et al., 2011). The preference-based block model and the overlapping SBM have some commonality in that each individual node can have distinct continuous preferences toward each community. However, most of overlapping SBMs estimate overlapping community memberships for an undirected network, while the preference-based block model estimates non-overlapping community memberships for one type of nodes and then estimates the preferences of each node using the estimated memberships for directed or bipartite networks. Non-overlapping community memberships for one type of nodes and the preferences for the other type makes the interpretation easier. The concept of preference in our model is analogous to overlapping community memberships. For example, one may interpret each node's propensity to communities as individual's preferences in directed or bipartite networks.

3.3 Theoretical results

In this section, we show theoretical properties of the proposed algorithm under the preference-based block model. The bounds on the number of mis-clustered nodes

can be obtained by adapting the results for undirected networks. This can be done by utilizing a larger block matrix $[0\ A; A^\intercal\ 0] \in \mathbb{R}^{2n\times 2n}$ for directed adjacency matrix $A \in \mathbb{R}^{n\times n}$ and then applying the results accordingly. See Rohe et al. (2016); Zhou and Amini (2019) for technical details for similar analyses. In this section, we will focus on directed networks. Generalization to the bipartite network is straightforward.

We first introduce definitions for average connectivity between and within communities and a community's variation from the average connectivity.

Definition III.4. Let $M \in \mathbb{R}^{K\times K}$ where $M_{rs} = \frac{1}{n_r}\sum_{i\in g_r}\tilde{\theta}_i w_{is}$ for $r, s = 1, \cdots, K$ and $S^{(r)} \in \mathbb{R}^{K\times K}$ for $r = 1, \cdots, K$ where $S^{(r)}_{st} = \sum_{i\in g_r}\left(\tilde{\theta}_i w_{is} - M_{rs}\right)\left(\tilde{\theta}_i w_{it} - M_{rt}\right)$ for $s, t = 1, \cdots, K$ where $\tilde{\theta}_i = \theta_i/\sqrt{\theta_i + \tau}$.

The matrices M and S are functions of w_{is}'s and $\tilde{\theta}_i$'s, which summarize connectivity between communities. The quantity M_{rs} is the average of $\tilde{\theta}_i w_{is}$ over the nodes in community r, which involves both transformed degree parameter $\tilde{\theta}_i$ and the preference parameter w_{is}. In other words, we can interpret M_{rs} as the average edge connectivity from community r to community s. M gives the average connectivity information for networks, which is often considered as the main signal in conventional networks literature. $S^{(r)}_{st}$ measures joint variability between the preferences to community s and to community t, from community r, which is related to the heterogeneity and the variation.

Recall in Lemma III.3, we introduced $\mathcal{H}$ that has the same singular values as $\mathcal{L}_\tau$. The decomposition of $\mathcal{H}^\intercal\mathcal{H}$ provides insight on the role of individual preferences in networks, and $\mathcal{H}^\intercal\mathcal{H}$ contains all the information about the right singular vectors and singular values of $\mathcal{L}_\tau$.

Lemma III.5. *Consider model (3.1). Let* $\mathcal{H} = \tilde{\Theta}W\tilde{\Psi}$ *with* $\tilde{\Theta}$ *being a diagonal*

matrix with $\tilde{\Theta}_{ii} = \theta_i/\sqrt{\theta_i + \tau}$ *and* $\tilde{\Psi}$ *being a diagonal matrix with* $\tilde{\Psi}_{tt} = \sum_{j \in g_t} \phi_j / \sqrt{\phi_j \sum_i \theta_i w_{ig_j} + \tau}$. *Then, we have*

1. $\mathcal{H}^\intercal \mathcal{H} = \mathcal{H}_M^\intercal \mathcal{H}_M + \mathcal{H}_S^\intercal \mathcal{H}_S$ *where* $\mathcal{H}_M = \left(Z(Z^\intercal Z)^{-1} Z^T\right) \tilde{\Theta} W \tilde{\Psi}$ *and* $\mathcal{H}_S = (I - Z(Z^\intercal Z)^{-1} Z^\intercal)\, \tilde{\Theta} W \tilde{\Psi}$.

2. $\mathcal{H}_M^\intercal \mathcal{H}_M = \tilde{\Psi} M^\intercal N M \tilde{\Psi}$ *and* $\mathcal{H}_S^\intercal \mathcal{H}_S = \tilde{\Psi} \sum_{r=1}^K S^{(r)} \tilde{\Psi}$ *for a diagonal matrix* $N \in \mathcal{R}^{K \times K}$ *with* $N_{rr} = n_r$.

3. *Let* $\sigma_k(B)$ *be the* k*th largest singular value of a matrix* B. *Then, we have*

$$\sigma_K(\mathcal{H}_M^\intercal \mathcal{H}_M) + \sigma_K(\mathcal{H}_S^\intercal \mathcal{H}_S) \leq \sigma_K(\mathcal{H}^\intercal \mathcal{H}) \leq \sigma_K(\mathcal{H}_M^\intercal \mathcal{H}_M) + \sigma_1(\mathcal{H}_S^\intercal \mathcal{H}_S).$$

Lemma III.5 shows that $\mathcal{H}^\intercal \mathcal{H}$ can be decomposed into two parts $\mathcal{H}_M^\intercal \mathcal{H}_M$ and $\mathcal{H}_S^\intercal \mathcal{H}_S$, which measure the average connectivity and the joint variability respectively. It also shows that $\mathcal{H}_M^\intercal \mathcal{H}_M$ and $\mathcal{H}_S^\intercal \mathcal{H}_S$ can be expressed with M and $S^{(r)}$'s in Definition III.4. In the planted partition model, which is a special type of the SBM, it is well known that the connectivity ratio between within communities and across communities plays an important role for community detection. Analogously, the joint variability contributes to the concentration of singular values given a fixed M. Intuitively, if there are more variation within community, S will contribute to the signal. The inequality in Lemma III.5 shows that the singular values of $\mathcal{H}_S^\intercal \mathcal{H}_S$ make a difference between $\sigma_K(\mathcal{H}_M^\intercal \mathcal{H}_M)$ and $\sigma_K(\mathcal{H}^\intercal \mathcal{H})$. If $\sigma_K(\mathcal{H}_M^\intercal \mathcal{H}_M)$ is fixed, increasing the smallest singular value of $\mathcal{H}_S^\intercal \mathcal{H}_S$ will also increase the lower bound. In the extreme case such as the SBM, there will be no variation of preferences at all. Thus, $\sigma_K(\mathcal{H}_S^\intercal \mathcal{H}_S)$ will be zero. Since the smallest singular values are important for singular vector concentration, we define λ_M and λ_S as follows.

Definition III.6. Define $\lambda_M = \sigma_K(\mathcal{H}_M^\intercal \mathcal{H}_M)$ and $\lambda_S = \sigma_K(\mathcal{H}_S^\intercal \mathcal{H}_S)$.

Assumption III.7. *Assume* $\lambda_M > 0$ *or* $\lambda_S > 0$.

Assumption III.7 is a sufficient condition for $\sigma_K(\mathcal{H}^\intercal\mathcal{H}) > 0$. Note that in the SBM and DC-SBM literature that consider the average connectivity, $\lambda_M > 0$ is often assumed. Next, we provide the concentration of singular vectors as a result of the concentration of the regularized Laplacian (Le et al., 2017).

Theorem III.8. *Consider an adjacency matrix* A *generated from the preference-based block model with* K *communities. Let* $d = n\max_{ij} p_{ij}$ *and* $m_r = \min_{j=1,\cdots,n} \|\mathcal{V}_j\|_2$. *Define* $\mathcal{L}_\tau$ *and* $\mathcal{H}$ *as in Lemma III.5. Choose a number* $\tau > 0$. *Then, for any* $r \geq 1$, *with probability at least* $1 - e^{-r}$, *we have*

$$\sqrt{\|U - \mathcal{U}\mathcal{Q}\|_F^2 + \|V - \mathcal{V}\mathcal{Q}\|_F^2} \leq \frac{4Cr^2\sqrt{K}}{\sqrt{\tau(\lambda_M + \lambda_S)}}\left(1 + \frac{d}{\tau}\right)^{5/2}$$

for some orthonormal matrix $\mathcal{Q} \in \mathbb{R}^{n\times n}$.

The presence of an orthogonal matrix $\mathcal{Q}$ is from the orthogonal Procrustes problem that solves $\min_{\mathcal{Q}} \|V - \mathcal{V}\mathcal{Q}\|$. The orthogonal matrix $\mathcal{Q}$ deals with the situation where the singular vectors are only determined up to rotations when some singular values have multiplicities. Theorem III.8 shows the convergence of the singular vectors. The upper bounds for both left and right singular vectors are inversely proportional to $\lambda_M + \lambda_S$.

For a degree heterogeneous network, row normalization is applied to the right singular vector V to remove the effect of heterogeneous degrees. Many studies have used row normalization of singular vectors in spectral clustering for DC-SBM (Lei et al., 2015; Jin et al., 2015). We normalize each row of V to have unit length. Denote V^* as the row normalized version of V, and $\mathcal{V}^*$ as the row normalized version of $\mathcal{V}$ which is the population version of V^*. We can check $\mathcal{V}^* = Z\mathcal{C}$ for some $\mathcal{C} \in \mathbb{R}^{K\times K}$

in Lemma III.3. Thus, nodes in the same community have the same row vectors in the normalized right singular vector matrix $\mathcal{V}^*$.

Applying the k-means algorithm to singular vectors is an important step of spectral clustering after singular value decomposition. The k-means algorithm minimizes $\|ZC - V^*\|_F^2$ over all $Z \in \mathbb{M}^{n\times K}$ and $C \in \mathbb{R}^{K\times K}$. Since solving the k-means problem is NP-hard, we consider the efficient approximate k-means algorithm Kumar et al. (2004), which is known to provide a solution $(\hat{Z}, \hat{C}) \in \mathbb{M}^{n\times K} \times \mathbb{R}^{K\times K}$ such that

$$\|\hat{Z}\hat{C} - V^*\|_F^2 \leq (1+\alpha) \min_{Z,C} \|ZC - V^*\|_F^2, \tag{3.7}$$

where $(\hat{Z}, \hat{C})$ is referred as a $(1+\alpha)$-approximate solution to the k-means problem and we denote $\bar{V}^* = \hat{Z}\hat{C}$. We define mis-clustered node set $\mathcal{S}$ as in Rohe et al. (2011) and Rohe et al. (2016). Since the observed singular vectors can be considered as noisy version of the population singular vectors, a node whose observed singular vector is far from the corresponding population singular vector can be considered as mis-clustered.

Definition III.9. Define the mis-clustered node set as $\mathcal{S} = \{i \in G(Z) : \|\bar{V}_i^* - \mathcal{V}^*\mathcal{Q}\| > 1/\sqrt{2}\}$.

If the corresponding centroid of a node as a result of the k-means algorithm is distant from the corresponding population centroid of the node, we define that node as mis-clustered. For mis-clustered nodes, we only consider the right singular vectors since it is not guaranteed for the left singular vectors to contain community structures in the preference model. We define $m_r = \min_{j=1,\cdots,n} \|\mathcal{V}_j\|_2$ to take into account the effect of low degree nodes after row normalization. Note if the in-degrees of receiving nodes are small, the clustering problem is difficult.

Theorem III.10. *Consider an adjacency matrix A generated from the preference-based block model with K communities. Let $d = n\max_{ij} p_{ij}$ and $m_r = \min_{j=1,\cdots,n} \|\mathcal{V}_j\|_2$ where $\mathcal{V}_j$ is the jth row of $\mathcal{V}$. Define $\mathcal{L}_\tau$ and $\mathcal{H}$ as in Lemma III.5. Choose a constant $\tau > 0$. Then, for any $r \geq 1$ there exists a constant $C(\alpha) > 0$ such that*

$$|\mathcal{S}|/n \leq \frac{C(\alpha) r^4 K}{n m_r^2 (\lambda_M + \lambda_S)\tau} \left(1 + \frac{d}{\tau}\right)^5.$$

with probability at least $1 - e^{-r}$.

Since λ_M and λ_S also depend on τ, it is not straightforward to look at the effect of λ_M and λ_S apart from τ. If we choose $\tau \sim d$, then $|\mathcal{S}|/n = O_p\left(\frac{K}{n m_r^2 (\lambda_M + \lambda_S) d}\right)$. If $\|\mathcal{V}_j\|_2$ is the same for all j, then $\|\mathcal{V}_j\|_2 = \sqrt{K/n}$ for all j due to $\|\mathcal{V}\|_F^2 = n$. Thus, if m_r is of order $\sqrt{K/n}$, $|\mathcal{S}|/n$ goes to zero as d increases. To provide an insight on the effect of λ_M and λ_S, we provide two examples below.

Example III.11. Let $A \in \mathbb{R}^{n\times n}$ be an adjacency matrix generated from the preference model with $\theta_i = c$ and $K = 2$ communities. Define $\mathcal{L}_\tau$, $\mathcal{H}_M$ and $\mathcal{H}_S$ as in Lemma III.5. Then, we have $\left(\mathcal{H}_M^T \mathcal{H}_M\right)_{s,t} = C_\tau^2 \tilde{\psi}_{ss} \left(\sum_{r=1}^K n_r \bar{w}_{rs} \bar{w}_{rt}\right) \tilde{\psi}_{tt}$ where $\bar{w}_{rs} = \frac{1}{n_r}\sum_i w_{is}$ for $C_\tau = c/\sqrt{c+\tau}$. After straightforward calculation, we have

$$\mathcal{H}_S^T \mathcal{H}_S = \tilde{\Psi} \begin{bmatrix} v & -v \\ -v & v \end{bmatrix} \tilde{\Psi}, \tag{3.8}$$

where $v = \sum_{i:g_i=1}(w_{i1} - \bar{w}_{11})^2 + \sum_{i:g_i=2}(w_{i1} - \bar{w}_{21})^2$. We can compare the effect from the variation of preferences given the same average preference and other parameters being fixed. Given $\bar{w}_{rs}$'s and ϕ_j's fixed, the lower bound for λ_S is zero if there is no variation in preferences within the same block (i.e. $w_{is} = w_{js}$ are the same if $g_i = g_j$), which is the special case of the DC-SBM assuming expected out-degrees are the same. As v increases, this will also increase the singular values of $\mathcal{H}_S^T \mathcal{H}_S$.

Therefore, given the same average preference in the community, the variation of the preferences makes community detection easier compared to no variation.

Example III.12. Let $A \in \mathbb{R}^{n\times n}$ be an adjacency matrix generated from the DC-SBM with K communities, which is a special case of the preference model. In the preference model, the DC-SBM is the case when $w_{is} = B_{rs}$ if $g_i = r$ for some block matrix $B \in \mathbb{R}^{K\times K}$. Then, we have

$$\mathcal{H}_M^T \mathcal{H}_M = \tilde{\Psi} B^\intercal L^{(M)} B \tilde{\Psi}$$

$$\mathcal{H}_S^T \mathcal{H}_S = \tilde{\Psi} B^\intercal L^{(S)} B \tilde{\Psi}$$

where $L^{(M)} \in \mathbb{R}^{K\times K}$ is a diagonal matrix with diagonal element $L^{(M)}_{ss} = n_s(\bar{\theta}_s)^2$ and $L^{(S)} \in \mathbb{R}^{K\times K}$ is a diagonal matrix with diagonal element $L^{(S)}_{ss} = \sum_{i:g_i=s}(\tilde{\theta}_i - \bar{\theta}_s)^2$ where $\bar{\theta}_r = \frac{1}{n_r}\sum_{i:g_i=r}\tilde{\theta}_i$ is the average $\tilde{\theta}_i$ of community r. It is clear that the variation of the out-degrees within each community contributes to the magnitude of the singular values of $\mathcal{H}_S^T\mathcal{H}_S$. Given other parameters the same except for the variation of the out-degree while we keep the average of out-degree within each community the same, larger variations of the out-degree will make community detection under the DC-SBM easier if we focus on the receiving nodes' information.

3.4 Simulation studies

In this section, we use simulation studies to illustrate how the variation of preferences and the variation of out-degrees affect the performance of spectral clustering. The performances of three spectral algorithms are reported: spectral clustering on the right singular vectors (SC-RSV), spectral clustering on the left singular vectors (SC-LSV), and spectral clustering for symmetrized network (SC-SYM). SC-RSV is the algorithm we propose for community detection under the preference model. SC-LSV is the algorithm similar to SC-RSV but with left singular vectors. For SC-SYM,

the k-means algorithm is applied to normalized singular vectors that are obtained from normalized Laplacian with symmetrized adjacency matrix. For symmetrized adjacency matrix A', we set $A'_{ij} = A'_{ji} = 1$ if either $A_{ij} = 1$ or $A_{ji} = 1$. We can check $\mathbb{E}[A'_{ij}] = p_{ij} + p_{ji} - p_{ij}p_{ji}$. SC-SYM is provided as the most common practice to deal with directed networks (Malliaros and Vazirgiannis, 2013).

To compare the performance, we use the error rate defined as

$$Error(\mathbf{z}, \hat{\mathbf{z}}) = n^{-1} \min_{\sigma} \sum_{i=1}^{n} I(z_i \neq \sigma(\hat{z}_i)), \tag{3.9}$$

where z_i is the true label for node i, $\hat{z}_i$ is the estimated label for node i, and σ is a permutation function.

To examine the effect of individual preference heterogeneity and individual degree heterogeneity, we present several simulations in this section. In Simulation 1, we investigate how preference variations affect the error rate of mis-clustered nodes given the same expected out-degree for all the nodes. In Simulation 2, we change the variation of out-degrees in the community while we fix the variation of preferences in the community. The effect of the variation of preferences while the out-degrees are heterogeneous is also examined (Simulation 3) and the results are given in the appendix. Overall, SC-RSV outperforms SC-LSV and SC-SYM when the variation of the preferences or out-degrees becomes large. Under the preference model, the right singular vectors are supposed to give correct information while the left singular vectors do not necessarily provide correct information as the variation of the preferences becomes large. All simulations were repeated 100 times for given parameters and the average results were reported.

Simulation 1

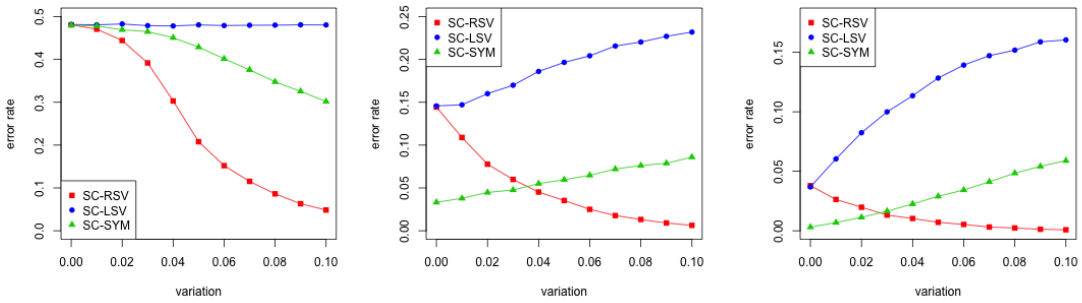
In this simulation, we vary the variation of the preferences given the same average preference within each block while we fix the other parameters. We generate networks

from the preference model with $K = 2$ and $n = 400$. Each community size is 200. To simplify the problem, we set the out-degrees to be the same for all sending nodes and in-degrees to be the same for all receiving nodes. To control the preference variation, we generated $(w_{i1}, w_{i2}) \sim Dirichlet(\gamma s, (1-\gamma)s)$ if $g_i = 1$ and $(w_{i1}, w_{i2}) \sim Dirichlet((1-\gamma)s, \gamma s)$ if $g_i = 2$, where γ is the community's average preference to it's own community. In the generation of $\mathbf{w}_i$'s, s controls the variation of the preferences. The role of γ and s is more clear if we look at $\mathbb{E}[W^\intercal W]$, which is given by

$$\frac{1}{n}\mathbb{E}[W^\intercal W] = \frac{1}{n}\left(\mathbb{E}[W]^\intercal \mathbb{E}[W]\right) + \frac{1}{n}\left(\mathbb{E}[W^\intercal W] - \mathbb{E}[W]^\intercal \mathbb{E}[W]\right)$$
$$\begin{bmatrix} \frac{1}{2} - \gamma(1-\gamma) & \gamma(1-\gamma) \\ \gamma(1-\gamma) & \frac{1}{2} - \gamma(1-\gamma) \end{bmatrix} + \frac{\gamma(1-\gamma)}{s+1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

When $s \to \infty$, it is equivalent to the SBM, which means there is no variation of the preferences. In this case, we set $\mathbf{w}$ in the same community the same instead of generating it from a Dirichlet distribution. As s decreases, the variation of the preferences increases. We can calculate the smallest singular value for the variation of the preferences as $\sigma_2(\mathbb{E}[W^\intercal W] - \mathbb{E}[W]^\intercal \mathbb{E}[W]) = 2\gamma(1-\gamma)/(s+1)$. We set $\gamma = \{1/2, 3/4, 5/6\}$ and set the variation term $\gamma(1-\gamma)/(s+1) \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\}$ by controlling s for fixed γ.

The results for Simulation 1 are shown in Figure 4.1. The variation of the preferences is on the $x-$axis and the error rate is on the $y-$axis. Three different colors correspond to different spectral clustering algorithms. When the out-degree is 8, as expected, the error rate of SC-RSV is always lower than that of SC-LSV under the preference model. SC-RSV performs better as the variation of the preferences increases, which is consistent with our theoretical results. The performance of SC-RSV decreases as the variation of the preferences increases, since the singular vectors become noisier. The performance of SC-SYM is somewhere between that of SC-LSV



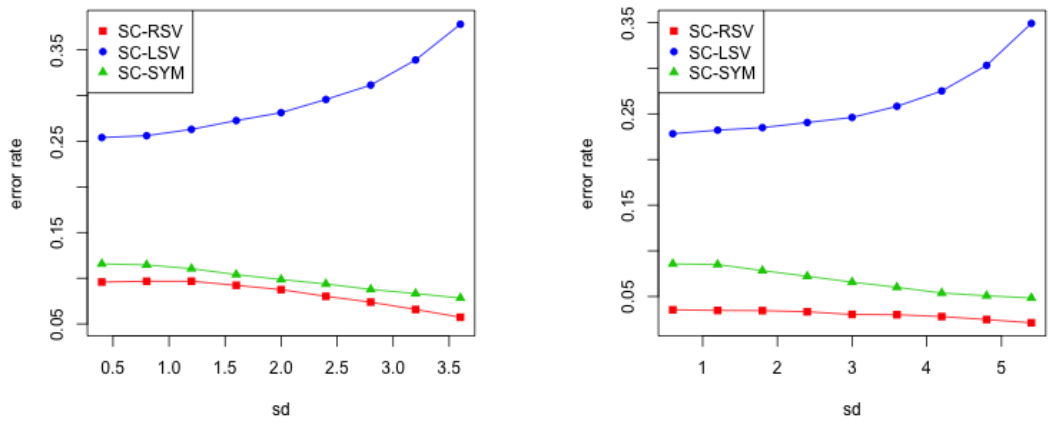
(a) $\gamma = 1/2$ (b) $\gamma = 3/4$ (c) $\gamma = 5/6$

Figure 3.1: The error rate of community detection as a function of the variation of the preferences $(\gamma(1-\gamma)/(s+1))$ for three values of γ given that the average out-degree is 8

and SC-RSV when the variation of the preferences is considerable. However, when the model is close to SBM (the variation ≈ 0), it performs better than other algorithms since symmetrization make a network denser, while the population version singular vectors of all three algorithms have correct information.

Simulation 2

In this simulation, we vary the out-degree heterogeneity given the variation of the preferences fixed. We set $K = 2$ and $n = 400$. Community size is 200 for each community. We set the parameters for in-degree the same for all the nodes in the network (i.e., ϕ_j are the same for all j). The out-degree θ_i parameters are generated from $\theta_i = (md - sd)B_i + (md + sd)B_i$ where $B_i \sim Binom(0.5)$. We have $\mathbb{E}[\theta_i] = md$ and $\sqrt{Var[\theta_i]} = sd$. We vary sd to see the effect of degree heterogeneity. As in Simulation 1, we control $\gamma(1-\gamma)/(s+1)$ to set the variation of the preferences. We set the $\gamma(1-\gamma)/(s+1)$ value to be 0.08. The simulation results for two different $md = 4, 6$ values are shown in Figure 3.2. As discussed in Example 2, the variation of out-degrees indeed helps to find communities given the other parameters are the same. SC-RSV performs better as standard deviation of θ increases. SC-RSV always performs better than SC-LSV as in Simulation 1. When the variation of the out-



(a) $\mathbb{E}[\theta_i] = 4$ (b) $\mathbb{E}[\theta_i] = 6$

Figure 3.2: The error rate of community detection as a function of the standard deviation of θ_i's given that $\gamma(1-\gamma)/(s+1) = 0.08$ and the average out-degree is 4 and 6 respectively

degree increases, the performance of SC-SYM improves as shown in 3.2. It is because the effect of degree heterogeneity is larger than the effect of noisy information from the left singular vectors.

3.5 Data examples

3.5.1 Political party examples

In this section, we examine the performance of the proposed algorithm in three networks: a US political blog network, a UK political Twitter network, and an Ireland political Twitter network. The US presidential political blog data were collected by Adamic and Glance (2005). In February 2005, they 1) retrieved the front page of each blog and 2) recorded the references (hyperlinks) to other blogs. The 1494 blogs were categorized into "liberal" and "conservative" based on the labels by self-report, automated categorization, or manual labeling. A directed interaction from node i to node j is created if blog i uploads a post with a hyperlink to blog j or blog i lists a link to blog j in sidebar. Many studies have used the data for clustering after

transforming the directed network to an undirected one. Since our main interest here is to investigate directed networks, we restrict our analysis to the 793 nodes in the largest strongly connected components.

The UK political Twitter network and the Ireland political Twitter network were curated by Greene and Cunningham (2013). These two data sets have three types of edges based on the sending node's activity: follows, mentions, and retweets. We only consider the "follows" activity, which has the most edges, for both networks. For the UK political Twitter network, we only consider the three largest political affiliations and extract the largest strongly connected component. This approach restricts our analysis to 388 out of 418 nodes in the network for the UK political Twitter network. Likewise, we consider the three largest political affiliations and extract the largest strongly connected component for the Ireland political Twitter network. Consequently, we restrict our analysis to 264 out of 348 nodes. Analyses with more political affiliations can be found in the appendix.

As in simulation studies, we have applied several spectral algorithms to community detection here. Specifically, we considered the SC-RSV as well as SC-LSV. In addition, we also considered spectral clustering with a symmetrized network (SC-SYM), transformed from a directed network. All the algorithms were applied with "true" number of communities (i.e. $K = 2$ for US political blogs network, $K = 3$ for UK and Ireland political Twitter networks). To measure the performance, we considered the number of mis-clustered nodes. The results are shown in Table 3.1. SC-RSV performs the best among the three methods. SC-LSV performs the worst. The performance of SC-SYM is in-between that of SC-SYM and SC-LSV. In addition to the quantitative measure, we also provided figures to show how the left and right singular vectors are positioned. The upper row of Figure 3.3 shows the first

Network	SC-LSV	SC-RSV	SC-SYM	Total number of nodes
US political blogs	35	**26**	28	793
UK political Twitter	4	**0**	1	388
Ireland political Twitter	1	**0**	**0**	264

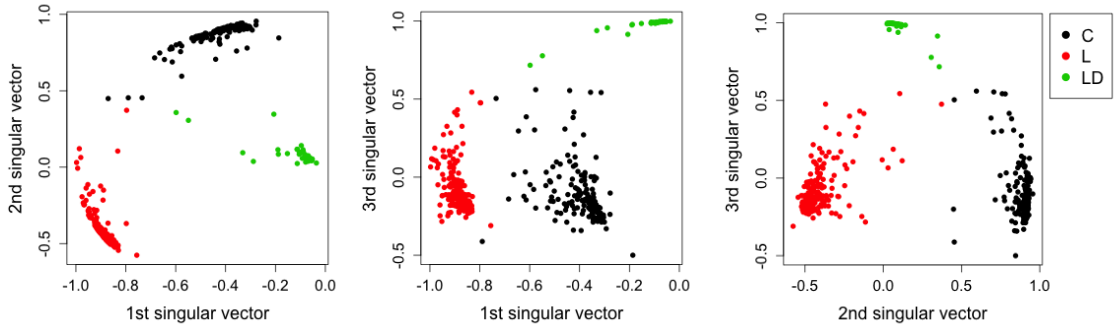
Table 3.1: Number of mis-clustered nodes for political networks

three normalized left singular vectors , while the lower row of Figure 3.3 displays the first three normalized right singular vectors. As can be seen in the six subfigures, clear separation is obvious for the right singular vectors. The left singular vectors are noisier than the right ones. This observation indicates that individual bloggers or Twitter accounts can have heterogeneous preferences to political affiliations. Not being aware of heterogeneous preferences may result in less accurate clustering results. Similar patterns for singular vectors have also been seen in other networks and the results are given in the appendix.

3.5.2 Author-paper citation network

Ji et al. (2016) have collected and curated coauthorship and citation networks (as well as other information such as paper titles, authors, citations, DOI, and abstracts) for statisticians and statistical papers. The data set is based on papers published in four prestigious statistics journals for the period of 2003 to the first half of 2012. The data set has been analyzed by many statisticians. Much work have focused on community detection of papers or community detection of authors to find research topics or collaboration groups.

In this section, we study the bipartite (directed) citation network of authors and papers. There is a directed edge between author i and paper j if author i cited paper j in one of i's papers. Instead of using the number of citations in the original weighted bipartite network, we use the binary version of the network for easier interpretation of the preferences. Using the original weighted network generates similar results as



(a) The 1st and the 2nd normalized left singular vectors

(b) The 1st and the 3rd normalized left singular vectors

(c) The 2nd and the 3rd normalized left singular vectors

(d) The 1st and 2nd normalized right singular vectors

(e) The 1st and the 3rd normalized right singular vectors

(f) The 2nd and the 3rd normalized left singular vectors

Figure 3.3: The normalized left singular vectors (in upper row) and right singular vectors (in lower row)

using the binary version. The original data set consists of 3607 authors and 3248 papers. Many networks have a core-periphery structure. A network has a core part where nodes are densely connected and periphery parts where nodes are sparsely connected. We focused on the core part of the bipartite network by following the procedure for finding the core of a directed network as in Wang and Rohe (2016). We removed all nodes of both types with degrees fewer than four with iterations until convergence. This pre-processing results in 780 authors and 742 papers for furthr analysis.

Then we applied the preference model to the bipartite network assuming authors have individual preferences to research topics. It is natural to assume that certain authors have broad research interests, not limited to one particular research topic. It is also possible that certain authors may have one focused research area. It is thus reasonable to consider that different authors have different research preferences whereas the papers form communities due to different research topics. In order to choose the number of communities, we looked at the scree plot of regularized Laplacian matrix and selected $K = 13$ since the elbow point of the scree plot seems to be at 14. We studied the communities of research topics using the word-of-bag method as in Wang and Rohe (2016) and displayed the research topics of several statisticians in Tables 3.3 and 3.4.

We restricted our analysis of the preferences to those who have cited more than 30 times in the period 2003-2012. Generally, the topics of papers an author cites show what topics the author is interested in. It provides more comprehensive information compared to only looking at the author's papers. We listed several authors and their estimated preferences as illustrative examples in Table 3.3. The authors are those who have cited more than two research topics with at least 0.3 preferences.

As can be seen, all of them had an interest in "variable selection/regularization". The table also shows that two of them were interested in "dimension reduction". Another two statisticians were interested in "semi/nonparametric" while the last two statisticians were interested in "covariance matrix". Table 3.4 shows individual preferences of three statisticians who have mainly focused on only one research topic, with a preference score at least 0.7. Each of them shows their main dominant research topics, which are "high-dimensional", "survival analysis" and "dimension reduction" respectively.

	Interpretation	Top seven representative words
1	high-dimensional (classification, regularization)	classification, learning, regularization, penalized, nonasymptotic, adaptive, sparsity
2	spatial statistics	spatial, predictive, feasible, temporal, domain, extreme, densities
3	dimension reduction	sliced, dimension, reduction, inverse, central, subspace, regression
4	randomized trial observational study	treatment, observational, sensitivity, randomized, biases, control, design
5	multiple testing	false, discovery, null, hypotheses, fdr, testing, rate
6	variable selection regularization	selection, lasso, algorithm, variable, oracle, penalty, path
7	functional data analysis	functional, smoothing, principal, component, function, variance, scalar
8	semiparametric/nonparametric	asymptotic, partially, semiparametric, backfitting, estimator, linear, varying
9	bayesian 1	prior, posterior, bayesian, wavelet, orthogonal, bayes, stationary
10	survival analysis	survival, censored, failure, proportional, hazard, consistent, censoring
11	bayesian 2	dirichlet, process, posterior, mixture bayesian, markov, chain
12	covariance matrix	matrix, covariance, volatility, matrices, diffusion, financial, sampled
13	analysis focusing on space	deconvolution, cluster, shape, distance, space, machine, size

Table 3.2: Top seven representative words and interpretation for $K = 13$ communities

	1	2	3	4	5	6	7	8	9	10	11	12	13
Chih-Ling Tsai	0.00	0.00	**0.47**	0.00	0.00	**0.30**	0.00	0.15	0.00	0.00	0.03	0.00	0.05
Hansheng Wang	0.02	0.00	**0.39**	0.00	0.00	**0.41**	0.00	0.13	0.00	0.00	0.00	0.04	0.00
Hua Liang	0.02	0.00	0.02	0.00	0.00	**0.43**	0.06	**0.42**	0.02	0.02	0.02	0.00	0.00
Runze Li	0.00	0.00	0.06	0.00	0.00	**0.47**	0.06	**0.35**	0.00	0.02	0.00	0.04	0.00
Ji Zhu	0.03	0.00	0.00	0.00	0.00	**0.57**	0.05	0.00	0.00	0.00	0.00	**0.35**	0.00
RJ Tibshirani	0.00	0.00	0.06	0.00	0.14	**0.44**	0.00	0.00	0.00	0.00	0.00	**0.31**	0.06

Table 3.3: Statisticians who have cited more than two research topics with at least 0.3 preferences

	1	2	3	4	5	6	7	8	9	10	11	12	13
AB Tsybakov	**0.74**	0.00	0.00	0.00	0.03	0.15	0.00	0.03	0.00	0.00	0.00	0.06	0.00
Donglin Zeng	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.03	0.00	**0.82**	0.00	0.00	0.00
Liping Zhu	0.00	0.00	**0.76**	0.00	0.00	0.18	0.03	0.00	0.00	0.03	0.00	0.00	0.00

Table 3.4: Statisticians who have cited one research topics with at least 0.7 preference

3.6 Discussion

In this chapter, we studied the effect of individual differences - the preference heterogeneity and out-degrees heterogeneity - in community detection in directed networks. We introduced a preference-based block model, where individual nodes can have different preferences to groups. The upper bound for the number of mis-clustered nodes was established which depends on the smallest singular values of two matrices: one related to the average connectivity and the other related to the variation of preferences or the variation of out-degrees. We have shown through numerical studies that it is possible to utilize the variation of the preferences and out-degrees to improve the performance of community detection given the same average connectivity between communities.

It would be interesting to design a model that considers not only the individual differences of the sender nodes but also those of the receiver nodes. There might be nodes whose popularities to nodes in different communities are different.

Another direction is to extend likelihood-based model selection criteria to the proposed model. Choosing the number of communities is also a relevant and challenge task. There has been much progress on likelihood-based model selection criteria (Wang et al., 2017; Hu et al., 2019) under the stochastic block model or degree-corrected stochastic block model. These methods can be good starting points to develop new methods with the additional consideration of the individual differences.

CHAPTER IV

Dyadic Latent Space Models for Directed Networks

4.1 Introduction

Networks have been an important representation of relationships among objects for a long time. It effectively represents relationships among objects inside complex systems such as social networks, brain networks, biological networks, to name a few (Malliaros and Vazirgiannis, 2013). In many cases, networks are directed, which means there is a directionality on the relationship, a distinct feature of directed networks different from undirected networks. Due to the directionality, the relationship between two nodes can be asymmetric. For example, node i sends an email to node j, but node j does not do the reverse. In addition, this property can result in significant difference between the in-degree and the out-degree of a node, which represents popularity and preference of the node respectively. A common approach to tackle directed networks in practice is to ignore the directionality and apply methods for undirected networks, causing potentially unsatisfactory results due to losing useful directional information (Leicht and Newman, 2008). Thus, recognizing and modeling the reciprocity in the edges of directed networks is of pressing need when analyzing directed networks.

Various models for directed networks have been developed to take into account

the tendency of edge reciprocity. For example, Holland and Leinhardt (1981) proposed the directed p_1 model, which concerns the distribution of edges with several parameters including a parameter for reciprocity, and they assumed the parameter for reciprocity is a constant over the network to reduce the number of parameters. Wang and Wong (1987) extended the p_1 model to consider subgraph information rather than just dyads. Holland et al. (1983) introduced a pair-dependent stochastic block model assuming the dyad probabilities can be different for different interactions between two communities. A mixture model of exponential families, including a mixture of p_1 models, was proposed in Vu et al. (2013a). They developed a variational EM algorithm to estimate the mixture model. For community detection, Li et al. (2012) introduced a spectral clustering method that only utilized mutual dyads to find communities. They argued mutual relations might be more stable and using only that relations can be of interest. However, all these existing methods in the literature fail to model reciprocity at the node level or the edge level, which requires a substantial number of parameters. In directed networks, it is likely that some nodes tend to reciprocate edges, while other nodes tend to only receive or send edges. In addition, some types of nodes might reciprocate edges more with the same type than other types. Thus, models ignoring individual reciprocity information in directed networks tend to be limited and fail to use holistic information in the network.

The latent space model for networks has also been widely used due to its flexibility, ease of interpretation and visualization. In such model, each node is represented as a vector z_i in a low-dimensional Euclidean space for parsimonious parameters. If two nodes are close in that space, they are more likely to connect to each other. Several distances can be used to measure the closeness between nodes. For example, Hoff et al. (2002) considered two types of latent space models, the distance model and the

projection model treating latent space positions as fixed effects. The original latent space model has been extended in the same author's later work (Hoff, 2015, 2005, 2009), in which the latent vectors were treated as random effects generated from a multivariate Gaussian distribution. Treating latent vectors as random effects allowed better modeling of key characteristics in networks such as transitivity, homophily, and degree heterogeneity. Model fitting and inference for latent space models have usually been carried out via Markov chain Monte Carlo. Due to the lack of scalability of MCMC, there is a limitation when applying these models to large network data. More recently, projected gradient descent algorithms to fit latent space models have been proposed (Wu et al., 2017; Ma et al., 2020), showing that the algorithms are scalable to large networks. Theoretical properties of the algorithms have also been established. A similar approach has been studied as node embedding (Grover and Leskovec, 2016; Tang et al., 2015; Perozzi et al., 2014), using stochastic gradient ascent with different neighborhood sampling schemes to make the algorithm scalable. To the best of our knowledge, there has been no latent space model for directed networks that takes into account diverse reciprocal relationships among the nodes. The model in Hoff (2015) is probably the closest one but it only considered a constant reciprocity parameter across the network.

In this chapter, we propose a new latent space model for directed networks. Unlike previous studies, we directly model the dyadic probability with low-dimensional latent vectors and three heterogeneity parameters. This enables the reciprocity to be different between different dyads as the reciprocity will depend on the interaction between two latent positions and each node's tendency to reciprocate, which was not considered in previous models. Thus, the proposed model allows for more flexibility for directed networks and enables to utilize overlooked differences in dyads, while

maintaining the advantages of the original latent space model. The model is promising for many tasks such as link prediction and community detection. In addition, by treating degree heterogeneity parameters and latent vectors as fixed effects, we use a projected gradient descent algorithm for model fitting, which is scalable and computationally efficient.

4.2 Models

In this section, we propose dyadic latent space models for directed networks using inner-product of latent vectors with degree heterogeneity parameters. Several conditions for the parameters are provided for identifiability. In addition, connections to previous models and the advantages over previous models are also discussed. The data we observe consists of a $n \times n$ adjacency matrix $A \in \{0,1\}^{n\times n}$ on n nodes, with entries A_{ij} equal to 1 if there is a link from node i to node j and 0 otherwise. This matrix is in general asymmetric since A_{ij} and A_{ji} are not necessarily the same in directed networks. We assume there is no self-loop, i.e. $A_{ii} = 0$ for all $i = 1, \cdots, n$. We denote the probability matrix $P = \mathbb{E}[A] \in \mathbb{R}^{n\times n}$. The model proposed in this section is based on $\binom{n-1}{2}$ random vectors $D_{ij} = (A_{ij}, A_{ji}) \in \{0,1\}^2,\ i < j$, each referred as a dyad. To simplify notation for dyadic outcomes, the observed cases can be denoted as, for $i \neq j$, $A_{ij}^{(s,t)} = I(A_{ij} = s, A_{ji} = t)$ for $s = 0, 1$ and $t = 0, 1$. Note that $A_{ij}^{(1,0)} + A_{ij}^{(0,1)} + A_{ij}^{(1,1)} + A_{ij}^{(0,0)} = 1$ and $A_{ij}^{(1,0)} + A_{ij}^{(1,1)} = A_{ij}$. Similarly, for $i \neq j$, $P_{ij}^{(s,t)} = P(A_{ij} = s, A_{ji} = t)$ for $s = 0, 1$ and $t = 0, 1$. Note that $P_{ij}^{(1,0)} + P_{ij}^{(0,1)} + P_{ij}^{(1,1)} + P_{ij}^{(0,0)} = 1$.

4.2.1 A dyadic latent space model

We consider a pair of edges (dyad) $D_{ij} = (A_{ij}, A_{ji}) \in \{(0,0), (0,1), (1,0), (1,1)\}$ for any $i < j$. We model the dyadic probability, which is a joint probability of A_{ij}

and A_{ji}, instead of assuming A_{ij} and A_{ji} are independent. The proposed model essentially uses multinomial distribution with 4 categories with shared latent variables. Assuming the D_{ij}'s (for any $i < j$) are mutually independent, we write the model as

$$
\begin{aligned}
D_{ij} &\overset{ind}{\sim} \text{Multinomial}(P_{ij}^{(1,0)}, P_{ij}^{(0,1)}, P_{ij}^{(1,1)}, P_{ij}^{(0,0)}) \\
\text{with } \Theta_{ij}^{(1,0)} &= \log(P_{ij}^{(1,0)}/P_{ij}^{(0,0)}) = a_i + b_j + u_i^\intercal R u_j \\
\Theta_{ij}^{(0,1)} &= \log(P_{ij}^{(0,1)}/P_{ij}^{(0,0)}) = b_i + a_j + u_i^\intercal R^\intercal u_j \\
\Theta_{ij}^{(1,1)} &= \log(P_{ij}^{(1,1)}/P_{ij}^{(0,0)}) = c_i + c_j + u_i^T S u_j,
\end{aligned} \tag{4.1}
$$

where $R, S \in \mathbb{R}^{K\times K}$, $a_i, b_i, c_i \in \mathbb{R}^1$ and $u_i \in \mathbb{R}^K$. Since $P_{ij}^{(0,1)} = P_{ji}^{(1,0)}$ and $P_{ij}^{(0,0)} = P_{ji}^{(0,0)}$, we have $\Theta_{ij}^{(0,1)} = \Theta_{ji}^{(1,0)}$. Note it is also possible to define D_{ij} for $i > j$ using the property $P_{ij}^{(0,1)} = P_{ji}^{(1,0)}$, even though D_{ij} and D_{ji} are basically the same random variable except for the order of categories. In total, there are $n(3 + K) + 2K^2$ parameters in the model.

The parameters $\{a_i : 1 \le i \le n\}$, $\{b_i : 1 \le i \le n\}$ and $\{c_i : 1 \le i \le n\}$ are used for modeling degree heterogeneity of each node. In directed networks, each node has two types of degrees: the in-degree of node j is the number of incoming edges, i.e, $\sum_{i=1}^n A_{ij}$, and the out-degree of node i is the number of outgoing edges, i.e, $\sum_{j=1}^n A_{ij}$. Both a_i and c_i parameters represent out-degree heterogeneity while both b_i and c_i parameters represent in-degree heterogeneity. a_i represent one-way sending tendency of node i and b_i represent one-way receiving tendency of node i. Specifically, c_i models both in-degree and out-degree which reflects the reciprocal tendency of node i. For instance, if c_i is large while a_i and b_i are small for node i, then it implies that node i is likely to have large out-degree and in-degree. In addition, the nodes who receive the edges from node i tend to have mutual relationships. Note that parameters a_i, b_i and c_i are not directly comparable to each other. When we interpret a_i, we need to compare a_i among a_i's.

The latent position vector $u_i \in \mathbb{R}^K$ plays an important role with u_j through $u_i^\intercal R u_j$ for $j \neq i$. Other functions of u_i and u_j that measure the similarity between u_i and u_j can be used instead of $u_i^\intercal R u_j$ to model the effect of latent vectors (Hoff et al., 2002; Ma et al., 2020). One reason we use $u_i^\intercal R u_j$ is because it is easy to take derivatives. Further, the representation $u_i^\intercal R u_j$ can be justified by singular value decomposition, which indicates that any $n \times n$ matrix can be expressed as $UV^\intercal$. Thus, if we take $v_j = Ru_j$ for parsimony, we have $u_i^\intercal R u_j$. An asymmetric matrix $R \in \mathbb{R}^{K\times K}$ is adopted to accommodate the asymmetric directional effect of $i \to j$ and $j \nrightarrow i$, i.e. $P_{ij}^{(1,0)}$ and $P_{ij}^{(0,1)}$ can be different. A symmetric matrix $S \in \mathbb{R}^{K\times K}$ is adopted to explain the effect of u_i and u_j on the probabilities of reciprocated edges; note $P^{(1,1)}$ is symmetric. In Section 4.2.2, a general model using the $UV^\intercal$ representation will be discussed. The $u_i^\intercal R u_j$ part can model transitivity and reciprocity. Take $S = R = I$ as a simple illustration. If $A_{ij} = 1$ and $A_{jk} = 1$ for $i \neq j$ and $j \neq k$, then both $u_i^\intercal u_j$ and $u_j^\intercal u_k$ are likely to be large, resulting $u_i^\intercal u_k$ likely to be large as well and increasing the chance of observing $A_{ik} = 1$. Similarly under the same condition, $u_j^\intercal u_i$ tends to be large, increasing the chance of observing $A_{ji} = 1$. Since the two matrices R and S need to be estimated, the estimates will accommodate with the transitivity and reciprocity in the network.

Matrix representations of $\Theta_{ij}^{(1,0)}$ and $\Theta_{ij}^{(1,1)}$ can be written as

$$\begin{aligned} \Theta^{(1,0)} &= \boldsymbol{a} 1_n^\intercal + 1_n \mathbf{b}^\intercal + URU^\intercal \\ \Theta^{(1,1)} &= \mathbf{c} 1_n^\intercal + 1_n \mathbf{c}^\intercal + USU^\intercal \end{aligned} \tag{4.2}$$

where $1_n \in \mathbb{R}^n$ is a vector with elements all equal to 1 and $U = (u_1, \cdots, u_n)^\intercal \in \mathbb{R}^{n\times K}$. Denote $\boldsymbol{a} = (a_1, \cdots, a_n) \in \mathbb{R}^n$, $\boldsymbol{b} = (b_1, \cdots, b_n) \in \mathbb{R}^n$ and $\boldsymbol{c} = (c_1, \cdots, c_n) \in \mathbb{R}^n$. Since we assume there is no self loop, diagonal elements of $\Theta^{(1,0)}$ and $\Theta^{(1,1)}$ could be ignored in the matrix form. Let $\Theta_0^{(1,0)}$ be a matrix that has the same value for the off-

diagonal elements as $\Theta^{(1,0)}$ and zero for diagonal elements, i.e., $\Theta^{(1,0)}_{0,ij} = \Theta^{(1,0)}_{ij} I(i \neq j)$, and $\Theta^{(1,1)}_0$ can be defined similarly, i.e., $\Theta^{(1,1)}_{0,ij} = \Theta^{(1,1)}_{ij} I(i \neq j)$. Without confusion, we assume the diagonal elements of $\Theta^{(1,0)}$ and $\Theta^{(1,1)}$ are zero.

Note that without any constraints, the model is not identifiable. To ensure identifiability, we impose certain conditions:

$$\begin{aligned} \mathbf{b}^\intercal 1_n = 0, \quad JU = U \\ \|UU^\intercal\|_F = n \end{aligned} \tag{4.3}$$

where $J = I_n - \frac{1}{n} 1_n 1_n^\intercal$ is a centering matrix. These conditions uniquely identify U up to an orthonormal matrix $\mathcal{Q} \in \mathbb{R}^{K \times K}$. In the model U, S and R are only identifiable upto $USU^\intercal$ and $URU^\intercal$. Multiplying any $c > 0$ to U results in $URU^\intercal = (cU)\left(\frac{1}{c^2} R\right)(cU)^\intercal$. Thus, $\|UU^\intercal\|_F = n$ is imposed. R and S also have to be scaled accordingly to keep $USU^\intercal$ and $URU^\intercal$ the same, respectively.

Rearranging model (4.1) gives us the log-odds ratio ρ_{ij}, that is

$$\begin{aligned} \rho_{ij} &= \log\left(\frac{P(A_{ij}=1|A_{ji}=1)P(A_{ij}=0|A_{ji}=0)}{P(A_{ij}=0|A_{ji}=1)P(A_{ij}=1|A_{ji}=0)}\right) \\ &= \log\left(\frac{P^{(1,1)}_{ij} P^{(0,0)}_{ij}}{P^{(1,0)}_{ij} P^{(0,1)}_{ij}}\right) \\ &= (c_i - a_i - b_i) + (c_j - a_j - b_j) + u_i^\intercal (S - R - R^\intercal) u_i. \end{aligned} \tag{4.4}$$

As in Holland and Leinhardt (1981), ρ_{ij} measures the tendency to reciprocate edges between two nodes i and j. We will have a positive ρ_{ij} if the probability for the two nodes to either have both edges or neither edge is relatively higher than the probability of having only one edge of the two possible edges. If the two random variables A_{ij} and A_{ji} are independent, then $\rho_{ij} = 0$. In this model, we have $\rho_{ij} = 0$ if $c_i = a_i + b_i$ and $S = R + R^\intercal$. Many previous work assumed a constant ρ in the network and then conducted hypothesis testing to test whether $\rho = 0$ or not. Unlike previous work, we do not impose any constraints on ρ. By using a flexible model for

dyadic probabilities, ρ can also be modeled flexibly, providing a way to understand reciprocity better.

Note that covariates can be naturally incorporated in the model as extra terms similar to previous work (Ma et al., 2020; Wu et al., 2017). However, we do not pursue that direction in this thesis.

Remark IV.1. A natural model for comparison to examine the effect of dyadic probabilities is a model that assumes all edge are independent. We thus define a pair-independent latent space model, which can be considered as a counterpart of the dyadic latent space model (4.1), as follows:

$$\begin{aligned} A_{ij} &\sim \text{Bernoulli}(P_{ij}), \quad \text{with} \\ logit(P_{ij}) &= \Theta_{ij} = a_i + b_j + u_i^T R u_j. \end{aligned} \tag{4.5}$$

In numerical studies, we will compare model (4.1) to model (4.5).

4.2.2 A general dyadic latent space model

In a directed network, many methods try to model the sender's behavior and the receiver's behavior separately. For example, in spectral clustering methods for directed networks, the left singular vectors and the right singular vectors are often utilized, with the left singular vectors responsible for senders' behavior and right singular vectors responsible for receivers' behavior (Rohe et al., 2016). Modeling latent vectors of outgoing nodes and incoming nodes separately has also been used in latent space models (Hoff, 2015). We also propose a general dyadic latent space model for directed networks that allows for two different latent vectors, for incoming and outgoing edges of a node respectively.

Assuming that D_{ij} (for any $i < j$) are mutually independent, we write the model

as

$$
\begin{aligned}
&D_{ij} \overset{ind}{\sim} \text{ Multinomial}(P_{ij}^{(1,0)}, P_{ij}^{(0,1)}, P_{ij}^{(1,1)}, P_{ij}^{(0,0)}) \\
&\text{with } \Theta_{ij}^{(1,0)} = \log(P_{ij}^{(1,0)}/P_{ij}^{(0,0)}) = a_i + b_j + u_i^\intercal v_j, \\
&\Theta_{ij}^{(0,1)} = \log(P_{ij}^{(0,1)}/P_{ij}^{(0,0)}) = b_i + a_j + v_i^\intercal u_j \\
&\Theta_{ij}^{(1,1)} = \log(P_{ij}^{(1,1)}/P_{ij}^{(0,0)}) = c_i + c_j + u_i^\intercal T v_j + v_i^\intercal T^\intercal u_j + v_i^\intercal S_v v_j + u_i^\intercal S_u u_j
\end{aligned} \tag{4.6}
$$

where $T \in \mathbb{R}^{K\times K}$ and symmetric matrices $S_v, S_u \in \mathbb{R}^{K\times K}$. The roles of $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are the same as in Section 4.2.1. Now we have two latent vectors $u_i \in \mathbb{R}^K$ and $v_j \in \mathbb{R}^K$, which influence $\Theta_{ij}^{(1,0)}$ through $u_i^\intercal v_j$, and they influence $\Theta_{ij}^{(1,1)}$ through 4 different terms: $u_i^\intercal T v_j$, $v_i^\intercal T^\intercal u_j$, $v_i^\intercal S_v v_j$ and $u_i^\intercal S_u u_j$. For the (1,1) pair, all combinations between (u_i, v_i) and (u_j, v_j) may have an influence on (1,1).

Again, to ensure identifiability, we impose the following conditions

$$
\begin{aligned}
\mathbf{b}^\intercal 1_n = 0, \quad JU = U \\
\|U\|_F = \sqrt{n} \quad \text{and} \quad \|V\|_F = \sqrt{n}
\end{aligned} \tag{4.7}
$$

where $J = I_n - \frac{1}{n} 1_n 1_n^\intercal$ is a centering matrix. These conditions uniquely identify U up to an orthonormal matrix $\mathcal{Q} \in \mathbb{R}^{K\times K}$.

Rearranging the model (4.6) gives us the log-odds ratio ρ_{ij}, that is

(4.8)

$$
\begin{aligned}
\rho_{ij} &= \log\left(\frac{P(A_{ij}=1|A_{ji}=1)P(A_{ij}=0|A_{ji}=0)}{P(A_{ij}=0|A_{ji}=1)P(A_{ij}=1|A_{ji}=0)}\right) \\
&= \log\left(\frac{P_{ij}^{(1,1)}P_{ij}^{(0,0)}}{P_{ij}^{(1,0)}P_{ij}^{(0,1)}}\right) \\
&= (c_i - a_i - b_i) + (c_j - a_j - b_j) + u_i^\intercal (T - I) v_j + v_i^\intercal (T - I)^\intercal u_j + v_i^\intercal S_v v_j + u_i^\intercal S_u u_j.
\end{aligned}
$$

Again, $\rho_{ij} = 0$ implies that A_{ij} and A_{ji} are independent, and it corresponds to $c_i = a_i + b_i$, $T = I$, $S_v = 0$ and $S_u = 0$.

Model (4.6) has more parameters and is more complicated than model (4.1). Unlike model (4.1), we need two times more latent variables in model (4.6) since we have

two types of latent spaces for the outgoing nodes and the incoming nodes respectively. This more complicated model will be suitable if the behaviors of outgoing nodes and incoming nodes are considerably different. However, if behaviors of outgoing nodes and incoming nodes are expected to be not that different, using both u_i and v_j may overparametrize the model. If we let $v_j = Ru_j$ and $S = TR + (TR)^\intercal + R^\intercal S_v R + S_u$, then model (4.6) reduces to model (4.1). In model (4.1), R compresses the difference between u_i and v_i via a matrix transformation. Since model (4.1) is simpler and easier to interpret than model (4.6), we will mainly use model (4.1).

4.2.3 Connections to other models

The models we propose are more general than previously studied models. Both directed and undirected networks can be understood under the proposed model. If the network is undirected, we will only observe (0,0) and (1,1) pairs. This can be considered as a special case of the directed model. Then, it is equivalent to using only the $\Theta^{(1,1)}$ part in model (4.1). Specifically, it is

$$\begin{aligned} A_{ij} = A_{ji} &\sim \text{Bernoulli}(P_{ij}), \quad \text{with} \\ logit(P_{ij}) &= \Theta_{ij} = c_i + c_j + u_i^T u_j \end{aligned} \tag{4.9}$$

after re-parametrization of u_i as $S^{1/2}u_i$. Model (4.9) has been studied for undirected networks in (Hoff, 2009; Ma et al., 2020).

A directed network assuming independent edges is a special case of model (4.1) or model (4.6), that is

$$\begin{aligned} A_{ij} &\sim \text{Bernoulli}(P_{ij}), \quad \text{with} \\ logit(P_{ij}) &= \Theta_{ij} = a_i + b_j + u_i^T v_j. \end{aligned} \tag{4.10}$$

Model (4.10) is the same as model (4.1) if we set $c_i = a_i + b_i$, $T = I$, $S_v = 0$ and $S_u = 0$. Model (4.10) has been studied in the literature (Wu et al., 2017; Hoff, 2015).

Hoff (2018) proposed an additive and multiplicative effects model (AME), in which $Cov[A_{ij}, A_{ji}]$ is a constant. In comparison, in model (4.1), we have $Cov[A_{ij}, A_{ji}] = P_{ij}^{(1,1)} - (P_{ij}^{(1,0)} + P_{ij}^{(1,1)})(P_{ij}^{(0,1)} + P_{ij}^{(1,1)})$. Since we model dyadic probabilities directly, which provides more information than just each edge, we do not impose strict constraints on model reciprocity. The main contribution of the proposed models is to model the dependency of edges in a pair by directly using latent space variables, with moderate increase in the number of parameters. This enables us to use reciprocal information, which was usually neglected in previous models for directed networks.

4.3 Algorithm

In this section, we develop algorithms for fitting model (4.1) by maximizing log-likelihood using a projected gradient descent approach. Denote $\mathbf{\Theta}_{ij} = \{\Theta_{ij}^{(1,0)}, \Theta_{ij}^{(0,1)}, \Theta_{ij}^{(1,1)}\}$. The link function is defined as

$$S^d(\mathbf{\Theta}_{ij}) = \exp(\Theta_{ij}^d) / \left(1 + \exp(\Theta_{ij}^{(1,0)}) + \exp(\Theta_{ij}^{(0,1)}) + \exp(\Theta_{ij}^{(1,1)})\right) \tag{4.11}$$

for $d \in \{(1,0),(0,1),(1,1)\}$. Let $P_{ij}^d = S^d(\mathbf{\Theta}_{ij})$ for $d \in \{(1,0),(0,1),(1,1)\}$. The log-likelihood function can be written as

$$\begin{aligned} l(U, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, S, R) &= \sum_{i<j} A_{ij}^{(1,0)} \log(P_{ij}^{(1,0)}) + A_{ij}^{(0,1)} \log(P_{ij}^{(0,1)}) + A_{ij}^{(1,1)} \log(P_{ij}^{(1,1)}) + A_{ij}^{(0,0)} \log(P_{ij}^{(0,0)}) \\ &= \sum_{i \neq j} A_{ij}^{(1,0)} \Theta_{ij}^{(1,0)} + \frac{1}{2} \sum_{i \neq j} A_{ij}^{(1,1)} \Theta_{ij}^{(1,1)} - \frac{1}{2} \sum_{i \neq j} \log\left(1 + e^{\Theta_{ij}^{(1,0)}} + e^{\Theta_{ij}^{(0,1)}} + e^{\Theta_{ij}^{(1,1)}}\right). \end{aligned} \tag{4.12}$$

Detailed derivation can be found in the appendix. Since this function is not convex with respect to u_i for all i, obtaining global optimum is not guaranteed. However, local minimum can be achieved by the gradient descent method with first-order derivative. To ensure identifiability conditions as in equation (4.3), we project parameter estimates into the constrained space after each step of iteration. Computational advantages of the gradient descent method for latent space models have been

demonstrated in Ma et al. (2020); Wu et al. (2017).

4.3.1 Projected gradient descent algorithm

We fit the model using a gradient descent algorithm by minimizing the negative log likelihood, which is

$$\min_{U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},S,R} -l(U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},S,R) = \min_{U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},S,R} f(U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},S,R), \tag{4.13}$$

where we set $f(\cdot) = -l(\cdot)$. The algorithm for fitting the model (4.1) using the projected gradient descent approach is given in Algorithm IV.1.

Algorithm IV.1 A projected gradient descent algorithm

Input: Adjacency matrix: A; latent space dimension: K;
initial estimates: $U^0, \boldsymbol{a}^0, \boldsymbol{b}^0, \boldsymbol{c}^0, S^0, R^0$;
step size: η_U, $\eta_{\boldsymbol{a}}$, $\eta_{\boldsymbol{b}}$, $\eta_{\boldsymbol{c}}$, η_R, η_S
Output: $\hat{U} = U^T, \hat{R} = R^T, \hat{S} = S^T, \hat{\boldsymbol{a}} = \boldsymbol{a}^T, \hat{\boldsymbol{b}} = \boldsymbol{b}^T, \hat{\boldsymbol{c}} = \boldsymbol{c}^T$
for t = 0,1,...,T-1 **do**
  $\tilde{U}^{t+1} = U^t + 2\eta_U \left((A^{(1,0)} - P^{(1,0)})U^t(R^t)^\intercal + (A^{(0,1)} - P^{(0,1)})U^t R^t + (A^{(1,1)} - P^{(1,1)})U^t S^t\right)$
  $\boldsymbol{a}^{t+1} = \boldsymbol{a}^t + 2\eta_{\boldsymbol{a}} \left(A^{(1,0)} - P^{(1,0)}\right) \mathbf{1}_n$
  $\tilde{\mathbf{b}}^{t+1} = \mathbf{b}^t + 2\eta_{\boldsymbol{b}} \left(A^{(0,1)} - P^{(0,1)}\right) \mathbf{1}_n$
  $\mathbf{c}^{t+1} = \mathbf{c}^t + 2\eta_{\boldsymbol{c}} \left(A^{(1,1)} - P^{(1,1)}\right) \mathbf{1}_n$
  $R^{t+1} = R^t + 2\eta_R (U^t)^\intercal (A^{(1,0)} - P^{(1,0)})U^t$
  $S^{t+1} = S^t + \eta_S (U^t)^\intercal (A^{(1,1)} - P^{(1,1)})U^t$
  $U^{t+1} = \mathcal{P}_U(\tilde{U}^{t+1}),\ \mathbf{b}^{t+1} = \mathcal{P}_b(\tilde{\mathbf{b}}^{t+1})$
  Update $P^{(1,0)}$ and $P^{(1,1)}$ using U^{t+1}, $\boldsymbol{a}^{t+1}$, $\boldsymbol{b}^{t+1}$,$\boldsymbol{c}^{t+1}$, R^{t+1} and S^{t+1}
end for

Algorithm IV.1 iteratively updates the estimates for the parameters U, $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, S and R. The estimates move along the direction of gradient descent by a small step size. In our implementation, we use an adaptive step size over iterations to prevent gradient from explosion and non-convergence. If $f(\cdot)$ with the updated parameters increases from the previous step, we reduce the step size until we have the updated parameters to decrease $f(\cdot)$. This prevents the objective function from getting to large. We stop the iteration if either the relative change of the object function is small enough, or the number of iterations reaches a pre-specified maximum value.

Step sizes are also important for algorithm convergence. We allow step sizes for U, $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, R, S to be different for better convergence.

For successful convergence of the algorithm IV.1, proper initialization is necessary. We take advantage of the parameter estimates from the pair-independent latent space model. Under the model, we estimate the parameters as $\boldsymbol{a}^0$,$\boldsymbol{b}^0$, U^0 and R^0, and we set $\boldsymbol{c}^0 = \boldsymbol{a}^0 + \boldsymbol{b}^0$ and $S^0 = R^0 + R^{0\intercal}$.

We suggest initialization using Algorithm IV.2. In our numerical studies, we found the performance of Algorithm IV.1 was robust to initialization even when we used randomly generated values. In practice, we suggest using multiple initial values and see which leads to the smallest objective value. A more principled way for initialization is developed in Algorithm C.2. Specifically, we separate $A^{(1,0)}$ and $A^{(1,1)}$ to estimate initial parameters. However, this algorithm did not necessarily perform better than other initialization algorithms. Thus, we used Algorithm IV.2 in our simulations for initialization.

Algorithm IV.2 Initialization algorithm

1: **Input**: A: Adjacency matrix; K: latent space dimension;
2: If dependent is TRUE, let $\boldsymbol{a}^0 = \left(\frac{1}{n(n-1)}\sum_{i\neq j} A^{(1,0)}_{i,j}\right)1_n$, $\boldsymbol{b}^0 = 0_n$ and $\boldsymbol{c}^0 = \left(\frac{1}{n(n-1)}\sum_{i\neq j} A^{(1,1)}_{i,j}\right)1_n$. Otherwise, let $\boldsymbol{a}^0 = \frac{1}{n(n-1)}\sum_{i\neq j} A^{(1,0)}_{i,j}$ and $\boldsymbol{b}^0 = 0_n$.
3: Find the singular vectors $U_K \in \mathbb{R}^{n\times K}$ of $\frac{A+A^\intercal}{2}$ corresponding to the K largest singular values $d_K \in \mathbb{R}^K$.
4: We have U^0 after centering and scaling U_K so that $JU = U$ and $\|UU^\intercal\|_F = n$.
5: Set $R^0 = diag(d_k)$ and $S^0 = 2 \cdot diag(d_k)$.
6: **end for**

4.3.2 Modified algorithm for link prediction

The proposed model can be used for link prediction, which predicts unobserved or missing edges. However, unlike the models assuming pair-independence, likelihood for missing edges is not clear in model (4.1) as we directly deal with dyads, which needs a pair of edges. Let $\mathcal{M} = \{(i,j) : A_{ij}$ exists and A_{ji} is missing$\}$ be the set

of pairs that one edge is missing and one edge is not missing. For edges in $\mathcal{M}$, we maximize the marginal likelihood where missing edges have been marginalized out. The marginalized log likelihood function for maximization is as follows

(4.14)
$$\sum_{i\neq j,(i,j)\in\mathcal{I}} A_{ij}^{(1,0)}\Theta_{ij}^{(1,0)} + \frac{1}{2}\sum_{i\neq j,(i,j)\in\mathcal{I}} A_{ij}^{(1,1)}\Theta_{ij}^{(1,1)} - \frac{1}{2}\sum_{i\neq j,(i,j)\in\mathcal{I}} \log\left(1+e^{\Theta_{ij}^{(1,0)}}+e^{\Theta_{ij}^{(0,1)}}+e^{\Theta_{ij}^{(1,1)}}\right) + \sum_{(i,j)\in\mathcal{M}} A_{ij}\log\left(P_{ij}^{(1,0)}+P_{ij}^{(1,1)}\right) + (1-A_{ij})\log\left(P_{ij}^{(0,0)}+P_{ij}^{(0,1)}\right)$$

Again, a projected gradient descent algorithm can be used to fit the model by maximizing the log likelihood function (4.14).

Once the model is fitted, we can predict the missing edges using the estimated probabilities. When two edges are missing in a dyad, we can predict the pair of edges using estimated probabilities $\hat{P}_{ij}^{(1,0)}, \hat{P}_{ij}^{(0,1)}$ and $\hat{P}_{ij}^{(1,1)}$. If A_{ij} is missing and $A_{ji}=1$, we can look at conditional probability $P(A_{ij}=1|A_{ji}=1)$ instead of $P(A_{ij}=1)$, which is more informative. That is

(4.15)
$$P(A_{ij}=1|A_{ji}=1) = \frac{P(A_{ij}=1,A_{ji}=1)}{P(A_{ij}=1,A_{ji}=1)+P(A_{ij}=0,A_{ji}=1)} = \frac{P_{ij}^{(1,1)}}{P_{ij}^{(1,1)}+P_{ij}^{(0,1)}}.$$

Similarly, if A_{ij} is missing and $A_{ji}=0$, we have

(4.16)
$$P(A_{ij}=1|A_{ji}=0) = \frac{P(A_{ij}=1,A_{ji}=0)}{P(A_{ij}=1,A_{ji}=0)+P(A_{ij}=0,A_{ji}=0)} = \frac{P_{ij}^{(1,0)}}{P_{ij}^{(1,0)}+P_{ij}^{(0,0)}}.$$

We obtain the estimates of conditional probabilities with estimated dyadic probabilities as in (4.15) and (4.16) for link prediction. Incorporating pair-dependence information can greatly improve the performance compared to other models that do not consider pair-dependence. In our numerical studies, we estimate the conditional probability (4.15) or (4.16) using estimated probabilities $\hat{P}_{ij}^{(1,0)}$, $\hat{P}_{ij}^{(1,0)}$ and $\hat{P}_{ij}^{(0,0)}$.

4.4 Theoretical results

In this section, we establish error bounds for Θ and P under the proposed model (4.1). First, we consider the following parameter space:

$$
\begin{aligned}
&\mathcal{F}(n,k,\alpha_l,\alpha_u) = \\
&\Big\{\mathbf{\Theta} | \Theta^{(1,0)} = \boldsymbol{a}1_n^\intercal + 1_n\mathbf{b}^\intercal + URU^\intercal \ , \\
&\quad \Theta^{(1,1)} = \mathbf{c}1_n^\intercal + 1_n\mathbf{c}^\intercal + USU^\intercal,\ JU = U,\ \boldsymbol{b}^\intercal 1_n = 0 \\
&\quad \min_{i\neq j}\Theta^{(1,1)}_{(i,j)}, \min_{i\neq j}\Theta^{(1,0)}_{(i,j)} \geq \alpha_l,\ \text{and}\ \max_{i\neq j}\Theta^{(1,1)}_{(i,j)}, \max_{i\neq j}\Theta^{(1,0)}_{(i,j)} \leq \alpha_u \leq 0\Big\}
\end{aligned} \tag{4.17}
$$

A constraint $\alpha_u \leq 0$ is given for simpler derivation, naturally making all probabilities less than 0.5. For simplicity, define $\mathbf{\Theta} = \left[\Theta^{(1,0)\intercal}, \Theta^{(1,1)\intercal}\right]^\intercal \in \mathbb{R}^{2n\times n}$. We also set $f(\mathbf{\Theta}) = -l(U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},S,R)$ as a function of $\mathbf{\Theta}$ for notational simplicity. Let ($\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}$, $\hat{\mathbf{c}}$, $\hat{R}$, $\hat{S}$, $\hat{U}$) be the optimal solution to (4.12). Denote $\hat{\Theta}^{(1,0)} = \hat{\boldsymbol{a}}1_n^\intercal + 1_n\hat{\boldsymbol{b}}^\intercal + \hat{U}\hat{R}\hat{U}^\intercal$ and $\hat{\Theta}^{(1,1)} = \hat{\mathbf{c}}1_n^\intercal + 1_n\hat{\mathbf{c}}^\intercal + \hat{U}\hat{S}\hat{U}^\intercal$. Similarly, let ($\boldsymbol{a}_*$, $\boldsymbol{b}_*$, $\mathbf{c}_*$, R_*, S_*, U_*) be the true parameters. Denote $\Theta^{(1,0)}_* = \boldsymbol{a}_*1_n^\intercal + 1_n\boldsymbol{b}_*^\intercal + U_*R_*U_*^\intercal$ and $\Theta^{(1,1)}_* = \mathbf{c}_*1_n^\intercal + 1_n\mathbf{c}_*^\intercal + U_*S_*U_*^\intercal$. We consider the upper bound for $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2 = \|\hat{\Theta}^{(1,0)} - \Theta^{(1,0)}_*\|_F^2 + \|\hat{\Theta}^{(1,1)} - \Theta^{(1,1)}_*\|_F^2$.

Theorem IV.2. *Suppose the network is generated from a dyadic latent space model (4.1) with parameter $\mathbf{\Theta}_*$. Let $\hat{\mathbf{\Theta}}$ be the global optimal solution to (4.12). Then, there exists constants $r, C > 0$ such that*

$$
\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2 \leq C(K+2)e^{-2\alpha_l}\max\{ne^{\alpha_u}, \log n\} \tag{4.18}
$$

with probability at least $1 - 2n^{-r}$.

Theorem IV.2 shows that the mean squared error, i.e. $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2/n^2$, is upper bounded by $CK\frac{e^{-2\alpha_l}}{n}\max(e^{\alpha_u}, \log n/n)$. If $e^{\alpha_u} \geq \log n/n$, the sparsity of the network influences the upper bound through e^{α_u}. If $\log n/n \leq e^{\alpha_u}$, $\log n/n$ will affect the rate. When e^{α_u} decreases, the bound increases. In addition, e^{α_u} also affect the upper

bound inversely, meaning the smallest probability makes the upper bound larger. For instance, if $e^{\alpha_u} = \Omega(\log n/n)$ and $e^{\alpha_l} = \Omega(\log n/n)$, then $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*\|_F^2/n^2 = O\left(\frac{1}{\log n}\right)$.

Denote $\boldsymbol{P} = [P^{(1,0)}, P^{(0,1)}, P^{(1,1)}, P^{(0,0)}] \in \mathbb{R}^{n\times 4n}$ and $\hat{\boldsymbol{P}} = [\hat{P}^{(1,0)}, \hat{P}^{(0,1)}, \hat{P}^{(1,1)}, \hat{P}^{(0,0)}] \in \mathbb{R}^{n\times 4n}$. We have the following result.

Theorem IV.3. *Suppose the network is generated from a dyadic latent space model with parameter $\boldsymbol{\Theta}_*$. Let $\hat{\boldsymbol{\Theta}}$ be the global optimal solution to (4.12). Let $\boldsymbol{P}_*$ and $\hat{\boldsymbol{P}}$ be obtained from $\boldsymbol{\Theta}_*$ and $\hat{\boldsymbol{\Theta}}$, respectively. Then, there exists constants $r, C > 0$ such that*

$$\|\hat{\boldsymbol{P}} - \boldsymbol{P}_*\|_F^2 \le C_1(K+2)e^{-\alpha_l}\max\left(ne^{\alpha_u}, \log n\right) \tag{4.19}$$

with probability at least $1 - 2n^{-r}$.

The rate of $\|\hat{\boldsymbol{P}} - \boldsymbol{P}_*\|_F^2$ is faster than that of $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*\|_F^2$ by $e^{-\alpha_l}$. The mean squared error, i.e. $\|\hat{\boldsymbol{P}} - \boldsymbol{P}_*\|_F^2/n^2$, is $O\left(\frac{K+2}{ne^{\alpha_l}}\max\left(e^{\alpha_u}, \log n/n\right)\right)$. The rate is reasonable since $\boldsymbol{P}_*$ is a transformation of $\boldsymbol{\Theta}_*$ via the link function $S^d(\cdot)$.

4.5 Simulation studies

In this section, we use three simulation studies to demonstrate the performance of the proposed method in different aspects: (1) estimation error, (2) community detection, and (3) link prediction. In Section 4.5.1, we study the estimation error using the proposed algorithm as we vary the size of the network and the dimension of the latent space. In Sections 4.5.2 and 4.5.3, we compare the proposed dyadic latent space model (DLSM) with pair-independent latent space model (PILSM). In PILSM, we assume all edges are independent Bernoulli random variables as in (4.5). Model fitting for PILSM is also implemented using a projected gradient descent algorithm and with initial values obtained from a similar initialization algorithm for

the independent edges case.

4.5.1 Estimation errors

We study how the estimation error by the proposed algorithm depends on the network size (n) and the dimension of the latent variables (K). We show that the proposed algorithm empirically converges well and estimates parameters well. As we have seen in Theorem IV.2, the estimation error depends on n and K. We set other parameters for any combination of $n \in \{250, 500, 1000\}$ and $K \in \{2, 4, 6\}$ as the following:

1. Generate degree heterogeneity parameters $\boldsymbol{a}$, $\mathbf{b}$ and $\mathbf{c}$. For all $i = 1, \cdots, n$,
$$\begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} \sim N_{[-1,1]} \left(\begin{bmatrix} -3 \\ 0 \\ -1.5 \end{bmatrix}, 0.1 \cdot \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \right)$$
2. Generate latent variables: For $k = 1, \cdots, K$, $u_{ik} \sim N_{[-2,2]}(0, 1)$ for $i = 1, \cdots, n$.
3. Set $\tilde{U} = JU$ as defined in (4.3) and U^* be the normalized version of $\tilde{U}$ such that $\|U^{*,\intercal} U^*\|_F = n$.
4. Set $R = diag(1, \cdots, 1) \in \mathbb{R}^{K \times K}$ and $S = diag(1, \cdots, 1) \in \mathbb{R}^{K \times K}$.

In these steps, $N_{[a,b]}(\cdot)$ is a truncated normal distribution by bounding the random variable from a to b. For each configuration of (n, K), the parameters generation is repeated 30 times, a corresponding adjacency matrix is generated, and the proposed algorithm is applied to the adjacency matrix. Initial estimates are also obtained from Algorithm **??**. We also found that convergence of algorithm is robust to other initial values.

We use relative error to compare the performances of the different methods, where the relative error of $(\hat{X}, X)$ is defined as $\|\hat{X} - X\|_F^2 / \|X\|_F^2$. Since U is identifiable up

to an orthonormal transformation, we measure the relative error of $UU^\intercal$. As can be seen in Figure 4.1, relative errors of $UU^\intercal$ and $P^{(1,0)}$ decrease as n increases and/or K decreases, which is consistent with the theoretical result. Relative errors of $URU^\intercal$, $USU^\intercal$ and $P^{(1,1)}$ show similar patterns as in Figure 4.1.

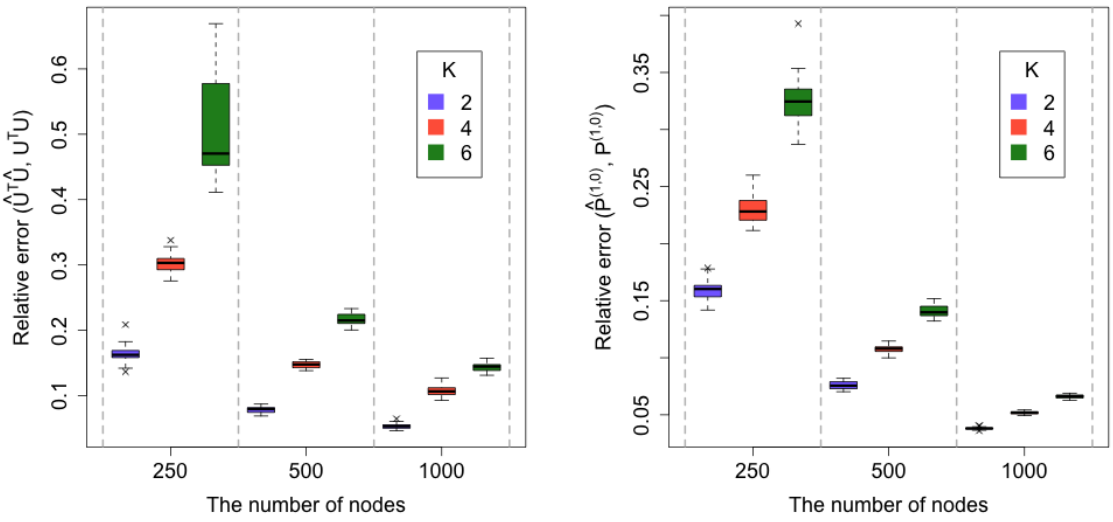


Figure 4.1: Boxplots of relative estimation error as we vary the number of nodes and the dimension of the latent space

4.5.2 Community detection

The stochastic block model (SBM) has been widely used as a standard model for community detection. Connectivity difference between communities in a network allows us to find communities. See Abbe (2018) for a review of the topic. Most models for community detection do not consider dyadic dependence. The degree of dyadic dependence can depend on the communities. In particular, it is possible that nodes in a community reciprocate more edges to nodes in the same community than to nodes in other communities. If reciprocity provides a strong signal for community information in directed networks, using reciprocity for community detection may

enable us to recover communities even if the parameters of the standard SBM are in a regime where community detection is difficult or even not possible.

To investigate the role of community-wise reciprocity for community detection, we vary the gap between the marginal probability within communities from the SBM and the marginal probability between communities while we fix the ratios between (1,0) probability and (1,1) probability of within and between communities and the average degree of the nodes. We set n=500 and K=2. Let $\gamma^{(1,0)}$ and $\gamma^{(1,1)}$ be the corresponding dyadic probabilities within communities and $\beta^{(1,0)}$ and $\beta^{(1,1)}$ be corresponding dyadic probabilities between communities. We set marginal probability within communities as $\gamma = \gamma^{(1,0)}+\gamma^{(1,1)}$ and marginal probability between communities as $\beta = \beta^{(1,0)} + \beta^{(1,1)}$. We set $\gamma^{(1,0)}/\gamma^{(1,1)} = 2$ for edges within the same communities and $\beta^{(1,0)}/\beta^{(1,1)} = 1/2$ for edges between communities. We vary marginal probabilities γ and β while keeping $(\gamma+\beta)/2 = 0.1$. To compare our results to true labels, we use the error rate which is defined as $n^{-1}\min_\sigma \sum_{i=1}^n I(z_i \neq \sigma(\hat{z}_i))$ where z_i is the true label, $\hat{z}_i$ is the estimated label and $\sigma(\cdot)$ is a permutation function. We also present the error rates obtained from traditional spectral clustering (SP) and PILSM. For SP, we perform singular value decomposition of A and apply the k-means algorithm to the singular vectors. SP and PILSM do not use pair dependence information, which are present in the simulation studies. For both the proposed model and PILSM, the k-means algorithm was applied to the fitted U to obtain cluster labels.

As expected, when $\gamma \approx \beta$, the SP and PILSM algorithms cannot distinguish the two communities, while the proposed DLSM can obtain much better results since the community information are available in dyadic dependence. Throughout the entire range of γ, the error rate of DLSM is always smaller than that of the other algorithms.

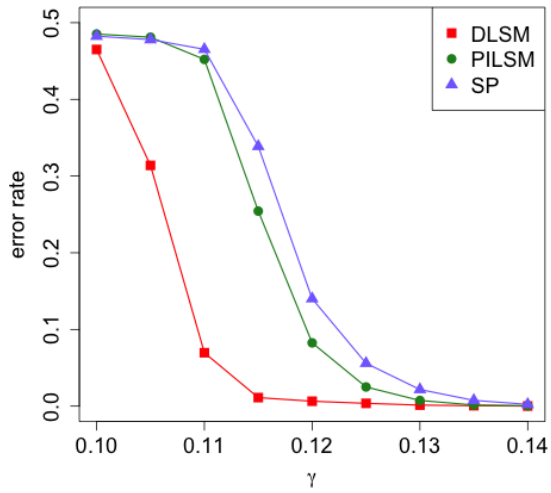


Figure 4.2: Error rates as γ varies

Especially from 0.1 to 0.12, where information of the dyadic probabilities is considerable while the signal is weak in the marginal connecting probabilities, the advantage of using DLSM is substantial. This also implies that neglecting pair-dependency in directed networks can result in significant information loss in community detection.

4.5.3 Link prediction

We also investigate how the proposed model can help to predict the missing edges in directed networks. Specifically, we vary the matrix S to change the tendency to reciprocate the links between nodes within close distance in the latent space. We set $n = 500$, $K = 2$ and $R = diag(0.5, 0.5) \in \mathbb{R}^{2\times 2}$. For a given s in $S = diag(s, s) \in \mathbb{R}^{2\times 2}$, we generated the data as follows:

1. Generate degree heterogeneity parameters $\boldsymbol{a}$, $\mathbf{b}$ and $\mathbf{c}$ as follows: For all $i = 1, \cdots, n$, $\begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} \sim N_{[-1,1]} \left(\begin{bmatrix} -3 \\ 0 \\ -1.5 \end{bmatrix}, 0.1 \cdot \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \right)$
2. Generate latent variables: For $k = 1, 2$, $u_{ik} \sim \mu_{1,k} + N_{[-2,2]}(0, 1)$ for $i = 1, \cdots, n$.

3. Set $\tilde{U} = JU$ as defined in (4.3) and U^* be the normalized version of $\tilde{U}$ such that $\|U^{*,\intercal}U^*\|_F = n$

For each model, the parameters generation is repeated 30 times. Note that varying s will also change the sparsity of networks. Among total $n \times (n-1)/2$ dyads, which are 124750 dyads, 10000 dyads are randomly sampled, which is around 8.02% of total dyads. Then, we randomly sample one edge from each sampled dyad and set it as missing. We apply the proposed algorithm and predict the link probability as described in Section 4.3.2. We compare the performance of the proposed model (DLSM) to that of PILSM in terms of the Area Under Curve (AUC), which is also known as the area under the ROC curve.

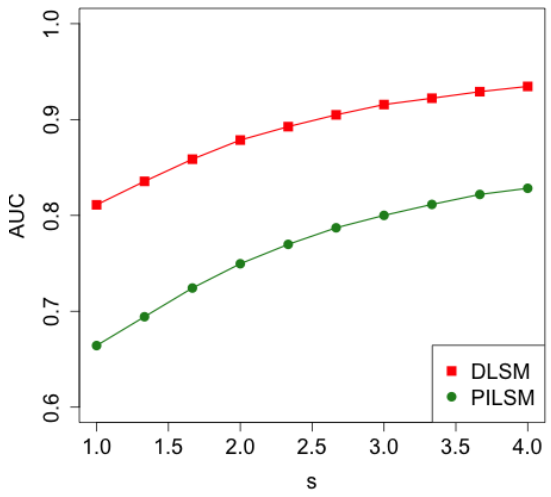


Figure 4.3: AUC for link prediction as s varies

Figure 4.3 shows that DLSM outperforms PILSM in terms of AUC in the entire range of s. When $s = 1$, a main difference between DLSM and PILSM comes exclusively from $\mathbf{c}$ parameters on (1,1) pairs. As we increase s, the effect of U on (1,1) pairs also increases. DLSM performs better than PLSM because the proposed model directly deals with the conditional probability of a edge in a dyad given the

other edge in the same dyad.

4.6 Data example

The data set we consider is a Twitter data of Rugby Union accounts, curated by Greene and Cunningham (2013). In this network, there is a collection of 854 international Rugby Union players, clubs, and organisations active on Twitter in 2012, who are linked to each other by several activities such as Follow, Mention and Retweet. Each activity can construct a directed network. There are overlapping communities corresponding to 15 different countries, which are considered to be ground truth groups. Some communities are dominantly large while some communities only have three to five accounts. Some largest communities are England, France, Ireland and Wales, etc. An adjacency matrix of only (1,0) pairs and an adjacency matrix of only (1,1) pairs are shown in Figure 4.4. When we made Figure 4.4, we ordered the nodes by countries to show existence of community structures. The accounts that have multiple affiliations are assigned to one of its affiliation. Clear community structures in both partial adjacency matrices are shown in Figure 4.4. In addition, different degree heterogeneity and different ratios of (1,0) and (1,1) pairs across countries can be observed, indicating there might be significant dependency within a pair.

In this analysis, we specifically focus on the Follow activity, which has the most edges among three activities. A directed edge from node i to node j implies that Twitter account i follows Twitter account j. We first perform the link prediction task with DLSM and PILSM and show how DLSM can achieve better results. We randomly choose $50,000$ dyads from $(848 \cdot 847)/2$ dyads and randomly set one of the edges in each sampled dyad as missing, which is around 13.9% of total dyads. We fit the model using only available links and perform link prediction with estimated

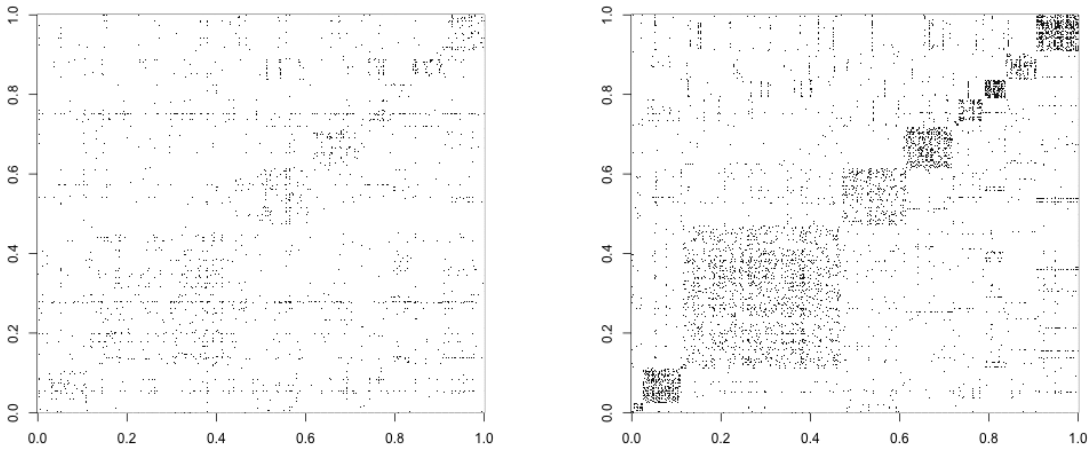


Figure 4.4: (1,0)-pair adjacency matrix (left) and (1,1)-pair adjacency matrix (right)

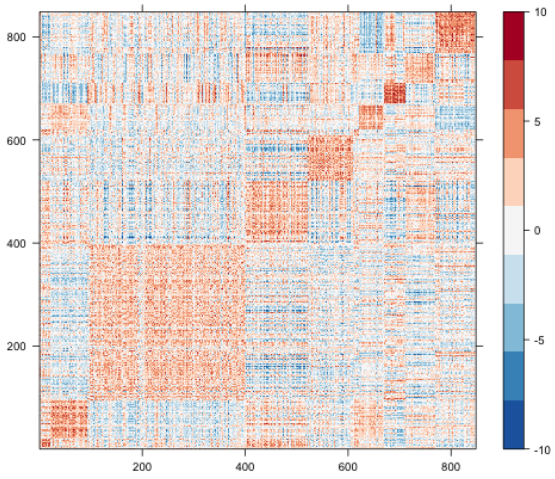


Figure 4.5: Estimated $\boldsymbol{\rho}$

	DLSM			PILSM		
K	Overall	$A_{ji}=1$	$A_{ji}=0$	Overall	$A_{ji}=1$	$A_{ji}=0$
2	0.963	0.833	0.904	0.861	0.695	0.834
4	0.970	0.852	0.924	0.927	0.766	0.902
6	0.972	0.862	0.927	0.944	0.806	0.913
8	0.977	0.876	0.939	0.953	0.825	0.924
10	0.977	0.882	0.942	0.955	0.822	0.925
12	0.980	0.885	0.948	0.963	0.842	0.930
14	0.981	**0.886**	0.952	**0.966**	**0.843**	**0.941**
16	**0.983**	0.884	**0.957**	0.965	0.838	0.933
18	0.977	0.876	0.943	0.952	0.795	0.891

Table 4.1: AUC for link prediction of the proposed model and PILSM using different K

conditional probabilities as in (4.15) and (4.16). Table 4.1 shows the result of AUC for several different K. To compare performance, we also present the result of PILSM as in Section 4.5. In Table 4.1, we calculated AUC with overall edges, AUC with edges A_{ij}s given $A_{ji}=1$, and AUC with edges A_{ij}s given $A_{ji}=0$. This illustrates how knowing one edge of a dyad helps to predict the other edge. It can be seen that the proposed model outperforms PILSM for all K, especially when the other edge in a dyad is one. The best performances were achieved when $K=14$ or $K=16$. When K is larger, both AUCs from the two models decrease because of overfitting.

The above link prediction task suggests using DLSM might be more suitable to this data set. However, one question about the data is which model should be used. One approach to tackle this is to use hypothesis testing with likelihood ratio based test statistics. Since PILSM is nested in DLSM, we can develop a hypothesis testing framework with hypotheses that"H_0: generated from PILSM" and "H_1 : generated from DLSM". We choose $K=8$ since AUC for link prediction was good when $K=8$ and $K=8$ gives a relatively parsimonious number of parameters. Since derivation of asymptotic distribution of the test statistics is not straightforward and asymptotic distribution may not reflect actual distribution unless n is very large, we choose to use parametric bootstrap (Paul et al., 2016). We fit the data with PILSM under

the null hypothesis. Then, we obtain fitted probability matrix $\hat{P}$. Considering $\hat{P}$ as the approximate probability under the null hypothesis, we then generate $\tilde{A}^{(b)}$ matrix from the fitted probability for $b = 1, \cdots, 1000$. For each b, we fit $\tilde{A}^{(b)}$ with DLSM and calculate log likelihood, which provide the empirical distribution of log likelihood with DLSM under the null hypothesis. Finally, we calculate the observed log likelihood with DLSM to the original adjacency matrix A and compare the observed value to the empirical distribution. The range of the empirical distribution of log likelihood was from -281445 to -273944, while the observed log likelihood with DLSM was -88199. The result shows we can reject the null hypothesis.

To show edge-wise different reciprocity, we first fit the model with $K = 8$ as before. Figure 4.5 displays the matrix $\hat{\boldsymbol{\rho}}$ where $\hat{\rho}_{ij} = \log(\hat{P}_{ij}^{(1,1)}\hat{P}_{ij}^{(0,0)} / \ \hat{P}_{ij}^{(1,0)}\hat{P}_{ij}^{(0,1)})$. After obtaining $\hat{\boldsymbol{\rho}}$, we threshold 44 dyads values that exceed the absolute value 10 to -10 or 10. As expected, we observe different structures of reciprocation across countries and node-wise heterogeneity of reciprocity. Within the same national team, the nodes tend to reciprocate the edges. In addition, some teams reciprocate edges more with some specific teams than with other teams, implying some teams are closer to each other. For instance, connections between New Zealand and Australia tend to be reciprocal. It is reasonable considering their geographical proximity.

4.7 Discussion

We have proposed a novel dyadic latent space model for directed networks by incorporating reciprocal relationships. The model directly targets the joint probability of a pair of edges with latent vectors to provide full information about pair dependency. We have also developed a projected gradient descent algorithm, which is computationally efficient.

The proposed model can be further improved. Though the proposed algorithm is computationally efficient as it is, the computation time can be further reduced at maybe the cost of reduced accuracy, even to the linear time of the size of a network, with a stochastic gradient descent approach. It would be interesting to evaluate the trade-off between the computing time and the accuracy.

One can further extend the ideas to weighted edges. Unlike binary edges, modeling the joint probability of a pair of weighted edges in a simple form is not a straightforward task. One promising candidate for the distribution in the continuous case will be the bivariate Gaussian distribution.

CHAPTER V

Discussion

In this thesis, we presented statistical methods and theories for directed networks and bipartite networks. First, we introduced one-to-one matched communities between two types of nodes and proposed a model with a two-stage spectral clustering algorithm, whose second stage adjusts the effect of community sizes. In the third chapter, we described a preference-based block model, where each node can have different preferences to groups. The spectral algorithm on the right singular vectors with weak consistency results on the number of mis-clustered nodes was presented as well. Finally, we demonstrated a model that can accommodate information from reciprocal relationships with latent positions. The proposed model, named a dyadic latent space model, showed excellent performance compared to the one that did not consider pair dependency.

The work presented in this thesis lays the foundation for future studies on directed and bipartite networks. The choice of the number of communities or the dimension of latent space is a critical part of analysis. There have been several studies on the topic (Li et al., 2020; Wang et al., 2017; Hu et al., 2019). One may adapt the proposed models into likelihood-based criteria developed for the stochastic block model (SBM) and the degree-corrected SBM. One can also attempt to develop a

statistical procedure or criterion for the choice of a proper model for specific data. Although there are many network models available, it is still difficult for one to choose a proper model for his or her data. A theoretically sound guidance would be greatly beneficial for users.

The dyadic latent space model, presented in the fourth chapter, can be extended for the application to weighted edges. For continuous edges, multivariate Gaussian distribution can be employed to model joint distribution of a pair of edges. Other types of edges such as counts will be more challenging to model jointly. It would also be an interesting task to adopt ideas from Bayesian methods to fit a model with Markov chain Monte Carlo. Finally, considering that the initialization and the setting of the step size are critical for the convergence of gradient descent algorithms solving non-convex problems, either a theoretical or at least an empirical study that can provide directions for the choice of initial values and step sizes would be greatly beneficial to the community.

APPENDIX

APPENDIX A

Appendix of Chapter II

A Proofs for lemmas in Section 2.2

Proof of Lemma II.3.

Let $\Delta_1 = diag(n_{1,1}, \cdots, n_{1,K})$, $\Delta_2 = diag(n_{2,1}, \cdots, n_{2,K})$, $\bigtriangledown_1 = diag(\mathcal{D}_{1,1}, \cdots, \mathcal{D}_{1,K})$ and $\bigtriangledown_2 = diag(\mathcal{D}_{2,1}, \cdots, \mathcal{D}_{2,K})$ where $\mathcal{D}_{1,s} = \sum_{a=1}^{K} B_{sa}N_a$ and $\mathcal{D}_{2,t} = \sum_{b=1}^{K} B_{tb}M_b$

$$\begin{aligned}\mathcal{L} = \mathcal{D}_1^{-1/2} P \mathcal{D}_2^{-1/2} &= \mathcal{D}_1^{-1/2} Z_1 B Z_2^\intercal \mathcal{D}_2^{-1/2} = Z_1 \bigtriangledown_1^{-1/2} B \bigtriangledown_2^{-1/2} Z_2^\intercal \\ &= Z_1 \Delta_1^{-1/2} \left(\Delta_1^{1/2} \bigtriangledown_1^{-1/2} B \bigtriangledown_2^{-1/2} \Delta_2^{1/2} \right) \Delta_2^{-1/2} Z_2^\intercal\end{aligned}$$

Let singular value decomposition of $\Delta_1^{1/2} \bigtriangledown_1^{-1/2} B \bigtriangledown_2^{-1/2} \Delta_2^{1/2}$ to be $XDY^\intercal$. Then, SVD of $\mathcal{L}$ is $(Z_1\Delta_1^{-1/2}X)D(Z_2\Delta_2^{-1/2}Y)^\intercal$. $\mathcal{U} = Z_1\Delta_1^{-1/2}X$ and $\mathcal{V} = Z_2\Delta_2^{-1/2}Y$. $C_\mathcal{U} = \Delta_1^{-1/2}X$ and $C_\mathcal{V} = \Delta_2^{-1/2}Y$. □

Proof of Lemma II.5

Let $\Delta_1 = diag(n_{1,1}, \cdots, n_{1,K})$, $\Delta_2 = diag(n_{2,1}, \cdots, n_{2,K})$, $\bigtriangledown_1 = diag(\mathcal{D}_{B,1}, \cdots, \mathcal{D}_{B,K})$ and $\bigtriangledown_2 = diag(\mathcal{D}_{B,1}, \cdots, \mathcal{D}_{B,K})$ where $\mathcal{D}_{B,s} = \sum_{a=1}^{K} B_{sa}$.

$$\begin{aligned}\tilde{\mathcal{L}} = \tilde{\mathcal{D}}_1^{-1/2} W_1^{-1/2} P W_2^{-1/2} \tilde{\mathcal{D}}_2^{-1/2} &= \tilde{\mathcal{D}}_1^{-1/2} W_1^{-1/2} Z_1 B Z_2^\intercal W_2^{-1/2} \tilde{\mathcal{D}}_2^{-1/2} \\ = Z_1 \Delta_1^{-1/2} \left(\bigtriangledown_1^{-1/2} B \bigtriangledown_2^{-1/2} \right) \Delta_2^{-1/2} Z_2^\intercal\end{aligned}$$

Let singular value decomposition of $\bigtriangledown_1^{-1/2} B \bigtriangledown_2^{-1/2}$ to be $X\tilde{D}X^\intercal$ (since $\bigtriangledown_1 = \bigtriangledown_2$).

Then, SVD of $\tilde{\mathcal{L}}$ is $(Z_1\Delta_1^{-1/2}X)\tilde{D}(Z_2\Delta_2^{-1/2}X)^\intercal$. $C_{\tilde{\mathcal{U}}} = \Delta_1^{-1/2}X$ and $C_{\tilde{\mathcal{V}}} = \Delta_2^{-1/2}X$. $\Delta_1^{1/2}C_{\tilde{\mathcal{U}}} = \Delta_2^{1/2}C_{\tilde{\mathcal{V}}}$. □

B Proof of Proposition II.7

To apply statistical techniques that have been used in (undirected) SBM, we can express a bipartite network as a symmetric network. Denote full graph versions of A and P from bipartite network as

$$A_f = \begin{bmatrix} 0 & A \\ A^\intercal & 0 \end{bmatrix}, \quad P_f = \begin{bmatrix} 0 & P \\ P^\intercal & 0 \end{bmatrix}$$

The next lemmas are slightly modified lemmas of Lei et al. (2015) to A_f and P_f.

Lemma A.1. *(concentration of Singular Space)*
Assume that $P \in \mathbb{R}^{n_1\times n_2}$ is a rank K matrix with smallest absolute non-zero singular value $\gamma_n > 0$. Let $A \in \{0,1\}^{n_1\times n_2}$ and U, $\mathcal{U} \in \mathbb{R}^{n_1\times K}$ contain the top K left singular vectors of A and P, respectively. Similarly, let V, $\mathcal{V} \in \mathbb{R}^{n_2\times K}$ contain the top K right singular vectors of A and P, respectively. Then there exists a $K \times K$ orthogonal matrix $\mathcal{Q} \in \mathbb{R}^{K\times K}$.

$$\|U - \mathcal{U}\mathcal{Q}\|_F \text{ or } \|V - \mathcal{V}\mathcal{Q}\|_F \leq \sqrt{\|U - \mathcal{U}\mathcal{Q}\|_F^2 + \|V - \mathcal{V}\mathcal{Q}\|_F^2} \leq \frac{4\sqrt{K}}{\gamma_n}\|A - P\|$$

Proof. Since P_f is a 2K rank matrix, P_f has 2K leading eigenvectors $\mathcal{W}_c = \begin{bmatrix} \mathcal{W} & \mathcal{W}' \end{bmatrix} \in \mathbb{R}^{n\times 2K}$, where $\mathcal{W} = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathcal{U} \\ \mathcal{V} \end{bmatrix}$ and $\mathcal{W}' = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathcal{U} \\ -\mathcal{V} \end{bmatrix}$ that have the same absolute eigenvalues. Similarly, A_f has 2K eigenvectors $W_c = \begin{bmatrix} W & W' \end{bmatrix} \in \mathbb{R}^{n\times 2K}$, where $W = \frac{1}{\sqrt{2}}\begin{bmatrix} U \\ V \end{bmatrix}$ and $W' = \frac{1}{\sqrt{2}}\begin{bmatrix} U \\ -V \end{bmatrix}$ that have the same absolute eigenvalues. By Proposition 2.2 in (Vu et al., 2013b), there exists a K-dimensional orthogonal

matrix $\mathcal{Q}$ such that

$$\frac{1}{\sqrt{2}}\sqrt{\|U-\mathcal{U}\mathcal{Q}\|_F^2+\|V-\mathcal{V}\mathcal{Q}\|_F^2}=\|W-\mathcal{W}\mathcal{Q}\|_F\leq\sqrt{2}\|(I-WW^\intercal)\mathcal{W}\mathcal{W}^\intercal\|_F$$
$$\leq\sqrt{2K}\|(I-WW^\intercal)\mathcal{W}\mathcal{W}^\intercal\|$$

Next, we establish that $\|(I-W_cW_c^\intercal)\mathcal{W}_c\mathcal{W}_c^\intercal\|\leq 2\frac{\|A_f-P_f\|}{\gamma_n}$. If $\|A_f-P_f\|\leq\frac{\gamma_n}{2}$, then by Davis-Kahan theorem, we have

$$\|(I-W_cW_c^\intercal)\mathcal{W}_c\mathcal{W}_c^\intercal\|\leq\frac{\|A_f-P_f\|}{\gamma_n-\|A_f-P_f\|}\leq 2\frac{\|A_f-P_f\|}{\gamma_n}$$

If $\|A-P\|>\frac{\gamma_n}{2}$, then

$$\|(I-W_cW_c^\intercal)\mathcal{W}_c\mathcal{W}_c^\intercal\|\leq 1\leq 2\frac{\|A_f-P_f\|}{\gamma_n}.$$

Using $\|(I-W_cW_c^\intercal)\mathcal{W}_c\mathcal{W}_c^\intercal\|=\|(I-WW^\intercal)\mathcal{W}\mathcal{W}^\intercal\|$ and $\|A_f-P_f\|=\|A-P\|$, we have

(A.1)

$$\|U-\mathcal{U}\mathcal{Q}\|_F \text{ or } \|V-\mathcal{V}\mathcal{Q}\|_F\leq\sqrt{\|U-\mathcal{U}\mathcal{Q}\|_F^2+\|V-\mathcal{V}\mathcal{Q}\|_F^2}\leq\frac{4\sqrt{K}}{\gamma_n}\|A-P\|. \quad \square$$

The following two lemmas are modified versions of Theorem 5.2 and Lemma 5.3 in Lei et al. (2015), showing concentration of A to P and approximate k-means error bound.

Lemma A.2. *(concentration of A to P)*

Let A_f be the adjacency matrix of a random graph on n nodes in which edges occur independently. Let $\mathbb{E}[A_f]=P_f=(p_{f,ij})_{i,j=1,\cdots,n}$ and assume that $n\cdot\max_{ij}p_{f,ij}\leq d$ for $d\geq c_0\log(n)$ and $c_0>0$. Then, for any $r>0$ there exists a constant $C=C(r,c_0)$ such that

$$\|A-P\|=\|A_f-P_f\|\leq C\sqrt{d} \tag{A.2}$$

with probability at least $1-n^{-r}$.

Lemma A.3. *(approximate k-means error bound)*

For $\varepsilon > 0$ and any two matrices U, $\mathcal{U}$ (V, $\mathcal{V}$) $\in \mathbb{R}^{n_1 \times K}$ ($\mathbb{R}^{n_2 \times K}$) such that $\mathcal{U} = Z_1 C_{\mathcal{U}}$ ($\mathcal{V} = Z_2 C_{\mathcal{V}}$) with $Z_1 \in \mathbb{M}_{n_1,K}$ ($Z_2 \in \mathbb{M}_{n_2,K}$), $C_{\mathcal{U}} \in \mathbb{R}^{K \times K}$ ($C_{\mathcal{V}} \in \mathbb{R}^{K \times K}$), let $[\hat{Z}_1, \hat{C}_U]$ $\left([\hat{Z}_2, \hat{C}_V]\right)$ be a $(1+\varepsilon)$-approximate solution to the k-means problem and $\bar{U} = \hat{Z}_1 \hat{C}_U$ ($\bar{V} = \hat{Z}_2 \hat{C}_V$). For any $\delta_{1,k} \leq \min_{l \neq k} \|C_{\mathcal{U},l*} - C_{\mathcal{U},k*}\|$ and $\delta_{2,k} \leq \min_{l \neq k} \|C_{\mathcal{V},l*} - C_{\mathcal{V},k*}\|$, define $\mathcal{S}_{1,k} = \{i \in G_{1,k}(Z_1) : \|\bar{U}_{i*} - \mathcal{U}_{i*}\| \geq \delta_{1,k}/2\|\}$ and $\mathcal{S}_{2,k} = \{j \in G_{2,k}(Z_2) : \|\bar{V}_{j*} - \mathcal{V}_{j*}\| \geq \delta_{2,k}/2\|\}$ then

$$\text{(A.3)} \quad \sum_{k=1}^{K} |\mathcal{S}_{1,k}| \delta_{1,k}^2 \leq 4(4+2\varepsilon)\|U - \mathcal{U}\|_F^2, \quad \sum_{k=1}^{K} |\mathcal{S}_{2,k}| \delta_{2,k}^2 \leq 4(4+2\varepsilon)\|V - \mathcal{V}\|_F^2$$

Moreover, if

$$\text{(A.4)} \quad (16+8\varepsilon)\|U - \mathcal{U}\|_F^2 / \delta_{1,k}^2 < n_{1,k} \text{ and } (16+8\varepsilon)\|V - \mathcal{V}\|_F^2 / \delta_{2,k}^2 < n_{2,k} \quad \text{for all } k$$

then there exists a $K \times K$ permutation matrix J_1 and J_2 such that $\hat{Z}_{1,G_{1*}} = Z_{1,G_{1*}} J_1$ and $\hat{Z}_{2,G_{2*}} = Z_{2,G_{2*}} J_2$, where $G_1 = \cup_{k=1}^{K} (G_{1,k}/\mathcal{S}_{1,k})$ and $G_2 = \cup_{k=1}^{K} (G_{2,k}/\mathcal{S}_{2,k})$.

Proof of Proposition II.7

Combining Lemma A.1 and Lemma A.2, we obtain that, for some K-dimensional orthogonal matrix $\mathcal{Q}$,

$$\sqrt{\|U - \mathcal{U}\mathcal{Q}\|_F^2 + \|V - \mathcal{V}\mathcal{Q}\|_F^2} \leq \frac{4\sqrt{K}}{\gamma_n} \|A - P\| \leq \frac{4\sqrt{K}}{\gamma_n} C \sqrt{n\alpha_n}$$

with probability at least $1 - n^{-1}$, where C is the absolute constant involved in Lemma A.2. Now, we apply Lemma A.3 to U and $\mathcal{U}\mathcal{Q}$. We can choose $\delta_{1,k} = \sqrt{1/n_{1,k}}$ and $\delta_{2,k} = \sqrt{1/n_{2,k}}$ in Lemma A.3 and hence $n_{1,k}\delta_{1,k}^2 = 1$ and $n_{2,k}\delta_{2,k}^2 = 1$ for all k. Using the above, a sufficient condition for Lemma A.3 to hold is

$$(16+8\varepsilon) 16 C^2 K \frac{n\alpha_n}{\gamma_n^2} \leq 1 = \min_k n_{2,k} \delta_k^2$$

Let $c^{-1} = 64C^2$. With the choice of $\delta_{1,k} = \sqrt{2/n_{1,max}}$ and $\delta_{2,k} = \sqrt{2/n_{2,max}}$, this yields that

$$\sum_{k=1}^{K} |\mathcal{S}_{1,k}| \leq c^{-1}(2+\varepsilon) n_{1,max} \frac{Kn\alpha_n}{\gamma_n^2}$$
$$\sum_{k=1}^{K} |\mathcal{S}_{2,k}| \leq c^{-1}(2+\varepsilon) n_{2,max} \frac{Kn\alpha_n}{\gamma_n^2}$$

□

C Proof of Theorem II.11

Proof of Lemma II.9

By the Assumption II.8, we have

$$1 - \eta \leq \frac{\hat{n}_{1,k}}{n_{1,k}} = 1 + \frac{\hat{n}_{1,k} - n_{1,k}}{n_{1,k}} \leq 1 + \eta$$

Therefore, we have

$$\left| \frac{\hat{n}_{1,k}}{n_{1,k}} - 1 \right| \leq \eta$$

Also for $\frac{n_{1,k}}{\hat{n}_{1,k}}$, we have inequality as follows by the Assumption II.8

$$\frac{1}{1+\eta} \leq \frac{n_{1,k}}{\hat{n}_{1,k}} = \frac{1}{1 + \frac{\hat{n}_{1,k} - n_{1,k}}{n_{1,k}}} \leq \frac{1}{1-\eta}$$

Therefore, we have

$$\left| \frac{n_{1,k}}{\hat{n}_{1,k}} - 1 \right| \leq \frac{\eta}{1-\eta}$$

Similarly for $n_{2,k}$. □

Lemma A.4. *(concentration of $\hat{A}$ to $\tilde{P}$)*

Let $\hat{A} = \hat{W}_1^{-1/2} A \hat{W}_2^{-1/2} \in \mathbb{R}^{n_1 \times n_2}$ and $\tilde{P} = W_1^{-1/2} P W_2^{-1/2} \in \mathbb{R}^{n_1 \times n_2}$ and $\max_{s,t} B_{st} \leq \alpha_n$ for some $\alpha_n \geq \log n / n$. Under the assumption II.8, there exists a constant C such that

$$\|\hat{A} - \tilde{P}\|^2 \leq C_{\eta,2}^2 \left(\frac{C\sqrt{n\alpha_n}}{\sqrt{n_{1,min} n_{2,min}}} + \alpha_n \sqrt{K\eta(K\eta+3)} \right)^2$$

with probability at least $1 - n^{-1}$.

Proof. Let $\hat{P} = \hat{W}_1^{-/2} P \hat{W}_2^{-1/2}$. From the inequality $\|\hat{A} - \tilde{P}\| \leq \|\hat{A} - \hat{P}\| + \|\hat{P} - \tilde{P}\|$. We bound two terms $\|\hat{A} - \hat{P}\|$ and $\|\hat{P} - \tilde{P}\|$ separately. From the Lemma A.2, we have

$$\mathbb{P}[\|A - P\| \leq C\sqrt{n\alpha_n}] \geq 1 - n^{-1} \tag{A.5}$$

where C is the constant in that Lemma A.2. For the first term, on the event (A.5)

$$\begin{aligned}\|\hat{A} - \hat{P}\| &= \|\hat{W}^{-1/2}(A - P)\hat{W}^{-1/2}\| \\ &\leq \|\hat{W}_1^{-1/2}\|\|(A - P)\|\|\hat{W}_1^{-1/2}\| \leq \frac{C\sqrt{n\alpha_n}}{\sqrt{\hat{n}_{1,min}\hat{n}_{2,min}}} \leq C_{\eta,1}\frac{C\sqrt{n\alpha_n}}{\sqrt{n_{1,min}n_{2,min}}}\end{aligned}$$

where we let $C_{\eta,1} = 1/(1 - C_1)$ for notational simplicity. Last inequality follows from assumption II.8.

For the second term also on the event (A.5), we bound $\|\hat{P} - \tilde{P}\|$ with the inequality $\|\hat{P} - \tilde{P}\| \leq \|\hat{P} - \tilde{P}\|_F$ with frobenius norm. When we compute the bound on the $\|\hat{P} - \tilde{P}\|_F$, we divide the set of edges into 4 non-overlapping cases; (1) $i \in \mathcal{S}_1^c, j \in \mathcal{S}_2^c$, (2) $i \in \mathcal{S}_1^c, j \in \mathcal{S}_2$, (3) $i \in \mathcal{S}_1, j \in \mathcal{S}_2^c$ and (4) $i \in \mathcal{S}_1, j \in \mathcal{S}_2$.

For the set of edges where the nodes are correctly clustered (1) $i \in \mathcal{S}_1^c, j \in \mathcal{S}_2^c$,

$$\begin{aligned}\sum_{i \in \mathcal{S}_1^c, j \in \mathcal{S}_2^c} \left(\hat{P}_{ij} - \tilde{P}_{ij}\right)^2 &= \sum_{i \in \mathcal{S}_1^c, j \in \mathcal{S}_2^c} \left(\frac{p_{ij}}{\sqrt{\hat{n}_{1,\hat{z}_{1i}}\hat{n}_{2,\hat{z}_j}}} - \frac{p_{ij}}{\sqrt{n_{1,z_{1i}}n_{2,z_{2j}}}}\right)^2 \\ &\leq \sum_{i \in \mathcal{S}_1^c, j \in \mathcal{S}_2^c} \left(\frac{p_{ij}}{\sqrt{n_{1,z_{1i}}n_{2,z_{2j}}}}\right)^2 \max_{i,j}\left(1 - \frac{\sqrt{n_{1,z_{1i}}n_{2,z_{2j}}}}{\sqrt{\hat{n}_{1,\hat{z}_{1i}}\hat{n}_{2,\hat{z}_{2j}}}}\right)^2 \\ &\leq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(\frac{p_{ij}}{\sqrt{n_{1,z_{1i}}n_{2,z_{2j}}}}\right)^2 \left((1 + \frac{\eta}{1 - \eta}) - 1\right)^2 \\ &\leq \left(\sum_{s=1}^{K}\sum_{t=1}^{K} B_{st}^2\right)\left(\frac{\eta}{1 - \eta}\right)^2 = C_{\eta,1}^2\left(\sum_{s=1}^{K}\sum_{t=1}^{K} B_{st}^2\right)\eta^2\end{aligned}$$

For the set of edges where the nodes of type-1 is correctly estimated and the nodes

of type-2 is mis-clustered (2) $i \in \mathcal{S}_1^c, j \in \mathcal{S}_2$,

$$\begin{aligned}
\sum_{i\in\mathcal{S}_1^c, j\in\mathcal{S}_2} \left(\hat{P}_{ij} - \tilde{P}_{ij}\right)^2 &= \sum_{i\in\mathcal{S}_1^c, j\in\mathcal{S}_2} \left(\frac{p_{ij}}{\sqrt{\hat{n}_{1,\hat{z}_i}\hat{n}_{2,\hat{z}_j}}} - \frac{p_{ij}}{\sqrt{n_{1,z_{1i}}n_{2,z_{2j}}}}\right)^2 \\
&\leq \sum_{i\in\mathcal{S}_1^c, j\in\mathcal{S}_2} \left(\frac{p_{ij}}{\sqrt{n_{1,z_{1i}}}}\right)^2 \max_{i,j} \left(\frac{\sqrt{n_{1,z_{1i}}}}{\sqrt{\hat{n}_{1,\hat{z}_{1i}}}}\frac{1}{\sqrt{\hat{n}_{2,\hat{z}_{2j}}}} - \frac{1}{\sqrt{n_{2,z_{2j}}}}\right)^2 \\
&\leq \left(\sum_{j\in\mathcal{S}_2}\sum_{s=1}^{K}\sum_{i:z_{1i}=s} \frac{B^2_{z_{1i}z_{2j}}}{n_{1,z_{1i}}}\right) \max_{i,j} \left(\frac{\sqrt{n_{1,z_{1i}}}}{\sqrt{\hat{n}_{1,\hat{z}_{1i}}}}\frac{1}{\sqrt{\hat{n}_{2,\hat{z}_{2j}}}} - \frac{1}{\sqrt{n_{2,z_{2j}}}}\right)^2 \\
&\leq \left(\max_t \sum_{s=1}^{K} B^2_{st}\right)\left(\sum_{j\in\mathcal{S}_2} 1\right)\frac{1}{n_{2,min}(1-\eta)^2} \\
&\leq C^2_{\eta,1}\left(\max_t \sum_{s=1}^{K} B^2_{st}\right)\frac{|\mathcal{S}_2|}{n_{2,min}} \leq C^2_{\eta,1}\left(\max_t \sum_{s=1}^{K} B^2_{st}\right)\eta
\end{aligned}$$

Similarly for the set of edges where the nodes of type-2 is mis-clustered and the nodes of type-1 is correctly estimated (3) $i \in \mathcal{S}_1, j \in \mathcal{S}_2^c$,

$$\sum_{i\in\mathcal{S}_1, j\in\mathcal{S}_2^c} \left(\hat{P}_{ij} - \tilde{P}_{ij}\right)^2 \leq C^2_{\eta,1}\left(\max_s \sum_{t=1}^{K} B^2_{st}\right)\eta$$

For the set of edges where both nodes are misclustered (4) $i \in \mathcal{S}_1, j \in \mathcal{S}_2$,

$$\begin{aligned}
\sum_{i\in\mathcal{S}_1, j\in\mathcal{S}_2} \left(\hat{P}_{ij} - \tilde{P}_{ij}\right)^2 &= \sum_{i\in\mathcal{S}_1, j\in\mathcal{S}_2} \left(\frac{p_{ij}}{\sqrt{\hat{n}_{1,\hat{z}_{1i}}\hat{n}_{2,\hat{z}_{2j}}}} - \frac{p_{ij}}{\sqrt{n_{1,z_{1i}}n_{2,z_{2j}}}}\right)^2 \\
&\leq \left(\sum_{i\in\mathcal{S}_1, j\in\mathcal{S}_2} B^2_{z_{1i}z_{2j}}\right) \max_{i,j} \left(\frac{1}{\sqrt{\hat{n}_{1,\hat{z}_{1i}}\hat{n}_{2,\hat{z}_{2j}}}} - \frac{1}{\sqrt{n_{1,z_{1i}}n_{2,z_{2j}}}}\right)^2 \\
&\leq \max_{s,t} B^2_{st}\left(\sum_{i\in\mathcal{S}_1, j\in\mathcal{S}_2} 1\right)\frac{1}{n_{1,min}n_{2,min}(1-\eta)^2} \\
&\leq \max_{s,t} B^2_{st}\frac{|\mathcal{S}_1||\mathcal{S}_2|}{n_{1,min}n_{2,min}(1-\eta)^2} \leq C^2_{\eta,1}\left(\max_{s,t} B^2_{st}\right)\eta^2
\end{aligned}$$

Therefore combining all those (i), (ii), (iii) and (iiii), we obtain

$$\begin{aligned}
\|\hat{P} - \tilde{P}\|_F^2 &\leq C^2_{\eta,1}\left(\left(\sum_{s=1}^{K}\sum_{t=1}^{K} B^2_{st}\right)\eta^2 + 2\left(\max_s \sum_{t=1}^{K} B^2_{st}\right)\eta + \max_{s,t} B^2_{st}\eta^2\right) \\
&\leq C^2_{\eta,1}\left(K^2\alpha_n^2\eta^2 + 2K\alpha_n^2\eta + \alpha_n^2\eta^2\right) \leq C^2_{\eta,1}\alpha_n^2\left(K\eta(K\eta+3)\right)
\end{aligned}$$

Therefore, finally we have

$$\|\hat{A} - \tilde{P}\|^2 \leq C_{\eta,2}^2 \left(\frac{C\sqrt{n\alpha_n}}{\sqrt{n_{1,min} n_{2,min}}} + \alpha_n \sqrt{K\eta(K\eta+3)} \right)^2 .$$

□

Proof of Theorem II.11

From (1+ε) approximate k-means algorithm, we have

$$\|\bar{T} - T\|_F^2 \leq (1+\varepsilon) \min_{Z,X} \|ZX - T\|_F^2 \leq (1+\varepsilon) \|\mathcal{T}\mathcal{Q}_2 - T\|_F^2$$

We further have $\|\bar{T} - \mathcal{T}\mathcal{Q}_2\|_F^2 \leq 2\left(\|\bar{T} - T\|_F^2 + \|T - \mathcal{T}\mathcal{Q}_2\|_F^2\right) \leq 2(2+\varepsilon)\|T - \mathcal{T}\mathcal{Q}_2\|_F^2$. Now, we can bound

$$\begin{aligned}
\|T - \mathcal{T}\mathcal{Q}_2\|_F^2 &= \left[\|\hat{W}_1^{1/2}\hat{U} - W_1^{1/2}\tilde{\mathcal{U}}\mathcal{Q}_2\|_F^2 + \|\hat{W}_2^{1/2}\hat{V} - -W_2^{1/2}\tilde{\mathcal{V}}\mathcal{Q}_2\|_F^2 \right] \\
&\leq 2\Big[\|\hat{W}_1^{1/2}\hat{U} - \hat{W}_1^{1/2}\tilde{\mathcal{U}}\mathcal{Q}_2\|_F^2 + \|(\hat{W}_1^{1/2} - W_1^{1/2})\tilde{\mathcal{U}}\mathcal{Q}_2\|_F^2 \\
&\quad + \|\hat{W}_2^{1/2}\hat{V} - \hat{W}_2^{1/2}\tilde{\mathcal{V}}\mathcal{Q}_2\|_F^2 + \|(\hat{W}_2^{1/2} - W_2^{1/2})\tilde{\mathcal{V}}\mathcal{Q}_2\|_F^2 \Big] \\
&\leq 2\hat{n}_{max} \left(\|\hat{U} - \tilde{\mathcal{U}}\mathcal{Q}_2\|_F^2 + \|\hat{V} - \tilde{\mathcal{V}}\mathcal{Q}_2\|_F^2 \right) \\
&\quad + \left(\|(\hat{W}_1^{1/2} - W_1^{1/2})\tilde{\mathcal{U}}\|_F^2 + \|(\hat{W}_2^{1/2} - W_2^{1/2})\tilde{\mathcal{V}}\|_F^2 \right).
\end{aligned}$$

$\|(\hat{W}_1^{1/2} - W_1^{1/2})\tilde{\mathcal{U}}\|_F^2$ can be re-expressed as

$$\begin{aligned}
\|(\hat{W}_1^{1/2} - W_1^{1/2})\tilde{\mathcal{U}}\|_F^2 &= tr(\mathcal{X}^\intercal \Delta_1^{-1/2} Z_1^\intercal (\hat{W}_1^{1/2} - W_1^{1/2})^2 Z_1 \Delta_1^{-1/2} \mathcal{X}) \\
&= tr(Z_1^\intercal (\hat{W}_1^{1/2} W_1^{-1/2} - I)^2 Z_1)
\end{aligned} \tag{A.6}$$

We first bound the (A.6),

$$\begin{aligned} tr(Z_1^\intercal(\hat{W}_1^{1/2}W_1^{-1/2} - I)^2 Z_1) &= \sum_i (\sqrt{\frac{\hat{n}_{1,\hat{z}_{1i}}}{n_{1,z_{1i}}}} - 1)^2 \\ &= \sum_{i\in\mathcal{S}_1^c} (\sqrt{\frac{\hat{n}_{1,\hat{z}_{1i}}}{n_{1,z_{1i}}}} - 1)^2 + \sum_{i\in\mathcal{S}_1} (\sqrt{\frac{\hat{n}_{1,\hat{z}_{1i}}}{n_{1,z_{1i}}}} - 1)^2 \\ &\leq \sum_{i\in\mathcal{S}_1^c} (\frac{\hat{n}_{1,\hat{z}_{1i}}}{n_{1,z_{1i}}} - 1)^2 + \sum_{i\in\mathcal{S}_1} (\frac{\hat{n}_{1,\hat{z}_{1i}}}{n_{1,z_{1i}}} - 1)^2 \\ &\leq \eta^2 \cdot n_1 + \frac{n_{1,max} + |\mathcal{S}_1|}{n_{1,min}} |\mathcal{S}_1| \\ &\leq (n_1\eta + n_{1,max} + |\mathcal{S}_1|)\eta \end{aligned}$$

In the fourth inequality, we used an inequality $(\sqrt{x} - 1)^2 \leq \max(1, x)$ for $x > 0$.

Similarly, $\|(\hat{W}_2^{1/2} - W_2^{1/2})\tilde{\mathcal{V}}\|_F^2 \leq (N\eta + n_{2,max} + |\mathcal{S}_2|)\eta$. Thus,

$$\begin{aligned} &\|(\hat{W}_1^{1/2} - W_1^{1/2})\tilde{\mathcal{U}}\|_F^2 + \|(\hat{W}_2^{1/2} - W_2^{1/2})\tilde{\mathcal{V}}\|_F^2 \\ &\qquad \leq (M\eta + n_{1,max} + |\mathcal{S}_1| + N\eta + n_{2,max} + |\mathcal{S}_2|)\eta \\ &\qquad \leq (n\eta + n_{1,max} + n_{2,max} + |\mathcal{S}_1| + |\mathcal{S}_2|)\eta \leq n(\eta + 3)\eta \end{aligned}$$

Then, the second term can be bounded by using Lemma A.1 and Lemma A.4,

$$\begin{aligned} 2\hat{n}_{max} \left(\|\hat{U} - \tilde{\mathcal{U}}\mathcal{Q}_2\|_F^2 + \|\hat{V} - \tilde{\mathcal{V}}\mathcal{Q}_2\|_F^2 \right) &\leq 2^2 n_{2,max} \left(\frac{2^4 K}{\alpha_n^2 \lambda_K^2} \|\hat{A} - \tilde{P}\|^2 \right) \\ &\leq \frac{2^6 n_{2,max} K}{\alpha_n \lambda_K^2} C_{\eta,1}^2 \left(\frac{C\sqrt{n}}{\sqrt{n_{1,min} n_{2,min}}} + \sqrt{\alpha_n K \eta (K\eta + 3)} \right)^2 \\ &\leq \frac{2^6 n_{2,max} K}{\alpha_n \lambda_K^2} C_{\eta,1}^2 \frac{C^2 n r K^2}{n_{1,min} n_{2,min}} \left(1/\sqrt{rK^2} + \sqrt{\frac{n_{1,min} n_{2,min}}{C^2 n r K^2} \alpha_n K \eta (K\eta + 3)} \right)^2 \\ &\leq C_{\eta,1}^2 2^6 C^2 n_{2,max} K^2 \frac{nrK}{n_{1,min} n_{2,min} \alpha_n \lambda_K^2} \left(1 + \sqrt{\frac{n_{1,min} n_{2,min}}{C^2 n r K^2} \alpha_n K \eta (K\eta + 3)} \right)^2 \\ &\leq C_{\eta,1}^2 \frac{n_{2,max} K^2}{(2+\varepsilon)} \beta \left(1 + \sqrt{2^6 \frac{(2+\varepsilon)}{\lambda_K^2} (K\eta + 3)} \right)^2 \end{aligned}$$

where $\beta = c^{-1}(2+\varepsilon) r \frac{Kn}{n_{1,min} n_{2,min} \lambda_K^2 \alpha_n}$ and $\eta \leq \beta$. First inequality follows from Lemma A.1 and second inequality is from Lemma A.4. For the fifth inequality, $\eta \leq c^{-1}(2+\varepsilon) \frac{Krn}{n_{1,min} n_{2,min} \lambda_K^2 \alpha_n}$ from Proposition II.7 with definition η is used.

Recall mis-clustered nodes are defined as in definition II.10.

$$
\begin{aligned}
|\mathcal{S}| &\leq \sum_{i\in\mathcal{S}} 2\|\bar{T}_{i*} - \mathcal{T}_{i*}\mathcal{Q}_2\|^2 \\
&\leq 2\cdot 2(2+\varepsilon)\|T - \mathcal{T}\mathcal{Q}_2\|_F^2 \\
&\leq 2^2(2+\varepsilon)\left[2\hat{n}_{max}\left(\|\hat{U} - \tilde{\mathcal{U}}\mathcal{Q}_2\|_F^2 + \|\hat{V} - \tilde{\mathcal{V}}\mathcal{Q}_2\|_F^2\right)\right. \\
&\quad \left. + \|(\hat{W}_1^{1/2} - W_1^{1/2})\tilde{\mathcal{U}}\|_F^2 + \|(\hat{W}_2^{1/2} - W_2^{1/2})\tilde{\mathcal{V}}\|_F^2\right] \\
&\leq 2^2(2+\varepsilon)\left[C_{\eta,1}^2\frac{n_{2,max}K^2}{(2+\varepsilon)}\beta\left(1+\sqrt{2^6\frac{(2+\varepsilon)}{\lambda_K^2}(K\eta+3)}\right)^2 + n(\eta+3)\eta\right]
\end{aligned}
$$

Therefore, we have

$$
|\mathcal{S}|/n \leq 2^2C_{\eta,1}^2\frac{n_{2,max}K^2}{n}\beta\left(1+\sqrt{2^6\frac{(2+\varepsilon)}{\lambda_K^2}(K\beta+3)}\right)^2 + 2^2(2+\varepsilon)(\beta+3)\beta.
$$

□

C.1 Algorithms for degree-corrected SBM

Algorithm A.1 Using adjacency matrix

1: Input: bipartite adjacency matrix $A \in \{0,1\}^{n_1\times n_2}$ and K
2: Compute K left and K right singular vectors $U \in \mathbb{R}^{n_1\times K}$ and $V \in \mathbb{R}^{n_2\times K}$ corresponding to the K largest singular values of A. Normalize each row of U and V to have unit length. Let normalized version of U and V be U^* and V^*, respectively. Run k-means separately on rows of U^* and rows of V^*.
3: Based on the result from 2, construct diagonal matrices $\hat{W}_1$ and $\hat{W}_2$, diagonal elements being the community size for each type where the node belongs.
4: Let $\hat{A} = \hat{W}_1^{-1/2}A\hat{W}_2^{-1/2}$. Compute K left and K right singular vectors $\hat{U} \in \mathbb{R}^{n_1\times K}$ and $\hat{V} \in \mathbb{R}^{n_2\times K}$ corresponding to the K largest singular values of $\hat{A}$. Normalize each row of $\hat{U}$ and $\hat{V}$ to have unit length. Let normalized version of $\hat{U}$ and $\hat{V}$ be $\hat{U}^*$ and $\hat{V}^*$, respectively.
5: Concatenate $\hat{U}^*$ and $\hat{V}^*$ and run k-means with K clusters at the same time.

Algorithm A.2 Using laplacian matrix

1: Input: bipartite adjacency matrix $A \in \{0,1\}^{n_1\times n_2}$ and K
2: Form $L = D_1^{-1/2}AD_2^{-1/2}$. Compute K left and K right singular vectors $U \in \mathbb{R}^{n_1\times K}$ and $V \in \mathbb{R}^{n_2\times K}$ corresponding to the K largest singular values of L. Normalize each row of U and V to have unit length. Let normalized version of U and V be U^* and V^*, respectively. Run k-means separately on rows of U^* and rows of V^*.
3: Based on the result from 2, construct diagonal matrices $\hat{W}_1$ and $\hat{W}_2$, diagonal elements being the community size for each type where the node belongs.
4: Let $\hat{A} = \hat{W}_1^{-1/2}A\hat{W}_2^{-1/2}$, $\hat{D}_1$=diag$(\sum_j A_{ij}/\hat{n}_{2,\hat{z}_{2j}}, i = 1,\cdots,n_1)$ and $\hat{D}_2$=diag$(\sum_i A_{ij}/\hat{n}_{1,\hat{z}_{1i}}, j = 1,\cdots,n_2)$.
5: Let $\hat{L} = \hat{D}_1^{-1/2}\hat{A}\hat{D}_2^{-1/2}$. Compute K left and K right singular vectors $\hat{U} \in \mathbb{R}^{n_1\times K}$ and $\hat{V} \in \mathbb{R}^{n_2\times K}$ corresponding to the K largest singular values of $\hat{L}$. Normalize each row of $\hat{U}$ and $\hat{V}$ to have unit length. Let normalized version of $\hat{U}$ and $\hat{V}$ be $\hat{U}^*$ and $\hat{V}^*$, respectively.
6: Concatenate $\hat{U}^*$ and $\hat{V}^*$ and run k-means with K clusters at the same time.

APPENDIX B

Appendix of Chapter III

A Proof for Section 3.2

Proof of Lemma III.3

$$\begin{aligned}
\mathcal{L}_\tau &= (\mathcal{D}_l + \tau I)^{-1/2} \Theta W Z^T \Phi (\mathcal{D}_r + \tau I)^{-1/2} \\
&= diag\left((\theta_i + \tau)^{-1/2}\right) \cdot \Theta W Z^T \Phi \cdot diag\left((\phi_j \sum_i \theta_i w_{ig_j} + \tau)^{-1/2}\right) \\
&= \tilde{\Theta} W Z^T \tilde{\Phi} \\
&= \left(\tilde{\Theta} W (Z^T \tilde{\Phi}^2 Z)^{1/2}\right) (Z^T \tilde{\Phi}^2 Z)^{-1/2} Z^T \tilde{\Phi} = \left(\tilde{\Theta} W \tilde{\Psi}\right) \tilde{\Psi}^{-1} Z^T \tilde{\Phi}
\end{aligned}$$

where $\tilde{\Theta}_{ii} = \frac{\theta_i}{\sqrt{\theta_i + \tau}}$ and $\tilde{\Phi}_{jj} = \frac{\phi_j}{\sqrt{\phi_j \sum_i \theta_i w_{ig_j} + \tau}}$. In the fourth equation, we can simplify the notation more by letting $\tilde{\Psi} = (Z^T \tilde{\Phi}^2 Z)^{1/2} \in \mathbb{R}^{K \times K}$. Let $\mathcal{H} = \tilde{\Theta} X \tilde{\Psi} \in \mathbb{R}^{n \times K}$. If we let singular value decomposition of $\mathcal{H}$ as $\mathcal{U}\mathcal{D}\mathcal{C}^\intercal$, singular value decomposition of $\mathcal{L}_\tau$ is $\mathcal{U}\mathcal{D}\mathcal{V}^\intercal$ where $\mathcal{V} = \tilde{\Phi} Z \tilde{\Psi}^{-1} \mathcal{C}$ is the right singular vector. After row normalization (row vector v divided by $\|v\|_2$) of $\mathcal{V}$, we get $Z\mathcal{C}$. We can easily check $\mathcal{V}^\intercal \mathcal{V} = I_K$ and $\mathcal{C}^\intercal \mathcal{C} = I_K$. □

Proof of Lemma III.5

1. To show equality, we will add and subtract $Z(Z^TZ)^{-1}Z^T$ matrix as follows.

$$\begin{aligned}
\mathcal{H}^T\mathcal{H} &= \tilde{\Psi}W^T\tilde{\Theta}^2W\tilde{\Psi} \\
&= \tilde{\Psi}W^T\tilde{\Theta}\left(Z(Z^TZ)^{-1}Z^T - Z(Z^TZ)^{-1}Z^T + I\right)\tilde{\Theta}W\tilde{\Psi} \\
&= \tilde{\Psi}(\tilde{\Theta}W)^T\left(Z(Z^TZ)^{-1}Z^T\right)\tilde{\Theta}W\tilde{\Psi} \\
&\quad + \tilde{\Psi}(\tilde{\Theta}W)^T\left(I - Z(Z^TZ)^{-1}Z^T\right)\tilde{\Theta}W\tilde{\Psi} \\
&= \mathcal{H}_M^T\mathcal{H}_M + \mathcal{H}_S^T\mathcal{H}_S
\end{aligned}$$

where $\mathcal{H}_M = \left(Z(Z^TZ)^{-1}Z^T\right)\tilde{\Theta}W\tilde{\Psi}$ and $\mathcal{H}_S = \left(I - Z(Z^TZ)^{-1}Z^T\right)\tilde{\Theta}W\tilde{\Psi}$. Fourth equality follows from the fact that $\left(Z(Z^TZ)^{-1}Z^T\right)^2 = Z(Z^TZ)^{-1}Z^T$ and $\left(I - Z(Z^TZ)^{-1}Z^T\right)^2 = \left(I - Z(Z^TZ)^{-1}Z^T\right)$.

2. After some calculation, we see that $(s,\, t)$ element of a matrix $\left((Z^TZ)^{-1}Z^T\right)\tilde{\Theta}W\tilde{\Psi} \in \mathbb{R}^{K\times K}$ is $M_{st}\tilde{\psi}_{tt}$ where $M_{st} = \frac{1}{n_s}\sum_{i\in g_s}\tilde{\theta}_i w_{it}$. Then, the $(i,\, t)$ element of $\mathcal{H}_M \in \mathbb{R}^{n\times K}$ is $M_{g_i t}$. Since $\left(\mathcal{H}_M^T\mathcal{H}_M\right)_{st}$ is the inner product of $\mathcal{H}_M$'s sth column and $\mathcal{H}_M$'s tth column, we have

$$\left(\mathcal{H}_M^T\mathcal{H}_M\right)_{st} = \tilde{\psi}_{ss}\left(\sum_{r=1}^{K} n_r M_{rs}M_{rt}\right)\tilde{\psi}_{tt}$$

Similarly, using the result from above, the $(i,\, t)$ element of $\mathcal{H}_S \in \mathbb{R}^{n\times K}$ is $\left(\tilde{\theta}_i w_{it} - M_{g_i t}\right)\tilde{\psi}_{tt}$. Since the element $\left(\mathcal{H}_S^T\mathcal{H}_S\right)_{st}$ is an inner product of $\mathcal{H}_S$'s sth column and $\mathcal{H}_S$'s tth column, we have

$$\left(\mathcal{H}_S^T\mathcal{H}_S\right)_{st} = \tilde{\psi}_{ss}\sum_{r=1}^{K} n_r\sigma_{r,st}^2\tilde{\psi}_{tt}$$

where $\sigma_{r,st}^2 = \frac{1}{n_r}\sum_{i\in g_r}\left(\tilde{\theta}_i w_{is} - M_{rs}\right)\left(\tilde{\theta}_i w_{it} - M_{rt}\right)$.

3. $\mathcal{H}_M^T\mathcal{H}_M$ and $\mathcal{H}_S^T\mathcal{H}_S$ are positive semi-definite matrix with eigenvalues $\sigma_1(\mathcal{H}_M^T\mathcal{H}_M) \geq \cdots \geq \sigma_K(\mathcal{H}_M^T\mathcal{H}_M) \geq 0$ and $\sigma_1(\mathcal{H}_S^T\mathcal{H}_S) \geq \cdots \geq \sigma_K(\mathcal{H}_S^T\mathcal{H}_S) \geq 0$. By applying weyl's inequality (Bhatia, 1987), we obtain the following inequality.

$$\sigma_K(\mathcal{H}_M^T\mathcal{H}_M) + \sigma_K(\mathcal{H}_S^T\mathcal{H}_S) \leq \sigma_K(\mathcal{H}^T\mathcal{H}) \leq \sigma_K(\mathcal{H}_M^T\mathcal{H}_M) + \sigma_1(\mathcal{H}_S^T\mathcal{H}_S). \quad \square$$

B Consistency

The matrix concentration inequalities results for symmetric matrices can be generalized for general matrices. We first introduce (Symmetric) dilation operator as in Tropp et al. (2015), which embeds matrices to larger block matrices.

Definition B.1. The Symmetric dilation $\mathcal{S} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{S}^{d_1+d_2}$ is the map from a general matrix B to an Symmetric matrix defined by

$$\mathcal{S}(B) = \begin{bmatrix} 0 & B \\ B^\intercal & 0 \end{bmatrix}$$

where $\mathbb{S}^n$ is the set of symmetric $n \times n$ matrices.

We can express singular decomposition of $\mathcal{S}(B)$ using that of $B \in \mathbb{R}^{d_1 \times d_2}$. The following lemma provides an important property of dilation.

Lemma B.2. *For any rank $K \leq \min(d_1, d_2)$ matrix $B \in \mathbb{R}^{d_1 \times d_2}$, let $\mathcal{U}\mathcal{D}\mathcal{V}^\intercal$ be the singular value decomposition of B, where $\mathcal{D} \in \mathbb{R}^{K \times K}$, $\mathcal{U} \in \mathbb{R}^{d_1 \times K}$ and $\mathcal{V} \in \mathbb{R}^{d_2 \times K}$. Then, it is easy to check singular value decomposition of $\mathcal{S}(B)$ is*

$$\mathcal{W} \cdot \begin{bmatrix} \mathcal{D} & 0 \\ 0 & -\mathcal{D} \end{bmatrix} \cdot \mathcal{W}^\intercal \quad \textit{where} \quad \mathcal{W} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathcal{U} & \mathcal{U} \\ \mathcal{V} & -\mathcal{V} \end{bmatrix}.$$

Moreover, $\|\mathcal{S}(B)\| = \|B\|$.

We also define diagonal matrices D and $\mathcal{D}$ from $\mathcal{S}(L)$ and $\mathcal{S}(\mathcal{L})$ as follows.

$$D = \begin{bmatrix} D_l & 0 \\ 0 & D_r \end{bmatrix}, \quad \mathcal{D} = \begin{bmatrix} \mathcal{D}_l & 0 \\ 0 & \mathcal{D}_r \end{bmatrix}$$

D is the diagonal matrix of out-going degree and in-coming degree. $\mathcal{D}$ is population version of D. We slightly modified the Theorem 4.1 from Le et al. (2017) to show the concentration inequality for a different version of regularized Laplacian matrix. Basic idea is to bound $\|(D_r+\tau I)^{-1/2}(A-P)(D_l+\tau I)^{-1/2}\|$ and $\|(D_r+\tau I)^{-1/2}P(D_l+\tau I)^{-1/2}-(\mathcal{D}_r+\tau I)^{-1/2}P(\mathcal{D}_l+\tau I)^{-1/2}\|$ separately. The Theorem 4.1 does not change with a different version of regularized Laplacian matrix since $0 \leq P_{ij} \leq \frac{d}{n}$ in the Step 3 of the Theorem 4.1's proof (Le et al., 2017).

Proposition B.3. *Consider an adjacency matrix A of a random graph with $\mathbb{E}[A] = P = (p_{ij})_{i,j=1,\cdots,n}$, and let $d = n\max_{ij} p_{ij}$. Choose a number $\tau > 0$. Then, for any $r \geq 1$, there exists a constant $C > 0$, with probability at least $1 - e^{-r}$,*

$$\|S(L_\tau) - S(\mathcal{L}_\tau)\| \leq \frac{Cr^2}{\sqrt{\tau}}\left(1+\frac{d}{\tau}\right)^{5/2}.$$

Here is the modified Lemma 5.1 from Lei et al. (2015).

Lemma B.4. *(concentration of singular space)*

Assume that $\mathcal{L} \in \mathbb{R}^{n_1\times n_2}$ is a rank K matrix with smallest non-zero singular value $\gamma_K > 0$. Let $L \in \mathbb{R}^{n_1\times n_2}$ and $U,\mathcal{U} \in \mathbb{R}^{n_1\times K}$ contain the top K left singular vectors of L and $\mathcal{L}$, respectively. Similarly, let $V,\mathcal{V} \in \mathbb{R}^{n_2\times K}$ contain the top K right singular vectors of L and $\mathcal{L}$, respectively. Then there exist $K \times K$ orthogonal matrix $\mathcal{Q} \in \mathbb{R}^{K\times K}$ such that

$$\|U - \mathcal{U}\mathcal{Q}\|_F \text{ or } \|V - \mathcal{V}\mathcal{Q}\|_F \leq \sqrt{\|U-\mathcal{U}\mathcal{Q}\|_F^2 + \|V-\mathcal{V}\mathcal{Q}\|_F^2} \leq \frac{4\sqrt{K}}{\gamma_K}\|L-\mathcal{L}\|$$

Proof.

$\mathcal{S}(\mathcal{L})$ has 2K leading eigenvectors $\mathcal{W} \in \mathbb{R}^{n\times 2K}$, where $\mathcal{W} = \frac{1}{\sqrt{2}}\begin{bmatrix}\mathcal{U} & \mathcal{U}\\ \mathcal{V} & -\mathcal{V}\end{bmatrix}$ and

$\mathcal{S}(L)$ has 2K leading eigenvectors $W \in \mathbb{R}^{n\times 2K}$, where $W = \frac{1}{\sqrt{2}}\begin{bmatrix}U & U\\ V & -V\end{bmatrix}$ and

$\|\mathcal{S}(L) - \mathcal{S}(\mathcal{L})\| = \|L - \mathcal{L}\|$ by lemma B.2.

We first establish that $\|(I - WW^T)\mathcal{W}\mathcal{W}^T\| \leq 2\frac{\|\mathcal{S}(L)-\mathcal{S}(\mathcal{L})\|}{\gamma_K}$. If $\|\mathcal{S}(L)-\mathcal{S}(\mathcal{L})\| \leq \frac{\gamma_K}{2}$, then by Davis-Kahan theorem, we have

$$\|(I - WW^T)\mathcal{W}\mathcal{W}^T\| \leq \frac{\|\mathcal{S}(L) - \mathcal{S}(\mathcal{L})\|}{\gamma_K - \|\mathcal{S}(L) - \mathcal{S}(\mathcal{L})\|} \leq 2\frac{\|L - \mathcal{L}\|}{\gamma_K}$$

If $\|\mathcal{S}(L) - \mathcal{S}(\mathcal{L})\| > \frac{\gamma_K}{2}$, then

$$\|(I - WW^T)\mathcal{W}\mathcal{W}^T\| \leq 1 \leq 2\frac{\|L - \mathcal{L}\|}{\gamma_K}.$$

By Proposition 2.2 in Vu et al. (2013b), there exists a K-dimensional orthogonal matrix $\mathcal{Q}$ such that

$$\begin{aligned}\frac{1}{\sqrt{2}}\sqrt{\|U - \mathcal{U}\mathcal{Q}\|_F^2 + \|V - \mathcal{V}\mathcal{Q}\|_F^2} = \|W - \mathcal{W}\mathcal{Q}\|_F &\leq \sqrt{2}\|(I - WW^T)\mathcal{W}\mathcal{W}^T\|_F \\ &\leq \sqrt{2K}\|(I - WW^T)\mathcal{W}\mathcal{W}^T\|.\end{aligned}$$

Thus, we have

$$\|U - \mathcal{U}\mathcal{Q}\|_F \text{ or } \|V - \mathcal{V}\mathcal{Q}\|_F \leq \sqrt{\|U - \mathcal{U}\mathcal{Q}\|_F^2 + \|V - \mathcal{V}\mathcal{Q}\|_F^2} \leq \frac{4\sqrt{K}}{\gamma_K}\|L - \mathcal{L}\|. \quad \square$$

Proof of Theorem III.8

From the Lemma B.4,

$$\begin{aligned}\sqrt{\|U - \mathcal{U}\mathcal{Q}\|_F^2 + \|V - \mathcal{V}\mathcal{Q}\|_F^2} &\leq \frac{4\sqrt{K}}{\gamma_K}\|L_\tau - \mathcal{L}_\tau\| \\ &\leq \frac{4\sqrt{K}}{\sqrt{\lambda_M + \lambda_S}}\|L_\tau - \mathcal{L}_\tau\| \\ &\leq \frac{4Cr^2\sqrt{K}}{\sqrt{\tau(\lambda_M + \lambda_S)}}\left(1 + \frac{d}{\tau}\right)^{5/2}.\end{aligned}$$

Second inequality follows from the Lemma III.5. Third inequality follows from the Proposition B.3. $\square$

Proof of Theorem III.10

$$
\begin{aligned}
|\mathcal{S}|/n &\leq \frac{1}{n}\sum_i 2\|\bar{V}_i^* - \mathcal{V}^*\mathcal{Q}\|^2 \\
&\leq \frac{2}{n}\|\bar{V}_i^* - \mathcal{V}^*\mathcal{Q}\|_F^2 \\
&\leq \frac{2(1+\alpha)}{n}\|V^* - \mathcal{V}^*\mathcal{Q}\|_F^2 \\
&\leq \frac{8(1+\alpha)}{nm_r^2}\|V - \mathcal{V}\mathcal{Q}\|_F^2 \\
&\leq \frac{2^7 C^2 r^4 (1+\alpha)K}{nm_r^2(\lambda_M+\lambda_S)\tau}\left(1+\frac{d}{\tau}\right)^5
\end{aligned}
$$

Fourth inequality follows from the fact that for any nonzero vectors v_1, v_2 of same dimension, we have $\|\frac{v_1}{\|v_1\|} - \frac{v_2}{\|v_2\|}\| \leq 2\frac{\|v_1 - v_2\|}{\max(\|v_1\|,\|v_2\|)}$. Last inequality is due to Theorem III.8.

APPENDIX C

Appendix of Chapter IV

A A derivation of likelihood in Section 4.3

$$
\begin{aligned}
&\sum_{i<j} A_{ij}^{(1,0)} \log(P_{ij}^{(1,0)}) + A_{ij}^{(0,1)} \log(P_{ij}^{(0,1)}) + A_{ij}^{(1,1)} \log(P_{ij}^{(1,1)}) + A_{ij}^{(0,0)} \log(P_{ij}^{(0,0)}) \\
&= \frac{1}{2} \sum_{i \neq j} A_{ij}^{(1,0)} \log(P_{ij}^{(1,0)}/P_{ij}^{(0,0)}) + A_{ij}^{(0,1)} \log(P_{ij}^{(0,1)}/P_{ij}^{(0,0)}) \\
&\quad + A_{ij}^{(1,1)} \log(P_{ij}^{(1,1)}/P_{ij}^{(0,0)}) + \log(P_{ij}^{(0,0)}) \\
&= \frac{1}{2} \sum_{i \neq j} A_{ij}^{(1,0)} \Theta_{ij}^{(1,0)} + A_{ij}^{(0,1)} \Theta_{ij}^{(0,1)} + A_{ij}^{(1,1)} \Theta_{ij}^{(1,1)} - \log\left(1 + e^{\Theta_{ij}^{(1,0)}} + e^{\Theta_{ij}^{(0,1)}} + e^{\Theta_{ij}^{(1,1)}}\right) \\
&= \sum_{i \neq j} A_{ij}^{(1,0)} \Theta_{ij}^{(1,0)} + \frac{1}{2} \sum_{i \neq j} A_{ij}^{(1,1)} \Theta_{ij}^{(1,1)} - \frac{1}{2} \sum_{i \neq j} \log\left(1 + e^{\Theta_{ij}^{(1,0)}} + e^{\Theta_{ij}^{(0,1)}} + e^{\Theta_{ij}^{(1,1)}}\right)
\end{aligned}
$$

B Algorithms for pair-independent latent space model

We also present projected gradient descent algorithm for pair-independent latent space model. Initialization algorithm for pair-independent latent space model is also introduced here.

Algorithm C.1 A projected gradient descent algorithm

Input: Adjacency matrix: A; latent space dimension: K;
initial estimates: $U^0, \boldsymbol{a}^0, \boldsymbol{b}^0, R^0$;
step size: η
Output: $\hat{U} = U^J, \hat{R} = R^J, \hat{\boldsymbol{a}} = \boldsymbol{a}^J, \hat{\boldsymbol{b}} = \boldsymbol{b}^J$
for t = 0,1,...,T-1 **do**
  $\tilde{U}^{t+1} = U^t + 2\eta\left((A-P)U^t(R^t)^\intercal + (A-P)^\intercal U^t R^t\right)$
  $\boldsymbol{a}^{t+1} = \boldsymbol{a}^t + 2\eta\,(A-P)\,1_n$
  $\tilde{\mathbf{b}}^{t+1} = \mathbf{b}^t + 2\eta\,(A-P)^\intercal\,1_n$
  $R^{t+1} = R^t + 2\eta(U^t)^\intercal(A-P)U^t$
  $U^{t+1} = \mathcal{P}_U(\tilde{U}^{t+1}),\ \mathbf{b}^{t+1} = \mathcal{P}_b(\tilde{\mathbf{b}}^{t+1})$
  Update P using U^{t+1}, $\boldsymbol{a}^{t+1}$, $\boldsymbol{b}^{t+1}$ and R^{t+1}
end for

C An Initialization Algorithm

Algorithm C.2 Initialization algorithm

Input: Adjacency matrix: A; latent space dimension: K;
Let $\tilde{P}^{(1,0)} = \sum_{i=1}^{K}\sigma_i u_i v_i^\intercal$ where $\sum_{i=1}^{n}\sigma_i u_i v_i^\intercal$ is the SVD of $A^{(1,0)}$. Then, project each element of $\tilde{P}^{(1,0)}$ onto the interval $[\varepsilon, 1-\varepsilon]$ for some small $0 < \varepsilon < 1/2$ to obtain $\hat{P}^{(1,0)}$. Let $\hat{\Theta}^{(1,0)} = logit(\hat{P}^{(1,0)})$.
Similarly construct $\hat{\Theta}^{(1,1)}$ with $A^{(1,1)}$.
Let $\boldsymbol{a}^0 = \frac{1}{n}\hat{\Theta}^{(1,0)}1_n$ and $\boldsymbol{b}^0 = \frac{1}{n}\hat{\Theta}^\intercal 1_n - \bar{\theta}^{(1,0)}1_n$ where $\bar{\theta}^{(1,0)} = \frac{1}{n^2}\sum_{i,j}\hat{\Theta}_{ij}^{(1,0)}$. Let $\boldsymbol{c}^0 = \frac{1}{n}\hat{\Theta}^{(1,1)}1_n - \bar{\theta}^{(1,0)}1_n$ where $\bar{\theta}^{(1,0)} = \frac{1}{2n^2}\sum_{i,j}\hat{\Theta}_{ij}^{(1,1)}$.
Let $\tilde{\Theta}^{(1,0)} = \left(\hat{\Theta}^{(1,0)} + \hat{\Theta}^{(1,0)\intercal} - \left(\boldsymbol{a}^0 + \boldsymbol{b}^0\right)\mathbf{1}^\intercal - 1_n\left(\boldsymbol{a}^0 + \boldsymbol{b}^0\right)^\intercal\right)/2$.
Also, let $\tilde{\Theta}^{(1,1)} = \left(\hat{\Theta}^{(1,1)} - \boldsymbol{c}^0 1_n^\intercal - 1_n \boldsymbol{c}^{0\intercal}\right)$.
Let $\tilde{\Theta} = \tilde{\Theta}^{(1,1)} + \tilde{\Theta}^{(1,0)}$.
Find the eigenvectors U_k of $\tilde{\Theta}$ corresponding to the k eigenvalues with largest magnitude. Set $U^0 = U_k$.
Set $R = U^\intercal\tilde{\Theta}^{(1,0)}U$ and $S = U^\intercal\tilde{\theta}^{(1,1)}U$.
end for

D Proofs of Theorems

Following lemma about concentration of a random (directed) graph adjacency matrix is from Le et al. (2017). Similar statement can be also found in Lei et al. (2015).

Lemma C.1. *Let A be the adjacency matrix of a random graph on n nodes in which edges occur independently. Let $\mathbb{E}[A_{ij}] = P_{ij}$ for all $i, j = 1, \cdots, n$. Assume that $n\max_{i,j} \leq d$ for $d \geq c_0 \log n$ and $c_0 > 0$. Then for any $r > 0$, there is a constant*

$C = C(r, c_0)$ *such that*

$$\|A - P\| \leq C\sqrt{d} \tag{C.1}$$

with probability at least $1 - n^{-r}$.

Lemma C.2. *There exist absolute constants* r, C *such that for any* $\mathbf{\Theta} \in \mathcal{F}$ *with probability at least* $1 - 2n^{-r}$, *the following inequality holds*

$$\max\left\{\|A^{(1,0)} - P^{(1,0)}\|, \|A^{(1,1)} - P^{(1,1)}\|\right\} \leq C\sqrt{\max\{ne^{\alpha_u}, \log n\}}.$$

Proof. For $\Theta^{(1,0)}$ and $\Theta^{(1,1)}$ in the parameter space, the off diagonal elements of two matrices are uniformly bounded above by α_u from our assumption. Thus, $\max_{ij}(P^{(1,0)}_{0,ij})$, $\max_{ij}(P^{(1,1)}_{0,ij}) \leq \sigma(\alpha_u)$. In our model, $\max_i P^{(1,0)}_{ii} \leq 1$ and $\max_i P^{(1,1)}_{ii} \leq 1$. We have $\|A^{(1,0)} - P^{(1,0)}\| \leq \|A^{(1,0)} - P^{(1,0)}_0\| + \|P^{(1,0)}_0 - P^{(1,0)}\| \leq \|A^{(1,0)} - P^{(1,0)}_0\| + 1$. Similarly, $\|A^{(1,1)} - P^{(1,1)}\| \leq \|A^{(1,1)} - P^{(1,1)}_0\| + 1$. Recall each variable $A^{(1,0)}_{ij}$ for any $i \neq j$ and $A^{(1,1)}_{ij}$ for any $i < j$ has a binomial distribution. By Lemma C.1, this implies that there exist absolute constants $r, C > 0$ such that

$$\mathbb{P}\left(\|A^{(1,0)} - P^{(1,0)}\| \leq C\sqrt{\max\{ne^{\alpha_u}, \log n\}}\right) \geq 1 - n^{-r}$$
$$\mathbb{P}\left(\|A^{(1,1)} - P^{(1,1)}\| \leq C\sqrt{\max\{ne^{\alpha_u}, \log n\}}\right) \geq 1 - n^{-r}$$

Thus, there exist some absolute constants $r, C > 0$ such that

$$\mathbb{P}\left(\max\left\{\|A^{(1,0)} - P^{(1,0)}\|, \|A^{(1,1)} - P^{(1,1)}\|\right\} \leq C\sqrt{\max\{ne^{\alpha_u}, \log n\}}\right) \geq 1 - 2n^{-r}.$$

□

Proof of Theorem IV.2.

$\hat{\mathbf{\Theta}}$ is the (global) optimal solution to (4.12) and the true parameter $\mathbf{\Theta}_*$ is feasible in the optimization. Let $f(\Theta) = -l(\Theta)$ be the negative log likelihood. Thus, we have the inequality,

$$f(\hat{\mathbf{\Theta}}) - f(\mathbf{\Theta}_*) \leq 0. \tag{C.2}$$

since $\hat{\mathbf{\Theta}}$ is the optimal solution.

For any $\Theta_{ij}^{(1,0)}, \Theta_{ij}^{(1,1)} \in \mathcal{F}$, $|\Theta_{ij}^{(1,0)}|, |\Theta_{ij}^{(1,1)}| \geq \alpha_l$ for all $i, j = 1, \cdots, n$. Let $\beta = \min_{x \in [\alpha_l, \alpha_u]} \sigma(x)\ (1 - \sigma(x)) = \sigma(\alpha_l)(1 - \sigma(\alpha_l)) \geq \frac{1}{4} e^{\alpha_l}$. The Hessian matrix of $f(\mathbf{\Theta})$ is

(C.3)

$$\nabla^2 f(\mathbf{\Theta}) = diag\left(vec\left(2\sigma(\Theta^{(1,0)}) \circ (1 - \sigma(\Theta^{(1,0)}))\right), vec\left(\sigma(\Theta^{(1,1)}) \circ (1 - \sigma(\Theta^{(1,1)}))\right)\right.$$
$$\geq \beta I_{2n^2 \times 2n^2}.$$

Here, $diag(a, b)$ is a diagonal matrix with elements a and b on its diagonals for any vectors a and b. For any matrix $B = [b_1, \cdots, b_n] \in \mathbb{R}^{n \times n}$, $vec(B) \in \mathbb{R}^{n^2}$ is obtained by vectorizing B matrix as $([b_1^\intercal, \cdots, b_n^\intercal])$. For notational simplicity, $\mathbf{\Theta} = [\Theta^{(1,0)\intercal}, \Theta^{(1,1)\intercal}]^\intercal \in \mathbb{R}^{2n \times n}$. Taylor expansion at Θ_* gives

$$f(\hat{\mathbf{\Theta}}) - f(\mathbf{\Theta}_*) \quad \geq \quad < \nabla_{\mathbf{\Theta}} f(\mathbf{\Theta}_*), \hat{\mathbf{\Theta}} - \mathbf{\Theta}_* > + \frac{\beta}{2} \|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2 \tag{C.4}$$

using Taylor's theorem since $f(\mathbf{\Theta})$ is convex function with respect to parameter $\mathbf{\Theta}$. Here $< M_1, M_2 >= tr(M_1^\intercal M_2)$ denotes inner product of two matrices M_1 and M_2. Together with (C.2) and (C.4) implies

(C.5)

$$\begin{aligned}
\frac{\beta}{2} \|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2 &= \frac{\beta}{2} \left(\|\hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)}\|_F^2 + \|\hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)}\|_F^2 \right) \\
&\leq \left| < \nabla_{\mathbf{\Theta}} f(\mathbf{\Theta}_*), \hat{\mathbf{\Theta}} - \mathbf{\Theta}_* > \right| \\
&= \left| < A^{(1,0)} - P^{(1,0)}, \hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)} > + \frac{1}{2} < A^{(1,1)} - P^{(1,1)}, \hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)} > \right| \\
&\leq \left| < A^{(1,0)} - P^{(1,0)}, \hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)} > \right| + \frac{1}{2} \left| < A^{(1,1)} - P^{(1,1)}, \hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)} > \right|.
\end{aligned}$$

Note that $|< M_1, M_2 >| \leq \|M_1\|_2 \|M_2\|_* \leq \|M_1\|_2 rank(M_2) \|M_2\|_F$ from matrix

norm inequalities. If we let $\lambda_n = \max\left\{\|A^{(1,0)} - P^{(1,0)}\|, \|A^{(1,1)} - P^{(1,1)}\|\right\}$, we have

$$\begin{aligned}
&\left|< A^{(1,0)} - P^{(1,0)}, \hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)} >\right| + \frac{1}{2}\left|< A^{(1,1)} - P^{(1,1)}, \hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)} >\right| \\
&\leq \left(\|A^{(1,0)} - P^{(1,0)}\|\sqrt{2(K+2)}\|\hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)}\|_F\right) \\
&\quad + \left(\frac{1}{2}\|A^{(1,1)} - P^{(1,1)}\|\sqrt{2(K+2)}\|\hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)}\|_F\right) \\
&\leq \frac{1}{2}\lambda_n\sqrt{2(K+2)}\left(2\|\hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)}\|_F + \|\hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)}\|_F\right) \\
&\leq \lambda_n\sqrt{(K+2)}\sqrt{4\|\hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)}\|_F^2 + \|\hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)}\|_F^2} \\
&\leq 2\lambda_n\sqrt{(K+2)}\sqrt{\|\hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)}\|_F^2 + \|\hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)}\|_F^2}
\end{aligned}$$

First inequality comes from the inequality that $\|M\|_* \leq \sqrt{rank(M)}\|M\|_F$ for any matrix M. Note that $rank\left(\hat{\Theta}^{(1,0)} - \Theta^{(1,0)}\right) \leq rank\left(\hat{\Theta}^{(1,0)}\right) + rank\left(\Theta^{(1,0)}\right) \leq 2(K+2)$. Similarly, $rank\left(\hat{\Theta}^{(1,1)} - \Theta^{(1,1)}\right) < 2(K+2)$. Third inequality comes from the fact $(a+b)^2 \leq 2(a^2+b^2)$ for any $a, b \geq 0$. Together with (C.5) and (D), we have

$$\|\hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)}\|_F^2 + \|\hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)}\|_F^2 \leq \frac{2^4(K+2)}{\beta^2}\lambda_n^2$$

By Lemma C.2, there exist constants r, C such that

$$\|\hat{\Theta}^{(1,0)} - \Theta_*^{(1,0)}\|_F^2 + \|\hat{\Theta}^{(1,1)} - \Theta_*^{(1,1)}\|_F^2 \leq C(K+2)e^{-2\alpha_l}\max\{ne^{\alpha_u}, \log n\}$$

with probability $1 - 2n^{-r}$. □

E Convergence of probability matrices

We first define Kullback-Leibler (KL) divergence for the probability distributions of dyads in a random matrix as

$$D_{KL}(f_P\|f_Q) = n^{-2}\sum_{i\neq j}\int_{-\infty}^{\infty} f_{P_{ij}}(x)\log\frac{f_{P_{ij}}(x)}{f_{Q_{ij}}(x)}dx$$

where f_P and f_Q are the probability distributions of $n \times n$ random matrices. In our case, f_P is multinomial distribution. In addition, we also define total variation

distance between probability distributions f_P and f_Q on a set $\mathcal{X}$ as

$$\begin{aligned}\|f_P - f_Q\|_{TV} &:= \sum_{i\neq j} \sup_{A\in\mathcal{X}} |P_{ij}(A) - Q_{ij}(A)| \\ &= \frac{1}{2}\sum_{i\neq j} \int |P_{ij}(x) - Q_{ij}(x)|dx.\end{aligned}$$

Since we have multinomial distribution for P_{ij} and Q_{ij}, $\int |P_{ij}(x) - Q_{ij}(x)|dx = \sum_{s,t=1}^{n} |P_{ij}^{(s,t)} - Q_{ij}^{(s,t)}|$ for our case.

Proof of Theorem IV.3.

Since $\hat{\mathbf{\Theta}}$ is optimal solution to maximum likelihood function, we have $l(\hat{\mathbf{\Theta}}) \geq l(\mathbf{\Theta})$.

$$\begin{aligned}\frac{n^2}{2} D_{KL}(\mathbf{P}\|\hat{\mathbf{P}}) &= \mathbb{E}_A[l(\mathbf{\Theta})] - \mathbb{E}_A[l(\hat{\mathbf{\Theta}})] \\ &\leq l(\hat{\mathbf{\Theta}}) - \mathbb{E}_A[l(\hat{\mathbf{\Theta}})] - l(\mathbf{\Theta}) + \mathbb{E}_A[l(\mathbf{\Theta})] \\ &\leq < A^{(1,0)} - P^{(1,0)}, \hat{\Theta}^{(1,0)} - \Theta^{(1,0)} > + \frac{1}{2} < A^{(1,1)} - P^{(1,1)}, \hat{\Theta}^{(1,1)} - \Theta^{(1,1)} > \\ &\leq \sqrt{2(K+2)}\|A^{(1,0)} - P^{(1,0)}\|\|\hat{\Theta}^{(1,0)} - \Theta^{(1,0)}\|_F \\ &\quad + \frac{1}{2}\sqrt{2(K+2)}\|A^{(1,1)} - P^{(1,1)}\|\|\hat{\Theta}^{(1,1)} - \Theta^{(1,1)}\|_F.\end{aligned}$$

Using Lemma C.1 and Theorem IV.2, we have

$$n^2 D_{KL}(\mathbf{P}\|\hat{\mathbf{P}}) \leq \frac{C'(K+2)}{\beta} \max\left(ne^{\alpha_u}, \log n\right)$$

with probability at least $1 - 2n^{-r}$. Next, we need to connect $D_{KL}(\mathbf{P}\|\hat{\mathbf{P}})$ with Frobenius norm.

$$n^2 D_{KL}(\mathbf{P}\|\hat{\mathbf{P}}) \geq 2\|\hat{\mathbf{P}} - \mathbf{P}\|_{TV}^2 \geq \frac{1}{2}\|\hat{\mathbf{P}} - \mathbf{P}\|_F^2$$

First inequality follows from Pinsker's inequality such thatt $D_{KL}(P\|Q) \geq 2\|P - Q\|_{TV}^2$. Second inequality follows from $2\|p-q\|_{TV}^2 \geq = \frac{1}{2}\left(\sum_{s=1}^{2}\sum_{t=1}^{2}|p^{(s,t)} - q^{(s,t)}|\right)^2 \geq \frac{1}{2}\left(\sqrt{\sum_{s=1}^{2}\sum_{t=1}^{2}(p^{(s,t)} - q^{(s,t)})^2}\right)^2 \geq \frac{1}{2}\|p-q\|_F^2$. Therefore, we have

$$\begin{aligned}\|\hat{\mathbf{P}} - \mathbf{P}\|_F^2 &\leq \frac{C_1(K+2)}{\beta} \max\left(ne^{\alpha_u}, \log n\right) \\ &\leq C_1(K+2)e^{-\alpha_l} \max\left(ne^{\alpha_u}, \log n\right)\end{aligned} \tag{C.6}$$

with probability at least $1 - 2n^{-r}$. □

BIBLIOGRAPHY

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

Kristen M Altenburger and Johan Ugander. Monophily in social networks introduces similarity among friends-of-friends. *Nature human behaviour*, 2(4):284, 2018.

Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

Arash A Amini, Elizaveta Levina, et al. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

Brian Ball, Brian Karrer, and Mark EJ Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Rajendra Bhatia. *Perturbation bounds for matrix eigenvalues*, volume 53. Siam, 1987.

Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

George T Cantwell and MEJ Newman. Mixing patterns and individual differences in networks. *Physical Review E*, 99(4):042306, 2019.

Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pages 35–1, 2012.

Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

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

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

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

Derek Greene and Pádraig Cunningham. Producing a unified graph representation from multiple social network views. In *Proceedings of the 5th annual ACM web science conference*, pages 118–121. ACM, 2013.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

John A Hartigan. Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129, 1972.

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

Peter D Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, 15(4):261, 2009.

Peter D Hoff. Dyadic data analysis with amen. *arXiv preprint arXiv:1506.08237*, 2015.

Peter D Hoff. Additive and multiplicative effects network models. *arXiv preprint arXiv:1807.08038*, 2018.

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

Paul W Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50, 1981.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Jianwei Hu, Hong Qin, Ting Yan, and Yunpeng Zhao. Corrected bayesian information criterion for stochastic block models. *Journal of the American Statistical Association*, pages 1–13, 2019.

Pengsheng Ji, Jiashun Jin, et al. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

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

Antony Joseph, Bin Yu, et al. Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791, 2016.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 454–462. IEEE, 2004.

Jérôme Kunegis and Andreas Lommatzsch. Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 561–568, 2009.

Daniel B Larremore, Aaron Clauset, and Abigail Z Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

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

Elizabeth A Leicht and Mark EJ Newman. Community structure in directed networks. *Physical review letters*, 100(11):118703, 2008.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 04 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa006. URL `https://doi.org/10.1093/biomet/asaa006`. asaa006.

Yanhua Li, Zhi-Li Zhang, and Jie Bao. Mutual or unrequited love: Identifying stable clusters in social networks with uni-and bi-directional links. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 113–125. Springer, 2012.

Zhuang Ma, Zongming Ma, and Hongsong Yuan. Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research*, 21(4):1–67, 2020.

Fragkiskos D Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95–142, 2013.

Mark Newman. *Networks*. Oxford university press, 2018.

Subhadeep Paul, Yuguo Chen, et al. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2): 3807–3870, 2016.

Leto Peel, Jean-Charles Delvenne, and Renaud Lambiotte. Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16):4057–4062, 2018.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

Zahra S Razaee, Arash A Amini, and Jingyi Jessica Li. Matched bipartite block model with covariates. *Journal of Machine Learning Research*, 20(34):1–44, 2019.

Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45):12679–12684, 2016.

Purnamrita Sarkar, Peter J Bickel, et al. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990, 2015.

Srijan Sengupta and Yuguo Chen. A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2): 365–386, 2018.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Duy Q Vu, David R Hunter, and Michael Schweinberger. Model-based clustering of large networks. *The annals of applied statistics*, 7(2):1010, 2013a.

Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013b.

Song Wang and Karl Rohe. Discussion of" coauthorship and citation networks for statisticians". *The Annals of Applied Statistics*, 10(4):1820–1826, 2016.

Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

YX Rachel Wang, Peter J Bickel, et al. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2):500–528, 2017.

Yun-Jhong Wu, Elizaveta Levina, and Ji Zhu. Generalized linear models with low rank effects for network data. *arXiv preprint arXiv:1705.06772*, 2017.

Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 25–32. ACM, 2001.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting overlapping communities in networks using spectral methods. *arXiv preprint arXiv:1412.3432*, 2014.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.

Yunpeng Zhao, Elizaveta Levina, Ji Zhu, et al. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Physical review E*, 76(4):046115, 2007.

Zhixin Zhou and Arash A Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20(47):1–47, 2019.