TECHNIQUES IN NETWORK EMBEDDING AND GAUSSIAN COMPARISON FOR HIGH-DIMENSIONAL STATISTICS

Aman Barot

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2021

Approved by:

Shankar Bhamidi

Amarjit Budhiraja

Peter J. Mucha

Marc Niethammer

Andrew B. Nobel

ABSTRACT

AMAN BAROT: Techniques in network embedding and Gaussian comparison for high-dimensional statistics
(Under the direction of Shankar Bhamidi and Andrew B. Nobel)

This dissertation consists of research on three high-dimensional statistical problems. In the first part of the dissertation, we study Gaussian comparison which is an important technique for comparing distributions and functionals of Gaussian random variables. We derive a Gaussian comparison result based on a smart-path argument. We show the significance of this result by an application to a problem of maximal correlations in high dimensions.

In the second part of the dissertation, we study brain connectivity data sets. Networks have emerged as an important tool to understand the complex structure and function of human brains. We analyze the structural connectivity structure of human brains using two network data sets.

In the third part of the dissertation, we focus on the problem of community detection on networks using node embedding methods. In recent decades, network data sets containing millions and billions of nodes have become available. This has necessitated the development of scalable methods for their analysis. One such class of methods are methods for node embedding. Node embedding methods encode nodes of a network in a low-dimensional Euclidean space which allows one to use well-known methods for Euclidean spaces for network analysis. In this dissertation we study the problem of community detection using two well-known node embedding methods: DeepWalk and node2vec. We describe the network sparsity regimes when the k-means algorithm applied to the node embeddings detects communities for graphs generated from the stochastic block model, and when such an approach might fail. We also describe how increasing the non-backtracking parameter in the node2vec method leads to provable improvements in community detection compared to DeepWalk.

To all my teachers.

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the incredible support of several people. First and foremost, I am extremely grateful to my dissertation advisors, Dr. Shankar Bhamidi and Dr. Andrew Nobel, for suggesting research problems, continuous guidance and being patient with me. A big thanks to Dr. Shankar Bhamidi for supporting me at every step of the dissertation and believing in my abilities even when I did not feel that way myself. I am very thankful to Dr. Andrew Nobel for closely working with me and supporting me on two research projects. Special thanks to Dr. Peter Mucha and Dr. Marc Niethammer for our weekly meetings, invaluable feedback, and constant encouragement. Many thanks to Dr. Souvik Dhara for all the advising and immense help during the writing process. I would like to thank Dr. Amarjit Budhiraja for his support, and his feedback during Preliminary and Final Oral Examinations.

Thanks to Dr. Shankar Bhamidi, Dr. Peter Mucha, Dr. Marc Niethammer and Dr. Andrew Nobel for supporting me with research grants. I am thankful for the NSF, NIH and ARO grants for several semesters of funding. I am also thankful to the Statistical and Applied Mathematical Sciences Institute and the ARPA grant for the funding during the current semester.

My mental health issues have played a big part in this dissertation. A special thanks to Dr. Matthew Munich for listening and support. I would also like to thank Patricia Catanio, Todd Colucci, and Avery Cook. I am grateful to Anie Barkley, Dr. Cheryl Mathews, and Charlie Schliesser for our valuable group.

I would like to thank Dr. Brian Rybarczyk for introducing me to evidence-based teaching, and for encouragement and support during teaching related classes and workshops. Thanks to Dr. Martin Styner for sharing brain data sets. Thanks to Chad Covin, Austin Ferguson, Matthew Ford, Dr. Roland Kwitt, Dr. Martin Styner, Dr. Natalie Stanley, and Md Asadullah Turja for being supportive and providing constructive feedback during our weekly meetings. Thanks to all the STOR support staff including Shayna Hill, Christine Keat, Alison Kieber, Ayla Ocasio, and Samantha Radel for all their work in the background. Thanks to ISSS and specifically Jessica

Larsen to whom I have sent countless emails. I would like to extend a warm thanks to fellow STOR graduate students Aditya Balaram, Michael Bostwick, Brendan Brown, Suman Chakraborty, Prabhanka Deka, Mark He, and Benjamin Leinwand for friendship and support.

I am deeply grateful to Samopriya Basu, Ajeenckya Chavan, Miheer Dewaskar, Manish Goyal, Yanan Li and Amrita Tembhe for our wonderful friendship. Thanks to Saumitra Sinha for being a great roommate. Finally, I am thankful to my family for their love and support, and for (unintentionally) reminding me where I come from.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
Introduction

This dissertation consists of research work on three research problems. We introduce each of these problems and describe our contributions in sections 1.1, 1.2 and 1.3 of this chapter. In Chapter 2, we describe our Gaussian comparison result and apply it to the problem of hypothesis testing for covariance structures in high dimensions. In Chapter 3, we describe our data analysis on structural brain connectivity data. In Chapter 4, we describe our work on community detection using low-dimensional node embeddings. In Chapter 5, we provide proofs of all the results in Chapter 4. In Chapter 6, we suggest directions for future work.

1.1 Hypothesis testing for covariance structures in high dimensions

Covariance structures in high-dimensions have been of interest in many applications such as gene co-expression analysis (D'haeseleer et al., 2000; Horvath and Dong, 2008) and brain network analysis (Stam, 2014; Teicher et al., 2016; van den Heuvel and Hulshoff Pol, 2010). These applications motivate the study of hypothesis testing for these structures.

In the small sample setting, sample covariance matrix is used for testing for the covariance structure. However, this is not a good estimate for the covariance structure in high dimensions. In the case of testing for the inverse of the covariance matrix, the precision matrix, the corresponding sample precision matrix is not well defined. In addition to high dimensionality, the dependency structure also makes the problem challenging.

There are two types of testing problems in high dimensions: global testing for the overall pattern of the covariance structure and simultaneous testing for a large collection of hypothesis for local covariance structures. We will focus on global testing. We first fix some notation. Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^T$ be a matrix consisting of n independent and identically distributed row vectors of dimension p each. We will further assume that each $\mathbf{X}_i$ has multivariate normal distribution unless

otherwise stated. Each $\mathbf{X}_i$ has mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$. The sample mean is given by $\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{X}_k$ and the sample covariance matrix is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{X}_k - \bar{\mathbf{X}})^T(\mathbf{X}_k - \bar{\mathbf{X}}).$$

We first discuss tests for $H_0 : \boldsymbol{\Sigma}_p = \boldsymbol{I}_p$. In the classical theory of statistics when $p < n$, the likelihood ratio test is used for testing $H_0 : \boldsymbol{\Sigma}_p = \boldsymbol{I}_p$. The likelihood ratio test statistic (LRT) in that setting is $L = n(tr(\hat{\boldsymbol{\Sigma}}) - \log|\hat{\boldsymbol{\Sigma}}| - p)$. For fixed p and $n \to \infty$, it can be shown that L converges in distribution to $\chi^2_{\frac{1}{2}p(p+1)}$ under H_0. This is useful in practice when $\frac{p}{n}$ is closer to 0. However, use of this for comparable n and p leads to a large Type I error. Bai & Silverstein (Bai and Silverstein, 2008) use random matrix theory (RMT) results to show that when $\frac{p}{n} \to (0, 1), \mu_p = 0$ and under moment bounds on the variables $\mathbf{X}_i$, a corrected LRT converges in distribution to $N(0, 1)$ distribution. Further extensions are proved in (Jiang et al., 2012; Zheng et al., 2015).

In addition to using LRT, the spectral norm and the Frobenius norm have also been used for testing. Johnstone (Johnstone et al., 2001) showed that when $\mathbf{X}_i \sim N(0, \boldsymbol{I}_p)$ and $\frac{n}{p}$ converges to a constant in $(0, \infty)$, the largest eigenvalue of $\mathbf{X}^T\mathbf{X}$ converges after recentering to the Tracy-Widom distribution of order 1. Soshnikov (Soshnikov, 2002) extended this result by showing that the joint distribution of the largest k eigenvalues converges to the Tracy-Widom distribution. He also extended results to the sub-gaussian case. Péché (Péché, 2009) further extended these results to the case of $\mathbf{X}_i$ having moment bounds and the limit $\frac{p}{n}$ possibly being equal to 0 or ∞. Use of Frobenius norm for testing was first done in (John, 1971) and (Nagao, 1973) in the small sample setting. They used the test statistic $\frac{1}{p}||\hat{\boldsymbol{\Sigma}} - I||_F$ where the subscript F denotes the Frobenius norm. Ledoit & Wolf (Ledoit et al., 2002) showed that when $p > n$, this test is not consistent. They introduced a correction term to form a consistent estimator. Further extensions are in (Srivastava, 2005; Birke and Dette, 2005; Chen et al., 2010).

Towards testing for more general covariance structures, the test for $H_0 : \boldsymbol{\Sigma}_p = \sigma^2 I_p$ was studied by John (John, 1971) and Ledoit & Wolfe (Ledoit et al., 2002). Ledoit & Wolfe (Ledoit et al., 2002) showed that the statistic $\frac{1}{p}tr\left\{\left(\frac{\hat{\boldsymbol{\Sigma}}}{p^{-1}tr(\hat{\boldsymbol{\Sigma}})} - I\right)^2\right\}$, used in the low dimension setting, is consistent even when p grows with n. Next, we discuss the case when $\boldsymbol{\Sigma}_p$ is diagonal under the null hypothesis. This case is equivalent to testing $H_0 : R = I_p$ where R is the correlation matrix. A natural test

statistic here is the largest off-diagonal value of the sample correlation matrix, $\hat{R}$, i.e.

$$L_{n,p} = \max_{1 \le i < j \le p} |\hat{R}_{i,j}|.$$

Cai & Jiang (Cai and Jiang, 2011) showed that when $\mathbf{X}_i$ are independent, $nL_{n,p}^2$ converges in distribution to the Gumbel distribution after recentering. Cai & Jiang (Cai and Jiang, 2011, 2012) extend these results under Gaussian assumption to the regimes when p, the number of variables, is exponential or super-exponential in n. Shao & Zhou (Shao and Zhou, 2014) provide conditions on the moments of $\mathbf{X}_{ij}$ for the convergence of $L_{n,p}$ in the sub-exponential and exponential setting. In recent work, Fan & Jiang (Fan and Jiang, 2019) study the case when the variables $\mathbf{X}_i$ are highly dependent. More specifically, the authors work under the case that correlations $R_{ij} = \rho, \forall 1 \le i \ne j \le p$. They show that the limit of $L_{n,p}$ is Gumbel, a convolution of Gumbel and normal, or a normal distribution depending on the growth of ρ with respect to $\frac{1}{\sqrt{\log p}}$.

In the case of more general covariance structures, methods used for the testing of $H_0 : \mathbf{\Sigma}_p = I_p$ are not directly applicable. All of the results above which are based on the maximum of a subset of the sample correlation or covariance matrix make use of the Chen-Stein method (Arratia et al., 1990) for Poisson approximation. Poisson approximation is a powerful technique to compare dependent vectors or processes. However the method requires sufficient amount of independence in order for it to be applicable. In the next chapter we introduce an alternative technique, namely Gaussian comparison, to approach the hypothesis testing problem.

Gaussian comparison techniques have been used in many research areas such as empirical processes (Boucheron et al., 2013) and extreme value theories (Leadbetter et al., 2012). These techniques help us compare probability distributions of Gaussian vectors or processes based on their covariance structures. Our Gaussian comparison result (Theorem 1) compares expectations of functionals of an arbitrary multivariate normal vector and a standard normal vector. In particular, this helps us compare the extremal processes derived from these vectors. Lemma 2 is another comparison result of a similar nature in Chapter 2.

In order to demonstrate how Gaussian comparison results such as Theorem 1 and Lemma 2 may be used, we apply these to the problem of hypothesis testing for covariance structures given by short-range dependence. In this problem, we test whether the covariance matrix $\mathbf{\Sigma}$ is banded

where $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ is said to be banded with bandedness $\tau \in \mathbb{N}$ if

$$\sigma_{kl} = 0 \quad \text{for all pairs } (k,l) \text{ such that } |k-l| \geq \tau.$$

We use the test statistic given by

$$U_{n,\tau} := \max_{1\leq i<j\leq p, |i-j|\geq\tau} \frac{|(\mathbf{X}_{\cdot i} \cdot \mathbf{X}_{\cdot j})|}{n^2},$$

which is the maximum over the pairs of columns or attributes which are independent under the null hypothesis. In addition to studying the maxima of the covariances, in Chapter 2 we also study the stochastic process derived from the full collection $\{(\mathbf{X}_{\cdot i} \cdot \mathbf{X}_{\cdot j})\}_{1\leq i<j\leq p, |i-j|\geq\tau}$.

1.2 Multilayer networks

Many real world systems consist of an interacting collection of entities. For example in the case of brain connectivity analysis, brain regions may be thought of as entities which interact with each other. A simplification of such complex systems is to only look at pairwise interactions between entities. This leads to what is known as a network which is a collection of entities and their pairwise relationships. One of the simplest examples of a network is a graph. A graph G is a pair (V, E) where V is the set of entities, referred to as nodes, in the network and $E \in \{0,1\}^{V\times V}$ is the set of edges which indicate the presence or absence of a relationship between pairs of nodes. More explicitly, an edge equal to 1 represents that there a link between two nodes and 0 indicates the absence of it. A weighted graph or network is defined similarly except $E \in \mathbb{R}^{V\times V}$. Edges in a weighted network represent the strength of the relationship between the nodes.

Networks have been useful in understanding structure, dynamics and function of complex systems. A few examples of such systems come from analysis of brain connectivity, gene co-expression (Stuart et al., 2003), protein-protein interactions (Schwikowski et al., 2000), social networks (Lusher et al., 2013) and transportation systems (Guimerà et al., 2005). Specifically in the context of brain connectivity analysis, network analysis has shown that many diseases are reflected as abnormal network organization between the cortical regions (Bassett et al., 2008; Bassett and Bullmore, 2009; He et al., 2008; Stam et al., 2007).

In order to understand the structure of complex networks, network modeling can be very useful to extract features that explain the network structure. Network models are also useful to understand whether certain features in real networks deviate from what we expect under a given model. In the following, we first discuss models for single networks followed by models for multilayer networks.

Starting with the work of Erdős & Rényi (Erdős and Rényi, 1960), a large body of work has been devoted to the study of random graph models for real world networks. Two properties were observed to be satisfied by many real world networks. The first property is the small world property which means that the shortest distance between any two nodes in the network is $O(\log n)$ where n is the number of nodes in the network. The second property is that the distribution of the node degrees (the number of edges starting at a node) follows a power law. To model these properties, the Watts-Strogatz model (Watts and Strogatz, 1998) and the Barabási-Albert model (Barabási and Albert, 1999) respectively were introduced.

Another feature of many real world networks is the presence of communities or clusters of nodes which have relatively distinct density of edges within the community compared to rest of the nodes in the network. This motivates a mixture model for random graph models referred to as the stochastic block model (SBM) (Holland et al., 1983a). In addition to modeling, SBMs can be used for community detection (Karrer and Newman, 2011) and as a benchmark for community detection algorithms. SBMs have been extended in various directions such as to allow for mixed memberships of nodes (Airoldi et al., 2005, 2008), latent class models (Hoff, 2008) and weighted networks (Aicher et al., 2015). Apart from these, two more important class of models are given by the exponential random graph models (Robins et al., 2007a,b) and latent feature models (Palla et al., 2012; Miller et al., 2009).

In addition to explaining features which account for properties in real networks, network models are also used for comparing features in real networks to features in network models satisfying constraints. Some examples of these are block models and its extensions which have been mentioned above. Two well-known types of these models given by the Chung-Lu model (Chung and Lu, 2002) and the configuration model (Molloy and Reed, 1995) model networks satisfying constraints relating to the distribution of the degree and the expected degree of nodes in the network.

Though the study of single networks has been very useful, a lot of networks today constitute what are known as multilayer networks. Multilayer networks can be loosely described as a collec-

tion of networks, also called layers in this context, which are related to each other in some way. For example, brain connectivity networks of a subject taken at several time points or different transportation networks such as subway and bus for a given city. Modeling each layer separately fails to borrow information across layers. Thus there is a need to model these networks jointly.

An analog of graphs can be defined for multilayer networks as a pair $(\mathbf{G}, \mathbf{E})$ where $\mathbf{G}$ is a set of graphs or layers i.e. $\mathbf{G} = \{G_k | 1 \leq k \leq n\}$, $G_k = (V_k, E_k)$ and $\mathbf{E} = \{\{0,1\}^{V_i \times V_j} | 1 \leq i \neq j \leq n\}$ is the set of edges or links between the layers. Therefore, in addition to links within a layer, there are links between nodes across layers as well. More generally, multilayer networks may have weighted edges and other attributes.

Several multilayer models have been proposed which model the layers with a fixed or expected degree sequence and model the links between layers (Leicht and D'Souza, 2009; Bianconi et al., 2015; Gao et al., 2012, 2013). Similar to single layer models, several multilayer models also model community structure in the networks. (Zhang et al., 2017) propose a null model for networks evolving over time. (Matias and Miele, 2017; Matias et al., 2018) define temporal stochastic models by modeling how the model parameters change over time. Several models have also been proposed for joint community detection for multilayer networks (Paul and Chen, 2018; Barbillon et al., 2017; Han et al., 2015; Stanley et al., 2016; Peixoto, 2015).

Towards jointly modeling networks, in Chapter 3 we start with describing two brain network data sets. The first data set consists of structural connectivity networks of human brains. The networks are of subjects over three time points from infancy to age 2. The networks in the data set are weighted. This is a period of immense brain development. We explore questions such as predicting whether the subject was preterm or not based on the networks, network normalization, effects of scanner, and analysis over time.

The second data set consists of subjects in the ages 50 and over. Most of these subjects have brain networks over $4 - 5$ time points and each of these networks are labeled as cognitively normal (CN) or mild cognitive impairment (MCI) or Alzheimer's disease (AD). The networks in this data set are weighted as well. We explore questions such as prediction of mini-mental state exam (MMSE) scores and predicting diagnosis labels.

In section 6.2 we describe a proposed method for joint modeling of multilayer networks. This approach makes use of optimal mass transport which may be thought of as a shortest path between

configurations or in our case, network models. We introduce and describe two methods and explain these methods in the context of the Erdős-Rényi model, the stochastic block model and the degree-corrected model.

1.3 Node embedding methods

In recent decades, very large network data sets have become available. This had led to the development of new methods for the analysis of such data sets. We review below one class of such methods called node embedding methods. Node embedding methods map nodes of a network into a low-dimensional Euclidean space. One can then use methods for Euclidean spaces for network analysis. Broadly, there are three types of node embedding methods in the literature. We discuss each of the three types of methods below. We first fix some notation. For an observed graph G, A denotes the adjacency matrix of the graph and D denotes the diagonal matrix with the degrees of the nodes, i.e. $D_{ii} = \sum_j A_{ij}$. $L = D - A$ denotes the Laplacian matrix for the graph. Z_i denotes the node embedding of node i in a low-dimensional Euclidean space. $A^{(k)}$ denotes the kth power of the matrix A.

The first type of methods are based on matrix factorization. The node embeddings for these methods are obtained by finding low-rank factorization of a matrix derived from the Laplacian or the adjacency matrix of the graph. Belkin and Niyogi (2002) compute the embeddings so that connected nodes tend to be close to each other in their embeddings. In particular, they minimize

$$\sum_{ij} \|Z_i - Z_j\|_{\mathrm{F}}^2 A_{ij}.$$

The node embeddings can then be obtained by the eigendecomposition of L. The Graph Factorization (GF) algorithm (Ahmed et al., 2013) approximate A by inner products of the embeddings and solve the following problem:

$$\sum_{(i,j)} (A_{ij} - \langle Z_i, Z_j \rangle)^2 + \frac{\lambda}{2} \sum_i \|Z_i\|_{\mathrm{F}}^2 .$$

GraRep (Cao et al., 2015) use higher powers of A to find the optimal embeddings. In particular, they compute the embeddings by finding the singular value decomposition (SVD) of the matrix

given by

$$(i,j) \mapsto \log\left(\frac{A^{(k)}_{ij}}{\sum_{i'} A^{(k)}_{i'j}}\right) - \log\left(\frac{c}{N}\right),$$

where c is a constant. HOPE (Ou et al., 2016) computes the node embeddings by the SVD of a node similarity matrix S. For this, S may be computed using any method of choice based on the application and the method allows for S to be asymmetric.

The second type of methods are based on random walks. The first step for these methods consists of running random walks on the underlying graph and creating a matrix of co-occurences. Pairs of nodes which are closer to each other on the random walks are assigned a higher co-occurence. The second step then consists of computing node embeddings so that pairs of nodes with higher co-occurences tend to be close to each other in terms of Euclidean inner product. These methods perform well for link prediction and node classification for large sparse graphs as the co-occurences provide a flexible measure of strength of the relationship between any two nodes as compared to deterministic measures. We discuss two methods of this type, DeepWalk (Perozzi et al., 2014) and node2vec (Grover and Leskovec, 2016), in more detail in Chapter 4. Another well-known method often compared to DeepWalk and node2vec is LINE (Tang et al., 2015). This solves the same optimization problem but uses the adjacency matrix or edge weights in place of co-occurences.

The third type of methods are based on deep neural networks. These methods allow for modeling highly non-linear relationships between the node features and node embeddings. One approach for this is based on applying the idea of deep autoencoder (Hinton and Salakhutdinov, 2006) to the setting of networks. Two methods using this approach are DNGR (Cao et al., 2016) and SDNE (Wang et al., 2016). Both of these algorithms have a shared set of parameters for embedding the nodes which means that the number of parameters do not scale linearly with n, the number of nodes in the network. Some of the other prominent deep neural network based methods use features from the neighborhoods of the nodes to construct node embeddings as opposed to using the whole network. This is an especially useful approach for very large graphs, but these are also flexible in the sense that they can be used to generate embeddings for new nodes using a previously trained model. Some examples of these methods are (Kipf and Welling, 2016a,b; Hamilton et al., 2017a).

Node embedding methods have been used successfully for tasks such as network visualization, node classification and link prediction. The node embeddings output from any one of these methods can be used for a variety of downstream tasks. However, it is not clear which of these methods are better suited for specific tasks than others. It is also not clear if certain graph structures are better encoded using specific embedding methods compared to others. Since most of the recent scalable node embedding methods lack interpretability, it is of interest to better understand the mechanisms underlying these methods to both understand the limitations of these methods and any in-built biases in these methods. This motivates the need for a theoretical study to answer these questions. Theoretical studies will also be helpful for future researchers to build upon when designing new methods. With this background, in this dissertation we study DeepWalk and node2vec algorithms. More specifically, we study how these methods perform for community detection. Towards this we use the stochastic block model, a well-known benchmark model for community detection, to generate graphs for our theoretical analysis. We find network sparsity levels when applying the k-means algorithm on the node embeddings detects communities. Our results also indicate the network sparsity levels when this approach will not work. Another important implication of our results is that node2vec performs provably better than DeepWalk when using a large backtracking parameter. The latter two results have significance for practitioners when choosing a node embedding method for community detection and also in choosing parameters for node2vec.

CHAPTER 2

Gaussian comparison for correlations in high-dimensions

2.1 Introduction

Gaussian comparison is an important technique to compare distributions and expectations of sets of Gaussian random variables. It is used in various applications such as empirical processes (Boucheron et al., 2013), extreme value theories (Leadbetter et al., 2012), high dimensional statistical inference (Chernozhukov et al., 2015) and several topics in probability theory (Adler and Taylor, 2007; Li, 1999; Li and Shao, 2001).

We discuss a couple of Gaussian comparison results. In (Slepian, 1962), D. Slepian proved an inequality which compares probabilities of sets of Gaussian vectors. In order to describe this more precisely, we fix some notation. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ be two Gaussian vectors. Suppose $\mathbb{E}\mathbf{X} = \mathbb{E}\mathbf{Y} = 0, \mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, and $\mathbb{E}X_iX_j \leqslant \mathbb{E}Y_iY_j$ for $i \neq j$. Then Slepian's inequality states that for all $u_1, u_2, \ldots, u_n \in \mathbb{R}$ we have

$$P(X_i \leqslant u_i, \forall i = 1, 2, \ldots, n) \leqslant P(Y_i \leqslant u_i, \forall i = 1, 2, \ldots, n).$$

In other words, an ordering between the covariances helps us establish stochastic domination between the two variables. An important extension of Slepian's inequality is the Sudakov-Fernique inequality which compares the expectations of maximum of random variables. The inequality says that provided $\mathbb{E}X = \mathbb{E}Y$ and $\mathbb{E}X_iX_j \leqslant \mathbb{E}Y_iY_j$ for $i \neq j$, we have

$$\mathbb{E}\max_i X_i \geqslant \mathbb{E}\max_i Y_i.$$

Several more important extensions of these two types of inequalities may be found in (Gordon, 1985, 1987; Kahane, 1986; Vitale, 2000). In this chapter we derive a Gaussian comparison result to compare expectations of functions of random variables based on a smart-path argument. The

significance of this result is demonstrated using an application to a problem of maximal covariances in high dimensions. We give a brief introduction to this problem in the next section. In section 2.3, we describe our results. In section 2.4, we give proofs of the Gaussian comparison results. In section 2.5.1, we give proofs of our results on maximal covariances.

2.2 Covariances in high dimensions

Correlation networks arise in applications such as gene co-expression analysis (D'haeseleer et al., 2000; Horvath and Dong, 2008) and brain network analysis (Stam, 2014; Teicher et al., 2016; van den Heuvel and Hulshoff Pol, 2010). Gene co-expression analysis reveals genes which are expressed similarly across experimental conditions. Similarly, functional brain network analysis identifies brain regions with similar activity. These methods are useful in understanding the gene regulatory networks and brain structure and function respectively.

Correlation network applications motivate the problem of hypothesis testing for correlation structures. Since the data sets in applications are large in dimension, we are interested in asymptotic results. We now set up the notation and describe our model for the data.

Let $\mathbf{Y}^n = \{Y_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}$ be a $n \times p$ dimensional matrix jointly Gaussian random variables. In the context of gene expression data, $\mathbf{Y}^n$ is the gene expression matrix constructed using p genes and n experimental conditions. Similarly in the context of functional connectivity data, $\mathbf{Y}^n$ describes the activation patterns of p brain regions over n time points. The rows are independent and identically distributed random vectors and $\mathbf{Y}_{1\cdot} = (Y_{11}, Y_{12}, \ldots, Y_{1p}) \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \{\rho_{kl} : 1 \leqslant k, l \leqslant p\}$ is the covariance matrix satisfying $\rho_{i,i} = 1$ for $1 \leqslant i \leqslant p$. We would like to test that the covariance matrix is banded with bandedness $\tau \in \mathbb{N}$, i.e.,

$$\rho_{kl} = 0 \quad \text{for all pairs } (k, l) \text{ such that } |k - l| \geqslant \tau.$$

A natural test statistic is the following

$$U_{n,\tau} := \max_{1 \leqslant i < j \leqslant p, |i-j| \geqslant \tau} \frac{|(Y_{\cdot i} \cdot Y_{\cdot j})|}{n^2},$$

where $(Y_{\cdot i} \cdot Y_{\cdot j})$ denotes the Euclidean inner product between vectors $Y_{\cdot i}$ and $Y_{\cdot j}$. This motivates the study of the extremal landscape of normalized sample covariances obtained from $\mathbf{Y}$. We describe our results towards this in the next section.

2.3 Results

In this section we present formal statements of our results. As the results involve point processes, we recall the necessary definitions in the next subsection. We refer interested readers to the comprehensive treatises (Kallenberg, 1973; Daley and Vere-Jones, 2003, 2008).

2.3.1 Point Process Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and for $d \geqslant 1$, let $\mathcal{B}(\mathbb{R}^d)$ denote the Borel subsets of $\mathbb{R}^d$. A *point process* with values in $\mathbb{R}^d$ is a map $\Pi : \Omega \times \mathcal{B}(\mathbb{R}^d) \to \{0, 1, \ldots\}$ such that (i) $\Pi(\cdot, B)$ is an $\mathcal{F}$-measurable function for every fixed $B \in \mathcal{B}(\mathbb{R}^d)$, and (ii) $\Pi(\omega, \cdot)$ is a counting measure on $\mathcal{B}(\mathbb{R}^d)$ for every $\omega \in \Omega$. In all our examples, Π will be simple almost surely. In what follows we will suppress the dependence of Π on ω, denoting $\Pi(\omega, B)$ as $\Pi(B)$. Let $\lambda : \mathbb{R}^d \to [0, \infty)$ be a Borel measurable function such that $\lambda(B) := \int_B \lambda(x)\, dx$ is finite for every bounded set $B \in \mathcal{B}(\mathbb{R}^d)$. A point process Π is said to be Poisson with rate function $\lambda(\cdot)$ if (i) $\Pi(B) \sim \text{Poisson}(\lambda(B))$ for each bounded set $B \in \mathcal{B}(\mathbb{R}^d)$, and (ii) for any $k \geqslant 2$ and disjoint Borel sets $B_1, \ldots, B_k$, $\Pi(B_1), \Pi(B_2), \ldots, \Pi(B_k)$ are independent. A sequence of point processes $\Pi_1, \Pi_2, \ldots$ converges to a point process Π in the vague topology, written $\Pi_n \stackrel{\text{d}}{\Longrightarrow} \Pi$, if for every compactly supported continuous function $f : \mathbb{R}^d \to [0, \infty)$ the (random) integral $\int f\, d\,\Pi_n$ converges to $\int f\, d\,\Pi$ in the usual sense of weak convergence.

2.3.2 Gaussian comparison

In this section we describe the main comparison result. We start by recalling extremal point process in the independent regime. Let $Z_1, Z_2, \ldots$ be independent $\mathcal{N}(0, 1)$ random variables. For each $N \geqslant 1$ let

$$a_N = \sqrt{2 \log N} \quad \text{and} \quad b_N = \sqrt{2 \log N} - \frac{\log(4\pi \log N)}{\sqrt{8 \log N}}, \tag{2.1}$$

be the standard extreme value scaling and centering constants, respectively, for the Gaussian. By the classical extreme value theorem for the Gaussian distribution, $a_N(\max\{Z_1,\dots,Z_N\} - b_N)$ converges in distribution to $-\log T$ where T is an Exp(1) random variable. Note that $-\log T$ has a standard Gumbel distribution. More generally, one may study the joint behavior of the order statistics of $Z_1,\dots,Z_N$ via the associated point process

$$\Gamma_N = \sum_{i=1}^{N} \delta\{a_N(Z_i - b_N)\}, \tag{2.2}$$

where $\delta\{x\}$ denotes the measure assigning mass one to the point x. It follows from standard results (cf. (Leadbetter et al., 1983, Theorem 5.7.2)) that as N tends to infinity Γ_N converges in the vague topology to the Poisson process Π_0 on $\mathbb{R}$ with intensity function $\gamma(x) = \exp(-x)$. In particular, if $\xi_1 > \xi_2 > \cdots$ denote the ordered points of Π_0, then ξ_r has the same distribution as $-\log(T_1 + \cdots + T_r)$ where $T_1,\dots,T_r$ are independent Exp(1) random variables.

The following result derives an explicit bound between the extremal process in the independent regime and the extremal process of an arbitrary multivariate normal vector. For this we need to setup notation. Let $\mathbf{Z} = (Z_1, Z_2, \dots, Z_N)$ be as above a vector of independent standard normal random variables, and let $\mathbf{X} = (X_1, X_2, \dots, X_N)$ be a multi-normal random vector such that $\mathbb{E}(X_i) = 0$, $\mathbb{E}(X_i^2) = 1$, and $\mathbb{E}(X_i X_j) := \sigma_{ij} \in (-1, 1)$ for $1 \leqslant i < j \leqslant N$.

Theorem 1 (**Gaussian comparison**). *Let $f : \mathbb{R} \to [0,1]$ be a compactly supported, twice differentiable function, and define $G_N : \mathbb{R}^N \to \mathbb{R}$ by*

$$G_N(\mathbf{x}) := \exp\left\{ -\sum_{i=1}^{N} f(a_N(x_i - b_N)) \right\} \tag{2.3}$$

where a_N and b_N are defined as in (2.1). Let $\theta \in \mathbb{R}$ be any number such that $f(x) = 0$ for all $x \leqslant \theta$ and define $u_N := b_N + \theta/a_N$. Then

$$|\mathbb{E}G_N(\mathbf{X}) - \mathbb{E}G_N(\mathbf{Z})| \;\leqslant\; \frac{2\,|f'|_\infty^2\, a_N^2}{u_N^4} \sum_{i\neq j, \sigma_{ij}\neq 0} \frac{e^{-u_N^2/(1+\sigma_{ij}^+)}}{(1-\sigma_{ij}^+)^{1/2}} \tag{2.4}$$

where $x^+ = \max\{x, 0\}$ and $|f'|_\infty = \sup_x |f'(x)|$.

Remark 2.3.1. Under the same assumptions on $\mathbf{X}$ and $\mathbf{Z}$ as those in Theorem 1, it is shown in (Bhamidi et al., 2012, Lemma 2.2) that for each $u \geqslant 1$,

$$\left|\mathbb{P}\left(\max_{1\leqslant i\leqslant N} X_i \geqslant u\right) - \mathbb{P}\left(\max_{1\leqslant i\leqslant N} Z_i \geqslant u\right)\right| \leqslant \sum_{i\neq j,\sigma_{ij}\neq 0} 2\,\bar{\Phi}^2(u)\sqrt{\frac{1+\sigma_{ij}^+}{1-\sigma_{ij}^+}}\cdot e^{-\sigma_{ij}^+ u^2/(1+\sigma_{ij}^+)}. \tag{2.5}$$

Using standard Gaussian tail bounds (see (2.8)), one may show that the bounds in (2.4) and (2.5) are of the same order when $u = u_N$.

2.3.3 Normalized covariances in high dimensions

Let $X_n = (X_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ be a $n \times p$ dimensional matrix with $X_{ij} \overset{iid}{\sim} N(0,1)$. We recall $\mathbf{Y}^n$ and $U_{n,\tau}$ from section 2.2. $\mathbf{Y}^n = \{Y_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}$ is a $n \times p$ dimensional matrix jointly Gaussian random variables and

$$U_{n,\tau} := \max_{1\leqslant i<j\leqslant p, |i-j|\geqslant\tau} \frac{|(Y_{\cdot i}\cdot Y_{\cdot j})|}{n^2}.$$

For $0 < \delta < 1$, let

$$\Gamma_{p,\delta} := \{1 \leqslant i \leqslant p : |\rho_{ij}| > 1-\delta \text{ for some } 1 \leqslant j \leqslant p \text{ with } j \neq i\}.$$

We now recall the following result in (Cai and Jiang, 2011).

Proposition 1 (Cai & Jiang 2011, Theorem 4). *Suppose as $n \to \infty$:*

1. *$p = p_n \to \infty$ with $\log p = o(n^{1/3})$;*
2. *$\tau = o(p^t)$ for any $t > 0$;*
3. *for some $\delta \in (0,1)$, $|\Gamma_{p,\delta}| = o(p)$, which is particularly true if $\max_{1\leqslant i<j\leqslant p}|\rho_{i,j}| \leqslant 1-\delta$.*

Then, $nU_{n,\tau}^2 - 4\log p + \log\log p$ converges weakly to an extreme distribution of type I with distribution function

$$F(y) = \exp\left\{-\frac{1}{\sqrt{8\pi}}e^{-y/2}\right\}, \quad y \in \mathbb{R}.$$

Motivated by the scaling in the above Proposition, we define

$$\begin{aligned}\boldsymbol{R}(U,V) &:= n\left(\frac{\langle U,V\rangle}{n}\right)^2 - 4\log p + \log\log p \\ &= \frac{(\langle U,V\rangle)^2}{n} - 4\log p + \log\log p\end{aligned}$$

Define two point processes by setting

$$\Gamma_{X_n}(B) = \sum_{1\leqslant a<b\leqslant p, |a-b|\geqslant\tau} \mathbb{1}_B\{\boldsymbol{R}(X_{\cdot a}, X_{\cdot b})\}, \tag{2.6}$$

$$\Gamma_{Y_n}(B) = \sum_{1\leqslant a<b\leqslant p, |a-b|\geqslant\tau} \mathbb{1}_B\{\boldsymbol{R}(Y_{\cdot a}, Y_{\cdot b})\} \tag{2.7}$$

for each Borel set $B \subset \mathbb{R}$. Then we have the following point process result.

Theorem 2. *Let Γ_0 be the Poisson point process with intensity function $\gamma(x) = \frac{1}{2\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)$. Let Γ_{Y_n} be the point process as defined above. Suppose as $n \to \infty$:*

1. *$p = p_n \to \infty$ with $\log p = o(n^{1/3})$;*
2. *$\tau = o(p^t)$ for any $t > 0$;*
3. *for some $\delta \in (0,1)$, there exists $\epsilon \in (0,1)$ such that $|\Gamma_{p,\delta}| = o(p^{1-\epsilon})$.*

Then $\Gamma_{Y_n} \to^d \Gamma_0$.

2.4 Proofs: Gaussian comparison

This section contains the proofs of the Gaussian comparison results. We begin with by showing Gaussian tail bounds needed for the proof of Theorem 1. Then, we provide a proof of Theorem 1.

2.4.1 Gaussian Tail Bounds

Let $\bar{\Phi}(x) = 1-\Phi(x)$ where $\Phi(x)$ is the cumulative distribution function of the standard Gaussian distribution. Recall that for $x > 0$,

$$\bar{\Phi}(x) \leqslant \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}. \tag{2.8}$$

The proof of Theorem 1 requires an inequality for the probability that two correlated Gaussian random variables each exceeds a common threshold.

Lemma 1. *Let (Z, Z_ρ) be jointly Gaussian random variables with mean 0, variance 1, and correlation $\mathbb{E}(ZZ_\rho) = \rho \in (-1, 1)$. Then for any $u > 0$,*

$$\mathbb{P}(Z > u, Z_\rho > u) \;\leqslant\; \frac{(1+\rho)^2}{2\pi u^2 \sqrt{1-\rho^2}} \exp\left(-u^2/(1+\rho)\right). \tag{2.9}$$

Proof of Lemma 1. Fix $u > 0$. When $\rho \geqslant 0$ the proof follows from inequality (2.8) and equation (1.2) in (Willink, 2004). Here we consider the case $\rho < 0$. Note that we may write $Z_\rho = \rho Z + \sqrt{1-\rho^2}Z'$, where Z' is a standard Gaussian random variable independent of Z. By conditioning on the value of Z, it is easy to see that

$$\mathbb{P}(Z > u, Z_\rho > u) \;=\; \int_u^\infty \bar{\Phi}\left(g(t)\right)\, \phi(t)\, dt \quad \text{where} \quad g(t) \,=\, \frac{u-\rho t}{\sqrt{1-\rho^2}}. \tag{2.10}$$

Now define

$$\eta = \sqrt{\frac{1-\rho}{1+\rho}} \;\text{ and }\; h(x) = e^{x^2/2}\, \bar{\Phi}(x).$$

As $h'(x) = x\, e^{x^2/2}\, \bar{\Phi}(x) - 1/\sqrt{2\pi}$, inequality (2.8) implies that $h(x)$ is decreasing for $x > 0$. It follows from equation (2.10) that

$$\begin{aligned} \mathbb{P}(Z > u, Z_\rho > u) \;&=\; \int_u^\infty e^{-g(t)^2/2}\, h(g(t))\, \phi(t)\, dt \qquad (2.11) \\ &\leqslant\; h(g(u)) \int_u^\infty e^{-g(t)^2/2}\, \phi(t)\, dt \\ &=\; h(\eta u) \int_u^\infty e^{-g(t)^2/2}\, \phi(t)\, dt, \end{aligned}$$

where in the last step we have used the fact that $g(u) = \eta u$. Routine algebra and a change of variables establishes that

$$\begin{aligned} \int_u^\infty e^{-g(t)^2/2}\,\phi(t)\,dt &= e^{-u^2/2}\int_u^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(t-\rho u)^2}{2(1-\rho^2)}\right)dt \\ &= \sqrt{1-\rho^2}\; e^{-u^2/2}\bar{\Phi}(\eta u). \end{aligned} \tag{2.12}$$

Combining (2.11), (2.13), and inequality (2.8) yields the bound (2.9), as desired. ■

2.4.2 Proof of Theorem 1

The proof of Theorem 1 requires a preliminary lemma. A version of the lemma appears in (Bhamidi et al., 2012), but as the proof is only sketched there, and as the lemma may be of independent interest, we provide a detailed statement and proof below. The proof relies on a smart-path argument and Gaussian integration by parts.

Lemma 2. *Let $G : \mathbb{R}^n \to \mathbb{R}$ be a bounded, twice continuously differentiable function with bounded derivatives*

$$G_i(x) = \frac{\partial G(x)}{\partial x_i} \quad 1 \leqslant i \leqslant n \qquad \textit{and} \qquad G_{ij} = \frac{\partial G(x)}{\partial x_i \partial x_j} \quad 1 \leqslant i,j \leqslant n.$$

If $\mathbf{X} \sim \mathcal{N}_n(0, \Sigma_X)$ and $\mathbf{Y} \sim \mathcal{N}_n(0, \Sigma_Y)$ are normal random vectors then

$$\mathbb{E}G(\mathbf{Y}) - \mathbb{E}G(\mathbf{X}) = \frac{1}{2}\sum_{i,j=1}^n \Delta_{ij}\int_0^1 \mathbb{E}G_{ij}(\mathbf{X}^t)\,dt$$

where $\Delta_{ij} = \mathbb{E}Y_iY_j - \mathbb{E}X_iX_j = (\Sigma_Y - \Sigma_X)_{ij}$ and $\mathbf{X}^t \sim \mathcal{N}_n(0, \Sigma_t)$ with $\Sigma_t := (1-t)\,\Sigma_X + t\,\Sigma_Y$.

Proof: Assume without loss of generality that $\mathbf{X}$ and $\mathbf{Y}$ are independent. For each $t \in [0,1]$ define the random vector

$$\mathbf{X}^t = (1-t)^{1/2}\,\mathbf{X} + t^{1/2}\,\mathbf{Y}$$

and the associated function $\varphi(t) = \mathbb{E}G(\mathbf{X}^t)$. Note that $\mathbf{X}^0 = \mathbf{X}$, $\mathbf{X}^1 = \mathbf{Y}$, and that $\mathbf{X}^t \sim \mathcal{N}_n(0, \Sigma_t)$, where Σ_t is defined as in the statement of the lemma. Thus

$$\mathbb{E}G(\mathbf{Y}) - \mathbb{E}G(\mathbf{X}) \;=\; \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\,dt,$$

and it suffices to show that for each $t \in (0,1)$

$$\varphi'(t) = \frac{1}{2}\sum_{i,j=1}^{n} \Delta_{ij}\mathbb{E}G_{ij}(\mathbf{X}^t). \tag{2.13}$$

To this end, fix $t \in (0,1)$ and note that $\mathbf{X}^t$ is distributed as $\Sigma_t^{1/2}\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(0, I)$ is a standard normal random vector with independent components. To simplify notation, let $A_t := \Sigma_t^{1/2}$. It follows from our regularity assumptions and the chain rule that

$$\begin{aligned}\varphi'(t) &= \frac{d}{dt}\mathbb{E}G(A_t\,\mathbf{Z}) \;=\; \mathbb{E}\left[\frac{d}{dt}G(A_t\,\mathbf{Z})\right] \;=\; \mathbb{E}\left[\sum_{i=1}^{n} G_i(A_t\,\mathbf{Z})\,\frac{d}{dt}\,(A_t\,\mathbf{Z})_i\right] \\ &= \sum_{i,j=1}^{n} (A_t')_{ij}\mathbb{E}\left(Z_j\,G_i(A_t\,\mathbf{Z})\right), \end{aligned} \tag{2.14}$$

where A_t' denotes the entry-by-entry derivative of the matrix A_t. Fix i, j for the moment and define the function

$$H_{ij}(s) \;:=\; \mathbb{E}G_i(A_t\,\mathbf{Z}_s) \quad \text{where} \quad \mathbf{Z}_s \;:=\; (Z_1, \cdots, Z_{j-1}, s, Z_{j+1}, \cdots, Z_n).$$

It follows from a simple conditioning argument and Gaussian integration by parts that

$$\mathbb{E}\left[Z_j\,G_i(A_t\mathbf{Z})\right] \;=\; \mathbb{E}\left[Z_j\,H_{ij}(Z_j)\right] \;=\; \mathbb{E}H_{ij}'(Z_j).$$

By another application of the chain rule,

$$\begin{aligned} H_{ij}'(s) &= \mathbb{E}\left[\frac{d}{ds}G_i(A_t\,\mathbf{Z}_s)\right] \;=\; \sum_{k=1}^{n}\mathbb{E}\left[G_{ik}(A_t\,\mathbf{Z}_s)\,\frac{d}{dt}(A_t\,\mathbf{Z}_s)_k\right] \\ &= \sum_{k=1}^{n}(A_t)_{jk}\mathbb{E}G_{ik}(A_t\,\mathbf{Z}_s). \end{aligned}$$

Thus, as $Z_1, \ldots, Z_n$ are independent,

$$\mathbb{E}H_{ij}'(Z_j) = \sum_{k=1}^{n}(A_t)_{jk}\mathbb{E}G_{ik}(A_t\,\mathbf{Z}).$$

Combining this last equation with (2.14), we find that

$$\begin{aligned}\varphi'(t) &= \sum_{i,k=1}^{n} \mathbb{E}G_{ik}(A_t\,\mathbf{Z}) \cdot \sum_{j=1}^{n} (A'_t)_{ij}(A_t)_{jk} \\ &= \sum_{i,k=1}^{n} \mathbb{E}G_{ik}(\mathbf{X}^t) \cdot (A'_t\,A_t)_{ik}. \end{aligned} \tag{2.15}$$

Recalling that $A_t = \Sigma_t^{1/2}$, it is easy to see that $(A_t^2)'_{ik} = (\Sigma_t)'_{ik} = \Delta_{ik}$. Furthermore, as A_t and A'_t are symmetric,

$$(A_t^2)' \;=\; A'_t\,A_t \;+\; A_t\,A'_t \;=\; A'_t\,A_t \;+\; (A'_t\,A_t)^T. \tag{2.16}$$

Fix $1 \leqslant i < k \leqslant n$. Continuity of the second partial derivatives ensures that $G_{ik} = G_{ki}$, and therefore

$$\begin{aligned}&\mathbb{E}G_{ik}(\mathbf{X}^t) \cdot (A'_t\,A_t)_{ik} \;+\; \mathbb{E}G_{ki}(\mathbf{X}^t) \cdot (A'_t\,A_t)_{ki} \\ &\quad= \;\mathbb{E}G_{ik}(\mathbf{X}^t)\,\big((A'_tA_t)_{ik} \;+\; (A'_tA_t)_{ki}\big) \\ &\quad= \;\mathbb{E}G_{ik}(\mathbf{X}^t)\,(A_t^2)'_{ik} \;=\; \mathbb{E}G_{ik}(\mathbf{X}^t)\,\Delta_{ik},\end{aligned}$$

where the penultimate equality follows from (2.16). A similar argument shows that $(A'_t\,A_t)_{ii} = \Delta_{ii}/2$. Thus (2.13) follows from (2.14), and the proof is complete.

Proof of Theorem 1: Fix $N \geqslant 1$. Let $f : \mathbb{R} \to [0,1]$ and $\theta \in \mathbb{R}$ be as in the statement of the theorem. To reduce notation, define $a := a_N$, $b := b_N$, and $u := b + \theta/a$, and let $G : \mathbb{R}^N \to \mathbb{R}$ be defined as in (2.3). Clearly, for $i \neq j$,

$$G_{ij}(\mathbf{x}) \;=\; a^2 f'(a(x_i - b))\,f'(a(x_j - b))\,G(\mathbf{x}).$$

Our assumptions concerning f and θ ensure that $f'(a(x-b)) = 0$ for $x \leqslant u$, and that $0 < G(\mathbf{x}) \leqslant 1$.

Now let $\mathbf{Z} \sim \mathcal{N}_N(0, I)$ and $\mathbf{X} \sim \mathcal{N}_N(0, \Sigma)$ be multinormal random vectors such that the covariance matrix $\Sigma = \{\sigma_{ij}\}$ of $\mathbf{X}$ satisfies $\sigma_{ij} = 1$ if $i = j$ and $\sigma_{ij} \in (-1, 1)$ for $i \neq j$. For $t \in [0,1]$ let $\mathbf{X}^t \sim \mathcal{N}_N(0, t\,\Sigma + (1-t)\,I)$. It follows from the properties of $G_{ij}(\mathbf{x})$ that for $i \neq j$ and each t,

$$|\mathbb{E}G_{ij}(\mathbf{X}^t)| \;\leqslant\; a^2\,|f'|_\infty^2\;\mathbb{P}(X_i^t \wedge X_j^t \geqslant u).$$

From this inequality and an application of Lemma 2, we obtain the bound

$$|\mathbb{E}G(\mathbf{X}) - \mathbb{E}G(\mathbf{Z})| \;\leqslant\; a^2\,|f'|_\infty^2 \sum_{i<j} |\sigma_{ij}| \int_0^1 \mathbb{P}(X_i^t \wedge X_j^t \geqslant u)\,dt$$

Applying the change of variables $t \to t/\sigma_{ij}$ for each i, j such that $\sigma_{ij} \neq 0$ yields the inequality

$$|\mathbb{E}G(\mathbf{X}) - \mathbb{E}G(\mathbf{Z})| \;=\; a^2\,|f'|_\infty^2 \sum_{i<j} \left| \int_0^{\sigma_{ij}} \mathbb{P}(Z \wedge Z_t \geqslant u)\,dt \right|$$

where Z, Z_t are jointly normal random variables with mean 0, variance 1, and correlation $\mathbb{E}(ZZ_t) = t$. By Lemma 1,

$$\mathbb{P}(Z \wedge Z_t \geqslant u) \;\leqslant\; \frac{(1+t)^2}{2\pi u^2\sqrt{1-t^2}} \exp\left(-\frac{u^2}{1+t}\right),$$

and therefore

$$|\mathbb{E}(G(\mathbf{X})) - \mathbb{E}(G(\mathbf{Z}))| \;\leqslant\; \frac{4\,|f'|_\infty^2\,a^2}{u^2} \sum_{i<j} \left| \int_0^{\sigma_{ij}} \frac{\exp\left(-u^2/(1+t)\right)}{2\pi\sqrt{1-t^2}}\,dt \right|.$$

Routine algebra implies that for $\sigma_{ij} \neq 0$

$$\left| \int_0^{\sigma_{ij}} \frac{\exp\left(-u^2/(1+t)\right)}{2\pi\sqrt{1-t^2}}\,dt \right| \;\leqslant\; 2\,\bar{\Phi}^2(u) \sqrt{\frac{1+\sigma_{ij}^+}{1-\sigma_{ij}^+}} \cdot e^{-\sigma_{ij}^+ u^2/(1+\sigma_{ij}^+)}$$

More details may be found in the latter part of proof of Lemma 2.2 from (Bhamidi et al., 2012). Applying the tail bound in (2.8) to the final expression above completes the proof. ■

2.5 Proofs: Maximal correlations

We begin this section with the proof of Theorem 2. We then state and prove the supplementary results.

2.5.1 Proof of Theorem 2

Proof of Theorem 2. Let $f : \mathbb{R} \to [0,1]$ be an arbitrary compactly supported function. For a $n \times p$ matrix Z, let

$$G_n^f(Z) := \prod_{1 \leqslant a < b \leqslant p, |a-b| \geqslant \tau} \exp\left\{-f(\boldsymbol{R}(Z_{\cdot a}, Z_{\cdot b}))\right\},$$

By Theorem 4.11 from (Kallenberg, 2017) and by Proposition 2 we have

$$\mathbb{E}G_n^f(X) = \mathbb{E}e^{-\Gamma_{X_n} f} \to \mathbb{E}e^{-\Gamma_0 f}.$$

In order to show $\Gamma_{Y_n} \to \Gamma_0$, it is enough to show that

$$|\mathbb{E}G_n^f(Y) - \mathbb{E}G_n^f(X)| \to 0.$$

By an approximation argument, it is enough to show this for functions f which are smooth. We will suppress the superscript f in G_n^f in the rest of the proof.

Let

$$X_{i\cdot}^t \sim N_p(0, \Sigma^t) \text{ with } (\Sigma^t)_{kl} := t\rho_{kl}, \quad 0 \leqslant t \leqslant 1, 1 \leqslant k, l \leqslant p \text{ and}$$

$$\boldsymbol{R}^{ij,t} := \boldsymbol{R}(X_{\cdot i}^t, X_{\cdot j}^t).$$

Note that the superscript t in the notation above does not stand for powers or exponentiation. Observe that $Y \stackrel{d}{=} X^1$. We will suppress the parameter t in the notation for $\boldsymbol{R}^{ij,t}$, $X_{\cdot i}^t$ and $X_{\cdot j}^t$, and only write $\boldsymbol{R}^{ij}$ for example.

We apply the Gaussian comparison to bound $|\mathbb{E}G_n(Y) - \mathbb{E}G_n(X)|$. Note that the Δ_{ij} factor in Lemma 2 is 0 for the summands involving independent coordinates since we compare to the independent setting. The subscripts in the notation below will denote matrix coordinates with respect to which we differentiate.

For $1 \leqslant a < b \leqslant p$ with $|a-b| < \tau$ and $1 \leqslant k \leqslant n$, we have

$$\frac{\partial G(X^t)}{\partial x_{ka}} = G(X^t) \cdot \sum_{i:|a-i|\geqslant\tau} -f'(\boldsymbol{R}^{ai})\boldsymbol{R}_{ka}^{ai} \text{ and,}$$

$$\frac{\partial G(X^t)}{\partial x_{ka}x_{kb}} = G(X^t) \cdot \left(\sum_{i:|a-i|\geqslant\tau} -f'(\boldsymbol{R}^{ai})\boldsymbol{R}^{ai}_{ka} \right) \cdot \left(\sum_{j:|b-j|\geqslant\tau} -f'(\boldsymbol{R}^{bj})\boldsymbol{R}^{bj}_{kb} \right).$$

Let $[u, v] \subset \mathbb{R}$ be the support of f. Let $u_n = n(u + 4\log p - \log\log p)$ and $v_n = n(v + 4\log p - \log\log p)$. The Gaussian comparison lemma gives the following equation:

$$\mathbb{E}G(Y) - \mathbb{E}G(X) = \sum_{1\leqslant a<b\leqslant p, |a-b|<\tau} \rho_{ab} \int_0^1 \left[\sum_{k=1}^{n} \mathbb{E}G_{ka,kb}(X^t) \right] dt. \tag{2.17}$$

The integrand above can expressed as follows.

$$\sum_{k=1}^{n} \mathbb{E}G_{ka,kb}(X^t) = \mathbb{E}\left[G(X^t) \cdot \sum_{i:|a-i|\geqslant\tau} \sum_{j:|b-j|\geqslant\tau} f'(\boldsymbol{R}^{ai}) f'(\boldsymbol{R}^{bj}) \left(\sum_{k=1}^{n} \boldsymbol{R}^{ai}_{ka}\boldsymbol{R}^{bj}_{kb} \right) \right] \tag{2.18}$$

Using Lemma 2 and equation 2.18 above, we have

$$|\mathbb{E}G(Y) - \mathbb{E}G(X)| \leqslant$$
$$\sum_{1\leqslant a<b\leqslant p, |a-b|<\tau} \int_0^1 \mathbb{E}\left[\sum_{i:|a-i|\geqslant\tau} \sum_{j:|b-j|\geqslant\tau} ||f'||_\infty^2 \mathbb{1}\{\boldsymbol{R}^{ai} \in [u, v]\} \mathbb{1}\{\boldsymbol{R}^{bj} \in [u, v]\} \left| \left(\sum_{k=1}^{n} \boldsymbol{R}^{ai}_{ka}\boldsymbol{R}^{bj}_{kb} \right) \right| \right] dt,$$

where $[u, v]$ is the support of f and hence contains the support of f'.

It can be verified that

$$\boldsymbol{R}^{ai}_{ka} = \frac{2(X_{\cdot a} \cdot X_{\cdot i})X_{ki}}{n}.$$

and similarly

$$\boldsymbol{R}^{bj}_{kb} = \frac{2(X_{\cdot b} \cdot X_{\cdot j})X_{kj}}{n}.$$

These two equations imply that

$$\sum_{k=1}^{n} \boldsymbol{R}^{ai}_{ka}\boldsymbol{R}^{bj}_{kb} = \frac{4}{n^2}(X_{\cdot a} \cdot X_{\cdot i})(X_{\cdot b} \cdot X_{\cdot j})(X_{\cdot i} \cdot X_{\cdot j}). \tag{2.19}$$

Now,

$$\mathbb{1}\{\boldsymbol{R}^{ai} \in [u, v]\} \implies |(X_{\cdot a} \cdot X_{\cdot i})| \leqslant \sqrt{n(4\log p - \log\log p + v)} \leqslant \sqrt{4n\log p},$$

for large n. And we have a similar inequality for $|(X_{\cdot b} \cdot X_{\cdot j})|$. This gives the following inequality

$$|\mathbb{E}G(Y) - \mathbb{E}G(X)| \leqslant$$
$$||f'||_\infty^2 \frac{16 \log p}{n} \sum_{1 \leqslant a < b \leqslant p, |a-b| < \tau} \int_0^1 \mathbb{E}\left[\sum_{i:|a-i| \geqslant \tau} \sum_{j:|b-j| \geqslant \tau} \mathbb{1}\{\boldsymbol{R}^{ai} \in [u,v]\} \mathbb{1}\{\boldsymbol{R}^{bj} \in [u,v]\} \, |(X_{\cdot i} \cdot X_{\cdot j})| \right] dt,$$

Let $\{q_n\} \subset \mathbb{R}$ be a sequence such that $q_n \to \infty$ and $q_n = o(n^t)$ for any $t > 0$. For any i, j write

$$|(X_{\cdot i} \cdot X_{\cdot j})| = |(X_{\cdot i} \cdot X_{\cdot j})| \, \mathbb{1}\{|(X_{\cdot i} \cdot X_{\cdot j})| \leqslant n q_n\} + |(X_{\cdot i} \cdot X_{\cdot j})| \, \mathbb{1}\{|(X_{\cdot i} \cdot X_{\cdot j})| > n q_n\}.$$

Using this we have the following inequality

$$|\mathbb{E}G(Y) - \mathbb{E}G(X)| \leqslant$$
$$||f'||_\infty^2 16 q_n \log p \sum_{1 \leqslant a < b \leqslant p, |a-b| < \tau} \int_0^1 \mathbb{E}\left[\sum_{i:|a-i| \geqslant \tau} \sum_{j:|b-j| \geqslant \tau} \mathbb{1}\{\boldsymbol{R}^{ai} \in [u,v]\} \mathbb{1}\{\boldsymbol{R}^{bj} \in [u,v]\} \right] dt$$
$$+ ||f'||_\infty^2 \frac{16 \log p}{n} \sum_{1 \leqslant a < b \leqslant p, |a-b| < \tau} \int_0^1 \mathbb{E}\left[\sum_{i:|a-i| \geqslant \tau} \sum_{j:|b-j| \geqslant \tau} \mathbb{1}\{|(X_{\cdot i} \cdot X_{\cdot j})| > n q_n\} \, |(X_{\cdot i} \cdot X_{\cdot j})| \right] dt$$

We first bound the second term. Note there are at most $2p^3\tau$ ways of choosing the four indices a, b, i and j. By Lemma 4, the second term is

$$\begin{aligned}
&\leqslant ||f'||_\infty^2 \frac{32 \log p}{n} p^3 \tau \exp\left\{ -\left(\frac{1}{8} q_n - \log C \right) n \right\}, \\
&= 16 ||f'||_\infty^2 \exp\left\{ -\left(\frac{1}{8} q_n - \log C \right) n + 3 \log p + \log \tau + \log \log p - \log n \right\}, \\
&\to 0 \text{ as } \log p = o(\sqrt[3]{n}), \tau = o(p^t) \text{ for any } t > 0,
\end{aligned}$$

where $0 < C < \infty$ is a constant. We now bound the first term. This can be rewritten as

$$\begin{aligned}
&||f'||_\infty^2 16 q_n \log p \sum_{1 \leqslant a < b \leqslant p, |a-b| < \tau} \sum_{i:|a-i| \geqslant \tau} \sum_{j:|b-j| \geqslant \tau} \int_0^1 P(\boldsymbol{R}^{ai} \in [u,v], \boldsymbol{R}^{bj} \in [u,v]) dt \\
&\leqslant ||f'||_\infty^2 16 q_n \log p \sum_{1 \leqslant a < b \leqslant p, |a-b| < \tau} \sum_{i:|a-i| \geqslant \tau} \sum_{j:|b-j| \geqslant \tau} \int_0^1 P(\boldsymbol{R}^{ai} \geqslant u, \boldsymbol{R}^{bj} \geqslant u) dt
\end{aligned}$$

To bound $P(\boldsymbol{R}^{ai} \geqslant u, \boldsymbol{R}^{bj} \geqslant u)$, we do case work. There are four column vectors involved.

1. One pair of the four columns are dependent. We know columns a and b are dependent already as $|a-b| < \tau$. So in this case only these two columns must be dependent. By Lemma 6.9 from (Cai and Jiang, 2011)

$$\sup_{|\rho_{ab}|} P(\boldsymbol{R}^{ai} \geqslant u, \boldsymbol{R}^{bj} \geqslant u) = O(p^{-4+\epsilon}),$$

for any $\epsilon > 0$. There are at most $2p^3\tau$ ways of choosing the four column vectors. Thus the bound in this case is

$$\leqslant (||f'||_\infty^2 16 q_n \log p)(2p^3\tau)(cp^{-4+\epsilon}),$$

for large n and some constant $c > 0$. The above product goes to 0 as $n \to \infty$.

2. Two pairs of columns are dependent. There are two subcases here:

2a. Columns a and b are dependent and either columns a and j are dependent or columns b and i are dependent. By Lemma 6.10 from (Cai and Jiang, 2011)

$$\sup_{|\rho_{ab}|,|\rho_{aj}|,|\rho_{bi}|} P(\boldsymbol{R}^{ai} \geqslant u, \boldsymbol{R}^{bj} \geqslant u) = O(p^{-\frac{8}{3}+\epsilon}),$$

for any $\epsilon > 0$. There are at most $8p^2\tau^2$ ways of choosing the four column vectors in this case. Thus the bound in this case is

$$(||f'||_\infty^2 16 q_n \log p)(8p^2\tau^2)(cp^{-\frac{8}{3}+\epsilon}),$$

for large n and some constant $c > 0$. This product goes to 0 as $n \to \infty$.

2b. Columns a and b are dependent and columns i and j are dependent. Let

$$S_1 := \{(s_1, s_2, s_3, s_4) | s_i \in \{1, 2, \ldots, p\} \text{ for } 1 \leqslant i \leqslant 4, |s_1 - s_2| < \tau, |s_3 - s_4| < \tau,$$
$$|s_i - s_j| \geqslant \tau \text{ for } (i,j) \neq (1,2), (3,4)\}$$

S_1 is a formal way of writing the set of all possible combinations for (a, b, i, j) under this subcase. Now let

$$S_2 := \{(s_1, s_2, s_3, s_4) \in S_1 | s_i \in \Gamma_{p,\delta} \text{ for some } 1 \leqslant i \leqslant 4\},$$

and let

$$S_3 := \{(s_1, s_2, s_3, s_4) \in S_1 | s_i \notin \Gamma_{p,\delta} \text{ for } 1 \leqslant i \leqslant 4\}.$$

We have $S_1 = S_2 \cup S_3$. We first bound for indices in S_3.

By Lemma 6.11 from (Cai and Jiang, 2011) $\exists \epsilon' > 0$ which is dependent on $\delta > 0$ such that

$$\sup_{|\rho_{ab}|,|\rho_{ij}| \leqslant 1-\delta} P(\boldsymbol{R}^{ai} \geqslant u, \boldsymbol{R}^{bj} \geqslant u) = O(p^{-2-\epsilon'}),$$

There are at most $4p^2\tau^2$ ways to choose the column vectors in this case. The bound in this case is

$$(||f'||_\infty^2 16 q_n \log p)(4p^2\tau^2)(cp^{-2-\epsilon'}),$$

for large n and some constant $c > 0$. This product goes to 0 as $n \to \infty$.

We now bound for indices in S_2. For this case we note that

$$P(\boldsymbol{R}^{ai} \geqslant u, \boldsymbol{R}^{bj} \geqslant u) \leqslant P(\boldsymbol{R}^{ai} \geqslant u) = \Theta(p^{-2}).$$

There are at most $8|\Gamma_{p,\delta}|p\tau^2$ ways of choosing the four indices in this case. Thus the bound in this case is

$$(||f'||_\infty^2 16 q_n \log p)(8|\Gamma_{p,\delta}|p\tau^2)(cp^{-2}),$$

for large n and some constant $c > 0$. This product goes to 0 as $n \to \infty$ as $\Gamma_{p,\delta}| = o(p^{1-\epsilon})$ for some $\epsilon \in (0, 1)$.

3. Three pairs of columns are dependent. Note that since columns a and i and columns b and j are independent as $|a - i| \geqslant \tau$ and $|b - j| \geqslant \tau$ respectively. If we think of the four columns as four nodes of a graph and draw an edge if the respective pair of columns are dependent, then this graph will be connected. Therefore, the number of ways to choose columns in this case is at most $8p\tau^3$. Note that

$$P(\boldsymbol{R}^{ai} \geqslant u, \boldsymbol{R}^{bj} \geqslant u) \leqslant P(\boldsymbol{R}^{ai} \geqslant u) = \Theta(p^{-2})$$

Thus the bound in this case is

$$(||f'||_\infty^2 16 q_n \log p)(8p\tau^3)(cp^{-2}),$$

for some $c > 0$ and large n. This product goes to 0 as well.

4. Four or more pairs of columns are dependent. This case is similarly handled as the previous one. There are at most $8p\tau^3$ ways of choosing the columns and the joint probability is bounded by cp^{-2} for some $c > 0$ and for large n.

Since, we have covered all the cases and the bound goes to 0 in each of the cases, we have shown that $|\mathbb{E}G(Y) - \mathbb{E}G(X)| \to 0$. Since $f \in C^\infty(\mathbb{R})$ was arbitrary we are done. ■

2.5.2 Proofs of supplementary results

We state and prove the supplementary results used in the proof of Theorem 2.

Lemma 3. *Let U and V be two $N(0,1)$ vectors and let $\mathbb{E}UV = \rho \in [-1,1]$. Then*

$$U^2 + V^2 \overset{d}{\leqslant} (1 + \rho^2 + \rho\sqrt{1-\rho^2})Z_1^2 + (1 - \rho^2 + \rho\sqrt{1-\rho^2})Z_2^2,$$

where Z_1 and Z_2 are independent $N(0,1)$ random variables (and they are also independent of U and V).

Proof of Lemma 3. Let Z_1 and Z_2 be two $N(0,1)$ random variables which are independent of each other and also independent of U and V. Then we have

$$U \overset{d}{=} Z_1 \text{ and,} \quad V \overset{d}{=} \rho Z_1 + \sqrt{1-\rho^2} Z_2.$$

Now $U_1^2 = Z_1^2$ and

$$\begin{aligned} V_1^2 &= \rho^2 Z_1^2 + (1-\rho^2)Z_2^2 + 2\rho\sqrt{1-\rho^2} Z_1 Z_2 \\ &\leqslant (\rho^2 + \rho\sqrt{1-\rho^2})Z_1^2 + (1 - \rho^2 + \rho\sqrt{1-\rho^2})Z_2^2. \end{aligned}$$

Adding expressions for U_1^2 and V_1^2 gives the result. ■

Lemma 4. *Assume $q_n \to \infty$ as $n \to \infty$. Then there exits n_0 such that for all $n \geqslant n_0$*

$$\mathbb{E}[|U \cdot V| \mathbb{1}\{|U \cdot V| > q_n n\}] \leqslant \exp\left\{-\left(\frac{1}{8} q_n - \log C\right) n\right\}$$

where U and V are $N(0, I_n)$ vectors and $0 < C < \infty$ is a constant not dependent on n.

Proof of Lemma 4. Let

$$S_n = U \cdot V = \sum_{k=1}^{n} U_k V_k.$$

Let $t_0 = \frac{1}{4}$. Then there exists $x_0 > 0$ such that for all $x \geqslant x_0$ we have $x < \exp\left\{\frac{t_0}{2} x\right\}$. Choose $n_0 \in \mathbb{N}$ such that $n q_n \geqslant x_0$ for all $n \geqslant n_0$. Then for $n \geqslant n_0$ we have

$$|S_n| \left(\mathbb{1}\{|S_n| > q_n n\} \exp\left\{\frac{t_0}{2} q_n n\right\}\right) \leqslant \exp\left\{\frac{t_0}{2} |S_n|\right\} \exp\left\{\frac{t_0}{2} |S_n|\right\} = \exp\{t_0 |S_n|\}.$$

Thus

$$\mathbb{E}|S_n| \mathbb{1}\{|S_n| > q_n n\} \leqslant \exp\left\{-\frac{t_0}{2} q_n n\right\} \cdot \mathbb{E} \exp\{t_0 |S_n|\}. \tag{2.20}$$

Now $|S_n| \leqslant \sum_{k=1}^{n} |U_k V_k|$ and therefore using independence we have that

$$\mathbb{E} \exp\{t_0 |S_n|\} \leqslant \mathbb{E} \prod_{k=1}^{n} \exp\{t_0 |U_k V_k|\} = (\mathbb{E} \exp\{t_0 |U_1 V_1|\})^n.$$

Further since $|U_1 V_1| \leqslant \frac{1}{2}(U_1^2 + V_1^2)$ we have

$$\mathbb{E} \exp\{t_0 |S_n|\} \leqslant \left(\mathbb{E} \exp\left\{\frac{t_0}{2}(U_1^2 + V_1^2)\right\}\right)^n$$

By Lemma 3 we have

$$\begin{aligned}\left(\mathbb{E}\exp\left\{\frac{t_0}{2}(U_1^2+V_1^2)\right\}\right)^n &\leqslant \left(\mathbb{E}\exp\left\{\frac{t_0}{2}((1+\rho^2+\rho\sqrt{1-\rho^2})Z_1^2)\right\}\right)^n \times \\ &\left(\mathbb{E}\exp\left\{\frac{t_0}{2}((1-\rho^2+\rho\sqrt{1-\rho^2})Z_2^2)\right\}\right)^n \\ &\leqslant \left(\mathbb{E}\exp\left\{\frac{t_0}{2}(3Z_1^2)\right\}\right)^n \times \left(\mathbb{E}\exp\left\{\frac{t_0}{2}(2Z_2^2)\right\}\right)^n \\ &= C^n, \text{ where } 0<C<\infty \text{ as } t_0=\frac{1}{4}.\end{aligned}$$

From (2.20) we have

$$\mathbb{E}|S_n|\mathbb{1}\{|S_n|>q_n n\} \leqslant \exp\left\{-\frac{1}{8}q_n n + n\log C\right\}.$$

■

Proposition 2. *Let Γ_0 be the Poisson point process with intensity function $\gamma(x)=\frac{1}{2\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)$. Let Γ_{X_n} be the point process as defined in 2.7. Suppose as $n\to\infty$:*

1. *$p=p_n\to\infty$ with $\log p = o(n^{1/3})$;*
2. *$\tau = o(p^t)$ for any $t>0$;*

Then $\Gamma_{X_n}\to^d\Gamma_0$.

Proof of Proposition 2. By Theorem 4.18 from (Kallenberg, 2017), it suffices to show that for every interval of the form $(a,b]\subset\mathbb{R}$

1. $P(\Gamma_{X_n}(a,b]=0)\to P(\Gamma_0(a,b]=0)$ and,
2. $\limsup_{n\to\infty}\mathbb{E}\Gamma_{X_n}(a,b]\leqslant\mathbb{E}\Gamma_0(a,b]<\infty$.

We first note that

$$\int_y^\infty\gamma(x)dx=\int_y^\infty\frac{1}{2\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)dx=\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{y}{2}\right).$$

This calculation implies that

$$\Gamma_0(y,\infty)\sim\text{Poisson}\left(\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{y}{2}\right)\right),$$

and so,

$$P(\Gamma_0(y,\infty)=0)=\exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{y}{2}\right)\right) \text{ and } \mathbb{E}\Gamma_0(y,\infty)=\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{y}{2}\right). \tag{2.21}$$

We now give proofs of the two statements at the start of the proof. We note that it suffices to show the above two statements for the the intervals of the form $(y,\infty)\subset\mathbb{R}$ for $y\in\mathbb{R}$.

1. Proof of the statement: $P(\Gamma_{X_n}(y,\infty)=0)\to P(\Gamma_0(y,\infty)=0)$.

$$\begin{aligned}
P(\Gamma_{X_n}(y,\infty)=0) &= P(\boldsymbol{R}(X_{\cdot a},X_{\cdot b})\leqslant y \text{ for } 1\leqslant a<b\leqslant p, |a-b|\geqslant\tau),\\
&= P\left(\max_{1\leqslant a<b\leqslant p,|a-b|\geqslant\tau}\boldsymbol{R}(X_{\cdot a},X_{\cdot b})\leqslant y\right),\\
&= P\left(\max_{1\leqslant a<b\leqslant p,|a-b|\geqslant\tau}(X_{\cdot a},X_{\cdot b})^2\leqslant n(4\log p-\log\log p+y)\right),\\
&\to \exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{y}{2}\right)\right),
\end{aligned}$$

   by Proposition 6.4 from (Cai and Jiang, 2011). The quantity in the last line above is equal to $P(\Gamma_0(y,\infty)=0)$ by 2.21.

2. Proof of the statement: $\lim_{n\to\infty}\mathbb{E}\Gamma_{X_n}(y,\infty)\to\mathbb{E}\Gamma_0(y,\infty)<\infty$.

$$\begin{aligned}
\mathbb{E}\Gamma_{X_n}(y,\infty) &= \sum_{1\leqslant a<b\leqslant p,|a-b|\geqslant\tau} P(\boldsymbol{R}(X_{\cdot a},X_{\cdot b})>y),\\
&= |D|\cdot P(\boldsymbol{R}(X_{\cdot 1},X_{\cdot\tau+1})>y),
\end{aligned}$$

   where $D=\{1\leqslant a<b\leqslant p,|a-b|\geqslant\tau\}$. Note that $|D^c|\leqslant 2p\tau \implies |D|\sim\frac{p^2}{2}$.

   By Lemma 6.8 from (Cai and Jiang, 2011) we have

$$\begin{aligned}
P(\boldsymbol{R}(X_{\cdot 1},X_{\cdot\tau+1})>y) &= P(|(X_{\cdot 1},X_{\cdot\tau+1})|>\sqrt{n(4\log p-\log\log p+y)})\\
&= 2P((X_{\cdot 1},X_{\cdot\tau+1})>\sqrt{n(4\log p-\log\log p+y)})\\
&= 2P\left(\frac{(X_{\cdot 1},X_{\cdot\tau+1})}{\sqrt{n\log p}}>y_n\right)
\end{aligned}$$

where $y_n \to 2$. Now,

$$P\left(\frac{(X_{\cdot 1}, X_{\cdot \tau+1})}{\sqrt{n \log p}} > y_n\right) \sim \frac{p^{-y_n^2/2}(\log p)^{-1/2}}{2\sqrt{2\pi}} \sim \frac{e^{-y/2}}{\sqrt{8\pi}p^2}.$$

Thus, $P(\boldsymbol{R}(X_{\cdot 1}, X_{\cdot \tau+1}) > y) \sim 2\frac{e^{-y/2}}{\sqrt{8\pi}p^2}$ and so, $\mathbb{E}\Gamma_{X_n}(y, \infty) \sim \frac{e^{-y/2}}{\sqrt{8\pi}}$. This implies $\lim_{n\to\infty} \mathbb{E}\Gamma_{X_n}(y, \infty) \to \frac{e^{-y/2}}{\sqrt{8\pi}} = \mathbb{E}\Gamma_0(y, \infty)$.

This completes the proof.

■

CHAPTER 3

Analysis of structural brain connectivity data sets

Networks are an important tool to analyze complex systems consisting of an interacting collection of entities. Though networks offer a simplification as they consist only of pairwise relationships between entities, networks have been extremely useful in understanding the structure and function of real world systems (Newman, 2010). A large body of network literature analyzes what are known as single layer networks which represent a single type of relationship. However several applications necessitate the study of multilayer networks which can represent multiple types of relationships. For example, snapshots of social media networks over time captures relationships between people at each time point but also how these relationships change over time (Borge-Holthoefer et al., 2011). Another example is of air transportation networks which captures whether there are direct flights between airports for a collection of airlines (Cardillo et al., 2013).

The analysis of multilayer networks has involved both generalizing methods for single layer networks and introduction of new methods suited for multilayer analysis. For example diagnostic measures such as node degree, clustering coefficient and centrality have been extended to multilayer networks (Battiston et al., 2014; Cozzo et al., 2013; Ng et al., 2011). Similarly for joint modeling of multilayer networks, several methods have been proposed (Bianconi, 2013; Stanley et al., 2016; Peixoto, 2015).

In this chapter we look at two multilayer network data sets. These data sets were provided to us by Martin Styner (personal communication). We start with a description of the data sets in section 3.1. We then describe the analysis of these data sets in section 3.2.

3.1 Description of data sets

We give a brief description of the two data sets we have worked with.

3.1.1 Infant data set

This data set consists of 617 networks. The networks are in the form of weighted adjacency matrices of dimension 78×78 for almost all the networks. The networks represent white matter connectivity between 78 regions in the brain. The regions are given by the Automated Anatomical Labeling (AAL 90) atlas. There are 219, 238 and 160 networks for ages 0, 1, and 2 respectively. There are multiple networks or scans for many of the subjects. The following table gives details about this:

	Years 0 and 1	Years 0 and 2	Years 1 and 2	Years 0, 1, and 2
Total	86	31	73	19
Twins	54	25	63	17

A key feature of the data set is the presence of many twins. For the infants, we also have information such as the date of birth, the gestation age at birth, gender and the time of brain scan. In addition, there is information about whether the parents have had PTSD, Schizophrenia or abuse.

3.1.2 ADNI data set

This data set consists of 514 networks. The networks are in the form of weighted adjacency matrices of dimension 148×148. The networks represent white matter connectivity between 148 regions in the brain. There are multiple networks or scans for the subjects. There is an average of 3.64 scans per subject and a median of 4. We have one of the three diagnosis labels for the scans: cognitively normal (CN), mild cognitive impairment (MCI), or Alzheimer's disease (AD). We know the scan date and the age at scan for most of the scans. For each of the subjects, we know the gender, education, ethnicity and whether or not the subject is married. In addition, we know the MMSE scores at the time of the first scan of the subjects.

3.2 Data analysis

In this section we give a summary of our data analysis on the two data sets described before.

3.2.1 Data analysis on infant data set

We had the following questions for the data analysis:

1. How to normalize the data to put networks coming from different ages at the same scale?
2. What are some ways to account for the scanner effect and noise in the data set?
3. Can we predict attributes such as gestation age at scan and age since birth from the networks?
4. Can we identify the twins as either preterm or not-preterm based on the networks?
5. What are some multilayer models for
   (a) networks over time?
   (b) networks in one year or networks coming from one scanner?

We now describe our exploratory data analysis towards answering these questions. There is a large amount of myelination in the human brain in the first few years. Thus, to compare all the networks taken at the various ages, we need to normalize the data. Based on domain experts, this can be done in two ways:

1. Matrix sum normalization: This is also referred to as the row-column normalization. Suppose we have a network matrix A on n nodes. This normalization is given by the mapping $A_{ij} \mapsto \frac{A_{ij}}{\sum_{k,l=1}^{n} A_{kl}}$. The interpretation is that after normalization the edge A_{ij} represents the strength of connectivity between node i and node j relative to the whole brain.
2. Row normalization: Given a network A on n nodes, the normalization is given by the mapping $A_{ij} \mapsto \frac{A_{ij}}{\sum_{k=1}^{n} A_{ik}}$. The interpretation is that after normalization the edge A_{ij} represents the strength of the connectivity relative to the connections originating from node i.

We also explored some other ways to normalize the data such as:

1. Normalization by surface area of the nodes in the brain: For a network A, this normalization sends $A_{ij} \mapsto \frac{A_{ij}}{s_i + s_j}$, where s_i and s_j are the surface areas of the nodes.
2. Row-column normalization: For a network A, this normalization sends $A_{ij} \mapsto \frac{A_{ij}}{\sum_{k=1}^{n} A_{ik} + \sum_{k=1}^{n} A_{kj}}$.

Among the four ways of normalizing the data, the broad summary of the data analysis is similar or no-better than that of the case when we use the matrix sum normalization. Therefore, we will mostly restrict to only the matrix sum normalization for the discussion below.

Identifying attributes from the networks To separate preterm and non-preterm among the twin subjects based on their year 0 networks, we used random forests. The random forests were trained on 50 random $70 - 30$ test-train sets. The mean accuracy was about 64% depending upon the normalization used. However, standard deviation of the accuracy was high to not be able to conclude if the prediction was better than random. Towards the same question, the node-layer community extraction in (Wilson et al., 2017) was applied to the data set. This does not separate the subjects as preterm or non-preterm either.

Next, to predict the gestation age at scan from the edge weights, random forests, LASSO and ridge regression was used. The mean squared predicted error (MSPE) was about 10^2 to 12^2 depending upon the normalization used or if the networks were thresholded at the 90% level. For reference, the prediction of reporting the mean age on the test set had a MSPE of 148.83. So, the methods performed a little better than random. Similar results were obtained for predicting age since birth.

kNN graphs, Principal component analysis and tSNE

kNN graphs were constructed as follows. Fix $k = 5$. Each network in our data set is considered to be a node in the kNN graph. The distance between two networks is given by the Frobenius norm of the difference between the respective adjacency matrices. A visualization of this graph shows 5 connected components. The value of k was varied around 5 and found the same connected components. We also used distances such as the l_1 norm and the cut metric and arrived at similar results. The year 0 networks form two of the five connected components. These connected components correspond to the two scanners, Allegra and TRIO, used for the brain scans. The year 1 and year 2 networks are spread over the other three components. One of these three connected components consists of only networks from the Allegra scanner.

Each network was flattened to a $\binom{78}{2}$-dimensional vector. A principal component analysis was performed using these vectors. The first three principal components (PCs) account for 50.86% of the variation in the data set. The rest of the PCs individually account for small amount of variation. We see 5 clusters in the PC1 vs. PC2 plot which are the same as the networks in the 5

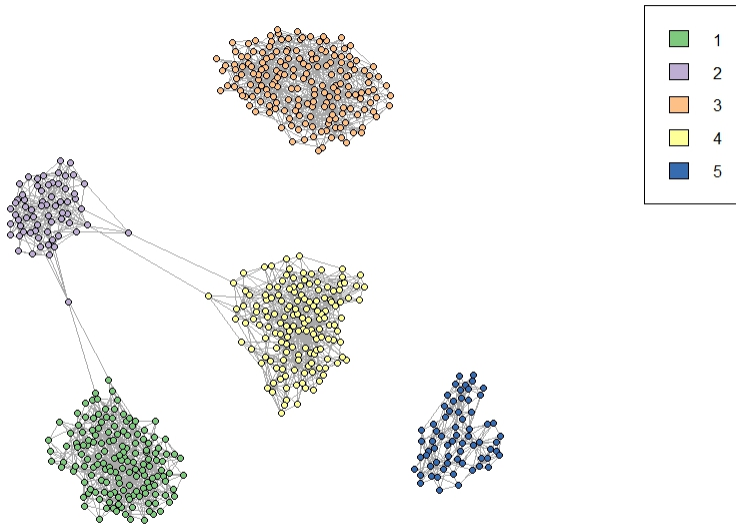

Figure 3.1: Visualization of the kNN graph over all networks. Here we take $k = 5$.

connected components in the kNN graphs. tSNE applied on the flattened networks also gives the same clusters.

Our unresolved questions emerging from this analysis are:

1. What property do the networks in the three connected components consisting of year 1 and year 2 networks satisfy which leads to their separation?

2. How to account for or adjust for the scanner effect observed in the data set?

Analyzing networks from the Allegra scanner

Since a scanner effect was observed in the data, analysis was restricted to networks from the Allegra scanner as these form a larger percentage of the data set. Further to investigate differences over time, we further restricted to Allegra networks in three clusters out of the five observed in PCA in Figure 3.2. The three groupings consist of year 0 networks, mostly year 1 networks, and mostly year 2 networks respectively. A principal component analysis leads to three clusters in the PC1 vs. PC2 plot. PC1 separates the three clusters. We observe that PC2 scores increase with age. A couple of natural questions are as follows:

1. Are the PC2 scores a measure of growth?

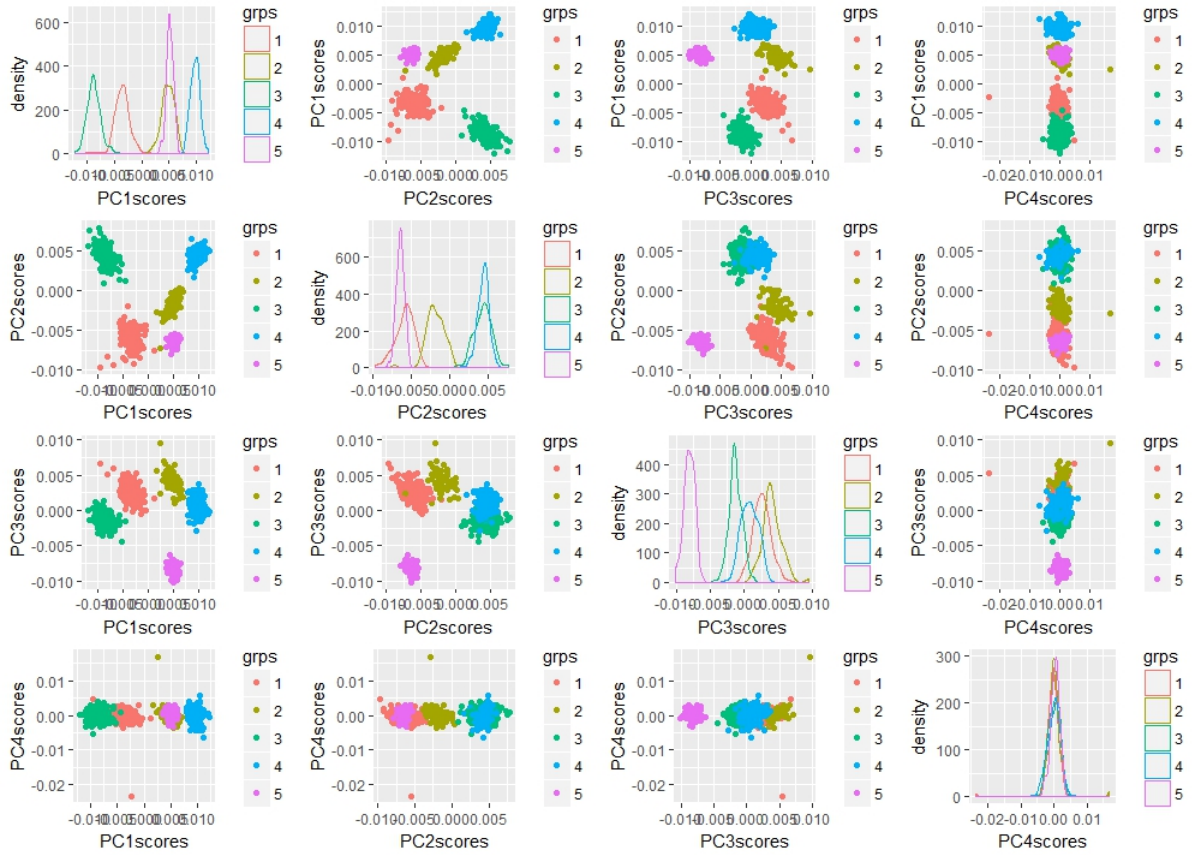

Figure 3.2: Visualization of the output of the PCA. Each point in scores plots is a network. The networks are colored by the connected component they belong to in kNN graphs

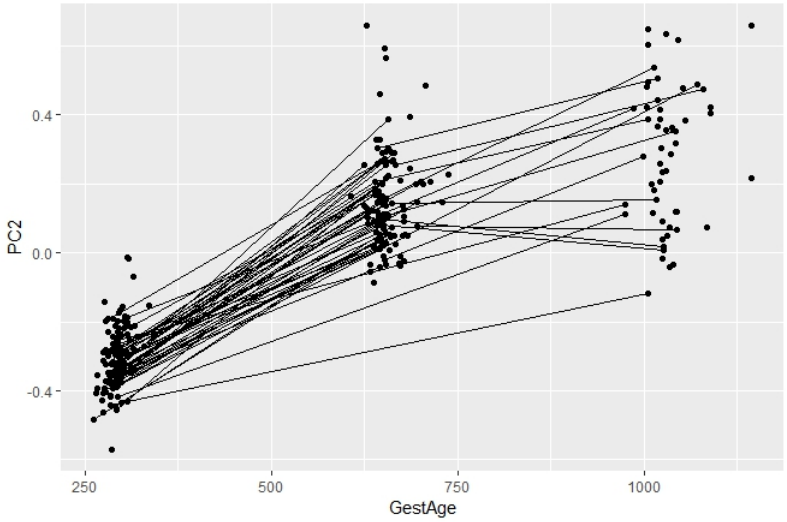

Figure 3.3: Line plot of PC2 values over the gestation age at scan. The lines connect the scans of the subjects over time

2. Is there consistency in how points(networks) move over time in the PC1-PC2 space?

For the first question, an ordinary least squares regression of PC2 over age at scan gave $R^2 = 72\%$(see Figure 3.3). Given the variation in PC2 values of subjects across time, one approach is to fit a random effects model with a random term in the slope. However, a least squares fit only for year 0 data produced $R^2 = 16.42\%$. One may hypothesize the length of the pregnancy to be a contributing factor to the development of the infant brain. However, adding an extra variable for the gestation age at birth only increases R^2 to 17.07%.

For the second question, Kendall's rank correlation test showed that only going from year 1 to year 2, the PC2 ranks were significantly concordant(p-value = 0.0018). Also, if we look at only the transitions between the three clusters, the PC2 ranks were not significantly concordant.

Fitting block models and edge distributions

The networks in the infant data set have a block structure. This suggests modeling them using the stochastic block model (SBM). Towards this, we applied the strata multilayer SBM (Stanley et al., 2016) to year 0 networks. The networks were thresholded at a some threshold levels such as $85\%, 90\%$ and 95% level and model parameters were checked. At the 85% threshold level, the model fit to two strata split the year 0 networks roughly into the networks scanned by Allegra and the TRIO scanner.

Next, we checked whether the model separates networks by year. For this, we collected the 56 subjects who have scans available at year 0 and year 1 and are scanned by Allegra scanner. This ensures there is no scanner effect. We then fit this subset of the data to two strata. However, the networks in both the strata have many year 0 and year 1 networks. We also fit SBM individually to each of the $56 * 2$ networks and clustered the parameters using the Euclidean distance between the fitted block probability matrices and the mutual information criterion on community labels. Both of these clustering approaches lead to clusters with both year 0 and year 1 networks. This analysis suggests that we may be losing structure in the data by thresholding. Alternatively, we need to model the networks in different ways. For the former, we looked at edge distributions and found that most of the high mean edges follow a Gamma distribution. This will be useful for future network models we propose.

3.2.2 Data analysis on ADNI data set

The following were some of our questions towards this data set.

1. How well can the MMSE scores be predicted from the networks?

2. Can the diagnosis labels be predicted from the networks with high accuracy?

3. What are some good models for the evolution of the networks over time?

We now describe our data analysis towards answering these questions. To begin with, a principal component analysis on the data set. For this, each network was flattened to a $\binom{148}{2}$-dimensional row vector. Thus each network is viewed as an observation and each edge is viewed as a variable. In the output, the first three PCs account for 35.38% of the variation in the data set. Rest of the PCs individually account for about 2% or less of the variation. Three clusters were observed in the PC1 vs. PC2 plot. The three groups given by the diagnosis labels are mixed in each of the clusters. It was also verified that the clusters do not align with age and MMSE scores. Since, a strong scanner effect was observed in the infant data set, it possible that these groupings may be related to the scanner used. However, the scanner information is not available for this data set to verify this.

kNN graphs were also constructed as follows. Choose $k = 5$. Each network is viewed as a node and the distance between the networks is given by the Frobenious norm of the difference of the network matrices. Using this setup, kNN graph over all the networks was constructed and visualized. The networks are roughly separated into three connected components as in the PC1 vs PC2 plot described earlier and do not correspond to the diagnosis labels, age or MMSE scores.

To find out well we can do with supervised learning methods, random forests and support vector machines were used. For the diagnosis labels, the mean accuracy was 66% with a standard deviation of 3.75% over 50 randomly generated test-train splits using 70% data on each split for the training. For the MMSE scores, the random forests achieves a mean square prediction error (MSPE) of 7.32 with standard deviation 1.24 over 50 random train-test splits of $70\% - 30\%$ ratio. The baseline prediction given by predicting the mean MMSE score on the test set achieves a MSPE of 9.14 with standard deviation of 1.47. We see that the MSPE for the predictor based on all edge weights is close to the baseline and not necessarily better given the standard deviations.

CHAPTER 4

Community detection using low-dimensional node embeddings

4.1 Introduction

Networks provide a useful framework to study interactions between entities, called nodes, in many complex systems such as social networks, protein-protein interactions, and citation networks. A number of important machine learning tasks such as clustering of nodes (community detection), node classification, link prediction, and network visualization of node interactions can be performed using network-based methods for these systems. In recent decades, network data sets containing millions and even billions of nodes have become available in areas such as social networks. This has necessitated the development of new methods which are scalable to very large networks. One major class of algorithms, often termed network representation learning or network embedding techniques, try to learn representations of network functionals (including nodes and in some cases edges) in a low-dimensional Euclidean space, thus making large-scale network-valued data amenable to well-known methods for data sets in Euclidean spaces. Typical applications of these methods include network visualization by using t-SNE or PCA on the network embeddings, clustering of related nodes by applying k-means on the network embeddings, and classification of nodes by applying machine learning methods for Euclidean spaces. The main advantage of these algorithms lies in the fact that they are scalable to very large networks even with millions of nodes; see (Chami et al., 2020; Zhang et al., 2018; Hamilton et al., 2017b) and the references therein for a comprehensive description of the multitude of methods now available and their applications in various domains of network science.

Network embedding methods can be broadly classified into three types: methods based on matrix factorization (Belkin and Niyogi, 2002; Cao et al., 2015; Ou et al., 2016; Ahmed et al., 2013), method based on random walks (Tang et al., 2015; Perozzi et al., 2014; Grover and Leskovec, 2016), and methods based on deep neural networks (Cao et al., 2016; Wang et al., 2016). In this chapter,

we will focus on methods based on random walks. The first step for these methods consists of running multiple random walks on the underlying graph and creating a matrix of *co-occurrences* which keeps track of how frequently random walks visit one node from another within certain time-span. Pairs of nodes which are closer to each other have a higher co-occurrence than pairs which are further apart. The second step then consists of optimizing for network embeddings so that the Euclidean inner products of the embeddings are proportional to the co-occurences (cf. Section 4.3.2). These algorithms turn out to be heavily used in practice owing to a number of reasons including: (a) network representations can be constructed using random walk based methods for large-scale networks efficiently (see e.g. (Perozzi et al., 2014; Grover and Leskovec, 2016) for applications on hundreds of thousands and in some cases million node networks); (b) these algorithms are easily parallelizable; and (c) they can be easily adapted for local changes in the graph and thus easy to use for streaming network data. These methods have been empirically observed to perform well for link prediction and node classification for large sparse graphs as co-occurences provide a flexible measure of strength of the relationship between any two nodes as compared to deterministic measures based on degrees and co-neighbors.

This chapter focuses on the problem of finding clustering of nodes, often referred to as the *community detection problem*. A well-known model for generating synthetic benchmarks for validating new community detection algorithms is the stochastic block model (SBM) (Holland et al., 1983b; Fortunato and Hric, 2016). The literature for theoretical studies for community detection on SBM is extensive and these existing methods primarily use spectral algorithms, semi-definite programming, and message passing approaches. We refer the reader to the survey (Abbe, 2018) for an overview. Since the more recent network embedding methods are designed to be scalable, it is desirable to rigorously study them for the community detection problem. In order to find K communities, we can apply *K-means clustering* on the embeddings from these algorithms to find communities. The primary scientific question we aim to address in this chapter is:

> *In settings when the underlying network is generated from SBM, when can a network-embedding followed by k-means clustering recover the community assignments? Since community detection is generally more difficult for sparse graphs, it is also of interest to know the relationship between sparsity levels and the co-occurrence length which result*

in meaningful guarantees as well as to understand regimes where these algorithms might fail.

Here, co-occurrence length being t means that the algorithm constructs the co-occurrence matrix by looking at the co-occurrence of nodes at t steps of the random walk. To answer the above questions, we look at perhaps the two most popular of random-walk based network embedding algorithms: (1) DeepWalk due to Perozzi, Al-Rfou, and Skiena (Perozzi et al., 2014), and (2) node2vec due to Grover and Leskovec (Grover and Leskovec, 2016). While DeepWalk uses simple random walks for the network embedding task, node2vec makes use of a weighted random walk with a different weight for backtracking to the last visited node. To understand our results from a high-level, let n denote the network size, ρ_n denote the sparsity level and t denote the co-occurrence length (these are defined more precisely in Sections 4.3.1, 4.3.2). We will establish the following:

(R1) There exists a $\phi = \phi(t)$ such that when $n^{t-1}\rho_n^t \times \frac{1}{(n\rho_n)^\phi} \gg (\log n)^C$, then DeepWalk recovers communities of all but $o(\sqrt{n})$ nodes with high probability (cf. Theorem 4).

(R2) If $n^{t-1}\rho_n^t \gg (\log n)^C$ and the backtracking probability is sufficiently small, then node2vec recovers communities of all but $o(\sqrt{n})$ nodes with high probability (cf. Theorem 6).

(R3) If $n^{t-1}\rho_n^t \ll 1$, then the co-occurrence matrix might be quite far from the ground truth, which provides strong evidence that DeepWalk, node2vec might end up misclassifying a positive fraction of nodes (cf. Theorem 3 (4.17), and Theorem 5 (4.23)).

The results show that if the random walks used for the embedding are close to being non-backtracking, then they can succeed up to the almost optimal sparsity level of the network. Intuitively, backtracks do not provide much new information about the network structure and their effect becomes more prominent as the network becomes sparser. The effect due to backtracks, as well as the interpretation of ϕ in (R1), is discussed in more detail in Remark 1. This phenomenon is in line with the effectiveness of spectral methods using non-backtracking matrices in the extremely sparse setting (networks with bounded average degree), which has been studied by Krzakala, Moore, Mossel, Neeman, Sly, Zdeborová, and Zhang (Krzakala et al., 2013) and Bordenave, Lelarge, and Massoulié (Bordenave et al., 2015) in the context of recovering communities partially. The closest paper to this work is the important recent result of Zhang and Tang (Zhang and Tang, 2021) which

proves an analogue of (R1) above for the DeepWalk. Our result (R1) further improves the regime of success for DeepWalk compared to (Zhang and Tang, 2021).

Organization. The rest of the chapter is organized as follows. Section 4.3 describes the background such as the stochastic block model, the DeepWalk and node2vec algorithms, and community detection using spectral clustering. Section 4.4 describes our main results in detail, followed by proof ideas. The next chapter is dedicated to proofs of these main results.

4.2 Background

We give a brief background to two methods which are useful in understanding the node embedding algorithms, DeepWalk and node2vec. We first discuss noise-contrastive estimation (NCE) in section 4.2.1. We then discuss the Skip-gram model from natural language processing (NLP) literature in section 4.2.2.

4.2.1 Noise-contrastive estimation

NCE (Gutmann and Hyvärinen, 2010) is a method to estimate parameters for unnormalized statistical models. We describe the method as follows. Suppose $x_1, x_2, \ldots, x_t$ are t data points where the data distribution is given by a probability density function, p_d. Further, suppose that p_d takes the following form

$$p_d(\cdot;\alpha) = \frac{p_d^0(\cdot;\alpha)}{Z(\alpha)},$$

where $Z(\alpha)$ is the normalization term. The goal is to estimate the parameters α. However, estimating the parameters by maximizing the likelihood is difficult if $Z(\alpha)$ is computationally intractable. As a solution to this problem, NCE considers a related optimization problem described as follows. First, an additional parameter c is introduced to replace the normalization term $Z(\alpha)$. Then, one generates t samples from a known *noise distribution*, p_N, which is easy to sample from. Then one maximizes J_t, as defined below, over the parameters $\theta = \{\alpha, c\}$.

$$J_t(\theta) = \sum_t \log \sigma\left(\log\left(\frac{p_d(x_t,\theta)}{p_N(x_t)}\right)\right) + \sum_t \log\left(1 - \sigma\left(\log\left(\frac{p_d(y_t,\theta)}{p_N(y_t)}\right)\right)\right),$$

where σ is the logistic function. This optimization problem can intuitively be seen to optimize for θ by discriminating between the two classes of samples: data samples and noise samples. If one generates b noise samples for every element in the data set, one optimizes an analogous optimization problem instead:

$$J_t(\theta) = \sum_t \log \sigma_b \left(\log \left(\frac{p_d(x_t, \theta)}{p_N(x_t)} \right) \right) + \sum_{t'} \log \left(1 - \sigma_b \left(\log \left(\frac{p_d(y_{t'}, \theta)}{p_N(y_{t'})} \right) \right) \right),$$
$$\text{where } \sigma_b(x) = (1 + be^{-x})^{-1},$$

the first sum is over data samples i.e. samples from the unknown distribution p_d and the second sum is over the noise samples i.e. samples from the (known) noise distribution p_N. Let θ_t be the parameters obtained by maximizing J_t. It is shown in (Gutmann and Hyvärinen, 2010) that as $t \to \infty$ and under certain conditions,

$$\theta_t \xrightarrow{\mathbb{P}} \theta^*,$$

where θ^* is the true set of parameters generating the data.

4.2.2 Skip-gram Model

Suppose we have a text data set such as a collection of sentences or text documents. The objective of the Skip-gram Model is to learn word representations which are useful for predicting surrounding words in a sentence. Given a data set with the sequence of words $(w_1, w_2, \ldots, w_n)$, the objective is to maximize the log-likelihood function

$$L_{SG}^{(1)} = \frac{1}{n} \sum_{t=1}^{n} \sum_{-c \leqslant j \leqslant c, j \neq 0} \log p(w_{t+j} | w_t), \tag{4.1}$$

where c is the number of words immediately preceding or succeeding the word w_t. In this context, the word w_t is also called the center word. The intuition is that we are optimizing for the probability of predicting nearby words and increasing c can lead to higher accuracy. The conditional probability $p(w_{t+j}|w_t)$ is given by the softmax function:

$$p(w_{t+j}|w_t) = \frac{\exp\left\{ \langle \boldsymbol{f}'_{w_{t+j}}, \boldsymbol{f}_{w_t} \rangle \right\}}{\sum_{i=1}^{n} \exp\left\{ \langle \boldsymbol{f}'_{w_i}, \boldsymbol{f}_{w_t} \rangle \right\}},$$

where $\boldsymbol{f}'_{w_i}$ is the vector representation of the predicted word, also called the *output representation* of w_i, and $\boldsymbol{f}_{w_t}$ is the vector representation of the center word, also called the *input representation* of the word $\boldsymbol{f}_{w_t}$. The denominator of the softmax function can be computationally intractable for large n. Motivated by NCE, (Mikolov et al., 2013) modify $L_{SG}^{(1)}$ in (4.1) to the following:

$$L_{SG}^{(2)} = \frac{1}{n}\sum_{t=1}^{n}\sum_{-c\leqslant j\leqslant c, j\neq 0}\left(\log\sigma(\langle \boldsymbol{f}'_{w_{t+j}}, \boldsymbol{f}_{w_t}\rangle) + \sum_{i=1}^{b}\mathbb{E}_{w_i\sim P_n}[\log\sigma(-\langle \boldsymbol{f}'_{w_i}, \boldsymbol{f}_{w_t}\rangle)]\right), \tag{4.2}$$

where P_n is the empirical distribution of words in the data set and we are sampling b words from P_n for every summand in (4.1). The optimization of $L_{SG}^{(2)}$ in (4.2) is much faster compared to optimization of $L_{SG}^{(1)}$. With the optimization as given in (4.2), this method to compute word vectors is called Skip-gram model with negative sampling.

4.3 Problem Setup

We begin this section by describing the notation used throughout this chapter and chapter 5, followed by the definition of the stochastic block model (SBM) in Section 4.3.1. The underlying graphs for our results will be generated by SBM. Next, we describe the DeepWalk and node2vec network embedding algorithms in Section 4.3.2. These algorithms will be run on graphs generated from the SBM. We end this section by describing the approach to recover communities from the solution of the DeepWalk and node2vec algorithms in Section 4.3.3.

Notation. For a graph $G = (V, E)$, A_G will denote the adjacency matrix representation of G. We will drop the subscript G on A_G and denote the adjacency matrix simply by A when the graph is clear from context. We write $S^{a\times b}$ for the set of $a \times b$ matrices with entries taking values in S. For any matrix X, $X_{i\star}$ denotes the i-th row and $X_{\star i}$ denotes the i-th column. Also, $|X|$ will denote the sum of its entries, and $\|X\|_{\mathrm{F}} = (\sum_{i,j} X_{ij}^2)^{1/2}$ will denote the Frobenius norm of X. Denote $[n] = \{1, 2, \ldots, n\}$. We often use the Bachmann–Landau notation $o(\cdot), O(\cdot)$, $\omega(\cdot), \Omega(\cdot)$ etc. For random variables $(Z_n)_{n\geqslant 1}$, we write $Z_n = o_{\mathbb{P}}(1)$ and $Z_n = O_{\mathbb{P}}(1)$ as a shorthand for $Z_n \to 0$ in probability, and $(Z_n)_{n\geqslant 1}$ is a tight sequence respectively.

4.3.1 Stochastic Block Model

We now describe the stochastic block model (Holland et al., 1983b). A random graph $G = (V, E)$ with $V = [n]$ from the SBM is generated as follows. Each node in the graph is assigned to one of K categories, called communities or blocks. We fix the set of communities to be $[K]$. Let $g(i) \in [K]$ denote the community of node i. We also define an $n \times K$ *membership matrix* Θ_0 which contains the communities of the nodes, defined as follows:

$$(\Theta_0)_{ir} := \mathbb{1}\{g(i) = r\}.$$

Let B be a $K \times K$ matrix of probabilities, i.e. $0 \leqslant B_{rs} \leqslant 1$. The matrix B will be called the matrix of block or community density parameters. Let

$$P := \Theta_0 B \Theta_0^T.$$

The edges of the graph G are generated using P as follows:

$$A_{ij} \sim \text{Bernoulli}(P_{ij}),\ A_{ij} := A_{ji} \quad \text{for } i < j, \qquad \text{and } A_{ii} := 0. \tag{4.3}$$

The graph generated from this model does not have self-loops. Note that the parameter of the Bernoulli distribution used for generating edge $\{i, j\}$ depends only on the communities of the nodes i and j.

For our results in this chapter, we are interested in the case when the number of communities, K, is fixed, and the number of nodes, n, tends to infinity. We will assume throughout that K is known. Then to generate a sequence of graphs, one graph for each $n \in \mathbb{N}$, we first fix a $K \times K$ matrix of probabilities B_0. Let $(\rho_n)_{n \geqslant 1} \subset \mathbb{R}$ be a sequence such that $0 \leqslant \rho_n \leqslant 1$. The sequence ρ_n will control the sparsity of the graphs as a function of n. The matrix B of block density parameters for the graph on $[n]$ is then given by

$$B = \rho_n B_0.$$

We make the following usual assumptions:

Assumption 1. *(1) If n_r denotes the size of community r, then $\frac{n_r}{n} \to \pi_r$ with $\pi_r > 0$ for all $r \in [K]$.*

(2) There exists constants c_L and c_U such that $0 < c_L \leqslant (B_0)_{ij} \leqslant c_U \leqslant 1$ and $\operatorname{rank}(B_0) = K$.

In our set up, the membership matrix Θ_0 is the unknown and we want to estimate it by observing one realization of A. Next, we discuss how to evaluate the accuracy of the predicted community assignments. Let $\hat{\Theta}$ be an estimator of the community assignments. Note that the communities are identifiable only upto a permutation of the community labels. Taking this into account, we measure the prediction error by

$$\operatorname{Err}(\hat{\Theta}, \Theta_0) := \frac{1}{n} \min_{J \in S_K} \sum_{i \in [n]} \mathbb{1}\big\{(\hat{\Theta} J)_{i\star} \neq (\Theta_0)_{i\star}\big\}, \tag{4.4}$$

where S_K denotes the set of all $K \times K$ permutation matrices. If J is the minimizer in (4.4), then we call a node to be *misclassified* under $\hat{\Theta}$ when $(\hat{\Theta} J)_{i\star} \neq (\Theta_0)_{i\star}$. Thus, we can note that $\operatorname{Err}(\hat{\Theta}, \Theta_0)$ just computes the proportion of misclassified nodes.

We conclude our description of the SBM by defining some notation used in the context of random-walks. Towards this, for a graph with adjacency matrix A, let D_A denote the diagonal matrix with the diagonal entries as the degrees of the nodes $[n]$, i.e., $(D_A)_{ii} = \sum_{j=1}^{n} A_{ij}$. Let W_A denote the one-step transition matrix of a simple symmetric random walk given by AD_A^{-1}. Similarly, we define $W_P = PD_P^{-1}$ where D_P is the diagonal matrix containing row sums of P.

4.3.2 DeepWalk and node2vec

We describe the two random-walk based network embedding algorithms, namely DeepWalk due to Perozzi, Al-Rfou and Skiena (Perozzi et al., 2014) and node2vec due to Grover and Leskovec (Grover and Leskovec, 2016). Let $G = (V, E)$ be an undirected, connected (and possibly weighted) graph with $V = [n]$, and let A_G be the adjacency matrix of G if G is unweighted. If G is weighted, we can take A_G to be the matrix of edge weights. Both the algorithms have two key steps. In step 1, the algorithm generates multiple random walks on G. In step 2, one uses a "word embedding" algorithm from the natural language processing literature such as word2vec (Mikolov et al., 2013) by interpreting these random walks as sequences of words.

Step 1: Generating walks. For both the methods, we generate r random walks, each of length l on G. In the case of DeepWalk, a simple random walk is performed which has the one-step transition probability given as

$$p(v_{i+1}|v_i) = \begin{cases} \frac{1}{|(A_G)_{i\star}|}, & \text{if } (v_i, v_{i+1}) \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{4.5}$$

In the case of node2vec, a second-order random walk with two parameters α and β is performed. If $\mathcal{N}(v)$ denotes the neighborhood of v, then the one-step transition probability in this case is given by

$$p(v_{i+1}|v_i, v_{i-1}) \propto \begin{cases} \alpha & \text{if } v_{i+1} = v_{i-1}, \\ 1 & \text{if } v_{i+1} \in \mathcal{N}(v_{i-1}) \cap \mathcal{N}(v_i), \\ \beta & \text{if } v_{i+1} \in \mathcal{N}(v_{i-1})^c \cap \mathcal{N}(v_i), \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

We consider $\beta = 1$ throughout this chapter. Following Qiu, Dong, Ma, Li, Wang, and Tang (Qiu et al., 2018), we initialize using the stationary distribution of the random walks. For DeepWalk, the initialization is given by

$$p(v) = \frac{|(A_G)_{v\star}|}{|A_G|}, \quad \forall v \in V. \tag{4.7}$$

And for node2vec, we initialize by

$$p(v_1, v_2) = \frac{(A_G)_{v_1 v_2}}{|A_G|}, \quad \forall v_1, v_2 \in V. \tag{4.8}$$

The initial distribution in (4.8) places equal probability on each ordered pair of nodes in the edge set E. This initial distribution also happens to be the invariant distribution for the second-order random walk if $\beta = 1$.

Step 2: Implementing word embedding algorithm. Next, the random walks are input to word2vec algorithm due to Mikolov, Sutskever, Chen, Corrado, and Dean (Mikolov et al., 2013).

This generates two node representations for each of the nodes. The latter is described, in the context of graphs, in steps 2a and 2b below.

Step 2a: Computing the co-occurence matrix C. In this step, we construct an $n \times n$ matrix C which counts how often two nodes appear in a certain distance of each other with respect to the observed random walks. Formally, let $t_L, t_U \in \mathbb{N}$ be such that $t_L \leqslant t_U$. Let $\boldsymbol{w}^{(m)}$ for $1 \leqslant m \leqslant r$ be the random walks from Step 1, each represented as a sequence of l nodes. Then

$$C_{ij} = \sum_{m=1}^{r} \sum_{t=t_L}^{t_U} \sum_{k=1}^{l-t} \big(\mathbb{1}\{\boldsymbol{w}_k^{(m)} = i, \boldsymbol{w}_{k+t}^{(m)} = j\} + \mathbb{1}\{\boldsymbol{w}_k^{(m)} = j, \boldsymbol{w}_{k+t}^{(m)} = i\} \big).$$

DeepWalk and node2vec was proposed to have $t_L = 1$. However, we will allow t_L to vary for our theoretical results. Also, the set up explained intuitively in the introduction corresponds to the particular case $t_L = t_U = t$.

Step 2b: Optimizing for node representations. The node representations are then computed using the Skip-gram model with negative sampling (SGNS) which was discussed in section 4.2.2. In the context of networks, nodes are the words and sentences are the random walks. We describe the details of the optimization in the context of networks. Let d be the embedding dimension. The algorithm takes the co-occurrence matrix C as an input, and then it outputs two node vectors $\boldsymbol{f}_i, \boldsymbol{f}_i' \in \mathbb{R}^d$ for each node $i \in [n]$. The vector $\boldsymbol{f}_i$ is the "input" representation of the node and the vector $\boldsymbol{f}_i'$ is the "output" representation of the node. We collect these node vectors in two $d \times n$ matrices F and F'. Although (Grover and Leskovec, 2016) initially proposed to have $F = F'$, in practice this requirement is often dropped. So we will not consider the assumption $F = F'$, which was also done in the theoretical analysis of (Zhang and Tang, 2021). The objective function of the optimization problem is described as follows. Let

$$P_C(j) := \frac{\sum_{i=1}^{n} C_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}} = \frac{|C_{\star j}|}{|C|},$$

be the empirical distribution on $[n]$ constructed using the column sums of C. In (Mikolov et al., 2013), a distribution proportional to $P_C^{3/4}$ was used instead, which performs better in practice. However, in theoretical analysis (Zhang and Tang, 2021; Qiu et al., 2018; Levy and Goldberg, 2014), one always considers P_C for simplicity. Then one computes (F, F') by maximizing the

objective function

$$L'_C(F,F') = \sum_{i,j=1}^{n} \Bigg(C_{ij} \log \sigma(\langle \boldsymbol{f}_i, \boldsymbol{f}'_j \rangle) + \sum_{\{l_m \overset{iid}{\sim} P_C | 1 \leqslant m \leqslant bC(i,j)\}} \log \sigma(-\langle \boldsymbol{f}_i, \boldsymbol{f}'_{l_m} \rangle) \Bigg), \tag{4.9}$$

where b samples are taken from P_C for each pair of nodes counted as a co-occurrence in C, and $\sigma(x) = (1 + \mathrm{e}^{-x})^{-1}$ denotes the sigmoid function. For our theoretical analysis, we again follow simplifications considered in earlier works (Levy and Goldberg, 2014; Qiu et al., 2018; Zhang and Tang, 2021). First, instead of (4.9), we look at a related maximization problem

$$L_C(F,F') = \sum_{i,j=1}^{n} C_{ij} \big(\log \sigma(\langle \boldsymbol{f}_i, \boldsymbol{f}'_j \rangle) + b \mathbb{E}_{l \sim P_C} [\log \sigma(-\langle \boldsymbol{f}_i, \boldsymbol{f}'_l \rangle)] \big). \tag{4.10}$$

Once we have an optimizer (F, F'), we embed node i in $\mathbb{R}^d$ using $F_{i\star}$. The following result due to Levy and Goldberg (Levy and Goldberg, 2014) gives us a way to compute the optimizer in (4.10) using factorization of a certain matrix:

Proposition 3. *Given a matrix $C \in \mathbb{R}^{n \times n}$, let $\bar{M}_C$ be a matrix with (i,j)-th entry given by*

$$(\bar{M}_C)_{ij} := \log \left(\frac{C_{ij} \cdot |C|}{|C_{i\star}||C_{\star j}|} \right) - \log b. \tag{4.11}$$

Let $F, F' \in \mathbb{R}^{n \times d}$ be such that $\bar{M}_C = FF'^T$. Then (F, F') maximizes (4.10).

We give a proof in APPENDIX 1 for completeness. To simplify the form of the optimizers, one either takes $r \to \infty$ or $l \to \infty$ (Qiu et al., 2018; Zhang and Tang, 2021). These assumptions ensure that the co-occurence of (i,j) is observed sufficiently many times for all (i,j) and also simplify the form of $\bar{M}_C$. For our results we will take $r \to \infty$ in which case we have the following result:

Proposition 4. *Let A be the adjacency matrix of a graph. For node2vec, we will assume A resulting from an unweighted graph. Let $\boldsymbol{w}^{(m)}$ for $1 \leqslant m \leqslant r$ be the random walks generated for DeepWalk or for node2vec with $\beta = 1$. If $\sum_{t=t_L}^{t_U} A_{ij}^{(t)} > 0$ then*

$$(\bar{M}_C)_{ij} \xrightarrow[r \to \infty]{a.s.} (M_C)_{ij} := \log \left(\frac{\sum_{t=t_L}^{t_U} (l-t) \big(\mathbb{P}(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = j) + \mathbb{P}(\boldsymbol{w}_1^{(1)} = j, \boldsymbol{w}_{1+t}^{(1)} = i) \big)}{2b\gamma(l, t_L, t_U) \mathbb{P}(\boldsymbol{w}_1^{(1)} = i) \mathbb{P}(\boldsymbol{w}_1^{(1)} = j)} \right), \tag{4.12}$$

where $\gamma(l, t_L, t_U) = \frac{(2l - t_L - t_U)(t_U - t_L + 1)}{2}$. *The limiting term is well-defined if* $\sum_{t=t_L}^{t_U} A_{ij}^{(t)} > 0$ *and the left hand side is also well-defined for large enough* r.

We give a proof in APPENDIX 1. If $\sum_{t=t_L}^{t_U} A_{ij}^{(t)} = 0$ then $(M_C)_{ij} := 0$. We will refer to M_C as the M *matrix* associated to DeepWalk or node2vec on a graph.

4.3.3 Clustering using factorization of M matrix

We simply write M to denote the M-matrix of SBM. The idea would be to apply spectral clustering to the M matrix in order to recover the communities, which considers the top K eigenvalues of M and applies an approximate K-means algorithm to recover the communities. Let us describe this more precisely below. Given a graph, we can always factorize $M = FF'^T$ such that $F, F' \in \mathbb{R}^{n \times d}$ for $d \geqslant n$. That would result in an n-dimensional embedding of the nodes. However, since the underlying graph has an approximate rank-K structure, it might make more sense to try to find an embedding in $\mathbb{R}^K$ using an approximate factorization of M. To that end, we can find

$$\underset{F, F' \in \mathbb{R}^{n \times K}}{\operatorname{argmin}} \left\| M - FF'^T \right\|_{\mathrm{F}}.$$

The solution can be obtained using the singular value decomposition of M. Indeed, if V (resp. U) is the matrix of top K left (resp. right) singular vectors, and S is the diagonal matrix with K largest singular values, then $F = V$ and $F' = SU$. To understand if F should preserve the community structure, let's look at the graph G_0 which is a weighted complete graph with weights given by P. Then the corresponding M-matrix, denoted as M_0, is computed using the expression of the limiting matrix in (4.12) with the initial distribution and transition probabilities given in Step 1. We note that M_0 is deterministic. Recall that Θ_0 is the matrix of community assignments. Also, let $V_0 \in \mathbb{R}^{n \times K}$ be the matrix of the top K left singular vectors of M_0. We will prove the following:

Proposition 5. *If* $\operatorname{rank}(M_0) = K$, *then there exists full-rank matrix* $X_0 \in \mathbb{R}^{K \times K}$ *such that* $V_0 = \Theta_0 X_0 + E_0$ *where* $(E_0)_{ij} = O(n^{-3})$.

The proof of this fact in provided in Section 5.2.1 for DeepWalk and in Section 5.4.1 for node2vec. This essentially shows that if $\operatorname{rank}(M_0)$ for DeepWalk there are K distinct rows in V_0 and two nodes have the same rows in V_0 if and only if they are in the same community. Thus one

can find the communities by applying K-means on the rows of V_0. For node2vec, we have a similar result except for a small noise term.

Next, we note that M_0 may not always have rank K even when B has rank K. This can happen as the rank of a matrix can drop after taking element-wise logarithm. To see this, consider the following simple counter-example:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & e \end{bmatrix}, \text{ and } \log A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

We provide the following lemma which says that such cases only happen on a set of matrices of measure zero. This implies that outside a set of measure zero, if B has rank K then M_0 also has rank K.

Lemma 5. *Let $X \in \mathbb{R}_+^{K\times K}$. Let $\bar{\log} : \mathbb{R}_+^{K\times K} \to \mathbb{R}^{K\times K}$ be the mapping given by taking the element-wise* log *of the matrix entries, and $\bar{\log}(x) = 0$ if $x = 0$. Then*

$$\operatorname{rank}(X) = K \implies \operatorname{rank}(\bar{\log}(X)) = K \quad a.s.\, \lambda,$$

where λ is the Lebesgue measure on $\mathbb{R}_+^{K\times K}$.

See APPENDIX 1 for a proof. For this reason, we assume the following throughout:

Assumption 2. $\operatorname{rank}(M_0) = K$.

Motivated by this, we compute the community assignments as a k-means algorithm and a $(1+\varepsilon)$-approximate solution to the same. The K-means algorithm on the rows of V solves the following optimization problem.

$$(\hat{\Theta}, \hat{X}) = \underset{\Theta\in\{0,1\}^{n\times K}, X\in\mathbb{R}^{K\times K}}{\operatorname{argmin}} \|\Theta X - V\|_{\mathrm{F}}^2. \tag{4.13}$$

Since it is NP-hard to find the minimizer for the above problem (Aloise et al., 2009), we can seek an approximate solution instead (Kumar et al., 2004). Let $\varepsilon > 0$ be given. Then we say $(\bar{\Theta}, \bar{X})$ is an $(1+\varepsilon)$-approximate solution to the K-means problem in (4.13) if $\bar{\Theta} \in \{0,1\}^{n\times K}, \bar{X} \in \mathbb{R}^{K\times K}$ and

$$\left\|\bar{\Theta}\bar{X} - V\right\|_{\mathrm{F}}^2 \leqslant (1+\varepsilon) \min_{\Theta \in \{0,1\}^{n\times K}, X \in \mathbb{R}^{K\times K}} \|\Theta X - V\|_{\mathrm{F}}^2 . \tag{4.14}$$

Thus the final community assignment is computed as follows:

Algorithm 1 (Low-dimensional embedding using DeepWalk and node2vec)**.**

(S1) Compute the M-matrix M and compute $V \in \mathbb{R}^{n\times K}$, the matrix containing the top K-eigenvectors (in absolute value of the eigenvalue) of M.

(S2) Compute $(\bar{\Theta}, \bar{X}) \in \{0,1\}^{n\times K} \times \mathbb{R}^{K\times K}$ as an $(1+\varepsilon)$-approximate solution to the k-means problem in (4.13) and output $\bar{\Theta}$.

The $\{0,1\}$-valued matrix $\bar{\Theta}$ has the predicted community assignments for each of the nodes. We note that each node is assigned exactly one community by the algorithm.

4.4 Main results

We describe our results in two parts. We first describe the results for the DeepWalk algorithm. We then describe the results for the node2vec algorithm. For both the algorithms, our approach to recovering communities from M proceeds by first showing that the Frobenius norm, $\|M - M_0\|_{\mathrm{F}}$, is $o_{\mathbb{P}}(n^2)$. Then, one uses the Davis-Kahan theorem (Yu et al., 2015), and shows that this implies that the proportion of misclassified nodes is $o_{\mathbb{P}}(n^{-1/2})$.

Following the intuitions from Section 4.3.3, the primary objective in our community detection task will be to bound $\|M - M_0\|_{\mathrm{F}}$. We first describe the results for DeepWalk in Section 4.4.1, and then state the results for node2vec in Section 4.4.2. We end this section with a proof outline.

4.4.1 Results for DeepWalk

We describe the following proposition for bounding the Frobenius norm $\|M - M_0\|_{\mathrm{F}}$. We will assume that $t_L \geqslant 2$, since if $t_L = 1$, then there are $O_{\mathbb{P}}(n^2)$ entries of M which are equal to 0. This makes $\|M - M_0\|_{\mathrm{F}}$ of the order n. The following result gives estimates on $\|M - M_0\|_{\mathrm{F}}$ based on t_L, t_U and the sparsity parameter ρ_n:

Theorem 3. *Fix $\eta > 0$. Suppose $t_L \geqslant 2$, and let $\phi = \phi(t_L) := \lfloor t_L/2 \rfloor$ if $t_L \geqslant 3$ and let $\phi = 0$ if $t_L = 2$. Also let $c_0 = c_0(t_L, \eta) = 4 + (t_L + 1)\eta$, and suppose that*

$$n^{t_L - 1}\rho_n^{t_L} \times \frac{1}{(n\rho_n)^{\phi}} \gg (\log n)^{c_0}. \tag{4.15}$$

Then

$$\|M - M_0\|_{\mathrm{F}} = O_{\mathbb{P}}\left(n(\log n)^{-\eta} + \mathbb{1}\{t_L = 2\}\left(n\sqrt{\frac{\log n}{n\rho_n^2}}\right)\right). \tag{4.16}$$

If ρ_n is such that $n^{t_U - 1}\rho_n^{t_U} \ll 1$ and $n\rho_n \gg 1$, then given $\varepsilon > 0$ there exists a constant $C_\varepsilon > 0$ such that

$$\mathbb{P}\left(\|M - M_0\|_{\mathrm{F}} \geqslant C_\varepsilon n\right) \geqslant 1 - \varepsilon. \tag{4.17}$$

The proposition says that $\|M - M_0\|_{\mathrm{F}} = o_{\mathbb{P}}(n)$ as long as $n^{t_L - 1 - \phi}\rho_n^{t_L - \phi}$ is growing faster than an appropriate power of $\log n$. The power of the $\log n$ term is required for concentration of the transition matrix of the random walk on SBM. The last part of the proposition says that the Frobenius norm is $\Omega(n^2)$ when $n^{t_U - 1}\rho_n^{t_U} \ll 1$. This suggests that spectral clustering on the M-matrix may not give good results at this level of sparsity. Thus in order to ensure a good low-dimensional embedding, we should take t_L suitably large.

Remark 1 (Effect of backtracking)**.** To understand the quantity ϕ intuitively, let us consider a simple case with $t_L = t_U = t \geqslant 4$ and t is even. If we were to compute the entry say M_{ii}, we need to understand how many walks of length t are possible from i to i. If we consider the edges in this walk to be distinct, then the expected number of paths is of order $n^{t-1}\rho_n^t$, as we need to choose $t - 1$ intermediate vertices and t specific edges need to appear. On the other hand, if we consider the path where the walk alternatively visits a new node and then backtracks to i, then the expected number of paths due to such backtracks is $(n\rho_n)^{t/2} = (n\rho_n)^{\phi}$. The condition in (4.22) is then just ensuring that the contribution due to the backtracking path is smaller than the contribution from the non-backtracking path. These kinds of backtracking walks do not contribute significantly for node2vec with α small.

With the bounds on the Frobenius norms in Theorem 3, we will conclude the following:

Theorem 4. *Suppose that* (4.15) *holds. Fix* $\varepsilon > 0$ *and let* $(\bar{\Theta}, \bar{X})$ *be a* $(1+\varepsilon)$*-approximate solution in Algorithm 1. Then*

$$\mathrm{Err}(\bar{\Theta}, \Theta_0) = o_{\mathbb{P}}(n^{-1/2}),$$

i.e., $\bar{\Theta}$ *misclassifies at most* $o_{\mathbb{P}}(n^{1/2})$ *many node-labels.*

4.4.2 Results for node2vec

We now describe our results for the node2vec algorithm. As mentioned earlier, for our theoretical analysis, we allow the parameter $\alpha = \alpha_n$ to vary with n and the parameter β to be fixed to be equal to 1. We will also consider the cases when $t_L > 2$. This is because when $t_L = t_U = 2$, M_0 may not have a block structure, even asymptotically, and so spectral clustering of M may not give us the communities of the nodes (cf. Lemma 17)

We look at three different regimes for the backtracking parameter α for the analysis of node2vec. To compare DeepWalk and node2vec, we can think of DeepWalk as a special case of node2vec when $\alpha = 1$ and $\beta = 1$. The first regime is for the case when the backtrack parameter is larger than 1 and potentially growing with n. This case assigns a higher weight to the backtracks on the random walks as compared to the DeepWalk algorithm. We do continue to have $\alpha \ll n\rho_n$ in this case which means that as a proportion of the degree of the node, the proportional weight assigned to the backtracking edge goes to zero. The second and third regimes are for the case when α goes to 0 which means that under the model it is less likely to backtrack on the random walks on the graphs. The latter two regimes are differentiated based on the rate at which $\alpha \to 0$.

The following proposition bounds the Frobenius norm $\|M - M_0\|_{\mathrm{F}}$ for each of the three regimes for node2vec.

Theorem 5. *Let* $3 \leqslant t_L \leqslant t_U$*. Fix* $\eta > 0$*. Consider three regimes given as follows:*

1. *Regime I: Let* $c_0 = c_0(t_L, \eta) = 1 + 2\eta(t_L - \lceil \frac{t_L - 1}{2} \rceil) + 1$ *and let* $c_1 = c_1(t_L) = \frac{2}{t_L - \lceil \frac{t_L - 1}{2} \rceil}$. *Let* α *be such that*

$$1 \leqslant \alpha \ll n^{-\frac{1}{t_L - \lceil \frac{t_L - 1}{2} \rceil} + 1} (\log n)^{-c_1}, \tag{4.18}$$

and assume that

$$n^{t_L - \lceil \frac{t_L - 1}{2} \rceil - 1} \rho_n^{t_L - \lceil \frac{t_L - 1}{2} \rceil} \gg \alpha^{t_L - \lceil \frac{t_L - 1}{2} \rceil} (\log n)^{c_0}. \tag{4.19}$$

2. *Regime II: Fix* $\delta > 0$. *Let*

$$c_0 = c_0(t_L, t_U, \eta, \delta) = t_L \cdot (2t_U(t_U + 1) + 2t_U + 1 + t_U\delta + 1) + 2\eta t_U + \frac{2t_L}{t_L - \lceil \frac{t_L - 1}{2} \rceil},$$

and let $c_1 = c_1(t_U, \delta) = \frac{2t_U(t_U+1)+2t_U+1}{t_U} + \delta$. *Let*

$$n^{-\frac{1}{t_U}} (\log n)^{c_1} \leqslant \alpha \leqslant 1, \tag{4.20}$$

and assume that

$$n^{t_L - 1} \rho_n^{t_L} \gg n^{\frac{t_L(t_U - t_L)}{t_U\left(t_L - \lceil \frac{t_L - 1}{2} \rceil\right)} + \frac{t_L}{t_U} - 1} (\log n)^{c_0}. \tag{4.21}$$

3. *Regime III: Let* $\alpha = O\left(\frac{1}{n}\right)$ *and let* $c_0 = c_0(t_L, \eta) = 4 + (t_L + 2)\eta$. *Assume that*

$$n^{t_L - 1} \rho_n^{t_L} \gg (\log n)^{c_0}. \tag{4.22}$$

Then under each of the three regimes,

$$\|M - M_0\|_{\mathrm{F}} = O_{\mathbb{P}}\left(n(\log n)^{-\eta}\right). \tag{4.23}$$

On the other hand, if ρ_n is such that $n^{t_U-1}\rho_n^{t_U} \ll 1$ and $n\rho_n \gg 1$, then given $\epsilon > 0$ there exists a constant $C_\epsilon > 0$ such that

$$\mathbb{P}\left(\|M - M_0\|_{\mathrm{F}} \geqslant C_\epsilon n\right) \geqslant 1 - \epsilon.$$

For the first regime, when α is growing with n, paths with backtracks have a larger contribution as compared to the case of DeepWalk. Therefore, although the communities can be recovered from the M-matrix, we need relatively dense graphs compared to DeepWalk. The second regime, when $\alpha \to 0$, says that the communities can be recovered even when the graphs are relatively sparse compared to the DeepWalk case. In particular if we take $t_L = t_U = t \geqslant 3$, then the result says that we can recover the communities even when close to the regime when $n^{t-1}\rho_n^t \gg 1$. In contrast, when $n^{t-1}\rho_n^t \ll 1$, we have $\|M - M_0\|_{\mathrm{F}} = \Omega(n)$ with probability bounded away from zero. The third regime concerns the case when $\alpha = O\left(\frac{1}{n}\right)$. In this case, again, the communities can be recovered near the regime $n^{t-1}\rho_n^t \gg 1$ when $t_L = t = t_U \geqslant 3$. The results in this third regime are the strongest among the three regimes. The intuitive reason for such good performance of node2vec is that, if α is small, then situations such as Remark 1 either have lower contribution to upper bounds, or do not arise for the biased random-walks in node2vec.

The bounds on the Frobenius norm in Theorem 5 leads to the following theorem about the proportion of misclassified nodes.

Theorem 6. *Fix $\varepsilon > 0$ and let $t_L \geqslant 3$. Suppose ρ_n satisfy the respective conditions for the three regimes as in Theorem 5, and let be a $(1+\varepsilon)$-approximate solution in Algorithm 1. Then*

$$\mathrm{Err}(\bar{\Theta}, \Theta_0) = o_{\mathbb{P}}(n^{-1/2}),$$

i.e., $\bar{\Theta}$ misclassifies at most $o_{\mathbb{P}}(n^{1/2})$ many node-labels.

Proof outline. Before going into the proofs of these results, let us give a brief outline. We will mainly provide a proof for the $t_L = t_U = t$ case and the general case can be reduced to this special case. The main idea for proving the upper bounds in Theorems 3, 5 is to get good estimates on the moments of the total number of paths of length t. We will count these paths for each of the possible community assignments of the intermediate vertices in the path. The goal is to show that the main

contribution on the k-moments come from k disjoint paths. Proving that the contribution due to rest of the paths is small requires a novel combinatorial analysis due to the possibilities of backtracks. We first develop this methodology for DeepWalk (cf. Section 5.1). The path counting estimates allow us to bound $\|M - M_0\|_{\mathrm{F}}$ using Proposition 4. To prove Theorem 4, we use perturbation analysis for the eigenspaces of $\|M - M_0\|_{\mathrm{F}}$. Due to Proposition 5, the eigenvectors of M_0 are such that two of its rows are the same if the nodes are in the same community and two rows are different if they are in different community. Applying the perturbation analysis, the same property remains approximately true for the top eigenvectors of M as well, which allows us to prove the success of Algorithm 1 (cf. Section 5.2). The proof for node2vec uses similar ideas though the different weights for the backtracking parameter of the random walk requires careful analysis (cf. Sections 5.3, 5.4).

CHAPTER 5

Proofs of DeepWalk and node2vec results

In this chapter we prove all the main results in Chapter 4. The proofs are organized as follows. In Section 5.1, we set up the path counting estimates that are crucially used in our proof. We complete analysis for DeepWalk in Section 5.2, and node2vec in Section 5.3 and Section 5.4.

5.1 Path counting for DeepWalk

In this section, we focus on computing the asymptotics for the number of paths having a some specified community assignments for the intermediate vertices. In Section 5.1.1, we bound its k-th moment and we end with a concentration condition in Section 5.1.2.

5.1.1 Bounding moments for paths of different type

Let us first set up some notation. Recall that $g(u)$ denotes the community assignment for vertex u. We say a path $(i_0, i_1, i_2, \ldots, i_t)$ has composition $(b_0, b_1, \ldots, b_t)$ if $g(i_l) = b_l$. Define the collection of path compositions for paths between two nodes i and j as

$$\mathcal{B}_{ij} := \{(b_0, b_1, \ldots, b_t) : b_l \in [K] \text{ for } 0 \leqslant l \leqslant t, b_0 = g(i), b_t = g(j)\}. \tag{5.1}$$

For $b = (b_0, b_1, \ldots, b_t) \in \mathcal{B}_{i,j}$, we define the collection of paths between i and j with composition b in the complete graph K_n as

$$\mathcal{P}_b := \{(i_0, i_1, \ldots, i_t) : i_l \in [n] \text{ for } 0 \leqslant l \leqslant t, i_0 = i, i_t = j, g(i_l) = b_l\}. \tag{5.2}$$

For any path $p = (i_0, i_1, \ldots, i_t) \in \mathcal{P}_b$, we associate the random variable

$$X_p := A_{i_0 i_1} A_{i_1 i_2} \cdots A_{i_{t-1} i_t}, \tag{5.3}$$

and define

$$Y_b := \sum_{p \in \mathcal{P}_b} X_p. \tag{5.4}$$

To each element $b = (b_0, b_1, \ldots, b_t) \in \mathcal{B}_{i,j}$ we associate the term

$$U_{(b_0,b_1,\ldots,b_t)} = U_b := \left(\prod_{i=1}^{t-1} n_{b_i} B_{b_{i-1} b_i} \right) B_{b_{t-1} b_t}. \tag{5.5}$$

We will upper bound the moments of Y_b in terms of U_b. Similarly for the lower bound, we define

$$L_{(b_0,b_1,\ldots,b_t)} = L_b := \left(\prod_{i=1}^{t-1} (n_{b_i} - (k(t-1)+1)) B_{b_{i-1} b_i} \right) B_{b_{t-1} b_t} \tag{5.6}$$

With this setup, we can state the following bounds on $\mathbb{E} Y_b^k$.

Proposition 6. *Let $t_L = t_U = t \geqslant 3$ be given and suppose that (4.15) holds. Then we have*

$$L_b^k \leqslant \mathbb{E} Y_b^k \leqslant U_b^k (1 + o(1)).$$

The idea is to show that the leading term for $\mathbb{E} Y_b^k$ is due to $\mathbb{E}\big(\prod_{\alpha=1}^k X_{p_\alpha} \big)$ of k ordered paths having kt distinct edges between them. The contribution of the rest of the terms are of a smaller order. We summarize the second claim below. For any path $p = (i_0, i_1, i_2, \ldots, i_t) \in \mathcal{P}_b$, let

$$e(p) := \{ \{i_l, i_{l+1}\} : 0 \leqslant l < t \},$$

be the set of edges in the path p. Let

$$E_m := \sum_{(p_1,p_2,\ldots,p_k) : p_i \in \mathcal{P}_b, |\cup_{\alpha \in [k]} e(p_\alpha)| = m} \mathbb{P}(X_{p_1} X_{p_2} \cdots X_{p_k} = 1). \tag{5.7}$$

We will show the following:

Proposition 7. *Under identical conditions as Proposition 6, we have $\sum_{m<kt} E_m = o(U_b^k)$.*

Proof of Proposition 6 using Proposition 7. Note that, we can write

$$\mathbb{E}Y_b^k = \mathbb{E}\Big(\sum_{p\in\mathcal{P}_b} X_p\Big)^k = \sum_{(p_1,p_2,\ldots,p_k):p_\alpha\in\mathcal{P}_b} \mathbb{P}(X_{p_1}X_{p_2}\cdots X_{p_k}=1). \tag{5.8}$$

For the upper bound, Proposition 7 shows that it is enough to bound the summands corresponding to sequences $(p_1, p_2, \ldots, p_k)$ that satisfy $|\cup_{\alpha=1}^k e(p_\alpha)| = kt$, i.e. sequences of paths consisting of kt distinct edges. In this case, $\mathbb{P}(X_{p_1}X_{p_2}\cdots X_{p_k}=1) = \prod_{\alpha=1}^k \mathbb{P}(X_{p_\alpha}=1)$. We note that each of the paths p_r has path type b, and we bound

$$\mathbb{P}(X_{p_\alpha}=1) \leqslant U_b = \Big(\prod_{i=1}^{t-1} n_{g_i} B_{g_{i-1}g_i}\Big) B_{g_{t-1}g_t},$$

and thus the upper bound follows using Proposition 7. For the lower bound, we can simply restrict the summands in (5.8) to the case $|\cup_{\alpha=1}^k e(p_\alpha)| = kt$. We compute

$$\begin{aligned}\mathbb{P}(X_{p_\alpha}=1) &\geqslant \Big(\prod_{i=1}^{t-1}(n_{b_i} - (2+(\alpha-1)(t-1)+(i-1)))B_{b_{i-1}b_i}\Big)B_{b_{t-1}b_t}\\ &\geqslant \Big(\prod_{i=1}^{t-1}(n_{b_i} - (k(t-1)+1))B_{b_{i-1}b_i}\Big)B_{b_{t-1}b_t} = L(b).\end{aligned}$$

Above for the marked vertices in path p_α, the term $(\alpha-1)(t-1)$ is to account for not choosing vertices seen in the first $\alpha-1$ paths, the summand 2 is for the nodes i and j, and $(i-1)$ is for the nodes upto index $(i-1)$ in path p_α. Hence the proof of the lower bound is also complete.

■

The rest of this section is devoted to the proof of Proposition 7. Let us start by setting up some definitions that will be useful for us to count the contributions coming from intersecting paths. All these definitions are demonstrated in Figure 5.1.

Definition 1 (Marked edge and marked vertex)**.** Let $(p_1, p_2, \ldots, p_k)$ be an ordered sequence of k paths in $\mathcal{P}_b$. Fix one of the paths $p_\alpha = (i_0, i_1, i_2, \ldots, i_t)$ and consider the directed edge (i_l, i_{l+1}) appearing at the l-th step. We will call (i_l, i_{l+1}) to be a *marked edge* if the undirected edge $\{i_l, i_{l+1}\}$ is not present in the paths $p_{\alpha'}, 1 \leqslant \alpha' < \alpha$ and also it is not equal to previous edges in the path p_α, i.e., $\{i_l, i_{l+1}\} \neq \{i_{l'}, i_{l'+1}\}$ for $0 \leqslant l' < l$. Intuitively, we call (i_l, i_{l+1}) a marked edge if it is the first

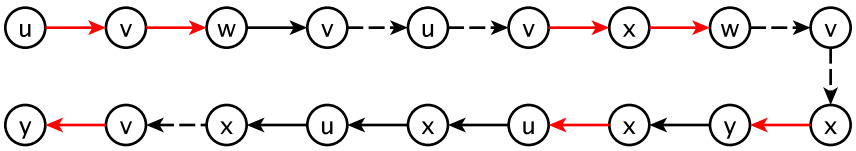

Figure 5.1: Illustrating marked edges (red), backtracks (black), and unmarked edges (dotted). The segments in this path are given by $S_1 = (u, v, w, v)$, $S_2 = (v, x, w)$, $S_3 = (x, y, x, u, x, u, x)$, $S_4 = (v, y)$. Here S_3 is a Type II segment with $k_2 - k_2 = 2$ and $k_3 - k_2 = 4$. The rest are Type I segments.

time we see it as we count the edges along the paths p_1 to p_k. For a marked edge (i_l, i_{l+1}), we will call i_{l+1} to be a *marked vertex* if it was not seen before in previous paths and also in $(i_0, \ldots, i_l)$.

Definition 2 (Backtrack)**.** A directed edge (i_l, i_{l+1}) in a path $p_\alpha = (i_0, i_1, i_2, \ldots, i_t)$ is called a *backtrack* if $i_{l+1} = i_{l-1}$.

Definition 3 (Segment)**.** Let $0 \leqslant l < l' \leqslant t$. We will say that $(i_l, i_{l'})$ is a *segment* in path p_α if the following conditions hold:

1. (i_l, i_{l+1}) is a marked edge, i.e., segments always start with a marked edge.
2. (i_j, i_{j+1}) is a marked edge or a backtrack for all $l \leqslant j < l'$.
3. There does not exist $0 \leqslant l'' < l$ such that (i_j, i_{j+1}) is a marked edge or a backtrack for all $l'' \leqslant j < l$ and $(i_{l''}, i_{l''+1})$ is a marked edge.
4. Either $l' = t$ or if $l' < t$ then $(i_{l'}, i_{l'+1})$ is neither a marked edge nor a backtrack.

Intuitively, segments are maximal parts of paths consisting of marked edges and their backtracks. The last two conditions ensure that segments cannot be extended to the left and to the right. The edges outside the segments will often be referred to as *unmarked* edges. Thus, an unmarked edge is an edge that was previously visited and it is not a backtrack of the last visited marked edge. In Figure 5.1, the dotted lines are unmarked edges, and we can note that the corresponding undirected edges had appeared previously in the path. Notice also that any two segments are separated by one or more unmarked edges. Finally, we remark that the edge preceding a segment may be a backtrack but it can only be a backtrack of an unmarked edge; see for example the second segment in Figure 5.1.

Definition 4 (Type I/II Segments and Paths)**.** Let $(i_l, i_{l'})$ be a segment with r marked edges. Suppose $l = k_1 < k_2 < \cdots k_{r+1} = l'$ be integers such that $(i_{k_{r'}}, i_{k_{r'}+1})$ for $1 \leqslant r' \leqslant r$ constitute the set of all marked edge. Then $(i_l, i_{l'})$ is said to be a *Type II segment* if $k_{r'+1} - k_{r'}$ is an even number for $1 \leqslant r' \leqslant r$. In all other cases, $(i_l, i_{l'})$ is said to be a *Type I segment*. Intuitively, a Type II segment just represents going back and forth on the same vertex. In Figure 5.1, the third segment is Type II and the rest of the segments are Type I. We say that a path p is a *Type I path* if there is at least one Type I segment in it. Otherwise, we call it a *Type II path*. Notice that Type II path may not have any segment.

Definition 5. We call a path p *saturated* if all the edges in p are part of some segment, i.e., there are no unmarked edges in a saturated path.

We now state the following elementary lemma which will be used throughout:

Lemma 6. *There are at most n^{r-1} ways to choose marked vertices for a Type I path with r marked edges, and there are at most n^r ways to do the same for a Type II path with r marked edges.*

Proof. Let us focus only on a path p that is saturated, since we want to maximize the choice of marked vertices and they can only be found by walking through a marked edge. Consider the path p, and note that the endpoints of p are fixed at i, j, and all the edges are either marked or is a backtrack. Let $(i_{k_\alpha}, i_{k_\alpha+1})$ with $1 \leqslant \alpha \leqslant r$ be the marked edges. We construct a graph H_p using only the marked edges (ignoring their orientation). Since p is saturated, we have H_p is connected, and also the vertices of H_p (except possibly i, j) are marked vertices. If p is Type II, then H_p is a star centered at i having r edges, and j can either be i or be one of the leaves. Thus the vertices except i, j can be chosen in at most n^r ways. If p is Type I, in order to get the maximum number of vertices in H_p, one can have H_p to be a tree if $i \neq j$ or a unicyclic graph if $i = j$. In both cases, there are at most $r - 1$ vertices (other than i, j) to choose, and these can be chosen in n^{r-1} ways. ■

To complete the proof of Proposition 7, let $E_{m,r}$ denote the summands in (5.7) restricted to the case that there are r Type I segments, so that

$$E_m = \sum_{r=r_*(m)}^{r^*(m)} E_{m,r},$$

where, given m, $[r_*(m), r^*(m)]$ denotes the possible range of r. The analysis will consist of two steps. In the first step, we analyze $E_{m,r_*(m)}$. Subsequently, we will show that $E_{m,r}$ is much smaller than $E_{m,r_*(m)}$ for $r > r_*(m)$.

Intuitively, $E_{m,r_*(m)}$ is the largest term as minimizing the number of Type I segments leads to maximizing the number of choices for marked vertices. This can be seen from Lemma 6 which says that there are at most $r-1$ marked vertices to choose for a Type I segment with r marked edges. In contrast, for a Type II segment we have an upper bound of r choices for r marked edges.

5.1.1.1 Computing $E_{m,r_*(m)}$.

We first find an expression of $r_*(m)$. Note the following:

Observation 1. A Type I path has maximum number of marked edges if there is only one Type I segment having t marked edges. We refer to these paths *maximal Type I* paths. (Thus, maximal paths are saturated).

Observation 2. For a Type II path, let $h = h(t, i, j)$ be the maximum number of marked edges. Note that each marked edge in a Type II segment has at least one backtrack. We refer to such paths as *maximal Type II* paths. Note that $h = t/2$ if t is even and $i = j$ as each marked edge in a Type II segment has at least one backtracking edge. If t is even and $i \neq j$ then $h = (t-2)/2$ as Type II segments must start and end at the same vertex and Type II segments are of an even length. If t is odd and $i \neq j$, then $h = (t-1)/2$ marked edges. If t is odd and $i = j$ then it may have a maximum of $(t-3)/2$ marked edges. In particular for the last case, we cannot have $(t-1)/2$ marked edges as if that were the case then the first or the last edge in the path would be a self-loop at node i which has probability 0 in our model. We can also note that as a result, we cannot have a Type II path when $t = 3$ and $i = j$ and we take $h = 0$ in the case. Note that $h \leqslant \phi(t) = \lfloor t/2 \rfloor$ for $t \geqslant 3$, where $\phi(t)$ is defined in Theorem 3.

We have that for the case of r Type I segments, the number of marked edges satisfy $m \leqslant rt + (k-r)h$. Let

$$f(r) := rt + (k-r)h, \quad 0 \leqslant r \leqslant k.$$

Then $r_*(m)$ is obtained by inverting f as described below:

Lemma 7. $r_*(m) = \max\{0, \lfloor \frac{m-kh}{t-h} \rfloor\}$.

Proof. Let $f(r_0 - 1) < m \leqslant f(r_0)$ for some $r_0 \geqslant 1$. We would like to put as many edges in Type I segments as possible to minimize the number of Type I segments. However, since each path has length t and $m > (r_0 - 1)t$, we need r_0 Type I segments, so that $r_*(m) \geqslant r_0$. Also, this lower bound is attained by taking $r_0 - 1$ many maximal Type I paths and another Type I path that is not maximal. If $m \leqslant f(0)$, then we can get away with having no Type I segments. This completes the proof. ■

Next we count the possible ways of rearranging the Type I segments in k paths.

Lemma 8. *Given m marked edges and $r_*(m)$ Type I segments, the number of configurations of Type I, II segments and unmarked edges $N_{m,r_*(m)}$ satisfies*

$$N_{m,r_*(m)} \leqslant \binom{k}{r_*(m)} k^{f(r_*(m))-m} 3^{k-r_*(m)} C^{f(r_*(m))-m}.$$

Proof. We treat four cases separately:

Case I: $m = f(r_0)$ for some $1 \leqslant r_0 \leqslant k$. $r_*(m) = r_0$. If $m = f(r_0)$, then we must take r_0 paths to place the $r_0 = r_*(m)$ Type I segments which can be chosen in $\binom{k}{r_0}$ ways. The rest of the Type II segments are placed in the remaining $k - r_0$ paths. The Type I, II paths have to be maximal in order to place these m marked edges. We note that there are at most $t - 2h$ unmarked edges in each of these Type II paths, which are not part of the segments, and these can be chosen in at most m ways each. This is because each unmarked edge is chosen to be one of the m marked edges. We also note that for each Type II path, there are $t - 2h \leqslant 3$ ways of placing the Type II segment in the path. So the overall bound for Case I is

$$\binom{k}{r_*(m)} 3^{(k-r_*(m))}.$$

Case II: $f(r_0-1) < m < f(r_0)$ for some $1 \leqslant r_0 \leqslant k$. Again, we have $r_*(m) = r_0$. Let $l = f(r_0)-m$. Then we need r_0 distinct paths to place the $r_0 = r_*(m)$ Type I segments, which again can be chosen in $\binom{k}{r_0}$. However, in this case, these paths might not be maximal. Take l_1 and l_2 such that $l_1+l_2 \leqslant l$,

all except l_1 of the r_0 Type I paths are maximal and and all except l_2 of the $k - r_0$ Type II paths are maximal. We can choose these $l_1 + l_2$ many non-maximal paths in at most $O(k^l)$ ways. Also, note that we can have at most $2l$ more unmarked edges compared to the case of $m = f(r_0)$. The total number of ways of arranging the segments and unmarked edges for the non-maximal paths is $O(C^l)$, where C is a constant that might depend on t. Thus the bound for this case is

$$\binom{k}{r_*(m)} k^l 3^{(k-r_*(m))} C^l.$$

Case III: $m = f(0)$. The same argument and the bound as Case I holds here. We note that this case does not occur if $h = 0$.

Case IV: $m < f(0)$. Thus $m = f(0) - l$ for $l > 0$. Note that this case also does not occur if $h = 0$. Recall that for $m = f(0)$ case, all the paths have Type II segments containing h marked edges. We would like to count the configurations for $m = f(0) - l$ marked edges with no Type I segments. We note that then there exist $1 \leqslant u \leqslant l$ paths such that they have less than h marked edges and the rest of the paths have h marked edges. We can choose the u paths in at most $O(k^l)$ ways. We can then arrange segments in these l paths in C^l ways. Note that we can have at most $2l$ more unmarked edges for this case compared to the $m = f(0)$ case. The overall bound for this case is $(Ck)^l 3^k$. ■

We are now ready to prove asymptotics of $E_{m,r_*(m)}$.

Lemma 9. $\sum_{m<kt} E_{m,r_*(m)} = o(U_b^k)$.

Proof. Let us start by noting that $U_b = \Theta(n^{t-1}\rho_n^t)$. The number of choices of segments is given by Lemma 8. Since we have $r_*(m)$ Type I segments, we must have at least $s(m) = \max\{r_*(m) - (f(r_*(m)) - m), 0\}$ maximal Type I segments, with each of them having probability at most U_b. In the rest, we have one Type I segment. Thus by Lemma 6, the vertices in the rest of the paths can be chosen in at most $n^{m-ts(m)-(r_*(m)-s(m))}$ ways. Also the $m-ts(m)$ many remaining marked edges give us a contribution of at most $(C\rho_n)^{m-ts(m)}$. The number of unmarked edges is

specified in each of the cases in the proof of Lemma 8 above. Combining all these, we get

$$\begin{aligned} E_{m,r_*(m)} \leqslant & \binom{k}{r_*(m)} k^{f(r_*(m))-m} 3^{k-r_*(m)} C^{f(r_*(m))-m} \\ & \times U_b^{s(m)} n^{m-ts(m)-(r_*(m)-s(m))} \rho_n^{m-ts(m)} \\ & \times m^{(k-r_*(m))\cdot(t-2h)+2(f(r_*(m))-m)}. \end{aligned}$$

Using these bounds we see that with the choice of $k = \lceil \log n \rceil$

$$\begin{aligned} \sum_{l=0}^{t-h-1} E_{f(r_0)-l,r_\star(f(r_0)-l)} &= E_{f(r_0)-l,r_\star(f(r_0)-l)} \left(1 + O\left(\frac{k^3}{n\rho_n}\right)\right), \quad r_0 \geqslant 1, \\ \sum_{l=0}^{f(0)-1} E_{f(0)-l,r_\star(f(0)-l)} &= E_{f(0)-l,r_\star(f(0)-l)} \left(1 + O\left(\frac{k^3}{n\rho_n}\right)\right), \\ \sum_{r_0=0}^{k} E_{f(r_0),r_\star(f(r_0))} &= E_{f(k),r_\star(f(k))} \left(1 + O\left(\frac{k^{1+t-2h}(n\rho_n)^h}{n^{t-1}\rho_n^t}\right)\right). \end{aligned}$$

These bounds in turn imply that

$$\sum_{m=1}^{kt-1} E_{m,r_*(m)} = o\left(U_b^k\right).$$

■

5.1.1.2 Computing $E_{m,r}$ for $r > r_*(m)$.

We start by noting the additional number of configurations with r Type I segments as compared to Lemma 6.

Lemma 10. *Given m marked edges and r Type I segments, let $N_{m,r}$ be the number of configurations of Type I, II segments and unmarked edges. Then, for any $r > r_*(m)$,*

$$N_{m,r} \leqslant N_{m,r_*(m)} \times (Ck)^{(r-r_*(m))(t+1)}.$$

Proof. Let T_r be the set of all configurations of m marked edges and r Type I segments. Note that a "configuration" here only specifies location of marked edges, segments, and the unmarked edges. The backtracks of the marked edges are already specified. We will inductively bound T_{r+1} in terms

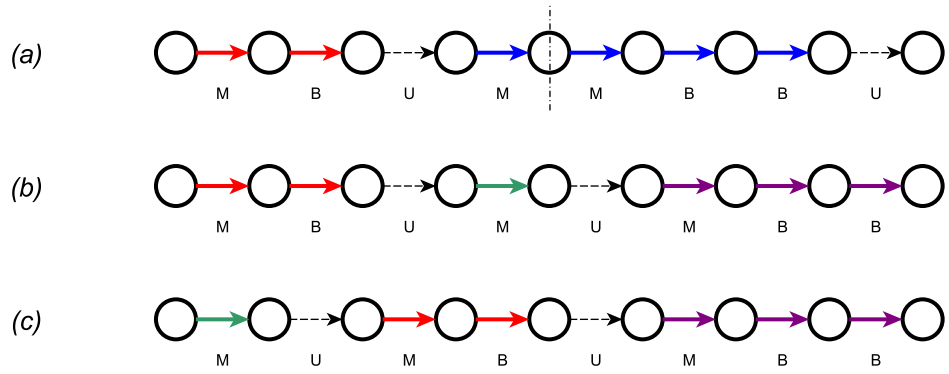

Figure 5.2: Example of a splitting a segment. (a) The original configuration of the segments in the chosen path. The edges in the two segments are colored red and blue respectively. The edges are also labeled by M, B and U if the edge is a marked edge, a backtrack of a marked edge and an unmarked edge respectively. The dotted line indicates the location of the split. $(b), (c)$ Two examples of configurations after the splitting the segment. The two new segments are colored by green and violet.

T_r. For that, we consider two cases depending on whether the elements of T_{r+1} has a Type I path with two Type I segments or not. In both cases, we will find a relation between T_r to T_{r+1}.

Case I. Consider elements of T_{r+1} that has a Type I path with two Type I segments. Let us call this subset T^I_{r+1}. We consider the elements in T_r which will be related to these elements in T^I_{r+1}. Let $T^I_r \subset T_r$ consisting of configuration such that there is at least one path p so that the following hold:

1. **Extra unmarked edges.** p has l segments and $l' \geqslant l$ unmarked edges for some $l, l' \geqslant 1$.

2. **Well-splittable.** Suppose p has a Type I segment $(i_l, i_{l+1}, \ldots, i_{l'})$ with marked edges given by (i_{k_s}, i_{k_s+1}) for $l \leqslant k_s < l'$, $1 \leqslant s \leqslant m'$, $m' \geqslant 1$, and the number of s with $k_{s+1} - k_s$ being odd is at least 2.

Note that a path with l segments can be formed by just putting $l-1$ unmarked edges between segments. The first condition ensures that we have extra $l' - l + 1$ of them. We will put these extra unmarked edges inside segments to split them. Regarding condition 2, notice that a Type I segment always has one of the $k_{s+1} - k_s$ being odd (by definition), but the well-splittable condition requires $k_{s+1} - k_s$ to be odd additionally in a second place. This allows us to split a well-splittable Type I segment into two Type I segments. For example, the blue segment in Figure 5.3 (a) is well-splittable and it can be split into two parts with $k_{s+1} - k_s$ being odd. We will split the segment by moving an unmarked edge in between as illustrated in Figure 5.3.

The general description for splitting is that given a path with extra marked edges and well-splittable property, we can think of unmarked edges as "bars", and segments as "labelled balls". The well-splittable segment is viewed as two Type I sub-segments and corresponds to two "labelled balls". Permute these bars and balls such that there is at least one bar between any two labelled balls (bars can be adjacent). Also, there may be multiple ways to split well-splittable segment and we consider all possible such splits. This creates multiple elements in T^I_{r+1} from an element in T^I_r. See Figure 5.2 (b), (c) for two possible elements. Moreover, we can get preimages of all the elements in T^I_{r+1},. To do this, we can first take a path with two Type I segments, rearranging the segments so that two Type I segments appear one after another. Then we can remove the unmarked edges between them, which joins the two Type I segments. The removed unmarked edges can be placed in other places adjacent to an unmarked edge. This is basically the inverse operation of the above splitting constructions.

We can note there are $O(k)$ preimages in total of an element in T^I_{r+1}, where the factor k comes from choice of the path p and the rest of the choices for permuting unmarked edges and segments are $O(1)$ as they are functions of only t which is bounded. Thus,

$$|T^I_{r+1}| \leqslant O(k) \times |T^I_r| \leqslant O(k) \times |T_r|.$$

Case II. We next relate arrangements of marked edges where there is at most one Type I segment per path. We denote such arrangements as $T^{II}_{r+1} \subset T_{r+1}$. Let T^{II}_r be the set of all ways of specifying locations of segments such that there are a total of r Type I segments and each of the Type I segments is placed on distinct paths. Suppose that $r+1 \leqslant \min(k, m)$ and $r \geqslant r_\star(m)$. Then we give a multi-map from T^{II}_r onto T^{II}_{r+1} using a construction. The condition $r+1 \leqslant m$ is necessary so that T^{II}_{r+1} is non-empty as we must have at least $r+1$ marked edges in order to have $r+1$ paths each having a Type I segment. The condition $r+1 \leqslant k$ is also necessary for T^{II}_{r+1} to be non-empty as we must have at least $r+1$ distinct paths to place the $r+1$ Type I segments. The last condition $r \geqslant r_\star(m)$ is to ensure that T^{II}_r is non-empty. To fix notation, let $S_1(l)$ be the set of all ways of arranging l marked edges in one path such that there is at least one Type I segment in the path. Similarly, let $S_2(l)$ be the set of all ways of arranging l marked edges in one path such that there are no Type I segments in the path. We now describe the construction. Let $A \in T^{II}_r$

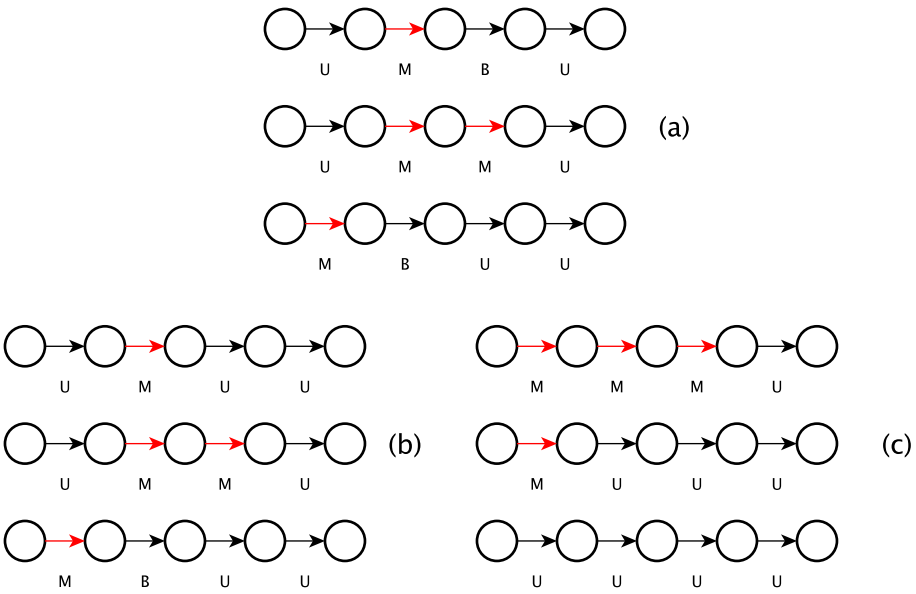

Figure 5.3: Example of the second construction for the case of $t = 4$, $m = 4$ and $s = 1$. The marked edges are colored red. The letters M, B and U denote a marked edge, a backtrack of a marked edge and an unmarked edge respectively. (a) The top path is the chosen path where we would place a Type I segment. The second path contains one Type I segment and the bottom path is a Type II path. $(b), (c)$ Two examples of new configurations from the construction. In both the cases the top path now has a Type I segment and the middle path continues to have a Type I segment.

and choose a path p not containing a Type I segment. We can do so as $r + 1 \leqslant k$. Suppose p has $l \geqslant 0$ marked edges. Choose $0 \leqslant u_1 + u_2 \leqslant t - l$ paths where u_1 of the paths are paths containing Type I segments with at least two marked edges and the rest u_2 paths are paths which are distinct from p, do not contain Type I segments, and contain at least one marked edge. If $l = 0$ we require $u_1 + u_2 > 0$. It is feasible to choose such path(s) as $r + 1 \leqslant m$. Suppose the u_1 paths are labeled as $q_1, q_2, \ldots, q_{u_1}$ and the u_2 paths indices $q'_1, q'_2, \ldots, q'_{u_2}$. Suppose these paths have $l_{q_1}, l_{q_2}, \cdots, l_{q_{u_1}}$ and $l_{q'_1}, l_{q'_2}, \cdots, l_{q'_{u_2}}$ marked edges respectively. Let $v_{q_1}, v_{q_2}, \cdots, v_{q_{u_1}}$ be such that $0 < v_{q_i} < l_{q_i}$. Similarly let $v_{q'_1}, v_{q'_2}, \cdots, v_{q'_{u_2}}$ be such that $0 < v_{q'_i} \leqslant l_{q'_i}$. We require $\sum_i v_{q_i} + \sum_i v_{q'_i} \leqslant t - l$. Then we modify the arrangements of marked edges in the paths so that the new arrangements for the sequence of paths $(p, q_1, q_2, \ldots, q_{u_1}, q'_1, q'_2, \ldots, q'_{u_2})$ is any element of

$$S_1\left(l + \sum_i v_{q_i} + \sum_i v_{q'_i}\right) \times \left(\prod_{i=1}^{u_1} S_1(l_{q_i} - v_{q_i})\right) \times \left(\prod_{i=1}^{u_2} S_2(l_{q'_i} - v_{q'_i})\right).$$

We keep the arrangements of marked edges in the rest $k-(1+u_1+u_2)$ paths unchanged. This leads to multiple images of A in T^{II}_{r+1}. We note that there are $O(k^{t+1})$ images of A due to the choice of the paths and since the number of ways of choosing the new arrangements for the $1+u_1+u_2$ paths is $O(1)$ as t is fixed. We also note that since we modify at most $t+1$ paths in this construction, the number of unmarked edges increases by at most $t(t+1)$.

Now we show that the multi-map given by the construction above is surjective onto T^{II}_{r+1}. For this, let $A' \in T^{II}_{r+1}$. Let p be a path containing a Type I segment. If p has $l \leqslant h$ edges we choose an element in $x \in S_2(l)$. We can then see that A' is in the image of the element $A \in T^{II}_r$ where the path p has the arrangement of marked edges given by x and the rest of the paths have the same arrangement of marked edges as A'. We note that since we have replaced the Type I segment in the path p with a Type II segment, the element A has one less Type I segment. Now suppose that p has $l > h$ edges. Choose $u_1 + u_2 \leqslant l - h$ paths so that u_1 paths contain Type I segments and these paths are not equal to p and, u_2 paths do not contain Type I segments segments. Suppose the u_1 paths are labeled as $q_1, q_2, \ldots, q_{u_1}$ and the u_2 paths are labeled as $q'_1, q'_2, \ldots, q'_{u_2}$. Suppose that these paths have $l_{q_1}, l_{q_2}, \cdots, l_{q_{u_1}}$ and $l_{q'_1}, l_{q'_2}, \cdots, l_{q'_{u_2}}$ marked edges respectively. We require that $l_{q_i} < t$ and $l_{q'_i} < h$ i.e. that these chosen paths are not saturated. Let $v_{q_1}, v_{q_2}, \cdots, v_{q_{u_1}}$ be such that $0 < v_{q_i} \leqslant t - l_{q_i}$. Similarly let $v_{q'_1}, v_{q'_2}, \cdots, v_{q'_{u_2}}$ be such that $0 < v_{q'_i} \leqslant h - l_{q'_i}$. We require that

$$\sum_{i=1}^{u_1} v_{q_i} + \sum_{i=1}^{u_2} v_{q'_i} = l - h.$$

This is feasible as long as $r \geqslant r_\star(m)$. The above condition says that the chosen paths have enough spaces to move $l - h$ edges from path p to the chosen paths. Choose a new arrangement x of the marked edges for the sequence of paths $(p, q_1, q_2, \ldots, q_{u_1}, q'_1, q'_2, \ldots, q'_{u_2})$ from the set

$$S_2(h) \times \left(\prod_{i=1}^{u_1} S_1(l_{q_i} + v_{q_i}) \right) \times \left(\prod_{i=1}^{u_2} S_2(l_{q'_i} + v_{q'_i}) \right).$$

Then keeping the arrangements of marked edges of the rest of the paths the same as in A' and choosing the arrangements for the chosen paths as x, we have a preimage $A \in T^{II}_r$ under the construction described above.

From the two constructions above we have

$$|T_{r+1}| \leqslant O(k)|T_r| + O(k^{t+1})|T_r|. \tag{5.9}$$

■

We can now compute asymptotics for $E_{m,r}$.

Lemma 11. $\sum_{r=r_*(m)}^{r^*(m)} E_{m,r} = E_{m,r_*(m)}(1+o(1))$.

Proof. We start by giving a bound for $E_{m,r}$. The probability of the m marked edges is bounded by ρ_n^m. By Lemma 6, the upper bound for the number of marked edges is $m-r$ as there are r Type I segments. Let

$$u(m) = (k - r_*(m)) \cdot (t-2h) + 2(f(r_*(m)) - m),$$

be the upper bound for the number of unmarked edges for the case of $r_*(m)$ segments obtained from the proof of Lemma 9. Then by the two constructions in Lemma 10, the number of unmarked edges for the case of r Type I segments is at most $u(m) + (r - r_*(m)) \cdot (t(t+1))$. Combining these we have

$$E_{m,r} \leqslant N_{m,r} n^{m-r} \rho_n^m m^{u(m)+(r-r_*(m))\cdot t(t+1)}.$$

Then by using Lemma 10 we have

$$\sum_{r=r_*(m)}^{r^*(m)} E_{m,r} = E_{m,r_*(m)} \sum_{r=r_*(m)}^{r^*(m)} \left(\frac{O(k^{t+1}) \cdot m^{t(t+1)}}{n} \right)^{r-r_*(m)} = E_{m,r_*(m)}(1+o(1)).$$

■

By Lemmas 9 and 11 we then have

$$\sum_{m<kt} E_m \leqslant C \sum_{m<kt} E_{m,r_*(m)} = o(U_b^k).$$

This completes the proof of Proposition 7.

5.1.2 Concentration of path counts

In this section, we prove concentration of Y_b defined in (5.4). Let us start with a general result which we will apply for indicator random variables appearing in paths:

Lemma 12. *Let $\mathcal{E}$ be a finite set and let ξ_e be indicator random variables for $e \in \mathcal{E}$ with $\pi_e = \mathbb{P}(\xi_e = 1) > 0$. For a subset $S \subset \mathcal{E}$, we define $\xi_S := \prod_{e\in S} \xi_e$. Suppose $S_1, S_2, \dots S_k$ be non-empty subsets of $\mathcal{E}$ and let $n_e := |\{j \in [k] : e \in S_j\}|$. Let $\mathcal{E}_1 := \{e \in \cup_{j=1}^k S_j : n_e = 1\}$ and $\mathcal{E}_2 = \left(\cup_{j=1}^k S_j\right) \setminus \mathcal{E}_1$. Then we have*

$$\left|\mathbb{E}\prod_{j=1}^{k}(\xi_{S_j} - \mathbb{E}\xi_{S_j})\right| \leqslant \Big(\prod_{e\in\mathcal{E}_1} \pi_e\Big) \times \Big(\prod_{e\in\mathcal{E}_2} \pi_e(1+\pi_e)\Big). \tag{5.10}$$

Proof. Let $S_j^{(1)} \subseteq S_j^{(2)} \subseteq S_j$ and $m = |\cup_{j=1}^k S_j^{(1)}|$. We define $n_e^{(1)}$ and $\mathcal{E}_1^{(1)}$ similarly as n_e, $\mathcal{E}_1$ but now with sets $S_j^{(1)}$. We prove by induction on m that, for all possible choices of $S_j^{(1)}, S_j^{(2)}$,

$$\left|\mathbb{E}\prod_{j=1}^{k}\big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big)\right| \leqslant \Big(\prod_{e\in\mathcal{E}_1^{(1)}} \pi_e\Big) \times \Big(\prod_{e\in\mathcal{E}_2^{(1)}} \pi_e(1+\pi_e)\Big). \tag{5.11}$$

Throughout, we use the convention that a product over an empty index set is 1. For the induction base case, suppose $m = 1$. Let e be the only element in $\cup_{i=1}^k S_j^{(1)}$ and without loss of generality let $e \in S_j^{(1)}$ for $1 \leqslant j \leqslant n_e^{(1)}$. If $n_e^{(1)} = 1$, then

$$\mathbb{E}\big(\xi_{S_1^{(1)}} - \mathbb{E}\xi_{S_1^{(2)}}\big) = \pi_e\big(1 - \mathbb{E}\xi'_{S_1^{(2)}}\big) \leqslant \pi_e,$$

where $\xi'_{S_1^{(2)}} := \xi_{S_1^{(2)}\setminus\{e\}}$. The terms with $S_j^{(1)} = \varnothing$ can be bounded by 1. This shows the base case for $m = 1$ and $n_e^{(1)} = 1$. If $n_e^{(1)} > 1$, then

$$\begin{aligned}
&\mathbb{E}\Big[\prod_{j=1}^{n_e^{(1)}}\big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big)\Big] = \mathbb{E}\Big[\prod_{j=1}^{n_e^{(1)}}\big(\xi_e - \mathbb{E}\xi_{S_j^{(2)}}\big)\Big] \\
&= \pi_e \cdot \prod_{j=1}^{n_e^{(1)}}\big(1 - \mathbb{E}\xi_{S_j^{(2)}}\big) + (1-\pi_e)\cdot\prod_{j=1}^{n_e^{(1)}}\big(-\mathbb{E}\xi_{S_j^{(2)}}\big) \leqslant \pi_e + (\pi_e)^{n_e^{(1)}} \leqslant \pi_e(1+\pi_e).
\end{aligned}$$

This completes the proof of the base case $m = 1$.

Next let $m > 1$. We will split in two cases depending on whether $\mathcal{E}_1^{(1)} \neq \varnothing$ and $\mathcal{E}_1^{(1)} = \varnothing$. First, if $\mathcal{E}_1^{(1)} \neq \varnothing$, then pick an element $e \in \mathcal{E}_1^{(1)}$. Note that e can only be in one of $S_j^{(1)}$'s, and without loss of generality let $e \in S_1^{(1)}$. Let $\mathscr{F}_e$ is the minimum sigma-algebra with respect to which $(\xi_f)_{f \neq e}$ are measurable, and analogously to before, define $\xi'_{S_1^{(1)}} := \xi_{S_1^{(1)} \setminus \{e\}}$ and $\xi'_{S_1^{(2)}} := \xi_{S_1^{(2)} \setminus \{e\}}$. Taking iterated conditional expectation with respect to $\mathcal{F}_e$, we get

$$\mathbb{E}\Big[\prod_{j \in [k] : S_j^{(2)} \neq \varnothing} \big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big) \Big] = \pi_e \cdot \mathbb{E}\Big[\big(\xi'_{S_1^{(1)}} - \mathbb{E}\xi'_{S_1^{(2)}}\big) \prod_{j \in \{2,3,\dots k\} : S_j^{(2)} \neq \varnothing} \big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big) \Big].$$

Now we can conclude (5.11) by induction step using $S_1^{(1)} \setminus \{e\}$, $S_1^{(2)} \setminus \{e\}$ and $S_j^{(1)}$, $S_j^{(2)}$ for $j \geqslant 2$, and noting that $\mathcal{E}_1^{(2)}$ is unchanged in this new set up.

Next suppose that $\mathcal{E}_1^{(1)} = \varnothing$. Pick any $e \in \mathcal{E}_2^{(1)}$. Then $n_e^{(1)} > 1$ and without loss of generality let $e \in S_j^{(1)}$ for $1 \leqslant j \leqslant n_e^{(1)}$. We again take iterated conditional expectation with respect to $\mathcal{F}_e$ to get

$$\begin{aligned} &\mathbb{E}\Big[\prod_{j \in [k] : S_j^{(2)} \neq \varnothing} \big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big) \Big] \\ &= \mathbb{E}\Big[\prod_{j \in \{n_e^{(1)}+1,\dots,k\} : S_j^{(2)} \neq \varnothing} \big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big) \mathbb{E}\Big[\prod_{j=1}^{n_e^{(1)}} \big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big) \,\Big|\, \mathcal{F}_e \Big] \Big]. \end{aligned} \tag{5.12}$$

We simplify

$$\mathbb{E}\Big[\prod_{j=1}^{n_e^{(1)}} \big(\xi_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big) \,\Big|\, \mathcal{F}_e \Big] = \pi_e \cdot \prod_{j=1}^{n_e^{(1)}} \big(\xi'_{S_j^{(1)}} - \mathbb{E}\xi_{S_j^{(2)}}\big) + (1 - \pi_e) \prod_{j=1}^{n_e^{(1)}} \big(- \mathbb{E}\xi_{S_j^{(2)}}\big). \tag{5.13}$$

Suppose that $m - l = |\cup_{j=n_e^{(1)}+1}^{k} S_j^{(2)}|$ for some $l \geqslant 1$ (we have $l \geqslant 1$ since e is not in the union). For the second term in (5.13), we have

$$\Big| (1 - \pi_e) \prod_{j=1}^{n_e^{(1)}} (-\mathbb{E}\xi_{S_j^{(2)}}) \Big| \leqslant \pi_e^{n_e^{(1)} - 1} \prod_{f \in \cup_{j=1}^{n_e^{(1)}} S_j^{(2)}} \pi_f =: Z_1,$$

as the term ξ_e is repeated r times. We bound the term in (5.12) outside conditional expectation by induction. If we consider the sets $S_j^{(1)}$, $S_j^{(2)}$ for $j > n_e^{(1)}$ with $S_j^{(2)} \neq \varnothing$ and create the sets $\tilde{\mathcal{E}}_1$ and

$\tilde{\mathcal{E}}_2'$ analogously to $\mathcal{E}_1^{(1)}$, $\mathcal{E}_1^{(2)}$ when restricted to this smaller class of subsets. Then

$$\left|\mathbb{E}\Big[\prod_{j\in\{n_e^{(1)}+1,\dots,k\}:S_j^{(2)}\neq\varnothing}\left(\xi_{S_j^{(1)}}-\mathbb{E}\xi_{S_j^{(2)}}\right)\Big]\right|\leqslant\prod_{f\in\tilde{\mathcal{E}}_1}\pi_f\prod_{f\in\tilde{\mathcal{E}}_2}\pi_f(1+\pi_f)=:Z_2.$$

Hence, the term in (5.12) is at most

$$\begin{aligned}&\pi_e\left|\mathbb{E}\Big[\prod_{j\in\{n_e^{(1)}+1,\dots,k\}:S_j^{(2)}\neq\varnothing}\left(\xi_{S_j^{(1)}}-\mathbb{E}\xi_{S_j^{(2)}}\right)\prod_{j=1}^{n_e^{(1)}}\left(\xi'_{S_j^{(1)}}-\mathbb{E}\xi_{S_j^{(2)}}\right)\Big]\right|+Z_1Z_2\\&\leqslant\pi_e\cdot\left(\prod_{f\in\mathcal{E}_1^{(1)}}\pi_f\cdot\prod_{f\in\mathcal{E}_2^{(1)}\setminus\{e\}}\pi_e(1+\pi_e)\right)+\pi_e^{n_e^{(1)}-1}\prod_{f\in\cup_{j=1}^{n_e^{(1)}}S_j^{(2)}}\pi_f\prod_{f\in\tilde{\mathcal{E}}_1}\pi_f\prod_{f\in\tilde{\mathcal{E}}_2'}\pi_f(1+\pi_f).\end{aligned}$$

The last term above is at most the bound in (5.11), which follows by noting that the second term has a factor of $(\pi_e)^{n_e^{(1)}}$ with $n_e^{(1)}\geqslant 2$, and for each of the variables $f\in\mathcal{E}_1^{(1)}, f\neq e$ the second term has a factor smaller or equal to π_f, and for each of the variables $f\in\mathcal{E}_1^{(1)}$ the second term has a factor smaller or equal to $\pi_f(1+\pi_f)$. This completes the proof. ■

We now prove the following concentration result for Y_b:

Proposition 8. *Let $t_L=t_U=t\geqslant 3$ be given and $k=\lceil\log n\rceil$. Suppose that (4.15) holds. Then we have*

$$\mathbb{P}\left(|Y_b-\mathbb{E}Y_b|>\delta\mathbb{E}Y_b\right)=O(n^{-c}),$$

where $\delta=\Theta((\log n)^{-\eta})$ for some $\eta>0$, and $c>0$ is any real number.

Proof. We will be using notation and terminology from proof of Proposition 7. By Markov's inequality, and using Proposition 6,

$$\mathbb{P}\left(|Y_b-\mathbb{E}Y_b|>\delta\mathbb{E}Y_b\right)\leqslant\frac{\mathbb{E}(Y_b-\mathbb{E}Y_b)^{2k}}{\delta^{2k}\left(\mathbb{E}Y_b\right)^{2k}}\leqslant\frac{\mathbb{E}(Y_b-\mathbb{E}Y_b)^{2k}}{\delta^{2k}(L_b)^{2k}},\tag{5.14}$$

and moreover,

$$\mathbb{E}(Y_b - \mathbb{E}Y_b)^{2k} = \mathbb{E}\Big(\sum_{p\in\mathcal{P}_b}(X_p - \mathbb{E}X_p)\Big)^{2k} = \sum_{(p_1,p_2,\ldots,p_{2k}):p_l\in\mathcal{P}_b}\mathbb{E}\prod_{l=1}^{2k}(X_{p_l} - \mathbb{E}X_{p_l}). \tag{5.15}$$

Fix an ordered sequence $(p_1, p_2, \ldots, p_{2k})$. We can first note that $\mathbb{E}\prod_{l=1}^{2k}(X_{p_l} - \mathbb{E}X_{p_l})$ is equal to 0 if there is a path p_l which does not have any edges in common with the other $2k-1$ paths. Suppose now that each of the paths share at least one edge with some other path. Then the minimum number of unmarked edges or repeats of edges is at least k. This minimum k arises from the worst case where we have k pairs of paths with each pair having one edge in common. Thus, the number of marked edges $m \leqslant 2kt - k$.

To bound $\mathbb{E}\prod_{l=1}^{2k}(X_{p_l} - \mathbb{E}X_{p_l})$, we use Lemma 12. For each marked edge e with endpoints in block b_i and $b_{i'}$ between two nodes blocks b_i and $b_{i'}$, we assign a weight $w(e) = B_{b_i,b_{i'}}$. For each unmarked edge e, we assign a weight $w(e) = (1 + B_{b_i,b_{i'}})$ instead. Then we can see by Lemma 12, $\mathbb{E}\prod_{l=1}^{2k}(X_{p_l} - \mathbb{E}X_{p_l})$ is bounded by the product of the weights on the edges in the $2k$ paths, i.e.,

$$\mathbb{E}\prod_{l=1}^{2k}(X_{p_l} - \mathbb{E}X_{p_l}) \leqslant \prod_{e:e \text{ is marked}} w(e) \times \prod_{e:e \text{ is unmarked}} (1 + w(e)). \tag{5.16}$$

We note that while bounding E_m in the proof of Proposition 7, we bounded the probability of each of the marked edges by $w(e)$ as well. Let E'_m be the sum of summands in (5.15) corresponding to paths with m marked edges. We use the bound $1 + w(e) \leqslant 2$ for the weight on the unmarked edges and proceed as in the proof of Proposition 7 to have for $m \leqslant 2kt - k$:

$$\begin{aligned} E'_m \leqslant C'\binom{2k}{r_*(m)}(2k)^{f(r_*(m))-m}3^{2k-r_*(m)}C^{f(r_*(m))-m} \\ \times U_b^{s(m)}n^{m-ts(m)-(r_*(m)-s(m))}\rho_n^{m-ts(m)} \\ \times (2m)^{(2k-r_*(m))\cdot(t-2h)+2(f(r_*(m))-m)}. \end{aligned}$$

Let m_0 be such that $m_0 = f(r_*(m))$ and define E'_{m_0} with the same expression as above. Then bounding as in the proof of Proposition 7 we have from (5.15)

$$\mathbb{E}(Y_b - \mathbb{E}Y_b)^{2k} \leqslant CE'_{m_0}.$$

Thus, (5.14) together with the fact that $\frac{U_b}{L_b} = 1 + O(\frac{k}{n})$ yields

$$\begin{aligned}
\mathbb{P}\left(|Y_b - \mathbb{E}Y_b| > \delta \mathbb{E}Y_b\right) &\leqslant \frac{\binom{2k}{r_*(m)} 3^{2k-r_*(m)} U_b^{r_*(m)} (n\rho_n)^{m-tr_*(m)} (2m)^{(2k-r_*(m))\cdot(t-2h)}}{\delta^{2k} L_b^{2k}} \\
&\leqslant \left(\frac{U_b}{L_b}\right)^{r_*(m)} \cdot \left(\frac{Ck^{1+t-2h}(n\rho_n)^h \delta^{-\frac{2k}{2k-r_*(m)}}}{n^{t-1}\rho_n^t}\right)^{2k-r_*(m)}, \\
&\leqslant C\left(\frac{C(\log n)^{1+t-2h-2\eta(t-h)}(n\rho_n)^h}{n^{t-1}\rho_n^t}\right)^{\lceil \log n \rceil (t-h)^{-1}} \leqslant n^{-c},
\end{aligned}$$

for any $c > 0$ when (4.15) holds. This completes the proof of Proposition 8.

■

5.2 Analysis of spectral clustering for DeepWalk

In this section, we first prove properties of the M_0 matrix in Section 5.2.1. In Section 5.2.2 we bound $\|M - M_0\|_{\mathrm{F}}$ and complete the proof of Theorem 3. Finally, we bound the number of misclassified nodes in Section 5.2.3 and complete the proof of Theorem 4.

5.2.1 Analysis of noiseless M-matrix

Recall the definition of the matrix M_0 from Section 4.3.3. Also recall that Θ_0 is the matrix of true community assignments. Let (λ_i, v_i), $1 \leqslant i \leqslant K$ be the K largest eigenvalue-eigenvector pairs of M_0, and let $V_0 = (v_1, \ldots, v_K)$. We will prove the following collection of claims for M_0 and its eigenspace:

Proposition 9. *(a) There exists a full-rank matrix $Z_0 \in \mathbb{R}^{K\times K}$ such that $M_0 = \Theta_0 Z_0 \Theta_0^T$. Moreover, $(M_0)_{ij} = \Theta(1)$ uniformly in $i, j \in [n]$.*

(b) We have $\lambda_i = \Theta(n)$ and there exists a full rank matrix $X_0 \in \mathbb{R}^{K\times K}$ such that $V_0 = \Theta_0 X_0$.

(c) If i and j are two nodes such that $g(i) \neq g(j)$, then we have

$$\|(V_0)_{i\star} - (V_0)_{j\star}\|_{\mathrm{F}} = \sqrt{\frac{1}{n_{g(i)}} + \frac{1}{n_{g(j)}}}. \tag{5.17}$$

Proof. Part a. Let $P^{(t)}$ be the transition matrix for a simple random walk on a complete graph with edge-weights P. Also, recall from (4.7) that the random walks are initialized with distribution $(|P_{i\star}|/|P|)_{i\in[n]}$. By Proposition 4, we have

$$(M_0)_{ij} = \log\left(\frac{\sum_{t=t_L}^{t_U}(l-t)\cdot(P_{ij}^{(t)}+P_{ji}^{(t)})}{2b\gamma(l,t_L,t_U)\frac{|P_{i\star}|}{|P|}\times\frac{|P_{j\star}|}{|P|}}\right).$$

So in order to show that $M_0 = \Theta_0 Z_0 \Theta_0^T$, it is enough to show that $(P^{(t)} + (P^{(t)})^T) = \Theta_0 Z \Theta_0^T$ for some matrix Z. Towards this we have

$$(P^{(t)})_{ij} = \frac{1}{|P|}\sum_{(i_0,i_1,\ldots,i_t):i_0=i,i_t=j} P_{i_0i_1}P_{i_1i_2}\cdots P_{i_{t-1}i_t}\left(\prod_{l=1}^{t-1}|P_{i_l\star}|^{-1}\right). \tag{5.18}$$

Similarly, we can compute $((P^{(t)})^T)_{ij}$, and it is clear from these expressions that $(P^{(t)} + (P^{(t)})^T)_{ij}$ only depends on i and j through their block labels $g(i)$ and $g(j)$. This shows that $(P^{(t)} + (P^{(t)})^T) = \Theta_0 Z \Theta_0^T$ for some appropriate matrix Z, and hence $M_0 = \Theta_0 Z_0 \Theta_0^T$. We now show that Z_0 has full rank. Since B has rank K, $P^{(t)}$ has rank K and has the same block structure as B. We note that $\operatorname{diag}\left(|P_{1\star}|^{-1}, |P_{1\star}|^{-1}, \ldots, |P_{n\star}|^{-1}\right)$ is an invertible matrix. This implies that $\left(\frac{\sum_{t=t_L}^{t_U}(l-t)\cdot(P_{ij}^{(t)}+P_{ji}^{(t)})}{2b\gamma(l,t_L,t_U)\frac{|P_{i\star}|}{|P|}\times\frac{|P_{j\star}|}{|P|}}\right)$ has rank K and has the same block structure as B. Then by Lemma 5, Z_0 has rank K a.s.

Next, we estimate the order of the coefficients $(M_0)_{ij}$. Note that

$$\sum_{(i_0,i_1,\ldots,i_t)|i_0=i,i_t=j} P_{i_0i_1}P_{i_1i_2}\cdots P_{i_{t-1}i_t} = \Theta(n^{t-1}\rho_n). \tag{5.19}$$

Indeed, the leading contribution is due to paths with t distinct edges and $t-1$ choices of intermediate vertices. If we have less distinct vertices among $(i_1,\ldots,i_{t-1})$, then the number of choices is at most $O(n^{t-2})$. This proves (5.19). Also, $|P| = \Theta(n^2\rho_n)$ and $|P_{i_l\star}| = \Theta(n\rho_n)$. Combining these orders implies that $(M_0)_{ij} = \Theta(1)$.

Part b. Note that

$$\lambda_i v_i = M_0 v_i = \Theta_0 Z_0 \Theta_0^T v_i \quad \Longrightarrow \quad v_i = \Theta_0\left(\lambda_i^{-1} Z_0 \Theta_0^T v_i\right).$$

Taking $X_0 = Z_0\Theta_0^T V_0\Lambda^{-1}$ with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$, we get $V_0 = \Theta_0 X_0$. Since Z_0 is full rank, $\operatorname{rank}(Z_0\Theta_0^T) = K$. Since $\operatorname{rank}(V_0\Lambda^{-1}) = K$, we have X_0 is full rank. Next we establish the order of the eigenvalues. For this, let $D = \operatorname{diag}\left(\sqrt{n_1}, \sqrt{n_2}, \ldots, \sqrt{n_K}\right)$. We write

$$M_0 = \Theta_0 D^{-1}\left(DZ_0D\right)D^{-1}\Theta_0^T.$$

This shows that M_0 and Z_0 have the same eigenvalues. Note that the entries of $(DZ_0D)_{ij} = \sqrt{n_i n_j}(Z_0)_{ij}$ and $(Z_0)_{ij} = O(1)$ as $(M_0)_{ij} = \Theta(1)$. Hence its non-zero eigenvalues of DZ_0D, and hence the non-zero eigenvalues of M_0 will be of order $\Theta(n)$.

Part (c). We start by noting that $V_0 = \Theta_0 X_0$, V_0 has K distinct rows, as X_0 has K distinct rows. Thus, $(V_0)_{i\star} = (V_0)_{j\star}$ whenever $g(i) = g(j)$ and $(V_0)_{i\star} \neq (V_0)_{j\star}$ whenever $g(i) \neq g(j)$. We now compute the inner products of rows of X_0. For this, we first note that

$$\begin{aligned}
\langle (V_0)_{\star i}, (V_0)_{\star j}\rangle &= \langle (\Theta_0 X_0)_{\star i}, (\Theta_0 X_0)_{\star j}\rangle, \\
&= \sum_r n_r((X_0)_{ri} \cdot (X_0)_{si}), \\
&= \langle D(X_0)_{\star i}, D(X_0)_{\star j}\rangle.
\end{aligned}$$

This shows that DX_0 has orthogonal columns and as a consequence, orthogonal rows. This shows that

$$\langle (X_0)_{r\star}, (X_0)_{s\star}\rangle = 0 \text{ if } r \neq s \quad \text{and} \quad \langle (X_0)_{r\star}, (X_0)_{s\star}\rangle = \frac{1}{n_r} \text{ if } r = s.$$

Thus, we have

$$\langle (V_0)_{i\star}, (V_0)_{j\star}\rangle = 0 \text{ if } g(i) \neq g(j) \quad \text{and} \quad \langle (V_0)_{i\star}, (V_0)_{j\star}\rangle = \frac{1}{n_i} \text{ if } g(i) = g(j).$$

Therefore, (5.17) follows immediately. ■

We finish this section with the following perturbation result about the eigenspace of M, which is a direct consequence of Davis-Kahan $\sin\theta$ theorem (Yu et al., 2015, Theorem 2):

Proposition 10. *Let V be the matrix of K largest eigenvectors of M. There exists an orthonormal matrix $O \in \mathbb{R}^{K\times K}$ such that*

$$\|V - V_0 O\|_{\mathrm{F}} \leqslant \frac{\sqrt{8K}\,\|M - M_0\|_{\mathrm{F}}}{\min_{1\leqslant r\leqslant K}|\lambda_r|}.$$

5.2.2 Bound on $\|M - M_0\|_{\mathrm{F}}$.

In this section, we prove Theorem 3. We start by proving it for a simple case $t = t_L = t_U \geqslant 2$.

Proposition 11. *The conclusion for Theorem 3 holds with $t = t_L = t_U \geqslant 2$.*

Proof. Let us first give a proof for $t = t_L = t_U \geqslant 3$ using the estimates from Section 5.1. The proof for $t = t_L = t_U = 2$ is similar and we will give the required modifications at the end. Throughout, the constants term C, C' may change from line to line in this proof. Let $a_n = 4n(\log n)^{-\eta}$. Recall the notation W_A and W_P from Section 4.3.1. Also, recall from Proposition 4 that

$$\begin{aligned} M_{ij} &= \log\left[\frac{2|A|}{b\gamma(l, t_L, t_U)}(l-t)\cdot(D_A^{-1}W_A^t)_{ij}\right]\mathbb{1}_{A_{ij}^{(t)}>0}, \\ (M_0)_{ij} &= \log\left[\frac{2|P|}{b\gamma(l, t_L, t_U)}(l-t)\cdot(D_P^{-1}W_P^t)_{ij}\right]. \end{aligned} \tag{5.20}$$

By Proposition 8 with $k = \lceil \log n \rceil$ we have for any $1 \leqslant i, j, \leqslant n$

$$\mathbb{P}\left(A_{ij}^{(t)} = 0\right) \leqslant O(n^{-3}),$$

And this implies that

$$\begin{aligned} &\mathbb{P}\left(A_{ij}^{(t)} = 0 \text{ for some } 1 \leqslant i, j \leqslant n\right) = o(1), \\ \implies &\mathbb{P}\left(M_{ij} = 0 \text{ for some } 1 \leqslant i, j \leqslant n\right) = o(1). \end{aligned} \tag{5.21}$$

Then we have

$$\mathbb{P}\left(\|M - M_0\|_{\mathrm{F}} \geqslant a_n\right) \leqslant o(1) + \mathbb{P}\Big(\sum_{(i,j)}(M - M_0)_{ij}^2 \geqslant a_n^2\Big). \tag{5.22}$$

Next we bound the second term in (5.22). By Chernoff bound, we have that for a sufficiently large constant $C_0 > 0$,

$$\begin{aligned}\mathbb{P}\bigg(\bigg|\frac{|A|}{|P|} - 1\bigg| \geqslant C_0 n^{-1/2}\bigg) &\leqslant 2\exp\{-C' n^2 \rho_n \times C_0^2/n\} = o(n^{-4}),\\ \mathbb{P}\bigg(\exists i \in [n] : \bigg|\frac{|A_{i\star}|}{|P_{i\star}|} - 1\bigg| > C_0\sqrt{\frac{\log n}{n\rho_n}}\bigg) &\leqslant 2n\exp(-C' n\rho_n \times C_0^2 \log n/n\rho_n) \leqslant o(n^{-4}).\end{aligned} \tag{5.23}$$

Using (5.23), we simplify the second term in (5.22) as

$$\begin{aligned}&\mathbb{P}\bigg(\sum_{(i,j)}(M - M_0)_{ij}^2 \geqslant a_n^2\bigg) \leqslant \sum_{(i,j)}\mathbb{P}\bigg((M - M_0)_{ij}^2 \geqslant \frac{a_n^2}{n^2}\bigg)\\ &\leqslant \sum_{(i,j)}\mathbb{P}\bigg(\max\bigg\{\frac{(D_A^{-1}W_A^t)_{ij}}{(D_P^{-1}W_P^t)_{ij}}, \frac{(D_P^{-1}W_P^t)_{ij}}{(D_A^{-1}W_A^t)_{ij}}\bigg\} \geqslant (1 + O(n^{-1/2}))\exp\Big(\frac{a_n}{n}\Big)\bigg) + o(n^{-4}).\end{aligned} \tag{5.24}$$

To analyze this, recall from (5.2) $\mathcal{P}_b$ is the set of paths with vertices having community assignment b for $b \in \mathcal{B}_{i,j}$. For $p = (i_0, \ldots, i_t) \in \mathcal{P}_b$, let

$$\begin{aligned}\bar{X}_p &= \frac{1}{|A_{i_0\star}|}\prod_{l=1}^{t}\frac{A_{i_{l-1}i_l}}{|A_{i_l\star}|}, \quad \text{and} \quad \bar{Y}_b = \sum_{p\in\mathcal{P}_b}\bar{X}_p,\\ \bar{X}_p^* &= \frac{1}{|P_{i_0\star}|}\prod_{l=1}^{t}\frac{P_{i_{l-1}i_l}}{|P_{i_l\star}|}, \quad \text{and} \quad \bar{Y}_b^* = \sum_{p\in\mathcal{P}_b}\bar{X}_p^*.\end{aligned}$$

Then we have $(D_A^{-1}W_A^t)_{ij} = \sum_{b\in\mathcal{B}_{i,j}}\bar{Y}_b$ and $(D_P^{-1}W_P^t)_{ij} = \sum_{b\in\mathcal{B}_{i,j}}\bar{Y}_b^*$. Now, for $b \in \mathcal{B}_{i,j}$, $\mathbb{P}(\bar{Y}_b = 0) = o(n^{-4})$ by Proposition 8, and on the set $\{\bar{Y}_b \neq 0\}$, we have

$$\frac{(D_P^{-1}W_P^t)_{ij}}{(D_A^{-1}W_A^t)_{ij}} \leqslant \sum_{b\in\mathcal{B}_{i,j}}\frac{\bar{Y}_b^*}{\bar{Y}_b}, \qquad \frac{(D_A^{-1}W_A^t)_{ij}}{(D_P^{-1}W_P^t)_{ij}} \leqslant \sum_{b\in\mathcal{B}_{i,j}}\frac{\bar{Y}_b}{\bar{Y}_b^*}.$$

Thus, in order to bound 5.24, it is enough to bound the probabilities for $\bar{Y}_b^*/\bar{Y}_b$ or $\bar{Y}_b/\bar{Y}_b^*$ being large. Since the row sums of A are concentrated by (5.23), we will bound Y_b^*/Y_b, Y_b/Y_b^* instead, where Y_b^* is as defined below:

$$X_p^* = \prod_{l=1}^{t}P_{i_{l-1}i_l}, \quad \text{and} \quad Y_b^* = \sum_{p\in\mathcal{P}_b}X_p^*.$$

Fix (i,j). We estimate the difference

$$|\mathbb{E}Y_b - Y_b^*| = \left|\mathbb{E}\sum_{p\in\mathcal{P}_b}(A_{i_0i_1}\cdots A_{i_{t-1}i_t} - P_{i_0i_1}\cdots P_{i_{t-1}i_t})\right|. \tag{5.25}$$

The summands in the equation above are equal to zero if the associated path $(i_0, i_1, \ldots, i_t)$ has t distinct edges and there are no self-loops. Consider the first set of summands in (5.25). By Proposition 7 the sum over summands with less than t distinct edges is $O(1/n\rho_n)\mathbb{E}[Y_b]$. We now give an upper bound on the summands $P_{i_0i_1}\cdots P_{i_{t-1}i_t}$ as follows. Suppose a path has less than t distinct edges. If the path is a Type I path then by Lemma 6 the number of choices of distinct vertices along the path is less than $t-1$. If the the path is a Type II path, then again by Lemma 6 the number of choices of distinct vertices along the path is at most $\lfloor t/2\rfloor$. Finally, if there are self-loops then the number of choices of vertices is less than $t-1$. This implies that the upper bound on the second set of summands is $O(1/n)\mathbb{E}[Y_b]$. Thus in summary we have

$$|\mathbb{E}Y_b - Y_b^*| = O\left(\frac{1}{n\rho_n}\right)\mathbb{E}Y_b. \tag{5.26}$$

These computations show that

$$\frac{Y_b}{Y_b^*} = \frac{Y_b}{\mathbb{E}Y_b\left(1+O((n\rho_n)^{-1})\right)}, \quad \text{and} \quad \frac{Y_b^*}{Y_b} = \frac{\mathbb{E}Y_b\left(1+O((n\rho_n)^{-1})\right)}{Y_b}. \tag{5.27}$$

Recall that $a_n = 4n(\log n)^{-\eta}$. To compute (5.24), we now use Proposition 8, (5.23) and (5.27) to obtain

$$\begin{aligned}\mathbb{P}\left(\frac{(D_P^{-1}W_P^t)_{ij}}{(D_A^{-1}W_A^t)_{ij}} \geqslant (1+O(n^{-1/2}))\exp\left(\frac{a_n}{n}\right)\right) &\leqslant \sum_{b\in\mathcal{B}_{i,j}}\mathbb{P}\left(\frac{\bar{Y}_b^*}{\bar{Y}_b} \geqslant \exp\{C(\log n)^{-\eta}\}\right) + o(n^{-3}) \\ &\leqslant \sum_{b\in\mathcal{B}_{i,j}}\mathbb{P}\left(\frac{Y_b^*}{Y_b} \geqslant 1+(\log n)^{-\eta}\right) + o(n^{-3}) = o(n^{-3}).\end{aligned}$$

A similar bound can be computed with $\frac{(D_A^{-1}W_A^t)_{ij}}{(D_P^{-1}W_P^t)_{ij}}$ as well repeating the same computations. Hence we have established that the term in (5.24) is at most $o(n^{-3})$, and thus combining (5.22) and (5.21), we conclude that $\|M - M_0\|_{\mathrm{F}} = O_{\mathbb{P}}(a_n)$, and thus (4.16) follows for $t_L = t_U = t \geqslant 3$. For $t_U = t = 2$, the argument is exactly similar except that we use (Janson et al., 2000, Theorem 2.8) for showing

(5.21), we use Bernstein's inequality in place of the concentration inequality in Proposition 8, we don't need (5.26) for $i \neq j$ case, and for the $i = j$ case we use the bound $\frac{\mathbb{E}Y_b}{Y_b^*} = O(\rho_n^{-1})$.

Next, we prove (4.17) in the case $t_L = t_U = t \geqslant 2$. Let $n^{t-1}\rho_n^t \ll 1$. We will show that, for any $\varepsilon > 0$,

$$\mathbb{P}\big(\|M - M_0\|_{\mathrm{F}} \geqslant C_\varepsilon n^2 \big) \geqslant \mathbb{P}\bigg(\sum_{(i,j)} (M_0)_{ij}^2 \mathbb{1}\{M_{ij} = 0\} \geqslant C_\varepsilon n^2 \bigg) \geqslant 1 - \varepsilon,$$

for some constant $C_\varepsilon > 0$ depending on ε and the last inequality holds for large enough n.

By Proposition 9, the entries of $(M_0)_{ij}$'s are constant over all i, j pairs such that $g(i) = r, g(j) = s$. Also, M_0 will have some non-zero entries since $\mathrm{rank}(M_0) = K$. Let r and s be such that $g(i) = r$ and $g(j) = s$ and $|(M_0)_{ij}| = C_1 > 0$ for all i, j such that $g(i) = r$ and $g(j) = s$, and the number of such pairs of i, j is at least $C_2 n^2$ for some $0 < C_2 < 1$. Let

$$S_{r,s} := \{(i,j) : A_{ij}^t = 0, g_i = r, g_j = s\}, \qquad T_{r,s} := \{(i,j) : A_{ij}^t > 0, g_i = r, g_j = s\}.$$

Then

$$\mathbb{P}\bigg(\sum_{(i,j)} (M_0)_{ij}^2 \mathbb{1}\{M_{ij} = 0\} \geqslant C_\varepsilon n^2 \bigg) \geqslant \mathbb{P}\bigg(\sum_{(i,j) \in S_{r,s}} C_1^2 \geqslant C_\varepsilon n^2 \bigg) = \mathbb{P}\bigg(|S_{r,s}| \geqslant \frac{C_\varepsilon n^2}{C_1^2} \bigg).$$

Next let i and j be any two nodes such that $g_i = r$, $g_j = s$, and $i \neq j$. Then by Proposition 6 we have

$$\mathbb{P}\big(A_{ij}^t > 0\big) \leqslant \sum_{b \in \mathcal{B}_{i,j}} \mathbb{P}\left(Y_b > 0\right) \leqslant C_3 n^{t-1} \rho_n^t,$$

for some constant $C_3 > 0$. This along with Markov inequality implies that

$$\mathbb{P}\left(|T_{r,s}| \geqslant C_3 \varepsilon^{-1} n^{t+1} \rho_n^t \right) \leqslant \frac{n^2 \cdot C_3 n^{t-1} \rho_n^t}{C_3 \varepsilon^{-1} n^{t+1} \rho_n^t} \leqslant \varepsilon.$$

This shows that for large enough n we have

$$\mathbb{P}\left(|S_{r,s}| \geqslant |S_{r,s}| + |T_{r,s}| - C_3 \varepsilon^{-1} n^{t+1} \rho_n^t \right) \geqslant 1 - \varepsilon \implies \mathbb{P}\left(|S_{r,s}| \geqslant C_2 n^2 - C_3 \varepsilon^{-1} n^{t+1} \rho_n^t \right) \geqslant 1 - \varepsilon.$$

Taking $C_\varepsilon = C_1^2 C_2/2$ and noting that $n^{t+1}\rho_n^t = o(n^2)$ completes the proof. ■

Next we complete the proof of Theorem 3 for general t_L, t_U.

Proof of Theorem 3. The general idea is to reduce the computations to the analogous computations for the $t_L = t_U$ case. If $t_L = 2$ and $i \neq j$ then by (Janson et al., 2000, Theorem 2.8),

$$\mathbb{P}\left(A_{ij}^{(2)} = 0\right) \leqslant \exp\{-\theta(n\rho_n^2)\}.$$

Similarly if $i = j$ (and $t_L = 2$) we have

$$\mathbb{P}\left(A_{ii}^{(2)} = 0\right) \leqslant \exp\{-\theta(n\rho_n)\}.$$

By the assumption of $n^{t_L-1}\rho_n^{t_L} \gg (\log n)$ when $t_L = 2$, we have

$$\mathbb{P}\left(A_{ij}^{(2)} = 0 \text{ for some } 1 \leqslant i, j \leqslant n\right) = o(1).$$

Next let $\max(3, t_L) \leqslant t \leqslant t_U$. Then by Proposition 8 with $k = \lceil \log n \rceil$ we have for any $1 \leqslant i, j, \leqslant n$

$$\mathbb{P}\left(A_{ij}^{(t)} = 0\right) \leqslant O(n^{-3}).$$

And this implies that

$$\mathbb{P}\left(A_{ij}^{(t)} = 0 \text{ for some } 1 \leqslant i, j \leqslant n \text{ and for some } t \in \{t_L, t_L + 1, \ldots, t_U\}\right) = o(1).$$

Then proceeding as in proof of Proposition 11 we have

$$
\begin{aligned}
&\mathbb{P}\left(\|M_0 - M\|_{\mathrm{F}} \geqslant a_n\right) \\
&\leqslant \mathbb{P}\left(\|M_0 - M\|_{\mathrm{F}} \geqslant a_n, A_{ij}^t > 0 \text{ for } 1 \leqslant i,j \leqslant n, t_L \leqslant t \leqslant t_U\right) + o(1), \\
&\leqslant o(1) + \sum_{1\leqslant i\neq j\leqslant n} \mathbb{P}\left(\frac{\sum_{t=t_L}^{t_U}(l-t)\left(D_P^{-1}W_P^t\right)_{ij}}{\sum_{t=t_L}^{t_U}(l-t)\left(D_A^{-1}\hat{W}_P{}^t\right)_{ij}} \geqslant \exp\left\{\frac{a_n}{\sqrt{2}n} - \theta(n^{-1/2})\right\}\right) \\
&\quad + \sum_{1\leqslant i\leqslant n} \mathbb{P}\left(\frac{\sum_{t=t_L}^{t_U}(l-t)\left(D_P^{-1}W_P^t\right)_{ii}}{\sum_{t=t_L}^{t_U}(l-t)\left(D_A^{-1}\hat{W}_P{}^t\right)_{ii}} \geqslant \exp\left\{\frac{a_n}{\sqrt{2}n} - \theta(n^{-1/2})\right\}\right) \\
&\quad + \sum_{1\leqslant i\neq j\leqslant n} \mathbb{P}\left(\frac{\sum_{t=t_L}^{t_U}(l-t)\left(D_A^{-1}\hat{W}_P{}^t\right)_{ij}}{\sum_{t=t_L}^{t_U}(l-t)\left(D_P^{-1}W_P^t\right)_{ij}} \geqslant \exp\left\{\frac{a_n}{\sqrt{2}n} - \theta(n^{-1/2})\right\}\right) \\
&\quad + \sum_{1\leqslant i\leqslant n} \mathbb{P}\left(\frac{\sum_{t=t_L}^{t_U}(l-t)\left(D_A^{-1}\hat{W}_P{}^t\right)_{ii}}{\sum_{t=t_L}^{t_U}(l-t)\left(D_P^{-1}W_P^t\right)_{ii}} \geqslant \exp\left\{\frac{a_n}{\sqrt{2}n} - \theta(n^{-1/2})\right\}\right). \qquad (5.28)
\end{aligned}
$$

We show how to bound the second term in the equation 5.28. For this we see that

$$
\begin{aligned}
&\mathbb{P}\left(\frac{\sum_{t=t_L}^{t_U}(l-t)\left(D_P^{-1}W_P^t\right)_{ij}}{\sum_{t=t_L}^{t_U}(l-t)\left(D_A^{-1}\hat{W}_P{}^t\right)_{ij}} \geqslant \exp\left\{\frac{a_n}{\sqrt{2}n} - \theta(n^{-1/2})\right\}\right) \\
&\leqslant \sum_{t=t_L}^{t_U} \mathbb{P}\left(\frac{\left(D_P^{-1}W_P^t\right)_{ij}}{\left(D_A^{-1}\hat{W}_P{}^t\right)_{ij}} \geqslant \exp\left\{\frac{a_n}{\sqrt{2}n} - \theta(n^{-1/2})\right\}\right).
\end{aligned}
$$

Then each of the probabilities for fixed t can be bounded as in the proof of Propositions 11. Analogously the rest of the terms in (5.28) can be bounded. This completes the proof. ■

5.2.3 Bounding the number of missclassified nodes

Proof of Theorem 4. This proof uses standard arguments to bound the proportion of misclassified nodes such as given in (Lei et al., 2015). For this proof, we choose $O \in \mathbb{R}^{K\times K}$ obtained by an application of Proposition 10 which satisfies

$$
\|V - V_0 O\|_{\mathrm{F}} \leqslant \frac{\|M - M_0\|_{\mathrm{F}}}{Cn} = o_{\mathbb{P}}(1), \qquad (5.29)
$$

where the last step follows using Theorem 3. To simplify notation, we denote $U = V_0 O$ in the rest of the proof. Let $n_{\max} = \max_{r\in[K]} n_r$. Recall $(\bar{\Theta}, \bar{X})$ from (4.14) and let $\bar{V} = \bar{\Theta}\bar{X}$. Then let

$$S = \left\{ i \in [n] : \left\|\bar{V}_{i\star} - U_{i\star}\right\|_{\mathrm{F}} \geqslant \frac{1}{5}\sqrt{\frac{2}{n_{\max}}} \right\}.$$

We will show that for $i \notin S$ the community is predicted correctly using $\bar{\Theta}$. The proof of the theorem is in two steps.

Step 1: Bounding $|S|$. By the definition of S we have

$$\sum_{i\in S} \sqrt{\frac{2}{n_{\max}}} \leqslant \sum_{i\in S} 5\left\|\bar{V}_{i\star} - U_{i\star}\right\|_{\mathrm{F}} \leqslant 5\left\|\bar{V} - U\right\|_{\mathrm{F}} \implies |S| \leqslant \frac{5}{\sqrt{2}}\sqrt{n_{\max}}\left\|\bar{V} - U\right\|_{\mathrm{F}}. \tag{5.30}$$

Next, we recall the optimization problem in (4.14) is given as follows.

$$\left\|\bar{\Theta}\bar{X} - V\right\|_{\mathrm{F}}^2 \leqslant (1+\varepsilon) \min_{\Theta\in\{0,1\}^{n\times K}, X\in\mathbb{R}^{K\times K}} \|\Theta X - V\|_{\mathrm{F}}^2.$$

We substitute U for ΘX to get the following upper bound:

$$\left\|\bar{V} - V\right\|_{\mathrm{F}}^2 \leqslant (1+\varepsilon)\|U - V\|_{\mathrm{F}}^2. \tag{5.31}$$

Then by (5.29) and (5.31) we have

$$\left\|\bar{V} - U\right\|_{\mathrm{F}} \leqslant \left\|\bar{V} - V\right\|_{\mathrm{F}} + \|V - U\|_{\mathrm{F}} \leqslant (1+\sqrt{1+\varepsilon})\|V - U\|_{\mathrm{F}} = o_{\mathbb{P}}(1).$$

This combined with equation 5.30 we have $|S| = o_{\mathbb{P}}(\sqrt{n})$.

Step 2: Bounding the prediction error. For any community r, there exists $i_r \in [n]$ such that $g_{i_r} = r$ and $i_r \notin S$ as $n_r = \theta(n)$ and $|S| = o_{\mathbb{P}}(\sqrt{n})$. By Proposition 9c, for $r \neq s$ we have

$$\begin{aligned}\left\|\bar{V}_{i_r\star} - \bar{V}_{i_s\star}\right\|_{\mathrm{F}} &\geqslant \|U_{i_r\star} - U_{i_s\star}\|_{\mathrm{F}} - \left\|\bar{V}_{i_r\star} - U_{i_r\star}\right\|_{\mathrm{F}} - \left\|\bar{V}_{i_s\star} - U_{i_s\star}\right\|_{\mathrm{F}} \\ &\geqslant \sqrt{\frac{1}{n_r} + \frac{1}{n_s}} - \frac{2}{5}\sqrt{\frac{2}{n_{\max}}} \geqslant \frac{3}{5}\sqrt{\frac{2}{n_{\max}}}. \end{aligned} \tag{5.32}$$

Next, let i be such that $g_i = r$ and $i \notin S$. We show that $\bar{V}_{i\star} = \bar{V}_{i_r\star}$. For this we note that $U_{i\star} = U_{i_r\star}$ and

$$\begin{aligned}\left\|\bar{V}_{i\star} - \bar{V}_{i_r\star}\right\|_{\mathrm{F}} &\leqslant \left\|\bar{V}_{i\star} - U_{i\star}\right\|_{\mathrm{F}} + \left\|U_{i_r\star} - \bar{V}_{i_r\star}\right\|_{\mathrm{F}} \\ &< \frac{1}{5}\sqrt{\frac{2}{n_{\max}}} + \frac{1}{5}\sqrt{\frac{2}{n_{\max}}} < \frac{2}{5}\sqrt{\frac{2}{n_{\max}}}. \end{aligned} \tag{5.33}$$

In view of (5.32) and (5.33) we must have $\bar{V}_{i\star} = \bar{V}_{i_r\star}$ as each node is assigned exactly one community by the $(1+\varepsilon)$-approximate k-means algorithm and there are exactly distinct K rows in $\bar{V}$ (again by (5.32)). Let C be a permutation matrix defined so that $\Theta_0 C$ assigns community r to node i_r, for $1 \leqslant r \leqslant K$. Then we have,

$$\sum_i \mathbb{1}\{\bar{\Theta}_{i\star} \neq (\Theta_0 C)_{i\star}\} \leqslant |S|.$$

From the bound on $|S|$ from Step 1, we have that $\mathrm{Err}(\bar{\Theta}, \Theta_0)$ is $o_{\mathbb{P}}\left(1/\sqrt{n}\right)$ and this completes the proof of Theorem 4. ■

5.3 Path counting for node2vec

In this section, we focus on computing the asymptotics for the sum of weighted paths having some specified community assignments for the intermediate vertices. In section 5.3.1 and 5.3.2, we bound its k-th moment and we end with a concentration inequality in section 5.3.3.

5.3.1 Bounding moments of path counts for Regime III

We compute upper bounds for the k-th moment and k-th centered moments for weighted paths between two nodes under conditions given by Regime III. We first fix some notation. Let $\mathcal{B}_{ij}$ and $\mathcal{P}_b$ be defined as in (5.1) and (5.2).

For any path $p = (i_0, i_1, i_2, \ldots, i_t) \in \mathcal{P}_b$, let

$$\mathfrak{N}((i_0, i_1, \ldots, i_t)) := \{l | 2 \leqslant l \leqslant t, i_{l-2} = i_l\},$$

be the set of locations of backtracks in the path. Let

$$\mathfrak{n} = \mathfrak{n}((i_0, i_1, \ldots, i_t)) = |\mathfrak{N}((i_0, i_1, \ldots, i_t))|,$$

be the number of backtracks in p. For any path $p = (i_0, i_1, i_2, \ldots, i_t) \in \mathcal{P}_b$ and $\alpha > 0$, we associate the random variable

$$X_{p,\alpha} := A_{i_0 i_1} A_{i_1 i_2} \cdots A_{i_{t-1} i_t} \alpha^{\mathfrak{n}((i_0, i_1, \ldots, i_t))}, \tag{5.34}$$

and let

$$Y_{b,\alpha} := \sum_{p \in \mathcal{P}_b} X_{p,\alpha}. \tag{5.35}$$

We note that when $\alpha = 1$, $X_{p,1} = X_p$ and $Y_{b,1} = Y_b$ where X_p and Y_b are as defined in (5.3) and (5.4) respectively. To simplify notation in this section, we will drop the subscript α and simply write X_p and Y_b in place of $X_{p,\alpha}$ and $Y_{b,\alpha}$ respectively. Let U_b and L_b be the upper and lower bounds for path type $b \in \mathcal{B}_{i,j}$ as defined in (5.5) and (5.6) respectively. Then we have the following bounds on $\mathbb{E}Y_b^k$

Proposition 12. *Let $t_L = t_U = t \geqslant 3$ be given and suppose that $\alpha = O\left(\frac{1}{n}\right)$ and (4.22) holds. Then we have*

$$L_b^k \leqslant \mathbb{E}Y_b^k \leqslant U_b^k(1 + o(1)).$$

Again the idea, as in section 5.1, is to show that the leading term for $\mathbb{E}Y_b^k$ is due to $\mathbb{E}\big(\prod_{\pi=1}^k X_{p_\pi}\big)$ of k ordered paths p_π having kt distinct edges between them. The contribution of the rest of the terms are of a smaller order. Similar to section 5.1, let

$$E_m = E_{m,\alpha} := \sum_{(p_1, p_2, \ldots, p_k): p_\pi \in \mathcal{P}_b, |\cup_{i \in [k]} e(p_\pi)| = m} \mathbb{E}(X_{p_1} X_{p_2} \cdots X_{p_k}). \tag{5.36}$$

We will show the following:

Proposition 13. *Under identical conditions as in Proposition 12, we have $\sum_{m<kt} E_m = o(U_b^k)$.*

Proof of Proposition 12 using Proposition 13. Note that, we can write

$$\mathbb{E}Y_b^k = \mathbb{E}\Big(\sum_{p\in\mathcal{P}_b} X_p\Big)^k = \sum_{(p_1,p_2,\ldots,p_k)|p_\pi\in\mathcal{P}_b} \mathbb{E}(X_{p_1}X_{p_2}\cdots X_{p_k}). \tag{5.37}$$

For the upper bound, Proposition 13 shows that it is enough to bound the summands corresponding to sequences $(p_1, p_2, \ldots, p_k)$ that satisfy $|\cup_{\pi=1}^k e(p_\pi)| = kt$, i.e. sequences of paths consisting of kt distinct edges. We note that there are no backtracks in this case and so, $\mathbb{E}(X_{p_1}X_{p_2}\cdots X_{p_k}) = \mathbb{P}(X_{p_1}X_{p_2}\cdots X_{p_k} = 1) = \prod_{\pi=1}^k \mathbb{P}(X_{p_\pi} = 1)$. Thus we have the same upper and lower bounds U_b^k and L_b^k as in Proposition 6.

■

The rest of this section is devoted to the proof of Proposition 13. Towards this let $E_{m,r}$ denote the summands in (5.36) restricted to the case that there are r segments, Type I or Type II, so that

$$E_m = \sum_{r=r_*(m)}^{r^*(m)} E_{m,r},$$

where, given m, $[r_*(m), r^*(m)]$ denotes the range of r. We note that this is in contrast to the proof of Proposition 7 where r was the number of Type I segments. We also note that we are reusing the notation $r_*(m)$ and $r^*(m)$ from the proof of Proposition 7 to simplify the notation but the values of $r_*(m)$ and $r^*(m)$ will be different in this proof.

The analysis will again consist of two steps. In the first step, we analyze $E_{m,r_*(m)}$ and in the second step, we will show that $E_{m,r}$ is much smaller than $E_{m,r_*(m)}$ for $r > r_*(m)$.

The following is the intuition for why $E_{m,r_*(m)}$ is the largest term. Due to the presence of backtracks, the contribution from Type II segments is of the same or a smaller order than Type I segments. By Lemma 6 minimizing Type I segments leads to the maximum number of choices of marked edges. Combining these two ideas, we see that we must minimize the number of segments. A formal proof is provided in the rest of the section.

5.3.1.1 Computing $E_{m,r_*(m)}$.

We begin by noting that $r_*(m)$ is given by

$$r_*(m) = \left\lceil \frac{m}{t} \right\rceil.$$

To see this we note that each path can have at most t marked edges. So we can place m marked edges in a minimum of $\lceil \frac{m}{t} \rceil$ paths. The next lemma counts the number of configurations of segments and unmarked edges. For this, let $N_{m,r_*(m)}$ be the number of configurations of m marked edges placed in r segments, Type I or II. Again, we are reusing the notation $N_{m,r_*(m)}$ from section 5.1.

Lemma 13.

$$N_{m,r_*(m)} \leqslant C \binom{k}{r_*(m)} k^{tr_*(m)-m}.$$

Proof. The proof is divided into two cases:

Case I: $\frac{m}{t} \in \mathbb{N}$. Note that $r_*(m) = \frac{m}{t}$ in this case. We can choose $\frac{m}{t}$ paths containing all the m marked edges in $\binom{k}{\frac{m}{t}}$ ways. Al the edges in the chosen $\frac{m}{t}$ paths are marked edges and all the edges in the rest of the paths are unmarked edges.

Case II: $\frac{m}{t} \notin \mathbb{N}$. We can first choose $\lceil \frac{m}{t} \rceil$ paths to place the marked edges. By pigeonhole principle, there are $0 < l \leqslant t\lceil \frac{m}{t} \rceil - m$ of the $\lceil \frac{m}{t} \rceil$ chosen paths which are not saturated. We can choose arrangements for these l paths in C^l ways where C is a constant that may depend on t. The chosen paths can have at most l unmarked edges. The rest $k - \lceil \frac{m}{t} \rceil$ all have only unmarked edges.

■

We now complete the proof with the following lemma.

Lemma 14. $\sum_{m<kt} E_{m,r_*(m)} = o(U_b^k)$.

Proof. We recall that $U_b = \Theta(n^{t-1}\rho_n^t)$. The number of choices of segments is given by Lemma 13. Let $s(m)$ be the number of maximal (or saturated) Type I paths when placing m marked edges in $r_*(m)$ paths. When $\frac{m}{t} \in \mathbb{N}$, all the $r_*(m)$ paths containing marked edges are maximal Type I paths and each of them have probability at most U_b as there are no backtracks. When $\frac{m}{t} \notin \mathbb{N}$ we

have $s \geqslant \max\left\{\left\lceil \frac{m}{t} \right\rceil - \left(t\left\lceil \frac{m}{t} \right\rceil - m\right), 0\right\}$ maximal Type I paths. The number of unmarked edges is at most $kt-m$. The number of marked edges in the non-maximal Type I paths is equal to $m-s(m)t$.

Choose a Type I segment with m' marked edges in a non-maximal Type I path. By Lemma 6, the vertices can be chosen in at most $n^{m'-1}$ ways. The number of backtracks in the Type I segment can be equal to 0 or larger than 0. So $\alpha^0 = 1$ is an upper bound for the factor coming from backtracks in (5.34). And so the upper bound for the contribution coming from the choice of vertices and backtracks is $n^{m'-1}\alpha^0 = n^{m'-1}$. Now for a Type II segment (in a non-maximal Type I path) with m' marked vertices, there must be at least m' backtracks. Thus the corresponding upper bound for a Type II segment is $n^{m'}\alpha^{m'} = O(1)$. We can note that we can have a Type II segment in only a non-saturated path and so the number of Type II segments are $O(1)$. Using this analysis, for any segment, Type I or II, with m' marked edges in a non-maximal (Type I) path, we upper bound the choices of marked vertices and the factors from backtracks by $Cn^{m'-1}$.

Combining all these, we get

$$E_{m,r_*(m)} \leqslant C\binom{k}{r_*(m)} k^{tr_*(m)-m} \times U_b^s n^{m-s(m)t-(r_*(m)-s(m))} \rho_n^{m-s(m)t} \times m^{kt-m}.$$

Using these bounds we see that with the choice of $k = \lceil \log n \rceil$

$$\sum_{l=0}^{t-1} E_{r_0t-l,r_\star(r_0t-l)} = E_{r_0t,r_\star(r_0t)}\left(1 + O\left(\frac{k^2}{n\rho_n}\right)\right), \quad r_0 \geqslant 1,$$
$$\sum_{r_0=1}^{k} E_{r_0t,r_\star(r_0t)} = E_{kt,r_\star(kt)}\left(1 + O\left(\frac{k^{t+1}}{n^{t-1}\rho_n^t}\right)\right).$$

These bounds in turn imply that

$$\sum_{m=1}^{kt-1} E_{m,r_*(m)} = o\left(U_b^k\right).$$

■

5.3.1.2 Computing $E_{m,r}$ for $r > r_*(m)$.

We start with a lemma to bound the number of configurations of segments and unmarked edges as in Lemma 13.

Lemma 15. *Given m marked edges and r segments, let $N_{m,r}$ be the number of configurations of segments and unmarked edges. Then, for any $r > r_*(m)$,*

$$N_{m,r} \leqslant N_{m,r_*(m)} \times O(k^{(r-r_*(m))(t+1)}).$$

Proof. Let T_r be the set of all configurations of m marked edges and r, Type I or Type II, segments. Note that we are reusing the notation T_r from proof of Proposition 10 but T_r is defined differently here. We will inductively bound T_{r+1} in terms T_r. For that, we consider two cases depending on whether the elements of T_{r+1} has a path with two segments or not. In both cases, we will find a relation between T_r to T_{r+1}.

Case I. Suppose that T_{r+1} has a path with two segments. Let us call this subset T^I_{r+1}. We consider the elements in T_r which will be related to these elements in T^I_{r+1}. Let $T^I_r \subset T_r$ consisting of configuration such that there is at least one path p so that the following condition holds:

- **Extra unmarked edges.** p has l segments and $l' \geqslant l$ unmarked edges for some $l, l' \geqslant 1$.

The first condition is the same as in the proof of Lemma 10. The second condition condition is absent as in this construction we will split any segment, Type I or II, as long as it has at least two marked edges. Similar to the proof of Lemma 10, we split a segment only at a marked edge. This is to ensure that splitting creates two segments. The rest of the details of this construction, and the proof that the relation given by the construction is surjective onto T^I_{r+1} are similar to Case I in the proof of Lemma 10. As in Lemma 10 we have

$$|T^I_{r+1}| \leqslant O(k) \times |T^I_r| \leqslant O(k) \times |T_r|.$$

Case II. We next relate arrangements of marked edges where there is at most one segment per path. We denote such arrangements as $T^{II}_{r+1} \subset T_{r+1}$. Let T^{II}_r be the set of all ways of specifying locations of segments such that there are a total of r segments and each of the segments is placed on distinct paths. We note that T^{II}_r is defined differently as compared to the construction for DeepWalk in Lemma 10. Suppose that $r + 1 \leqslant \min(k, m)$ and $r \geqslant r_\star(m)$. Then we give a multi-map from T^{II}_r onto T^{II}_{r+1} using a construction. The condition $r + 1 \leqslant m$ is necessary so that T^{II}_{r+1} is non-empty as we must have at least $r + 1$ marked edges in order to have $r + 1$ paths each having

a Type I segment. The condition $r+1 \leqslant k$ is also necessary for T^{II}_{r+1} to be non-empty as we must have at least $r+1$ distinct paths to place the $r+1$ Type I segments. The last condition $r \geqslant r_\star(m)$ is to ensure that T^{II}_r is non-empty. To fix notation, let $S(l)$ be the set of all ways of arranging l marked edges in one path. We now describe the construction. Let $A \in T^{II}_r$ and choose a path p not containing a segment. We can do so as $r+1 \leqslant k$. Suppose p has $l \geqslant 0$ marked edges. Choose $0 \leqslant u \leqslant t-l$ paths where the u paths are distinct from p, and contain at least one marked edge. If $l=0$ we require $u>0$. It is feasible to choose such path(s) as $r+1 \leqslant m$. Suppose the u paths are labeled as $q_1, q_2, \ldots, q_u$. Suppose these paths have $l_{q_1}, l_{q_2}, \cdots, l_{q_u}$ marked edges respectively. Let $v_{q_1}, v_{q_2}, \cdots, v_{q_u}$ be such that $0 < v_{q_i} < l_{q_i}$. We require $\sum_i v_{q_i} \leqslant t-l$. Then we modify the arrangements of marked edges in the paths so that the new arrangements for the sequence of paths $(p, q_1, q_2, \ldots, q_u)$ is any element of

$$S\left(l + \sum_i v_{q_i}\right) \times \left(\prod_{i=1}^{u} S(l_{q_i} - v_{q_i})\right).$$

We keep the arrangements of marked edges in the rest $k-(1+u)$ paths unchanged. This leads to multiple images of A in T^{II}_{r+1}. We note that there are $O(k^{t+1})$ images of A due to the choice of the paths and since the number of ways of choosing the new arrangements for the $1+u$ paths is $O(1)$ as t is fixed. We also note that since we modify at most $t+1$ paths in this construction, the number of unmarked edges increases by at most $t(t+1)$.

Now we show that the multi-map given by the construction above is surjective onto T^{II}_{r+1}. For this, let $A' \in T^{II}_{r+1}$. Let p be a path containing a Type I segment. Suppose p has l marked edges. Choose $u \leqslant l$ paths so that u paths contain marked edges and these paths are not equal to p. Suppose the u paths are labeled as $q_1, q_2, \ldots, q_u$. Suppose that these paths have $l_{q_1}, l_{q_2}, \cdots, l_{q_u}$ marked edges respectively. We require that $l_{q_i} < t$ i.e. that these chosen paths are not saturated. Let $v_{q_1}, v_{q_2}, \cdots, v_{q_u}$ be such that $0 < v_{q_i} \leqslant t - l_{q_i}$. We require that

$$\sum_{i=1}^{u} v_{q_i} = l.$$

This is feasible as long as $r \geqslant r_\star(m)$. The above condition says that the chosen paths have enough spaces to move l edges from path p to the chosen paths. Choose a new arrangement x of the marked

edges for the sequence of paths $(p, q_1, q_2, \ldots, q_u)$ from the set

$$S(0) \times \left(\prod_{i=1}^{u} S(l_{q_i} + v_{q_i}) \right).$$

Then keeping the arrangements of marked edges of the rest of the paths the same as in A' and choosing the arrangements for the chosen paths as x, we have a preimage $A \in T_r^{II}$ under the construction described above.

From the two constructions above we have

$$|T_{r+1}| \leqslant O(k)|T_r| + O(k^{t+1})|T_r|. \tag{5.38}$$

■

We can now compute asymptotics for $E_{m,r}$.

Lemma 16. $\sum_{r=r_*(m)}^{r^*(m)} E_{m,r} = E_{m,r_*(m)}(1 + o(1))$.

Proof. We start by giving a bound for $E_{m,r}$. The probability of the m marked edges is bounded by ρ_n^m. The upper bound for the unmarked edges is $kt - m$. We now compute a bound for the choices of marked vertices and factors arising from backtracks. For this we note that for both the constructions in the proof of Lemma 15 we modify at most $t + 1$ paths and create an additional segment in the modified paths. By similar reasoning as in the proof of Lemma 14, for any Type I segment with m' marked edges in the modified paths we have an upper bound of $n^{m'-1}$ and for a Type II segment (in the modified paths) we have an upper bound of $O(1)$. Since we create an additional segment in the modified paths, we have an associated factor of $O(n^{-1})$. Combining these we have

$$E_{m,r} \leqslant N_{m,r} n^{m-r} \rho_n^m m^{kt-m}.$$

This implies that

$$\sum_{r=r_*(m)}^{r^*(m)} E_{m,r} = E_{m,r_*(m)} \sum_{r=r_*(m)}^{r^*(m)} \left(\frac{O(k^{t+1})}{n} \right)^{r-r_*(m)} = E_{m,r_*(m)}(1 + o(1)).$$

■

By Lemmas 14 and 16 we have

$$\sum_{m<kt} E_m \leqslant C \sum_{m<kt} E_{m,r_*(m)} = o(U_b^k).$$

This completes the proof of Proposition 13.

5.3.2 Bounding moments of path counts for Regimes I and II

In this section we bound the kth moment $\mathbb{E}Y_{b,\alpha}^k$ for regimes I and II.

Proposition 14. *Let $t_L = t_U = t \geqslant 3$ be given and suppose that either* (4.18) *and* (4.19) *hold or* (4.20) *and* (4.21) *hold. Then we have*

$$L_b^k \leqslant \mathbb{E}Y_{b,\alpha}^k \leqslant U_b^k(1 + o(1)).$$

Proof of Proposition 14. Let $X_{p,\alpha}$ and $Y_{b,\alpha}$ be defined as in (5.34) and (5.35). We will suppress the dependence on α in the notation. We first provide a proof for Regime I and then provide a proof for Regime II.

Proof for Regime I. We note that if there are m marked edges, there can be at most $kt - m$ backtracks. So for this regime, we add an additional factor of α^{kt-m} (i.e. a factor of α for each unmarked edge) to our bounds for an upper bound on the contribution from backtracking edges (in the case of m marked edges). This shows that

$$\begin{aligned} E'_m \leqslant C' \binom{2k}{r_*(m)} (2k)^{f(r_*(m))-m} 3^{2k-r_*(m)} C^{f(r_*(m))-m} \\ \times U_b^{s(m)} n^{m-ts(m)-(r_*(m)-s(m))} \rho_n^{m-ts(m)} \\ \times (2m)^{(2k-r_*(m))\cdot(t-2h)+2(f(r_*(m))-m)} \times \alpha^{kt-m}. \end{aligned}$$

Note that the upper bound on the marked edges remains the same for summands with m marked edges and $r > r_*(m)$ Type I segments.

$$\sum_{r=r_*(m)}^{r^*(m)} E_{m,r} = E_{m,r_*(m)} \sum_{r=r_*(m)}^{r^*(m)} \left(\frac{O(k^{t+1})}{n}\right)^{r-r_*(m)} = E_{m,r_*(m)}(1+o(1)).$$

This implies that

$$\sum_{m<kt} E_m \leqslant C \sum_{m<kt} E_{m,r_*(m)} = o(U_b^k).$$

This completes the proof for Regime I.

Proof for Regime II. Suppose we place m marked edges using r Type I segments. Then the minimum number, $b = b(r,m)$, of marked edges in Type II segments is given by

$$b := \max(m - rt, 0). \tag{5.39}$$

For the lower bound above, we subtract rt as each of the r Type I paths can have at most t marked edges each. Since each marked edge is followed by a backtrack in a Type II segment, there are a minimum of b backtracking edges when using r Type I segments to place m marked edges. Then we can use a similar proof as the proof of Proposition 6 and Proposition 7 to establish an upper bound for the kth moment of $Y_{b,\alpha}$. Towards the same, we use an additional factor of α for each backtracking edge in the upper bounds, and we use the lower bound on such edges from (5.39). This gives us the following bound for the summands corresponding to paths with m marked edges and $r_*(m)$ Type I segments:

$$\begin{aligned} E'_m \leqslant C' \binom{2k}{r_*(m)} (2k)^{f(r_*(m))-m} 3^{2k-r_*(m)} C^{f(r_*(m))-m} \\ \times U_b^{s(m)} n^{m-ts(m)-(r_*(m)-s(m))} \rho_n^{m-ts(m)} \\ \times (2m)^{(2k-r_*(m))\cdot(t-2h)+2(f(r_*(m))-m)} \times \alpha^b. \end{aligned}$$

where $s(m)$ is the number of saturated paths with t marked edges.

Next, we show that $E_{m,r}$ is much smaller than $E_{m,r_*(m)}$. For this, from (5.39) we note that if we increase the number of Type I segments by $(r - r_*(m))$, then the lower bound on the number of backtracking edges reduces by $(r - r_*(m))t$ edges. With this observation and the proof of Proposition 7 in Section 5.3.1.2 we have

$$\sum_{r=r_*(m)}^{r^*(m)} E_{m,r} = E_{m,r_*(m)} \sum_{r=r_*(m)}^{r^*(m)} \left(\frac{O(k^{t+1})}{n\alpha^t}\right)^{r-r_*(m)} = E_{m,r_*(m)}(1 + o(1)).$$

This implies that

$$\sum_{m<kt} E_m \leqslant C \sum_{m<kt} E_{m,r_*(m)} = o(U_b^k).$$

This completes the proof for Regime II.

■

5.3.3 Concentration of path counts

We will prove the following concentration result for $Y_{b,\alpha}$.

Proposition 15. *Let $t_L = t_U = t \geqslant 3$ be given and $k = \lceil \log n \rceil$. Suppose that* (4.18) *and* (4.19) *holds for Regime I,* (4.20) *and* (4.21) *holds for Regime II, and $\alpha = O(1/n)$ and* (4.22) *holds for Regime III. Then we have*

$$\mathbb{P}\left(|Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha}| > \delta \mathbb{E}Y_{b,\alpha}\right) = O(n^{-3}),$$

where $\delta = \Theta((\log n)^{-\eta})$ for some $\eta > 0$.

Proof. Recall the definition of $X_{p,\alpha}$ from (5.34). By Markov's inequality, and using Proposition 12,

$$\mathbb{P}\left(|Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha}| > \delta \mathbb{E}Y_{b,\alpha}\right) \leqslant \frac{\mathbb{E}(Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha})^{2k}}{\delta^{2k}\left(\mathbb{E}Y_{b,\alpha}\right)^{2k}} \leqslant \frac{\mathbb{E}(Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha})^{2k}}{\delta^{2k}(L_b)^{2k}}, \tag{5.40}$$

and moreover,

$$\mathbb{E}(Y_{b,\alpha}-\mathbb{E}Y_{b,\alpha})^{2k}=\mathbb{E}\Big(\sum_{p\in\mathcal{P}_b}(X_{p,\alpha}-\mathbb{E}X_{p,\alpha})\Big)^{2k}=\sum_{(p_1,p_2,\ldots,p_{2k}):p_l\in\mathcal{P}_b}\mathbb{E}\prod_{l=1}^{2k}(X_{p_l,\alpha}-\mathbb{E}X_{p_l,\alpha}),$$
$$=\sum_{(p_1,p_2,\ldots,p_{2k}):p_l\in\mathcal{P}_b}\prod_{l=1}^{2k}\alpha^{N(p_l)}\cdot\mathbb{E}\prod_{l=1}^{2k}(X_{p_l,1}-\mathbb{E}X_{p_l,1}). \tag{5.41}$$

where $X_{p_l,1}$ is obtained by plugging $\alpha=1$. We note that for any path p, $X_{p,1}=X_p$ where X_p is as defined in (5.3). With the observation (5.41), we bound as in the proof of Proposition 8 to complete the proof. There are three regimes. We compute the bounds for each of the three regimes below.

Proof for Regime I. We use the bounds in the proof of Proposition 14 to have the following bounds for $m\leqslant 2kt-k$ marked edges:

$$\begin{aligned}E'_m&\leqslant C'\binom{2k}{r_*(m)}(2k)^{f(r_*(m))-m}3^{2k-r_*(m)}C^{f(r_*(m))-m}\\&\quad\times U_b^{s(m)}n^{m-ts(m)-(r_*(m)-s(m))}\rho_n^{m-ts(m)}\\&\quad\times(2m)^{(2k-r_*(m))\cdot(t-2h)+2(f(r_*(m))-m)}\times\alpha^{2kt-m}.\end{aligned}$$

Let m_0 be such that $m_0=f(r_*(m))$ and define E'_{m_0} with the same expression as above. Then bounding as in the proof of Proposition 14 we have from (5.40)

$$\mathbb{E}(Y_{b,\alpha}-\mathbb{E}Y_{b,\alpha})^{2k}\leqslant CE'_{m_0}.$$

Thus, (5.41) together with the fact that $\frac{U_b}{L_b}=1+O(\frac{k}{n})$ yields

$$\begin{aligned}\mathbb{P}\left(|Y_{b,\alpha}-\mathbb{E}Y_{b,\alpha}|>\delta\mathbb{E}Y_{b,\alpha}\right)&\leqslant\frac{\binom{2k}{r_*(m)}3^{2k-r_*(m)}U_b^{r_*(m)}(n\rho_n)^{m-tr_*(m)}(2m)^{(2k-r_*(m))\cdot(t-2h)}\alpha^{2kt-m}}{\delta^{2k}L_b^{2k}}\\&\leqslant C\left(\frac{C(\log n)^{1+t-2h-2\eta(t-h)}(n\rho_n)^h\alpha^{t-h}}{n^{t-1}\rho_n^t}\right)^{\lceil\log n\rceil(t-h)^{-1}}\leqslant n^{-c},\end{aligned}$$

for any $c>0$ when (4.18) and (4.19) hold. This completes the proof.

Proof for Regime II. We use the bounds in the proof of Proposition 14 to have the following bounds for $m \leqslant 2kt - k$ marked edges:

$$\begin{aligned} E'_m &\leqslant C' \binom{2k}{r_*(m)} (2k)^{f(r_*(m))-m} 3^{2k-r_*(m)} C^{f(r_*(m))-m} \\ &\quad \times U_b^{s(m)} n^{m-ts(m)-(r_*(m)-s(m))} \rho_n^{m-ts(m)} \\ &\quad \times (2m)^{(2k-r_*(m))\cdot(t-2h)+2(f(r_*(m))-m)} \times \alpha^b. \end{aligned}$$

Let m_0 be such that $m_0 = f(r_*(m))$ and define E'_{m_0} with the same expression as above. Then bounding as in the proof of Proposition 14 we have from (5.40)

$$\mathbb{E}(Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha})^{2k} \leqslant C E'_{m_0}.$$

Thus, (5.41) together with the fact that $\frac{U_b}{L_b} = 1 + O(\frac{k}{n})$ yields

$$\begin{aligned} \mathbb{P}\left(|Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha}| > \delta \mathbb{E}Y_{b,\alpha}\right) &\leqslant \frac{\binom{2k}{r_*(m)} 3^{2k-r_*(m)} U_b^{r_*(m)} (n\rho_n)^{m-tr_*(m)} (2m)^{(2k-r_*(m))\cdot(t-2h)} \alpha^b}{\delta^{2k} L_b^{2k}}, \\ &\leqslant C \left(\frac{C(\log n)^{1+t-2h-2\eta(t-h)} (n\rho_n)^h \alpha^h}{n^{t-1}\rho_n^t} \right)^{\lceil \log n \rceil (t-h)^{-1}} \leqslant n^{-c}, \end{aligned}$$

for any $c > 0$ when (4.20) and (4.21) hold. This completes the proof.

Proof for Regime III. We use the bounds in the proof of Proposition 13 to have the following bounds for $m \leqslant 2kt - k$ marked edges:

$$E'_m \leqslant C \binom{2k}{r_*(m)} (2k)^{tr_*(m)-m} \times U_b^{s(m)} n^{m-s(m)t-(r_*(m)-s(m))} \rho_n^{m-s(m)t} \times (2m)^{2kt-m}.$$

Let m_0 be such that $m_0 = f(r_*(m))$ and define E'_{m_0} with the same expression as above. Then bounding as in the proof of Proposition 13 we have from (5.40)

$$\mathbb{E}(Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha})^{2k} \leqslant C E'_{m_0}.$$

Thus, (5.41) together with the fact that $\frac{U_b}{L_b} = 1 + O(\frac{k}{n})$ yields

$$\begin{aligned}
\mathbb{P}\left(|Y_{b,\alpha} - \mathbb{E}Y_{b,\alpha}| > \delta \mathbb{E}Y_{b,\alpha}\right) &\leqslant \frac{\binom{2k}{r_*(m)} U_b^{r_*(m)} (2m)^{2kt-m_0}}{\delta^{2k} L_b^{2k}}, \\
&\leqslant \left(\frac{U_b}{L_b}\right)^{r_*(m)} \cdot \left(\frac{Ck \cdot k^t \delta^{-\frac{2k}{2k-r_*(m)}}}{n^{t-1}\rho_n^t}\right)^{2k-r_*(m)}, \\
&\leqslant C\left(\frac{C(\log n)^{1+t-2\eta t}}{n^{t-1}\rho_n^t}\right)^{\lceil \log n \rceil (t)^{-1}} \leqslant n^{-c},
\end{aligned}$$

for any $c > 0$ when (4.22) holds. This completes the proof.

■

5.4 Analysis of spectral clustering for node2vec

We analyze the matrix M_0 and the eigendecomposition of M in section 5.4.1. We then prove Theorem 5 in section 5.4.2. We then end with the proof of Theorem 6.

5.4.1 Analysis of M-matrix

We start with the proof of Lemma 17 which shows that M_0 has an approximate block structure. This then leads to the proof of Proposition 5. Using these two results, in Lemma 18 we then provide bounds for the inner products of rows of V_0 similar to the bounds given for DeepWalk in Proposition 9c.

Now we show that M_0 has a block structure when $t_L \geqslant 3$ but not when $t_L = 2$. We also show that the entries of M_0 are $O(1)$ when $t_L \geqslant 3$.

Lemma 17. *Suppose that $t_L \geqslant 2$. Then we have*

$$M_0 = M_0(\alpha_n) = \log\left(\sum_{t=t_L}^{t_U} \Theta_0 G_{t,n} \Theta_0^T + R_{t,n}\right),$$

where $G_{t,n}$ is a $K \times K$ matrix and $R_{t,n}$ is a diagonal matrix. The following hold for the decomposition above

1. *$R_{t,n} = 0$ when t is odd, and $(R_{t,n})_{ii} = O(n^{-\frac{t}{2}})$ when $t > 2$ and t is even.*

2. *If $t_L > 2$ or if $t_L = 2$ and $i \neq j$: $(G_{t,n})_{ij} = O(1)$ and $M_{ij} = O(1)$ (as a function of n).*

3. *If $t_L = 2$ and $i = j$: $M_{ii} \to \infty$ if $\alpha \to \infty$, $M_{ii} \to -\infty$ if $\alpha \to 0$, and $M_{ii} = O(1)$ if $\alpha = \theta(1)$. Further, in this case we also have $(G_{t,n})_{ii} = 0$.*

Proof of Lemma 17. Let

$$P_{ij}^{(t)} = \sum_{(i_0,i_1,\ldots,i_t):i_0=i,i_t=j} P_{i_0 i_1} P_{i_1 i_2} \cdots P_{i_{t-1} i_t} \frac{1}{|P|} \left(\prod_{l=1}^{t-1} \frac{1}{|P_{i_l \star}| - 1 + \alpha} \right) \cdot \alpha^{\mathfrak{n}((i_0,i_1,\ldots,i_t))}, \tag{5.42}$$

be the t-step transition probability for node2vec. Then

$$(M_0)_{ij} = \log \left(\frac{\sum_{t=t_L}^{t_U} (l-t) \cdot (P_{ij}^{(t)} + P_{ji}^{(t)})}{2b\gamma(l, t_L, t_U) \frac{|P_{i\star}|}{|P|} \times \frac{|P_{j\star}|}{|P|}} \right). \tag{5.43}$$

And we define M_0' as follows.

$$(M_0')_{ij} := \left(\frac{\sum_{t=t_L}^{t_U} (l-t) \cdot (P_{ij}^{(t)} + P_{ji}^{(t)})}{2b\gamma(l, t_L, t_U) \frac{|P_{i\star}|}{|P|} \times \frac{|P_{j\star}|}{|P|}} \right). \tag{5.44}$$

We first describe a decomposition of M_0'. Towards this, for any sequence $k = (k_0, k_1, \ldots, k_{m+1})$ such that $0 = k_0 < k_1 < \cdots < k_m < k_{m+1} = t$, and any path type $b = (b_0, b_1, \ldots, b_t)$ such that $b_0 = g(i)$ and $b_t = g(j)$ we define

$$\begin{aligned} E_{b,k,i,j,t} := \sum_{(i_0,i_1,\ldots,i_t)|i_0=i,i_t=j} & \mathbb{1}\{(g(i_0), g(i_1), \ldots, g(i_t)) = (b_0, b_1, \ldots, b_t)\} \\ & \mathbb{1}\{(i_l, i_{l+1}) \text{ is a backtrack for } k_r < l < k_{r+1}, (i_{k_r}, i_{k_{r+1}}) \text{ is not a backtrack for } 1 \leqslant r \leqslant m\} \\ & P_{i_0 i_1} P_{i_1 i_2} \cdots P_{i_{t-1} i_t} \frac{1}{|P|} \left(\prod_{l=1}^{t-1} \frac{1}{|P_{i_l \star}| - 1 + \alpha} \right) \cdot \alpha^{\mathfrak{n}((i_0,i_1,\ldots,i_t))}. \end{aligned} \tag{5.45}$$

$E_{b,k,i,j,t}$ is a sum over paths of length t between nodes i and j with locations of edges which are not backtracks given by the sequence k and the block labels of the vertices along the paths given by the sequence b. We can note that

$$P_{ij}^{(t)} = \sum_k \sum_b E_{b,k,i,j,t}. \tag{5.46}$$

For a given t, consider a sequence $k = (k_0, k_1, \ldots, k_{m+1})$ such that $k_{r+1} - k_r$ is odd for some r where $1 \leqslant r \leqslant m$. This implies that the endpoints of the path do not have an equality constraint between them. From this we can see that the summands $E_{b,k,i,j}$ in (5.45) depend on i and j only through the block types $g(i)$ and $g(j)$.

On the other hand, consider a sequence k such that $k_{r+1} - k_r$ is even for $1 \leqslant r \leqslant m$. In this case we have the equality constraint $i = j$. We also have that t is even and $m \leqslant t/2$. By Lemma 6, there are $O(n^{\frac{t}{2}})$ paths associated with the sequence k.

Then we define

$$
\begin{aligned}
(N_{t,n})'_{ij} &:= \sum E_{b,k,i,j,t} \mathbb{1}\{k = (k_0, k_1, \ldots, k_{m+1}), k_{r+1} - k_r \text{ is odd for some } 1 \leqslant r \leqslant m, k_{m+1} = t\}, \\
(R_{t,n})'_{ij} &:= \sum E_{b,k,i,j,t} \mathbb{1}\{k = (k_0, k_1, \ldots, k_{m+1}), k_{r+1} - k_r \text{ is even for } 1 \leqslant r \leqslant m, k_{m+1} = t\}, \\
(N_{t,n})_{ij} &:= \left(\frac{\sum_{t=t_L}^{t_U} (l-t) \cdot \left((N_{t,n})'_{ij} + (N_{t,n})'_{ji} \right)}{2b\gamma(l, t_L, t_U) \frac{|P_{i\star}|}{|P|} \times \frac{|P_{j\star}|}{|P|}} \right), \\
(R_{t,n})_{ij} &:= \left(\frac{\sum_{t=t_L}^{t_U} (l-t) \cdot \left((R_{t,n})'_{ij} + (R_{t,n})'_{ji} \right)}{2b\gamma(l, t_L, t_U) \frac{|P_{i\star}|}{|P|} \times \frac{|P_{j\star}|}{|P|}} \right).
\end{aligned} \tag{5.47}
$$

Then $M'_0 = N_{t,n} + R_{t,n}$ by (5.44) and (5.46). Further, by the discussion in the previous paragraph $N_{t,n}$ is a block matrix with the same block structure as P, $R_{t,n} = 0$ when t is odd and when t is even and $i \neq j$, and $(R_{t,n})_{ij} = O(n^{-\frac{t}{2}})(N_{t,n})_{ij}$ when t is even, $t > 2$ and $i = j$. In the case when $t = 2$ and $i = j$, we have $(N_{t,n})_{ij} = 0$.

Now we describe the order of the coefficients of M_0 and N. We note that if $\alpha = 1$, we can see that $M_0(1)$ is the matrix in the DeepWalk in which case by Proposition 9 we have $(M_0(1))_{ij} = O(1)$. Thus more generally $(M_0)_{ij} = O(1)$ iff for all $t_L \leqslant t \leqslant t_U$

$$
\frac{\sum_{(i_0, i_1, \ldots, i_t): i_0 = i, i_t = j} P_{i_0 i_1} P_{i_1 i_2} \cdots P_{i_{t-1} i_t} \frac{1}{|P|} \left(\prod_{l=1}^{t-1} \frac{1}{|P_{i_l \star}| - 1 + \alpha} \right) \cdot \alpha^{\mathfrak{n}((i_0, i_1, \ldots, i_t))}}{\sum_{(i_0, i_1, \ldots, i_t): i_0 = i, i_t = j} P_{i_0 i_1} P_{i_1 i_2} \cdots P_{i_{t-1} i_t} \frac{1}{|P|} \left(\prod_{l=1}^{t-1} \frac{1}{|P_{i_l \star}|} \right)} = \theta(1). \tag{5.48}
$$

Towards this note that $\frac{p_{i_l}}{p_{i_l} - 1 + \alpha} \to 1$ as $p_{i_l} = \theta(n\rho_n)$. Next consider paths between i and j without any backtracks i.e. $\mathfrak{n}((i_0, i_1, \ldots, i_t)) = 0$. There are $\theta(n^{t-1})$ such paths if $t > 2$ and if $t = 2$ and $i \neq j$. Now consider paths with $\mathfrak{n}((i_0, i_1, \ldots, i_t)) = s > 0$ and consider the case $t > 2$. If we consider Type I paths with s backtracks, then there are $O(n^{t-s-1})$ such paths and combined

with the factor of α^s from backtracks, the total contribution is $O(n^{t-s-1}\alpha^s) = o(n^{t-1})$ by (4.18). If we consider Type II paths with s backtracks, then there are $O(n^h)$ such paths and combined with the factor from backtracks, the total contribution is $O(n^h\alpha^h)$ which is $o(n^{t-1})$ by (4.18). This implies that the fraction in (5.48) tends to 1. We can note that the leading term, i.e. paths with $\mathfrak{n}((i_0, i_1, \ldots, i_t)) = 0$, is a summand in the definition of $(N')_{t,n}$ in (5.47). Thus $(N'_{t,n})_{ij} = \theta(1)$.

Now consider the case when $t = 2$. If $t = 2$ and $i \neq j$, there cannot be a backtrack and so the leading contribution is from the case when $\mathfrak{n}((i_0, i_1, \ldots, i_t)) = 0$. Thus, the fraction in (5.48) tends to 1. In this case $(R_{2,n})'_{ij} = 0$ and $(N_{2,n})'_{ij} = O(1)$. If $t = 2$ and $i = j$ then the fraction in (5.48) tends to 0 when $\alpha \to 0$, tends to 0 if $\alpha \to 0$, and is $\theta(1)$ if $\alpha = \theta(1)$. In this case $(N_{2,n})'_{ii} = 0$. This completes the proof.

■

Now we are ready to prove Proposition 5.

Proof of Proposition 5. Let $\tilde{\log} : [0, \infty] \to \mathbb{R}$ be defined by

$$x \mapsto \log x, \quad x > 0,$$
$$x \mapsto 0, \quad x = 0.$$

To simplify notation, we will write log for $\tilde{\log}$ at all places in the proof. By Lemma 17, $(\log(Z_0))_{ij}$ is $O(1)$. So by approximating log around 1 we have

$$\begin{aligned} M_0 &= \log\left(\Theta Z_0 \Theta^T + R\right), \\ &= \Theta \log Z_0 \Theta^T + R', \end{aligned}$$

where R' is a diagonal matrix and $R'_{ii} = O(n^{-2})$ as we consider $t_L > 2$ for node2vec. By Proposition 9b (applied to $\Theta \log Z_0 \Theta^T$) and since $\log Z_0$ has rank K, $\Theta \log Z_0 \Theta^T$ has K non-zero eigenvalues, $\tilde{\lambda}_i$, where $\tilde{\lambda}_i = \theta(n)$. By Weyl's theorem on eigenvalues, M_0 has rank K and each of the K non-zero eigenvalues is $\theta(n)$. Now let v_i, $1 \leqslant i \leqslant K$ be the columns (i.e. eigenvectors of M_0) in V_0.

Then

$$\lambda_i v_i = M_0 v_i = \Theta \log Z_0 \Theta^T v_i + R' v_i, \tag{5.49}$$
$$\implies v_i = \Theta \left(\lambda_i^{-1} \log Z_0 \Theta^T v_i\right) + \lambda_i^{-1} R' v_i.$$

Let $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_K)$. Then taking $X_0 = \log Z_0 \Theta^T V_0 \Lambda^{-1}$ and $E_0 = R' V \Lambda^{-1}$ completes the proof. ■

Next, we compute bounds on the inner products of rows of V_0. The bounds are similar to the DeepWalk case except that we have small error terms.

Lemma 18. *Consider the decomposition $M_0 = \log\left(\Theta Z_0 \Theta^T + R\right)$. Let $V_0 \in \mathbb{R}^{n \times K}$ be the matrix of top K left singular vectors of M_0 and let $V_0 = \Theta X_0 + R_0$ be the decomposition from Proposition 5. Then*

$$\langle (V_0)_{i\star}, (V_0)_{j\star} \rangle = \mathbb{1}\{g_i = g_j\} \left(\Theta \left(\frac{1}{\sqrt{n_1}}, \frac{1}{\sqrt{n_2}}, \ldots, \frac{1}{\sqrt{n_K}} \right)^T \right)_i + O(n^{-2.5}).$$

If i and j are two nodes such that $g_i \neq g_j$, then we have

$$\left\| (V_0)_{i\star} - (V_0)_{j\star} \right\|_{\mathrm{F}} = \sqrt{\frac{1}{n_{g_i}} + \frac{1}{n_{g_j}}} + O(n^{-3}).$$

Proof of Lemma 18. As shown in proof of Proposition 5

$$M_0 = \Theta \log Z_0 \Theta^T + R',$$

where R' is a diagonal matrix and $R'_{ii} = O(n^{-2})$. Let $V \in \mathbb{R}^{n \times K}$ be matrix of top K left singular vectors of $N = \Theta \log Z_0 \Theta^T$. Then by Proposition 9bc (applied to N), $V = \Theta X$ satisfying the following:

1. $X \in \mathbb{R}^{K \times K}$ and has K distinct rows. And so, V has K distinct rows.
2. Rows of X are orthogonal and the row norms are given by $\Theta \left(\frac{1}{\sqrt{n_1}}, \frac{1}{\sqrt{n_2}}, \ldots, \frac{1}{\sqrt{n_K}} \right)^T$.
3. Let $\tilde{\lambda}_i$ for $1 \leqslant i \leqslant K$ be the top K eigenvalues. Then $\tilde{\lambda}_i = \theta(n)$.

Let λ_i be the top K eigenvalues of V_0. By Weyl's theorem on eigenvalues, $|\lambda_i - \tilde{\lambda}_i| \leqslant O(n^{-1.5})$. Let $\hat{E}_i$ (resp. E_i) be the eigenspace corresponding to $\tilde{\lambda}_i$ (resp. λ_i). Let $V_{\hat{E}}$ and $(V_0)_E$ be the eigenvectors corresponding to the eigenspaces. Then by Davis-Kahan theorem there exists O such that

$$\left\|(V_0)_E - V_{\hat{E}}O\right\|_{\mathrm{F}} \leqslant \frac{\|R'\|_{\mathrm{F}}}{\theta(n)} = O(n^{-2.5}).$$

We can note that if we replace $V_{\hat{E}}$ by $V_{\hat{E}}O$, then we continue to have the three properties for V listed above. As a consequence, $\|V - V_0\|_{\mathrm{F}} = O(n^{-2.5})$ which implies the first part of the result.

For the second part we have for any i, j such that $g(i) \neq g(j)$

$$\begin{aligned}
\|(V_0)_{i\star} - (V_0)_{j\star}\|_{\mathrm{F}} &= \sqrt{\|(V_0)_{i\star}\|_{\mathrm{F}}^2 + \|(V_0)_{j\star}\|_{\mathrm{F}}^2 - 2\langle (V_0)_{i\star}, (V_0)_{j\star}\rangle}, \\
&= \sqrt{\frac{1}{n_{g_i}} + \frac{1}{n_{g_j}} + O(n^{-2.5})}, \\
&= \sqrt{\frac{1}{n_{g_i}} + \frac{1}{n_{g_j}}} + O(n^{-3}).
\end{aligned}$$

■

5.4.2 Bound on $\|M - M_0\|_{\mathrm{F}}$.

We provide a proof of Theorem 5 in this section.

Proof of Theorem 5. To begin the proof we can note that $M_{ij} = 0$ iff $\sum_{t=t_L}^{t_U} A_{ij}^{(t)} = 0$. Using the same arguments as in the proof of Proposition 11 and Theorem 3 along with the lower tail inequality in Proposition 8 we can conclude that

$$\mathbb{P}\left(M_{ij} = 0 \text{ for some } i, j\right) = o(1).$$

Thus we can assume that $M_{ij} \neq 0$ for the rest of the proof.

Let $a_n = 4n(\log n)^{-\eta}$ and let $b_{n,1} = \exp\left\{\frac{a_n}{\sqrt{2n}}\right\}$ and $b_{n,2} = \exp\left\{\frac{a_n}{\sqrt{2n}}\right\}$. Recall from (5.42) that $P_{ii}^{(t)}$ is the t-step transition probability for node2vec, and recall from (5.43) the form of M_0 for

node2vec. Then we have

$$\begin{aligned}
&\mathbb{P}\left(\|M - M_0\|_{\mathrm{F}} \geqslant a_n\right) = \mathbb{P}\left(\sum_{1\leqslant i,j\leqslant n} \log^2\left(\frac{M_{ij}}{(M_0)_{ij}}\right) \geqslant a_n^2\right), \\
&\leqslant o(1) + \sum_{1\leqslant i\neq j\leqslant n}\sum_{t=t_L}^{t_U} \mathbb{P}\left(\frac{\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = j\right)\frac{|P_{i\star}|}{|P|}\times\frac{|P_{j\star}|}{|P|}}{P_{ij}^{(t)}\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right)\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = j\right)} \geqslant b_{n,1}\right) \\
&\quad + \sum_{1\leqslant i\neq j\leqslant n}\sum_{t=t_L}^{t_U} \mathbb{P}\left(\frac{P_{ij}^{(t)}\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right)\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = j\right)}{\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = j\right)\frac{|P_{i\star}|}{|P|}\times\frac{|P_{j\star}|}{|P|}} \geqslant b_{n,1}\right) \\
&\quad + \sum_{1\leqslant i\leqslant n}\sum_{t=t_L}^{t_U} \mathbb{P}\left(\frac{\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = i\right)\frac{|P_{i\star}|}{|P|}\times\frac{|P_{i\star}|}{|P|}}{P_{ii}^{(t)}\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right)\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right)} \geqslant b_{n,2}\right) \\
&\quad + \sum_{1\leqslant i\leqslant n}\sum_{t=t_L}^{t_U} \mathbb{P}\left(\frac{P_{ii}^{(t)}\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right)\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right)}{\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = i\right)\frac{|P_{i\star}|}{|P|}\times\frac{|P_{i\star}|}{|P|}} \geqslant b_{n,2}\right). \qquad (5.50)
\end{aligned}$$

Fix $t_L \leqslant t \leqslant t_U$ and $i \neq j$. We show how to bound a typical term in the third summand in (5.50). Recall from (5.2) that $\mathcal{P}_b$ is the set of paths with vertices having community assignment b for $b \in \mathcal{B}_{i,j}$. Similar to the proof of Proposition 11, for $p = (i_0, \ldots, i_t) \in \mathcal{P}_b$, let

$$\begin{aligned}
\bar{X}_{p,\alpha} &= \frac{A_{i_0 i_1}}{|A_{i_0\star}|}\frac{1}{|A_{i_t\star}|}\left(\prod_{l=1}^{t-1}\frac{A_{i_l i_{l+1}}}{|A_{i_l\star}| - 1 + \alpha}\right)\cdot\alpha^{\mathfrak{n}(p)}, \quad \text{and} \quad \bar{Y}_{b,\alpha} = \sum_{p\in\mathcal{P}_b}\bar{X}_{p,\alpha}, \qquad (5.51)\\
\bar{X}^*_{p,\alpha} &= \frac{P_{i_0 i_1}}{|P_{i_0\star}|}\frac{1}{|P_{i_t\star}|}\left(\prod_{l=1}^{t-1}\frac{P_{i_l i_{l+1}}}{|P_{i_l\star}| - 1 + \alpha}\right)\cdot\alpha^{\mathfrak{n}(p)}, \quad \text{and} \quad \bar{Y}^*_{b,\alpha} = \sum_{p\in\mathcal{P}_b}\bar{X}^*_{p,\alpha}.
\end{aligned}$$

Then we can write

$$\begin{aligned}
&\mathbb{P}\left(\frac{P_{ij}^{(t)}\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right)\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = j\right)}{\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = j\right)\frac{|P_{i\star}|}{|P|}\times\frac{|P_{j\star}|}{|P|}} \geqslant b_{n,1}\right) \\
&= \mathbb{P}\left(\frac{|P|}{|A|}\frac{\sum_{b\in\mathcal{B}_{i,j}}\bar{Y}^*_{b,\alpha}}{\sum_{b\in\mathcal{B}_{i,j}}\bar{Y}_{b,\alpha}} \geqslant b_{n,1}\right) \leqslant \sum_{b\in\mathcal{B}_{i,j}}\mathbb{P}\left(\frac{|P|}{|A|}\frac{\bar{Y}^*_{b,\alpha}}{\bar{Y}_{b,\alpha}} \geqslant b_{n,1}\right) + o(n^{-4}). \qquad (5.52)
\end{aligned}$$

The last term $o(n^{-4})$ above, obtained by an application of Proposition 15, is a bound on the probability $\mathbb{P}(X_b = 0)$ so that $\frac{\bar{Y}^*_{b,\alpha}}{\bar{Y}_{b,\alpha}}$ is well-defined for the rest of the arguments below. We now show how to bound the typical summand in (5.52). Towards this we use the Chernoff bound as in the proof of Proposition 11 to bound $\frac{|P|}{|A|}$. We next show how to bound the fractions coming from

the degree terms in (5.51). Towards this we note that for any $0 < \delta < 1$ and for any $l \in [n]$ we have

$$\begin{aligned}\mathbb{P}\left(|A_{i_l\star}| - 1 + \alpha \leqslant (1-\delta)(|P_{i_l\star}| - 1 + \alpha)\right) &= \mathbb{P}\left(|A_{i_l\star}| \leqslant \left((1-\delta) + \frac{\delta(1-\alpha)}{|P_{i_l\star}|}\right)|P_{i_l\star}|\right), \\ &\leqslant \mathbb{P}\left(|A_{i_l\star}| \leqslant \left(1 - \frac{\delta}{2}\right)|P_{i_l\star}|\right),\end{aligned}$$

as $|P_{i_l\star}| = \theta(n\rho_n)$. By similar reasoning we also have for $0 < \delta < 1$ and for any $l \in [n]$

$$\mathbb{P}\left(|A_{i_l\star}| - 1 + \alpha \geqslant (1+\delta)(|P_{i_l\star}| - 1 + \alpha)\right) \leqslant \mathbb{P}\left(|A_{i_l\star}| \geqslant \left(1 + \frac{\delta}{2}\right)|P_{i_l\star}|\right).$$

These inequalities combined with the Chernoff bound help bound $\mathbb{P}\left(\frac{|P|}{|A|}\frac{\bar{Y}^*_{b,\alpha}}{\bar{Y}_{b,\alpha}} \geqslant b_{n,1}\right)$ as follows. Define

$$X^*_{p,\alpha} = \prod_{l=1}^{t} P_{i_{l-1}i_l}\alpha^{\mathfrak{n}}, \quad \text{and} \quad Y^*_{b,\alpha} = \sum_{p\in\mathcal{P}_b} X^*_{p,\alpha}.$$

Then we have

$$\begin{aligned}\mathbb{P}\left(\frac{|P|}{|A|}\frac{\bar{Y}^*_{b,\alpha}}{\bar{Y}_{b,\alpha}} \geqslant b_{n,1}\right) &\leqslant \mathbb{P}\left(\frac{Y^*_{b,\alpha}}{Y_{b,\alpha}} \geqslant \exp\left\{\frac{a_n}{\sqrt{2}n} - \theta\left(\frac{1}{\sqrt{n}}\right) - \theta\left(\sqrt{\frac{\log n}{n\rho_n}}\right)\right\}\right) + O(n^{-4}), \\ &\leqslant \mathbb{P}\left(\frac{Y^*_{b,\alpha}}{Y_{b,\alpha}} \geqslant 1 + (\log n)^{-\eta}\right) + O(n^{-4}), \qquad (5.53)\end{aligned}$$

as $a_n = 4n(\log n)^{-\eta}$.

Fix (i, j) such that $i \neq j$. We estimate the difference

$$\left|\mathbb{E}Y_{b,\alpha} - Y^*_{b,\alpha}\right| = \left|\sum_{p\in\mathcal{P}_b}\left(\mathbb{E}A_{i_0i_1}\cdots A_{i_{t-1}i_t} - P_{i_0i_1}\cdots P_{i_{t-1}i_t}\right)\alpha^{\mathfrak{n}}\right|. \qquad (5.54)$$

If $\alpha \leqslant 1$, we have the right hand side in (5.54) to be $O\left(\frac{1}{n\rho_n}\right)\mathbb{E}Y_{b,\alpha}$ from (5.26) in the proof of Proposition 11 for DeepWalk, and as both $\mathbb{E}Y_{b,\alpha}$ and $\mathbb{E}Y_{b,1}$ are $\theta(n^{t-1}\rho_n^t)$ by Proposition 12 and Proposition 14. We now look at the case when $\alpha \geqslant 1$. The summands in (5.54) are equal to zero if the associated path $(i_0, i_1, \ldots, i_t)$ has t distinct edges and there are no self-loops. Consider the first set of summands in (5.54). Consider Type I paths with $s < t$ marked edges. Then there are

$O(n^{s-1})$ such paths and their total contribution is $O(n^{s-1}\rho_n^s\alpha^{t-s}) = O\left(\frac{\alpha}{n\rho_n}\right)U_b$. Now consider Type II paths with s marked edges. Recall that the path starts at node i and ends at node j. Then let $l > 0$ be the first index such that $i_l = j$. Then i_l must be an endpoint of a marked edge. We note that i_l cannot be the last vertex of a segment as Type II segments start and end at the same vertex and l is the first index such that $i_l = j$. This implies that the choice of the vertex i_l is determined to be equal to j. Thus there are $O(n^{s-1})$ paths of this type and by a similar calculation as for Type I paths, the overall contribution is $O\left(\frac{\alpha}{n\rho_n}\right)U_b$.

We now give an upper bound on the summands $P_{i_0 i_1}\cdots P_{i_{t-1}i_t}$ as follows. Suppose a path has less than t distinct edges. If the path is a Type I path with s backtracks then by Lemma 6 the number of choices of distinct vertices along the path is at most $t-s-1$. Thus the contribution from such paths is $O\left(\frac{\alpha}{n}\right)\mathbb{E}Y_{b,\alpha}$. If the the path is a Type II path with s marked edges, then by reasoning as in the previous paragraph the number of paths of this type are $O(n^{s-1})$ and so the $O\left(\frac{\alpha}{n\rho_n}\right)U_b$.

Finally, if there are self-loops then the number of choices of vertices is less than $t-1$. This implies that the upper bound on the second set of summands is $O(1/n)\mathbb{E}Y_{b,\alpha}$. Thus in summary we have

$$\begin{aligned}\left|\mathbb{E}Y_{b,\alpha} - Y^*_{b,\alpha}\right| &= O\left(\frac{1}{n\rho_n}\right)\mathbb{E}Y_{b,\alpha}, \quad i \neq j, \alpha = O(1),\\ \left|\mathbb{E}Y_{b,\alpha} - Y^*_{b,\alpha}\right| &= O\left(\frac{\alpha}{n\rho_n}\right)\mathbb{E}Y_{b,\alpha}, \quad i \neq j, \alpha \gg 1.\end{aligned} \tag{5.55}$$

These computations show that in particular for $i \neq j$ we have

$$\frac{Y_{b,\alpha}}{Y^*_{b,\alpha}} = \frac{Y_{b,\alpha}}{\mathbb{E}Y_{b,\alpha}\left(1+O((n\rho_n)^{-1})\right)}, \quad \text{and} \quad \frac{Y^*_{b,\alpha}}{Y_{b,\alpha}} = \frac{\mathbb{E}Y_{b,\alpha}\left(1+O((n\rho_n)^{-1})\right)}{Y_{b,\alpha}}. \tag{5.56}$$

With this bound we complete our calculation from (5.53). We use Proposition 15 and (5.56) to conclude that

$$\mathbb{P}\left(\frac{Y^*_{b,\alpha}}{Y_{b,\alpha}} \geqslant 1 + (\log n)^{-\eta}\right) = O(n^{-3}).$$

This completes the argument to show that the last term in (5.50) goes to 0. The second term in (5.50) can be bounded similarly. The third and the fourth term in (5.50) are for the case when $i = j$. Again, the proof is similar except we use the following computations to bound $\frac{\mathbb{E}Y_{b,\alpha}}{Y^*_{b,\alpha}}$. Recall that $\mathbb{E}Y_{b,\alpha} = \theta(n^{t-1}\rho_n^t)$. We estimate $Y^*_{b,\alpha}$. Suppose a path has $m < t$ distinct edges and b backtracks. Then by Lemma 6, the number of such paths is $O(n^m)$ and so the contribution from such paths is $O(n^m\rho_n^t\alpha^b) \ll O((n\rho_n)^t)$ as $m + b \leqslant t$ and $\alpha \ll n\rho_n$. This shows that

$$\frac{1}{n} \ll \frac{\mathbb{E}Y_{b,\alpha}}{Y^*_{b,\alpha}} \ll n, \quad i = j. \tag{5.57}$$

This completes the first part of the proof.

The second part of the proof, for the range $n^{t_U-1}\rho_n^{t_U} \ll 1$, is similar to the analogous proof in the proof of Theorem 3 except the following two changes:

1. Lemma 17 is used to show that the entries of M_{ij} are $O(1)$.
2. We use Proposition 12 to bound $\mathbb{E}Y_{b,\alpha}$.

■

5.4.3 Bounding the number of missclassified nodes

We provide a proof of Theorem 6 in this section.

Proof of Theorem 6. For this proof, we choose $O \in \mathbb{R}^{K\times K}$ obtained by an application of Proposition 10 which satisfies

$$\|V - V_0O\|_{\mathrm{F}} \leqslant \frac{\|M - M_0\|_{\mathrm{F}}}{Cn} = o_{\mathbb{P}}(1), \tag{5.58}$$

where the last step follows using Theorem 5. To simplify notation, we denote $U = V_0O$ in the rest of the proof. Let $n_{\max} = \max_{r\in[K]} n_r$. Recall $(\bar{\Theta}, \bar{X})$ from (4.14) and let $\bar{V} = \bar{\Theta}\bar{X}$. Then let

$$S = \left\{ i \in [n] : \left\|\bar{V}_{i\star} - U_{i\star}\right\|_{\mathrm{F}} \geqslant \frac{1}{5}\sqrt{\frac{2}{n_{\max}}} \right\}.$$

We will show that for $i \notin S$ the community is predicted correctly using $\bar{\Theta}$. The proof of the theorem is in two steps.

Step 1: Bounding $|S|$. By the definition of S we have

$$\sum_{i\in S}\sqrt{\frac{2}{n_{\max}}} \leqslant \sum_{i\in S} 5\left\|\bar{V}_{i\star}-U_{i\star}\right\|_{\mathrm{F}} \leqslant 5\left\|\bar{V}-U\right\|_{\mathrm{F}} \implies |S| \leqslant \frac{5}{\sqrt{2}}\sqrt{n_{\max}}\left\|\bar{V}-U\right\|_{\mathrm{F}}. \tag{5.59}$$

Next, we recall the optimization problem in (4.14) is given as follows.

$$\left\|\bar{\Theta}\bar{X}-V\right\|_{\mathrm{F}}^2 \leqslant (1+\varepsilon)\min_{\Theta\in\{0,1\}^{n\times K}, X\in\mathbb{R}^{K\times K}}\|\Theta X-V\|_{\mathrm{F}}^2.$$

We substitute U for ΘX to get the following upper bound:

$$\left\|\bar{V}-V\right\|_{\mathrm{F}}^2 \leqslant (1+\varepsilon)\|U-E_0-V\|_{\mathrm{F}}^2. \tag{5.60}$$

Then by (5.58) and (5.60) we have

$$\left\|\bar{V}-U\right\|_{\mathrm{F}} \leqslant \left\|\bar{V}-V\right\|_{\mathrm{F}} + \|V-U\|_{\mathrm{F}} \leqslant (1+o(n^{-1.5})+\sqrt{1+\varepsilon})\|V-U\|_{\mathrm{F}} = o_{\mathbb{P}}(1).$$

This combined with (5.59) we have $|S| = o_{\mathbb{P}}(\sqrt{n})$.

Step 2: Bounding the prediction error. For any community r, there exists $i_r \in [n]$ such that $g_{i_r} = r$ and $i_r \notin S$ as $n_r = \theta(n)$ and $|S| = o_{\mathbb{P}}(\sqrt{n})$. By Lemma 18, for $r \neq s$ we have

$$\begin{aligned}\left\|\bar{V}_{i_r\star}-\bar{V}_{i_s\star}\right\|_{\mathrm{F}} &\geqslant \|U_{i_r\star}-U_{i_s\star}\|_{\mathrm{F}} - \left\|\bar{V}_{i_r\star}-U_{i_r\star}\right\|_{\mathrm{F}} - \left\|\bar{V}_{i_s\star}-U_{i_s\star}\right\|_{\mathrm{F}},\\ &\geqslant \sqrt{\frac{1}{n_r}+\frac{1}{n_s}} + O(n^{-3}) - \frac{2}{5}\sqrt{\frac{2}{n_{\max}}},\\ &\geqslant \frac{2.9}{5}\sqrt{\frac{2}{n_{\max}}}.\end{aligned} \tag{5.61}$$

Next, let i be such that $g_i = r$ and $i \notin S$. We show that $\bar{V}_{i\star} = \bar{V}_{i_r\star}$. For this we note that $U_{i\star} = U_{i_r\star}$ and

$$\begin{aligned}\left\|\bar{V}_{i\star}-\bar{V}_{i_r\star}\right\|_{\mathrm{F}} &\leqslant \left\|\bar{V}_{i\star}-U_{i\star}\right\|_{\mathrm{F}} + \left\|U_{i_r\star}-\bar{V}_{i_r\star}\right\|_{\mathrm{F}},\\ &< \frac{1}{5}\sqrt{\frac{2}{n_{\max}}} + \frac{1}{5}\sqrt{\frac{2}{n_{\max}}} + O(n^{-3}),\\ &< \frac{2.1}{5}\sqrt{\frac{2}{n_{\max}}}.\end{aligned} \tag{5.62}$$

In view of (5.61) and 5.62 we must have $\bar{V}_{i\star} = \bar{V}_{i_r\star}$. Let C be a permutation matrix defined so that $\Theta_0 C$ assigns community r to node i_r, for $1 \leqslant r \leqslant K$. Then we have,

$$\sum_i \mathbb{1}\{\bar{\Theta}_{i\star} \neq (\Theta_0 C)_{i\star}\} \leqslant |S|.$$

From the bound on $|S|$ from Step 1, we have that $\mathrm{Err}(\bar{\Theta}, \Theta_0)$ is $o_{\mathbb{P}}\left(1/\sqrt{n}\right)$ and this completes the proof of Theorem 6. ■

CHAPTER 6
Future work

In this chapter we discuss future extensions of our work. In section 5.1 we discuss extensions of our work in Chapter 2 to problems in correlations in high dimensions. In section 5.2 we discuss an approach to modeling multilayer networks which can be applied to brain network data sets discussed in Chapter 3. In section 5.3 we discuss extensions of our work in Chapter 4 to problems in network embedding.

6.1 Correlations in High Dimensions

There are several extensions to explore connected to our work in Chapter 2. We first recall some notation. Let $\mathbf{X}$ be a $n \times p$ matrix. This is to be viewed as n observations of p variables. Suppose the entries of $\mathbf{X}$ are normally distributed and let $\mathbf{\Sigma} = (\sigma_{ij})_{1 \leqslant i,j \leqslant p}$ be the $p \times p$ matrix of covariances of columns of $\mathbf{X}$ and let $\mathbf{R} = (\rho_{ij})_{1 \leqslant i,j \leqslant p}$ be the matrix of correlations of columns of X. Let

$$\hat{\rho}_{ij} = \frac{(X_{\cdot i} - \bar{X}_{\cdot i})^T (X_{\cdot j} - \bar{X}_{\cdot j})}{\|(X_{\cdot i} - \bar{X}_{\cdot i})^T\| \cdot \|(X_{\cdot j} - \bar{X}_{\cdot j})\|}, \quad 1 \leqslant i, j \leqslant p$$

be the sample correlations of the columns of $\mathbf{X}$. Let $L_n = \max_{1 \leqslant i < j \leqslant p} \hat{\rho}_{ij}$. Then distributional results for L_n can be used to test $H_0 : \mathbf{R} = \boldsymbol{I}$ (Cai and Jiang, 2012) i.e. whether the covariance matrix $\mathbf{\Sigma}$ is diagonal. Another covariance structure of interest from applications in time series and econometrics is the banded covariance structure. We say a covariance matrix $\mathbf{\Sigma}$ is banded with bandedness τ if $\sigma_{ij} = 0$ whenever $|i - j| > \tau$. In (Cai and Jiang, 2011) a modified form of L_n is used to test if $\mathbf{\Sigma}$ has a banded structure under the assumption that $\log p = o(n^{\frac{1}{3}})$. In extension of these results, it is of interest to explore the following question.

- Can we test whether $\mathbf{\Sigma}$ has a banded structure when $p \gg n$? In (Cai and Jiang, 2012), it is shown that asymptotically the distribution of L_n shows a phase transition depending on whether $\frac{\log p}{n}$ converges to $\gamma = 0$, or $\gamma \in (0, \infty)$ or to $\gamma = \infty$. Do similar results hold under the three regimes in the presence of a banded structure? Further how does the size of the band τ need to vary with respect to n and p for the results to continue to hold?

Another direction to explore is to study related covariance structures which capture dependence between the columns of $\mathbf{X}$. More explicitly, what the are the distributional limits of maximal correlation under the following conditions:

- The correlation $\rho_{i,j} \to 0$ as $|i - j| \to \infty$.
- The variables X_{ki} and X_{kj} are strongly dependent when $|i - j| \leqslant \tau$ and weakly dependent when $|i - j| \geqslant \tau$ where τ is the bandedness of $\mathbf{\Sigma}$.

Many of the distributional results for L_n or modifications of it assume that $\mathbf{X}$ is a Gaussian matrix. This motivates the following question:

- To what extent can the limiting results for the maximal correlation be extended to sub-Gaussian distributions both with and without the banded condition?

(Fan and Jiang, 2019) describe limits of L_n when $\rho_{ij} = \rho, \quad \forall i \neq j$. It is of interest to understand the limiting regimes under more general dependence structures. This motivates the following question.

- (Low dependence) In (Fan and Jiang, 2019) it is shown that a recentered and scaled version of L_n converges to the Gumbel distribution under the assumption that $\rho_{i,j} = \rho, 1 \leqslant i \neq j \leqslant p$ and $\rho = o\left(\frac{1}{\sqrt{\log p}}\right)$. Does a similar result hold when $\rho_{i,j} = o\left(\frac{1}{\sqrt{\log p}}\right)$ but the correlations, $\rho_{i,j}$, are not assumed to be equal to each other?

One of the challenges to the previous question posed above is to define a common recentering since $\rho_{i,j}$ can vary with i and j. For example, a large recentering may force the smaller correlations to $-\infty$.

In addition to the maximal correlation, it is of interest to understand the full process of correlations and understand how it behaves for large n and p. As a consequence, this helps us to

understand the second largest correlation, the third largest correlation and so on. This motivates the following question.

- How to compute functionals of the point process of sample covariances and sample correlations?

6.2 Multilayer Networks

In this section we propose a method for modeling multilayer networks. The data sets, as described in Chapter 3, consist of networks with weights and networks over time. Therefore our goal is a modeling framework which encompasses those two properties. We describe our modeling approach which uses the framework of optimal mass transport. This framework is flexible in various ways. For example it can include weighted networks, temporal networks and networks of different sizes. We describe the approach in more detail below.

In many real world contexts, it is of interest to find an shortest or minimum cost method to move between two configurations. Optimal mass transport is a way to formalize this problem. This problem was first formulated by G. Monge in 1781 in the context of computing the cost of moving a mound of sand from a source to a destination. The problem has since been generalized and has applications in imaging (Haker et al., 2004), astrophysics (Frisch et al., 2002), meteorology (Cullen et al., 1991) and seismology (Métivier et al., 2016).

Optimal mass transport has also been used in the context of networks for graph matching (Xu et al., 2019). However to the best of my knowledge, no multilayer network models uses optimal transport. Towards this, we can formulate the optimal distance between either

1. the random graph models which approximate the observed graphs or,

2. between two fixed observed networks.

The work in (Xu et al., 2019; Peyré and Cuturi, 2019) uses the latter framework. We describe the modeling approach using the former framework in three steps below.

Define a cost function between networks. Suppose we have two networks G and H consisting of n nodes each. Then one can define a wide range of similarity measures to compare G and H. We will think of these measures as costs to transform the network G to network H. For example,

the Hamming distance computes the number of edges which are present in G but not in H or vice versa. Another example is any measure which compares the number of motifs such as triangles, stars etc. between two graphs. If G and H have weights and other attributes, then these can be compared and included in the overall measure as well. Many of these costs can also be defined when G and H have differing number of nodes.

In order to keep the notation and description easy to follow, we fix n the number of nodes and focus on $\mathcal{G}_n$ the set of all unweighted simple symmetric graphs on n nodes. Let $c : \mathcal{G}_n \times \mathcal{G}_n \to \mathbb{R}$ be a cost function comparing pairs of graphs. If c is the Hamming distance then for any two graphs G and H

$$c(G, H) = \sum_{i=1}^{n} \sum_{j=1}^{n} |G_{i,j} - H_{i,j}|,$$

where $G_{i,j}$ (resp. $H_{i,j}$) is 1 is there is an edge between node i and node j and 0 otherwise.

Computing transportation cost between models. We begin by describing how to compute transportation cost between two sets of points. We will then describe how this can be used to compute costs between models as well. To fix notation, let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$ be two finite sets. Let

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \text{ and } \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$$

be two probability measures on $\mathcal{X}$ and $\mathcal{Y}$. Let $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_+$ be a cost function which assigns a cost $c(x_i, y_j)$ of transporting a unit mass from location x_i to location y_j. Let π be a coupling of the measures α and β. In other words, π is a joint distribution on $\mathcal{X} \times \mathcal{Y}$ with marginals α and β. Then $\pi(x_i, y_j)$ can be interpreted as the amount of mass transported from location x_i to location y_j. The total cost of transportation with respect to c is given by

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_j)\pi(x_i, y_j) \text{ or } \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} c(x, y)\pi(x, y) \text{ or } \mathbb{E}_{\pi}(c).$$

Alternatively if X and Y are two random variables with $(X, Y) \sim \pi$, the cost of transportation is given by

$$\mathbb{E}_{(X,Y)}(c(X, Y)).$$

Let $\Pi(\alpha, \beta)$ be the set of all couplings between α and β. Then the optimal transport cost between α and β with respect to c is given by

$$\mathcal{L}_c(\alpha, \beta) := \min_{\pi \in \Pi(\alpha,\beta)} \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} c(x, y)\pi(x, y) \text{ or,}$$

$$\min_{\pi \in \Pi(\alpha,\beta)} \mathbb{E}_\pi(c) \text{ or,}$$

$$\min_{(X,Y)} \mathbb{E}_{(X,Y)}(c(X, Y)),$$

where (X, Y) is a couple of random variables $X \sim \alpha$ and $Y \sim \beta$.

In the context of graphs, let P_1 and P_2 be two random graph models on $\mathcal{G}_n$, the set of all graphs of size n. For example, $P_1 = ER(p_1)$ and $P_2 = ER(p_2)$ where $ER(p)$ is the Erdős-Renyi random graph model with parameter p. Note that P_1 and P_2 are probability distributions on $\mathcal{G}_n$. Let π be a coupling of the P_1 and P_2. Then the transport cost between the two models with respect to a cost function is c is given by

$$\sum_{G \in \mathcal{G}_n} \sum_{H \in \mathcal{G}_n} c(G, H)\pi(G, H) \text{ or } \sum_{(G,H) \in \mathcal{G}_n \times \mathcal{G}_n} c(G, H)\pi(G, H) \text{ or } \mathbb{E}_\pi(c).$$

$\mathbb{E}_\pi(c)$ can be thought of as the average cost of moving from model P_1 to model P_2 with respect to cost c.

We give an example when $P_1 = ER(p_1)$ and $P_2 = ER(p_2)$ and c is the Hamming distance. Let π be a coupling of P_1 and P_2. Suppose (G, H) is a sample from π. Let $q_{i,j} = \mathbb{P}_\pi(G_{ij} = 1, H_{ij} = 1)$ i.e. the probability that both the graphs have ijth edge is equal to $q_{i,j}$. Then

$$\mathbb{E}_\pi(c) = \sum_{i,j \in [n]: i<j} (p_1 + p_2 - 2q_{i,j}),$$

is the transportation cost between the two models.

Network models using optimal mass transport. Above we described how to find optimal cost or distance between two single layer network models. However in practice we do not know the underlying model which best approximates the observed networks. Therefore we would like to fit the best model for each of the individual observed networks in addition to the best coupling

between them. We propose two methods towards this goal. We describe these methods for jointly modeling two graphs. In future work, we will analyze how to extend these two more than two graphs and networks with weights.

Let G_1 and G_2 be two observed graphs. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be families of single layer models we would like for modeling G_1 and G_2 separately. For example, $\mathcal{P}_1$ and $\mathcal{P}_2$ may be Erdős-Renyi network models or stochastic block models (SBM). Then

$$\mathcal{F}(\mathcal{P}_1, \mathcal{P}_2) := \bigcup_{\alpha_1 \in \mathcal{P}_1, \alpha_2 \in \mathcal{P}_2} \Pi(\alpha_1, \alpha_2).$$

is the set of all coupling we would like to optimize over to find the best fit for (G_1, G_2). In applications it would be of interest to use a smaller subset of $\mathcal{F}(\mathcal{P}_1, \mathcal{P}_2)$ for modeling or for computational purposes. To simplify notation if π is a coupling, we denote the marginal distributions by π_1 and π_2. We propose two ways to jointly model networks

a) $\min_{\pi \in \mathcal{F}(\mathcal{P}_1, \mathcal{P}_2)} - \log \pi(G_1, G_2) + \lambda \mathcal{L}_c(\pi_1, \pi_2)$.

b) $\min_{\pi \in \mathcal{F}(\mathcal{P}_1, \mathcal{P}_2)} - \log \pi(G_1, G_2) + \lambda \mathbb{E}_\pi(c)$.

The first term in both the methods is the log-likelihood of the two graphs. The first method penalizes the optimal cost between the marginal distributions. This method will choose a coupling for which the marginals are close with respect to the optimal transport cost. The second method penalizes a coupling by the transportation cost associated to it. We describe these two methods in three examples.

6.2.1 Erdős-Renyi Models

Let $ER(p)$ be the Erdős-Renyi random graph model on $[n]$ with parameter p. Let $\mathcal{P}_1 = \mathcal{P}_2 = \{ER(p) | 0 \leqslant p \leqslant 1\}$. We consider the following types of couplings π.

$$\mathcal{F}_{ER} = \{\pi \in \mathcal{F}(\mathcal{P}_1, \mathcal{P}_2) | \mathbb{P}_\pi(X_{ij} = 1, Y_{ij} = 1) = \mathbb{P}_\pi(X_{kl} = 1, Y_{kl} = 1), \text{ where } (X, Y) \sim \pi\}.$$

These are the set of couplings where we have a single parameter describing the dependence between the two graphs. The the optimization problems simplify as follows

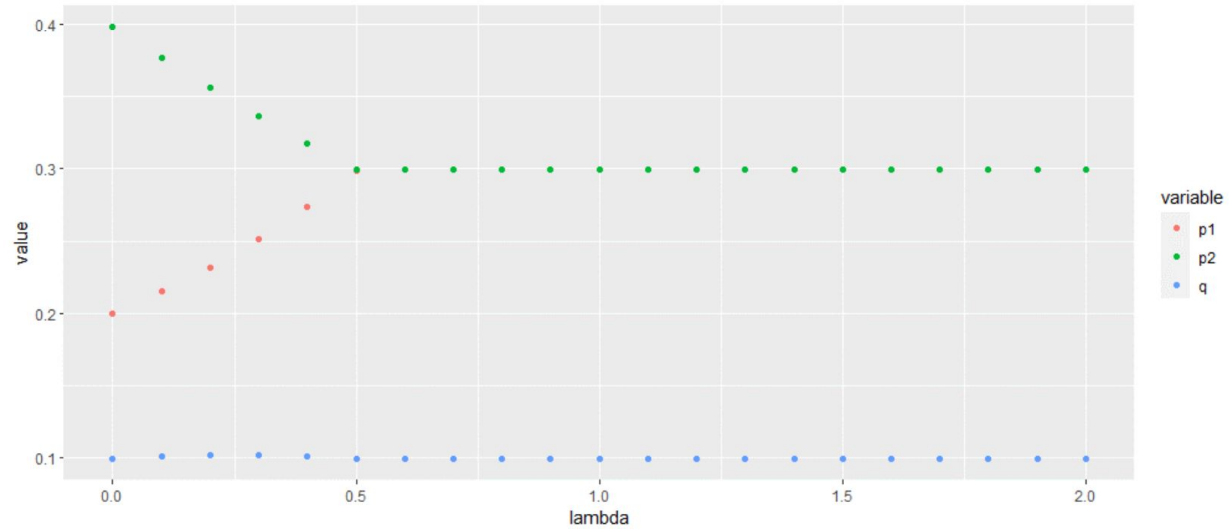

Figure 6.1: This is a simulation for optimization problem a). It describes the effect of λ on the fitted parameters p_1, p_2 and q. For each point in the plot, 10 Erdős-Renyi graphs on 10 nodes were generated with parameters $p_1 = 0.2, p_2 = 0.4$ and the dependence parameter $q = 0.1$. The 10 fitted parameters were then averaged. An increase in λ forces the parameters p_1 and p_2 to be equal to each other.

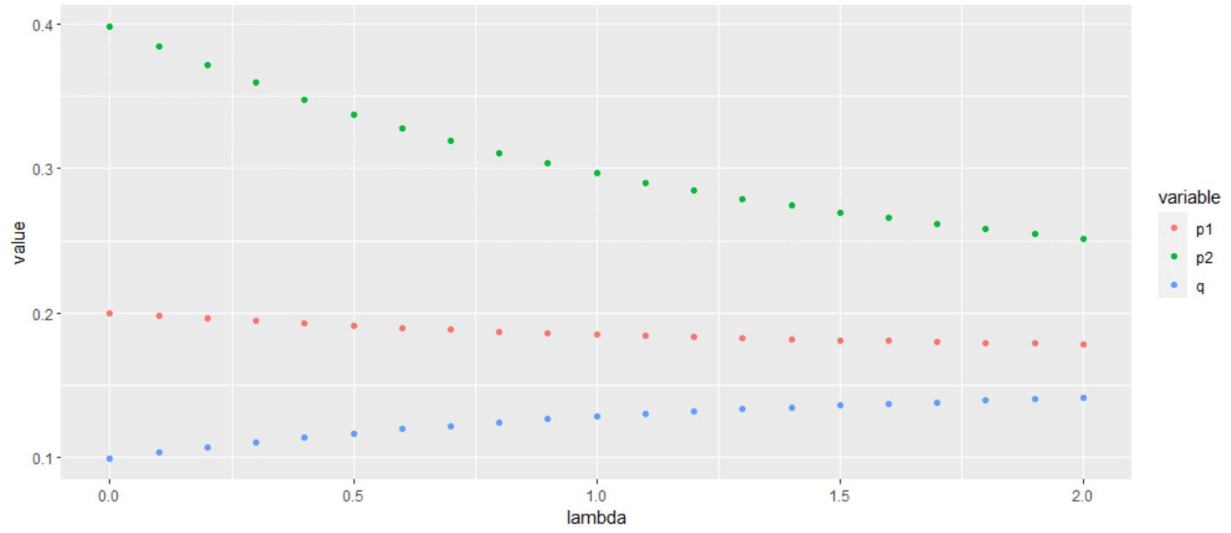

Figure 6.2: This is a simulation for optimization problem b). It describes the effect of λ on the fitted parameters p_1, p_2 and q. For each point in the plot, 10 Erdős-Renyi graphs on 10 nodes were generated with parameters $p_1 = 0.2, p_2 = 0.4$ and the dependence parameter $q = 0.1$. The 10 fitted parameters were then averaged. An increase in λ forces the parameter q to increase and the probabilities $P_\pi(X_{ij} = 1, Y_{ij} = 0)$ and $P_\pi(X_{ij} = 0, Y_{ij} = 1)$ to decrease. Here π refers to the fitted coupling or joint model.

a) $\min_{(p_1,p_2)} \min_{\pi \in \Pi(ER(p_1),ER(p_2)) \cap \mathcal{F}_{ER}} \{-\log \pi(G_1, G_2) + \lambda \binom{n}{2} |p_1 - p_2|\}$.

b) $\min_{(p_1,p_2)} \min_{\pi \in \Pi(ER(p_1),ER(p_2)) \cap \mathcal{F}_{ER}} \left\{-\log \pi(G_1, G_2) + \lambda(\sum_{i,j \in [n]: i<j} p_1 + p_2 - 2q_\pi)\right\}$, where $q_\pi = \mathbb{P}_\pi(X_{ij} = 1, Y_{ij} = 1)$.

Note that for the first method, penalty term is the difference between the density parameters for the Erdős-Renyi models. For the second method, the penalty term is the transportation cost for the coupling. The latter may be interpreted as following. Suppose $(X, Y) \sim \pi$ and we observe $X_{ij} = 1$ and the conditional probability that $Y_{ij} = 1$ is high, then we have a smaller penalty term. While if this conditional probability is lower, the penalty will be higher. Figures 6.1 and 6.2 show effect of the parameter λ on the fitted parameters p_1, p_2 and q. This is informative about the fitted models from the above two methods.

6.2.2 Stochastic Blocks Models

Let K be the number of blocks. Assume the blocks are known and let g_i denote the block for node i. Also for $1 \leqslant k \leqslant l \leqslant K$ let

$$N_{k,l} = \sum_{1 \leqslant i < j \leqslant n} \delta_{\{} g_i = k, g_j = l\},$$

be the number of pairs of nodes between groups k and l.

Let $\mathbf{p} = (p_{k,l})$ with $0 \leqslant p_{k,l} \leqslant 1, 1 \leqslant k \leqslant l \leqslant K$ be a matrix of block probabilities and let $\mathbf{g} = (g_i)_i$ be the block labels. Let $SBM(\mathbf{p}, \mathbf{g})$ be the stochastic block model with the given parameters. We will suppress notation and just write $SBM(\mathbf{p})$ below. Similar to the Erdős-Renyi case, let $\mathcal{P}_1 = \mathcal{P}_2 = \{SBM(\mathbf{p}) | 0 \leqslant p_{k,l} \leqslant 1, 1 \leqslant k \leqslant l \leqslant K\}$ and let

$$\mathcal{F}^1_{SBM} = \{\pi \in \mathcal{F}(\mathcal{P}_1, \mathcal{P}_2) | \mathbb{P}_\pi(X_{ij} = 1, Y_{ij} = 1) = \mathbb{P}_\pi(X_{kl} = 1, Y_{kl} = 1), \text{ where } (X, Y) \sim \pi\}.$$

In words, this means that we would like to model the individual networks using a stochastic block model with K blocks and the dependence between the networks is given by a single parameter (which is to be fitted). The two methods simplify as follows.

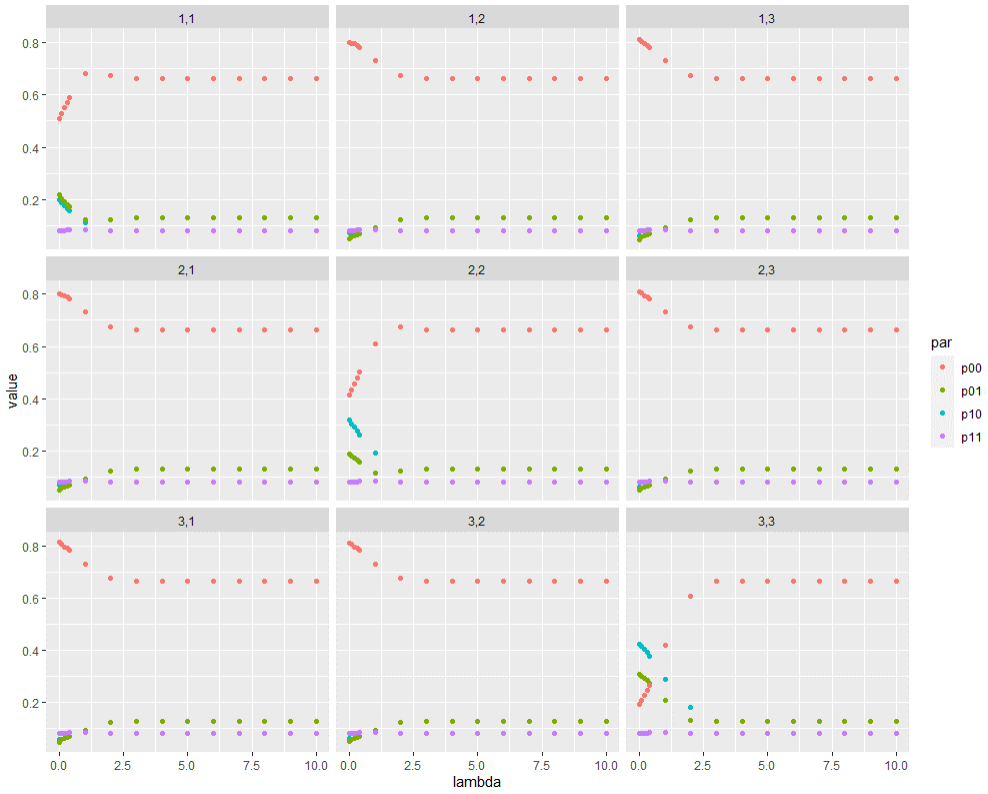

Figure 6.3: This is a simulation for optimization problem a). It describes the effect of λ on the fitted parameters $\mathbf{p}_{10} = P_\pi(X_{ij} = 1, Y_{ij} = 0)$, $\mathbf{p}_{01} = P_\pi(X_{ij} = 0, Y_{ij} = 1), \mathbf{p}_{11} = P_\pi(X_{ij} = 1, Y_{ij} = 1)$ and $\mathbf{p}_{00} = P_\pi(X_{ij} = 0, Y_{ij} = 0)$ where π is the fitted coupling. The plot consists of a symmetric matrix of 9 subplots, one for each pair of blocks. For each point in the plot, 10 graphs were generated from SBM on 10 nodes were generated with parameters. The 10 fitted parameters were then averaged. At $\lambda = 0$, we can observe the original parameters. An increase in λ forces the parameters $\mathbf{p}_{10}$ and $\mathbf{p}_{01}$ to be equal to each other.

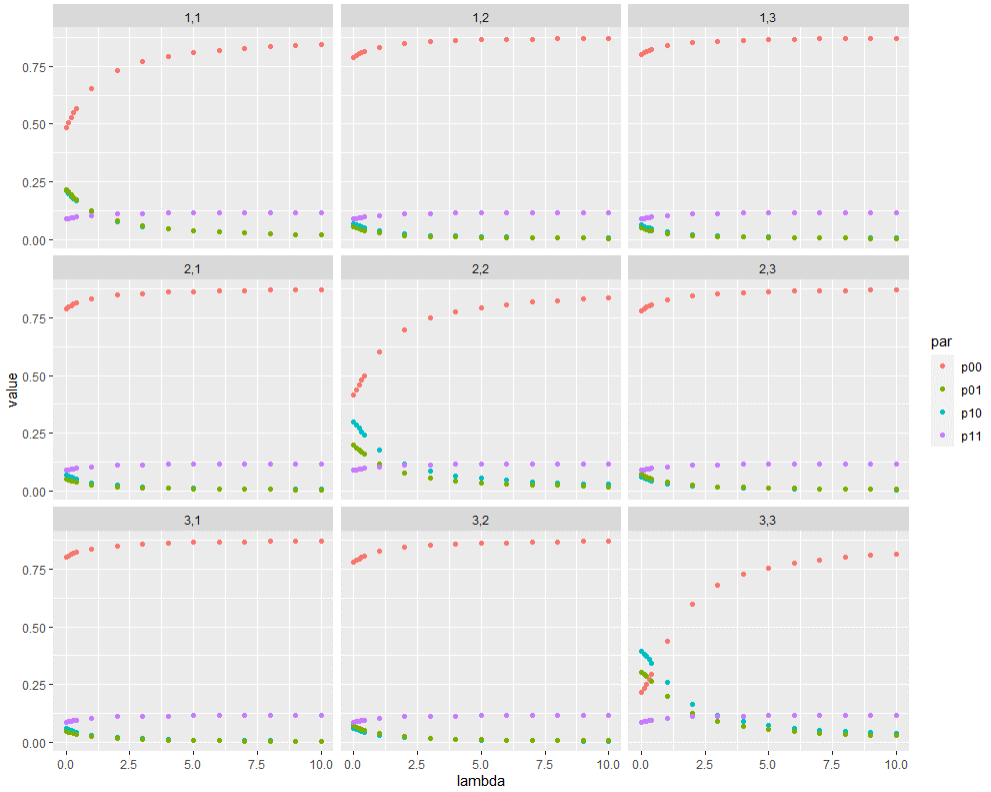

Figure 6.4: This is a simulation for optimization problem b). It describes the effect of λ on the fitted parameters $\mathbf{p}_{10} = P_\pi(X_{ij} = 1, Y_{ij} = 0)$, $\mathbf{p}_{01} = P_\pi(X_{ij} = 0, Y_{ij} = 1), \mathbf{p}_{11} = P_\pi(X_{ij} = 1, Y_{ij} = 1)$ and $\mathbf{p}_{00} = P_\pi(X_{ij} = 0, Y_{ij} = 0)$ where π is the fitted coupling. The plot consists of a symmetric matrix of 9 subplots, one for each pair of blocks. For each point in the plot, 10 graphs were generated from SBM on 10 nodes were generated with parameters. The 10 fitted parameters were then averaged. At $\lambda = 0$, we can observe the original parameters. An increase in λ forces the parameters $\mathbf{p}_{10}$ and $\mathbf{p}_{01}$ to decrease.

a) $\min_{(\mathbf{p}^1,\mathbf{p}^2)} \min_{\pi\in\Pi(SBM(\mathbf{p}^1),SBM(\mathbf{p}^2))\cap\mathcal{F}^1_{SBM}}$
$\Big\{-\log\pi(G_1,G_2) + \lambda\sum_{1\leqslant k\leqslant l\leqslant K} N_{k,l}(p^1_{k,l}+p^2_{k,l}-2\min\{\mathbf{p}^1,\mathbf{p}^2\})\Big\}$.

b) $\min_{(\mathbf{p}^1,\mathbf{p}^2)} \min_{\pi\in\Pi(SBM(\mathbf{p}^1),SBM(\mathbf{p}^2))\cap\mathcal{F}^1_{SBM}}$
$\Big\{-\log\pi(G_1,G_2) + \lambda\sum_{1\leqslant k\leqslant l\leqslant K} N_{k,l}(p^1_{k,l}+p^2_{k,l}-2q_\pi)\Big\}$, where $q_\pi = \mathbb{P}_\pi(X_{ij}=1,Y_{ij}=1)$.

The first term in both the optimization problems above is the log-likelihood of the observed networks as in the Erdős-Renyi case. The penalty term in b) is also similar as before. In a), the penalty term is somewhat different but it again penalizes the difference between the density parameters. Figures 6.3 and 6.4 describe the effect of λ on the fitted parameters which in turn helps us understand how the models are fitted.

If we instead chose to restrict couplings to

$$\mathcal{F}^2_{SBM} = \{\pi\in\mathcal{F}(\mathcal{P}_1,\mathcal{P}_2) | \mathbb{P}_\pi(X_{ij}=1,Y_{ij}=1) = \mathbb{P}_\pi(X_{kl}=1,Y_{kl}=1), \\ \text{where } (X,Y)\sim\pi \text{ and } g_i = g_k \,\&\, g_j = g_l\},$$

the two optimization problems can be written as follows.

a) $\min_{(\mathbf{p}^1,\mathbf{p}^2)} \min_{\pi\in\Pi(SBM(\mathbf{p}^1),SBM(\mathbf{p}^2))\cap\mathcal{F}^2_{SBM}}$
$\Big\{-\log\pi(G_1,G_2) + \lambda\sum_{1\leqslant k\leqslant l\leqslant K} N_{k,l}|p^1_{k,l}-p^2_{k,l}|\Big\}$.

b) $\min_{(\mathbf{p}^1,\mathbf{p}^2)} \min_{\pi\in\Pi(SBM(\mathbf{p}^1),SBM(\mathbf{p}^2))\cap\mathcal{F}^2_{SBM}}$
$\Big\{-\log\pi(G_1,G_2) + \lambda\sum_{1\leqslant k\leqslant l\leqslant K} N_{k,l}(p^1_{k,l}+p^2_{k,l}-2q^\pi_{k,l})\Big\}$, where $q^\pi_{k,l} = \mathbb{P}_\pi(X_{ij}=1,Y_{ij}=1)$ for (i,j) such that $g_i = k$ and $g_j = l$ and $i<j$.

This case is identical to the Erdős-Renyi case. The only difference being that there are $\binom{K}{2}$ block parameters compared to 1 density parameter in the Erdős-Renyi case. The optimization can be done separately for each of the pairs of blocks.

6.2.3 Degree Corrected Models

For simplicity, we only discuss degree corrected Erdős-Renyi models. Analogously, optimization problems may be written for degree corrected stochastic block models. Fix two sequences $\mathbf{d}^a = (d^a_i)_i$ and $\mathbf{d}^b = (d^b_i)_i$ satisfying $0\leqslant d^a_i, d^b_j \leqslant 1$, $\sum_{i=1}^n d^a_i = 1$ and $\sum_{i=1}^n d^b_i = 1$. We write $DC-ER(\mathbf{d}^a, p)$

for the degree corrected Erdős-Renyi model. To be specific, if $X \sim DC-ER(\mathbf{d}^a, p)$, $P(X_{i,j}=1) = pd_i^a d_j^a$ for $1 \leqslant i < j \leqslant n$. Now let $\mathcal{P}_1 = \mathcal{P}_2 = DC-ER(\mathbf{d}^a, p)$ and

$$\mathcal{F}_{DC-ER} = \{\pi \in \mathcal{F}(\mathcal{P}_1, \mathcal{P}_2) | \mathbb{P}_\pi(X_{ij}=1, Y_{ij}=1) = \mathbb{P}_\pi(X_{kl}=1, Y_{kl}=1), \text{ where } (X,Y) \sim \pi\}.$$

This means that we would like to model both the layers using $DC-ER$ models and the dependence between the networks is given by a single parameter. If we know the degree parameters, then the optimization problems become

a) $\min_{(p^a,p^b)} \min_{\pi \in \Pi(DC-ER(\mathbf{d}^a,p^a), DC-ER(\mathbf{d}^b,p^b)) \cap \mathcal{F}_{DC-ER}}$
$\Big\{ -\log \pi(G_1,G_2) + \lambda \Big\{ \frac{p^a}{2}(1-\sum_i (d_i^a)^2) + \frac{p^b}{2}(1-\sum_i (d_i^b)^2) - q^*(r^2 - \sum_i (d_i^a)^2 (d_i^b)^2) \Big\} \Big\}$ where
$q^* = \min_{i<j}\{p^a d_i^a d_j^a, p^b d_i^b d_j^b\}$ and $r = \sum_i d_i^a d_i^b$.

b) $\min_{(p^a,p^b)} \min_{\pi \in \Pi(DC-ER(\mathbf{d}^a,p^a), DC-ER(\mathbf{d}^b,p^b)) \cap \mathcal{F}_{DC-ER}}$
$\Big\{ -\log \pi(G_1,G_2) + \lambda \Big\{ \frac{p^a}{2}(1-\sum_i (d_i^a)^2) + \frac{p^b}{2}(1-\sum_i (d_i^b)^2) - q(r^2 - \sum_i (d_i^a)^2 (d_i^b)^2) \Big\} \Big\}$, where
$q_\pi = \mathbb{P}_\pi(X_{ij}=1, Y_{ij}=1)$ and $r = \sum_i d_i^a d_i^b$.

The expressions may look complicated due to the presence of the degree parameters. However the optimization problems are similar in nature. The method a) penalizes by the difference between the marginal parameters given constraints and method b) penalizes using the transportation cost of the coupling.

We review our discussion in this section. We introduced two ways to jointly model networks. We used transport cost and optimal transport cost between distributions towards this. We described these methods in the context of three examples of random graph models. The methods were described for modeling two networks. In future work we will analyze how to extend these models to more than two networks and to weighted networks.

6.3 Network Embeddings

Our work in Chapter 4 can be extended in various directions described as follows.

Interplay between node2vec parameters. Our work explores the effect of the non-backtracking parameter α in node2vec on community detection. It would be of interest to explore the effect of the

other parameter β as well. In particular, it would be interesting to explore the interplay between the two parameters on its effects on community detection. Answers to these questions can inform practitioners when choosing parameters for node2vec for the purposes of community detection.

Degree-corrected stochastic block model. Our work uses graphs generated from the stochastic block model (SBM) and studies the community detection problem using node embeddings from node2vec and DeepWalk. However given the community labels, nodes have identical degree distributions under SBM. This is unlike real-world networks where degrees follow a power-law distribution. To generalize our work, one could analyze the same community detection problem using node embeddings but now with graphs generated from the degree-corrected stochastic block model.

Weighted networks. Another direction of research is to consider the problem in Chapter 4 for weighted networks. node2vec algorithm allows for weights in the model and so one could start with a weighted stochastic block model and analyze the corresponding community detection problem.

Time-series of networks. All of the suggested directions of research above are for studying one network. However, in many real-world data sets we have a collection of networks related to each other. One example is a time-series of networks. In such a setting one would start with a model for the time-series of networks. Then one would apply the node embedding method to the networks at each time slice to obtain a time-series of network embeddings. One would then analyze the relationships between the embeddings.

Application to brain networks. One could apply the node embedding algorithms, DeepWalk and node2vec, to brain network data sets discussed in Chapter 3. For both the infant data set and the ADNI data set discussed in Chapter 3, one could study the time series of network embeddings to gain insight about brain development or degeneration over time. It would also be interesting to see how network embeddings vary with groups. For example, for the infant data set, one could compare network embeddings for years 0, 1 and 2. One could also compare network embeddings for preterm and non-preterm subjects. Another interesting grouping is by the scanner, Allegra and TRIO, used in creating the data sets. This may help in gaining insight into the effects of the scanner in the networks and potentially help in removing some of this effect. Similarly for the ADNI data set, one could compare network embeddings for the three groups: cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD).

APPENDIX 1: PROPERTIES OF OPTIMIZERS FOR EMBEDDING ALGORITHMS

Proof of Proposition 3. Note that

$$\mathbb{E}_{l\sim P_C}\big[\log\sigma(-\langle \boldsymbol{f}_i, \boldsymbol{f}'_l\rangle))\big] = \sum_{j'=1}^{n} \frac{|C_{\star j'}|}{|C|}\log\sigma(-\langle \boldsymbol{f}_i, \boldsymbol{f}'_{j'}\rangle)).$$

Thus, (4.10) reduces to

$$L_C(F,F') = \sum_{i,j=1}^{n}\left[C_{ij}\log\sigma(\langle \boldsymbol{f}_i, \boldsymbol{f}'_j\rangle) + b\frac{|C_{i\star}||C_{\star j}|}{|C|}\log\sigma(-\langle \boldsymbol{f}_i, \boldsymbol{f}'_j\rangle)\right]. \tag{1}$$

Let $\ell_C(\langle \boldsymbol{f}_i, \boldsymbol{f}'_j\rangle)$ be the (i,j)-th summand above. Defining $x = \langle \boldsymbol{f}_i, \boldsymbol{f}'_j\rangle$, we optimize $\ell_C(x)$. Taking partial derivative with respect to x,

$$\frac{\partial \ell_C}{\partial x} = \frac{C_{ij}}{1+\mathrm{e}^x} - b\frac{|C_{i\star}||C_{\star j}|}{|C|}\times\frac{\mathrm{e}^x}{1+\mathrm{e}^x}.$$

Equating the derivative to zero, we have $x = \log\left(\frac{C_{ij}\cdot|C|}{|C_{i\star}||C_{\star j}|}\right) - \log b$, and thus,

$$\langle \boldsymbol{f}_i, \boldsymbol{f}'_j\rangle = \log\left(\frac{C_{ij}\cdot|C|}{|C_{i\star}||C_{\star j}|}\right) - \log b = (M_C)_{ij}. \tag{2}$$

Thus, if $\bar{M}_C = FF'^T$, then $\langle F_{i\star}, F'_{j\star}\rangle$ satisfies (2). ■

Proof of Lemma 4. We recall that co-occurences are given by

$$C_{ij} = \sum_{t=t_L}^{t_U}\sum_{m=1}^{r}\sum_{k=1}^{l-t}\mathbb{1}\left\{\boldsymbol{w}_k^{(m)} = i, \boldsymbol{w}_{k+t}^{(m)} = j\right\} + \mathbb{1}\left\{\boldsymbol{w}_k^{(m)} = j, \boldsymbol{w}_{k+t}^{(m)} = i\right\}.$$

By the strong law of large numbers,

$$\frac{\sum_{m=1}^{r}\mathbb{1}\left\{\boldsymbol{w}_k^{(m)} = i, \boldsymbol{w}_{k+t}^{(m)} = j\right\}}{r} \xrightarrow[r\to\infty]{a.s.} \mathbb{P}\left(\boldsymbol{w}_k^{(1)} = i, \boldsymbol{w}_{k+t}^{(1)} = j\right). \tag{3}$$

The initial distribution for DeepWalk is an invariant distribution for each of the random walks. We now note that the initial distribution for node2vec is also invariant for the walks for node2vec, i.e.:

$$\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_1^{(2)} = j\right) = \mathbb{P}\left(\boldsymbol{w}_k^{(1)} = i, \boldsymbol{w}_{k+1}^{(1)} = j\right), \quad 1 \leqslant k < l. \tag{4}$$

The computation below shows that this follows by induction. Suppose that $(i_{k+1}, i_{k+2}) \in E$.

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{w}_{k+1}^{(1)} = i_{k+1}, \boldsymbol{w}_{k+2}^{(1)} = i_{k+2}\right) \\
&= \sum_{i_k | (i_k, i_{k+1}) \in E} \mathbb{P}\left(\boldsymbol{w}_k^{(1)} = i_k, \boldsymbol{w}_{k+1}^{(1)} = i_{k+1}, \boldsymbol{w}_{k+2}^{(1)} = i_{k+2}\right), \\
&= \sum_{i_k | (i_k, i_{k+1}) \in E} \mathbb{P}\left(\boldsymbol{w}_{k+2}^{(1)} = i_{k+2} | \boldsymbol{w}_k^{(1)} = i_k, \boldsymbol{w}_{k+1}^{(1)} = i_{k+1}\right) \cdot \mathbb{P}\left(\boldsymbol{w}_k^{(1)} = i_k, \boldsymbol{w}_{k+1}^{(1)} = i_{k+1}\right), \\
&= \frac{\alpha A_{i_{k+1} i_k}}{d_{i_{k+1}} + (-1+\alpha) A_{i_{k+1} i_k}} \cdot \frac{A_{i_k i_{k+1}}}{|A|} + \sum_{i_k \neq i_{k+2}} \frac{A_{i_{k+1} i_{k+2}}}{d_{i_{k+1}} + (-1+\alpha) A_{i_{k+1} i_k}} \cdot \frac{A_{i_k i_{k+1}}}{|A|} \mathbb{1}\left\{(i_k, i_{k+1}) \in E\right\}, \\
&= \frac{1}{|A|}.
\end{aligned}$$

In view of (3) and (4) we have

$$\begin{aligned}
\frac{C_{ij}}{r} &\xrightarrow[r\to\infty]{a.s.} \sum_{t=t_L}^{t_U} (l-t) \left(\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = j\right) + \mathbb{P}\left(\boldsymbol{w}_1^{(1)} = j, \boldsymbol{w}_{1+t}^{(1)} = i\right)\right), \\
\frac{\sum_{i'} C_{i'j}}{r} &\xrightarrow[r\to\infty]{a.s.} 2\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = j\right) \cdot \sum_{t=t_L}^{t_U} (l-t) = 2\gamma(l, t_L, t_U) \mathbb{P}\left(\boldsymbol{w}_1^{(1)} = j\right), \\
\frac{\sum_{j'} C_{ij'}}{r} &\xrightarrow[r\to\infty]{a.s.} 2\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right) \cdot \sum_{t=t_L}^{t_U} (l-t) = 2\gamma(l, t_L, t_U) \mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i\right), \\
\frac{\sum_{i'j'} C_{i'j'}}{r} &\xrightarrow[r\to\infty]{a.s.} 2\gamma(l, t_L, t_U).
\end{aligned}$$

The result follows from these equations. We note that if, $\sum_{t=t_L}^{t_U} A_{ij}^{(t)} > 0$ then,

$$\sum_{t=t_L}^{t_U} (l-t) \left(\mathbb{P}\left(\boldsymbol{w}_1^{(1)} = i, \boldsymbol{w}_{1+t}^{(1)} = j\right) + \mathbb{P}\left(\boldsymbol{w}_1^{(1)} = j, \boldsymbol{w}_{1+t}^{(1)} = i\right)\right) > 0,$$

and so the logarithm is well-defined for the limiting term. In this case, we also have $C_{ij} > 0$ for large r and so $\log\left(\frac{C_{ij}\cdot\left(\sum_{i'j'} C_{i'j'}\right)}{b\sum_{j'} C_{ij'}\sum_{i'} C_{i'j}}\right)$ is well-defined for large enough r. ■

Proof of Lemma 5. Let

$$Z = \{X \in \mathbb{R}^{K\times K} : \operatorname{rank}(X) < K\},$$

be the set of all matrices which are not of full rank. Note that $\lambda(Z) = 0$ as Z is the zero set of the determinant function on $\mathbb{R}^{K\times K}$ which is a polynomial function of the matrix entries. Next, let

$$S = \{X \in \mathbb{R}_+^{K\times K} : \operatorname{rank}(X) = K,\ \operatorname{rank}(\bar{\log} X) < K\},$$

be the set of all matrices in $\mathbb{R}_+^{K\times K}$ for which the rank drops after applying $\bar{\log}$. Let $\bar{\exp}$ be the mapping given by applying the exponential function element-wise to the matrix entries. Then we can note that $S \subset \bar{\exp}(Z)$. Thus to complete the proof, it is enough to show that $\lambda(\bar{\exp}(Z)) = 0$. This final assertion follows by (Rudin, 1987, Lemma 7.25) as $\bar{\exp}$ is a smooth function and so it must map null sets to null sets. ■

APPENDIX 2: PATH COUNTING WITH CONSTRAINTS

For the analysis of paths with backtracks, it is useful to have an accurate estimate of the number of paths between two nodes satisfying constraints specifying the locations of the backtracking edges along the paths. Lemma 19 below counts the number of vertices that need to be chosen along the paths under such constraints. This is a more general version of Lemma 6 which may be of independent interest.

Lemma 19 is stated in terms of sets and equivalence classes. For ease of reading, we describe the setup in the context of graphs. In the graphs context, the set $\mathcal{A}$ in Lemma 19 is the set of vertices, the sequence $(i_0, i_1, \ldots, i_s)$ is a path on the graph that does not intersect itself, and the sequence k_i gives all the non-backtracking edges. The relation R simply says that two vertices in the path have to be equal when there is a backtracking edge. The equivalence relation $\sim$ is defined to keep track of distinct vertices under the constraints given by backtracking edges. We now state and prove the lemma.

Lemma 19. *Let $\mathcal{A}$ be a set and let $(i_0, i_1, \ldots, i_s)$ be a sequence of distinct elements from $\mathcal{A}$, i.e. $i_l \neq i_{l'}, 0 \leqslant l \neq l' \leqslant s$. Let $I = \{i_0, i_1, \ldots, i_s\}$. Let $0 = k_1 < k_2 < \cdots < k_m < k_{m+1} = s$ be a subsequence of $(0, 1, \ldots, s)$ of length $m + 1$. Let R be a relation defined on I by*

$$i_{l-1} \, R \, i_{l+1}, \quad k_r < l < k_{r+1}, 1 \leqslant r \leqslant m.$$

Let $\sim$ be an equivalence relation defined on I using the following:

1. *$i \sim i$*
2. *$i \sim j$ if $i \, R \, j$ or $j \, R \, i$*
3. *$i \sim j$ if there is a sequence of elements $l_1, l_2, \ldots, l_r$ such that $i \, R \, l_1$, $l_{r'} \, R \, l_{r'+1}$ for $1 \leqslant r' < r$, and $l_r \, R \, j$.*

Let $[i_l]$ be the equivalence class of i_l for $0 \leqslant l \leqslant s$ and let

$$E := \{\, [i_l] \,|\, 0 < l < s\}.$$

If $k_{r+1} - k_r$ is an even number for $1 \leqslant r \leqslant m$, then $|E| = m$ and $[i_0] = [i_s]$. In all other cases, $|E| = m - 1$ and $[i_0] \neq [i_s]$.

Proof of Lemma 19. We prove the result by induction on m. For the induction base case, suppose $m = 1$. If s is an even number, then $k_2 - k_1 = s$ is even. We have $[i_0] \sim [i_2]$, $[i_2] \sim [i_4], \ldots, [i_{s-2}] \sim [i_s]$, and so we have $i_0 \sim i_s$. All the elements indexed by the even numbers are in one equivalence class and the ones indexed by odd numbers are in a different equivalence. Since there are only two equivalence classes and one of them is $[i_0]$, $E = [i_1]$. Now suppose that s is an odd number. Then we have $[i_0] \sim [i_2], [i_2] \sim [i_4], \ldots, [i_{s-3}] \sim [i_{s-1}]$ and $[i_1] \sim [i_3], [i_3] \sim [i_5], \ldots, [i_{s-2}] \sim [i_s]$. There are two equivalence classes but $E = \phi$. Further $[i_0] \neq [i_s]$ as s is odd. This proves the base case.

Now suppose $m > 1$. Suppose there exists k_r such that $k_{r+1} - k_r$ is even. Then with similar reasoning as in the base case above, we have $[i_{k_r}] = [i_{k_{r+1}}]$ and $[i_{k_r}] \neq [i_{k_r+1}]$. Further by the definition of the relation, there is no sequence of elements such that $i_{k_r+1} \, R \, i_{l_1}, i_{l_1} \, R \, i_{l_2}, \ldots, i_{l_{j-1}} \, R \, i_{l_j}$ where $l_j < k_r$ or $l_j > k_{r+1}$. Thus $[i_{k_r+1}] \neq [i_l]$ for $l < k_r$ and $l > k_{r+1}$. Thus we have

$$E = [i_{k_r+1}] \bigsqcup \{[i_l] \,|\, 0 < l \leqslant k_r \text{ or } k_{r+1} < l < s\}. \tag{5}$$

Consider the shortened sequence $(i_0, i_1, \ldots, i_{k_r}, i_{k_{r+1}+1}, i_{k_{r+1}+2}, \ldots, i_s)$ along with the relation R defined using the subsequence $0 = k_1 < k_2 < \cdots < k_r < k_{r+2} < k_{r+3} < \cdots < k_{m+1} = s$. There are $m - 1$ elements in the subsequence and so by induction, $\{[i_l] \,|\, 0 < l \leqslant k_r \text{ or } k_{r+1} < l < s\}$ has $m - 2$ or $m - 1$ elements depending upon whether $k_{l+1} - k_l$ is even for $1 \leqslant r' \leqslant m, r' \neq r + 1$. By equation 5, this shows the result.

Suppose now instead that there does not exist k_r such that $k_{r+1} - k_r$ is even. Then with similar reasoning as in the base case, for any $1 \leqslant r \leqslant m$ we have for $k_r < l < k_{r+1}$ that $[i_l] = [i_{k_r}]$ or $[i_l] = [i_{k_{r+1}}]$. Further $[i_{k_r}] \neq [i_{k_{r+1}}]$ as $k_{r+1} - k_r$ is odd and by the definition of R, elements can only be equivalent to each other at an even distance from each other. Next we claim $[i_{k_r}] \neq [i_{k_{r'}}]$ for $r \neq r'$. Suppose on the contrary that it were the case that $[i_{k_r}] = [i_{k_{r'}}]$ for some $r \neq r'$. Then let $i_{k_r} \, R \, i_{l_1} \, R \, i_{l_2} \cdots R \, i_{l_j} \, R \, i_{k_{r'}}$ and assume that this is the shortest sequence of elements relating i_{k_r} and $i_{k_{r'}}$. Let $l_{j'}$ be the first index in the sequence such that $l_{j'} = k_{r''}$ for some r''. By definition of R, if $i \, R \, i'$ then $k_p \leqslant i, i' \leqslant k_{p+1}$ for some p. Thus $l_{j'} = k_r$ or $l_{j'} = k_{r+1}$ or $l_{j'} = k_{r-1}$. We cannot have the latter two cases as we have shown that $[i_{k_r}] \neq [i_{k_{r+1}}]$ and $[i_{k_r}] \neq [i_{k_{r-1}}]$. Thus $l_{j'} = k_r$. This implies that $l_{j'} \, R \, l_{j'+1} \, R \cdots R \, i_{k_{r'}}$ is a shorter sequence of elements relating i_{k_r} and $i_{k_{r'}}$. But this is a contradiction to the fact that we started with the shortest sequence with such a property. Thus we conclude that

$$E = \{[i_l] \,|\, l = k_r, 1 < r \leqslant m\}, \quad |E| = m - 1.$$

This completes the proof.

■

BIBLIOGRAPHY

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

Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer-Verlag, New York.

Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., and Smola, A. J. (2013). Distributed large-scale natural graph factorization. In *Proceedings of the 22nd international conference on World Wide Web*, pages 37–48.

Aicher, C., Jacobs, A. Z., and Clauset, A. (2015). Learning Latent Block Structure in Weighted Networks. *Journal of Complex Networks*, 3(2):221–248.

Airoldi, E., Blei, D., Xing, E., and Fienberg, S. (2005). A latent mixed membership model for relational data. In *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, pages 82–89, Chicago, Illinois. ACM Press.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009). Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248.

Arratia, R., Goldstein, L., and Gordon, L. (1990). Poisson approximation and the chen-stein method. *Statistical Science*, pages 403–424.

Bai, Z. D. and Silverstein, J. W. (2008). Clt for linear spectral statistics of large-dimensional sample covariance matrices. In *Advances In Statistics*, pages 281–333. World Scientific.

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

Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2017). Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):295–314.

Bassett, D. S., Bullmore, E., Verchinski, B. A., Mattay, V. S., Weinberger, D. R., and Meyer-Lindenberg, A. (2008). Hierarchical Organization of Human Cortical Networks in Health and Schizophrenia. *Journal of Neuroscience*, 28(37):9239–9248.

Bassett, D. S. and Bullmore, E. T. (2009). Human Brain Networks in Health and Disease. *Current opinion in neurology*, 22(4):340–347.

Battiston, F., Nicosia, V., and Latora, V. (2014). Structural measures for multiplex networks. *Physical Review E*, 89(3):032804. Publisher: American Physical Society.

Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Bhamidi, S., Dey, P. S., and Nobel, A. B. (2012). Energy landscape for large average submatrix detection problems in gaussian random matrices. *arXiv preprint arXiv:1211.2284*.

Bianconi, G. (2013). Statistical mechanics of multiplex networks: Entropy and overlap. *Physical Review E*, 87(6):062806. Publisher: American Physical Society.

Bianconi, G., Dorogovtsev, S. N., and Mendes, J. F. F. (2015). Mutually connected component of networks of networks with replica nodes. *Physical Review E*, 91(1):012804.

Birke, M. and Dette, H. (2005). A note on testing the covariance matrix for large dimension. *Statistics & probability letters*, 74(3):281–289.

Bordenave, C., Lelarge, M., and Massoulié, L. (2015). Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS'15)*, pages 1347–1357. IEEE.

Borge-Holthoefer, J., Rivero, A., García, I., Cauhé, E., Ferrer, A., Ferrer, D., Francos, D., Iñiguez, D., Pérez, M. P., Ruiz, G., Sanz, F., Serrano, F., Viñas, C., Tarancón, A., and Moreno, Y. (2011). Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study. *PLOS ONE*, 6(8):e23883. Publisher: Public Library of Science.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford. Google-Books-ID: koNqWRluhP0C.

Cai, T. T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39:1496–1525.

Cai, T. T. and Jiang, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39.

Cao, S., Lu, W., and Xu, Q. (2015). Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 891–900.

Cao, S., Lu, W., and Xu, Q. (2016). Deep neural networks for learning graph representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Cardillo, A., Gómez-Gardeñes, J., Zanin, M., Romance, M., Papo, D., Pozo, F. d., and Boccaletti, S. (2013). Emergence of network features from multiplexity. *Scientific Reports*, 3(1):1344. Number: 1 Publisher: Nature Publishing Group.

Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., and Murphy, K. (2020). Machine learning on graphs: A model and comprehensive taxonomy. *arXiv preprint arXiv:2005.03675*.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162(1):47–70.

Chung, F. and Lu, L. (2002). Connected Components in Random Graphs with Given Expected Degree Sequences. *Annals of Combinatorics*, 6(2):125–145.

Cozzo, E., Arenas, A., Moreno, Y., Gómez, S., Porter, M. A., De Domenico, M., Kivelä, M., and Solé, A. (2013). Clustering Coefficients in Multiplex Networks. Number: arXiv:1307.6780 Pages: 073029 Volume: 7.

Cullen, M. J. P., Norbury, J., and Purser, R. J. (1991). Generalised Lagrangian Solutions for Atmospheric and Oceanic Flows. *SIAM Journal on Applied Mathematics*, 51(1):20–31. Publisher: Society for Industrial and Applied Mathematics.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition. Elementary theory and methods.

Daley, D. J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes. Vol. II.* Probability and its Applications (New York). Springer, New York, second edition. General theory and structure.

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726. Publisher: Oxford Academic.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Fan, J. and Jiang, T. (2019). Largest entries of sample correlation matrices from equi-correlated normal populations. *Annals of Probability*.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics reports*, 659:1–44.

Frisch, U., Matarrese, S., Mohayaee, R., and Sobolevski, A. (2002). A reconstruction of the initial conditions of the Universe by optimal mass transportation. *Nature*, 417(6886):260–262. Number: 6886 Publisher: Nature Publishing Group.

Gao, J., Buldyrev, S. V., Havlin, S., and Stanley, H. E. (2012). Robustness of a network formed by n interdependent networks with a one-to-one correspondence of dependent nodes. *Physical Review E*, 85(6):066134.

Gao, J., Buldyrev, S. V., Stanley, H. E., Xu, X., and Havlin, S. (2013). Percolation of a general network of networks. *Physical Review E*, 88(6):062816.

Gordon, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289.

Gordon, Y. (1987). Elliptically contoured distributions. *Probability Theory and Related Fields*, 76(4):429–438.

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

Guimerà, R., Mossa, S., Turtschi, A., and Amaral, L. a. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Haker, S., Zhu, L., Tannenbaum, A., and Angenent, S. (2004). Optimal Mass Transport for Registration and Warping. *International Journal of Computer Vision*, 60(3):225–240.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017a). Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017b). Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 40(3):52–74.

Han, Q., Xu, K., and Airoldi, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning*, pages 1511–1520.

He, Y., Chen, Z., and Evans, A. (2008). Structural Insights into Aberrant Topological Patterns of Large-Scale Cortical Networks in Alzheimer's Disease. *Journal of Neuroscience*, 28(18):4756–4766.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*, pages 657–664.

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

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

Horvath, S. and Dong, J. (2008). Geometric Interpretation of Gene Coexpression Network Analysis. *PLOS Computational Biology*, 4(8):e1000117. Publisher: Public Library of Science.

Janson, S., Luczak, T., and Rucinski, A. (2000). *Random Graphs*. John Wiley & Sons.

Jiang, D., Jiang, T., and Yang, F. (2012). Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference*, 142(8):2241–2256.

John, S. (1971). Some optimal multivariate tests. *Biometrika*, 58(1):123–127.

Johnstone, I. M. et al. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327.

Kahane, J.-P. (1986). Une inegalite du type de Slepian et Gordon sur les processus gaussiens. *Israel Journal of Mathematics*, 55(1):109–110.

Kallenberg, O. (1973). Characterization and convergence of random measures and point processes. *Probability Theory and Related Fields*, 27(1):9–21.

Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Springer.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., and Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940.

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

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer-Verlag, New York.

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (2012). *Extremes and Related Properties of Random Sequences and Processes*. Springer Science & Business Media.

Ledoit, O., Wolf, M., et al. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, 30(4):1081–1102.

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

Leicht, E. and D'Souza, R. M. (2009). Percolation on interacting networks. *arXiv preprint arXiv:0907.0894*.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27:2177–2185.

Li, W. (1999). A Gaussian Correlation Inequality and its Applications to Small Ball Probabilities. *Electronic Communications in Probability*, 4:111–118. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Li, W. V. and Shao, Q. M. (2001). Gaussian processes: Inequalities, small ball probabilities and applications. In *Handbook of Statistics*, volume 19 of *Stochastic Processes: Theory and Methods*, pages 533–597. Elsevier.

Lusher, D., Koskinen, J., and Robins, G. (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

Matias, C., Rebafka, T., and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Miller, K., Jordan, M. I., and Griffiths, T. L. (2009). Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, pages 1276–1284.

Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016). Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophysical Journal International*, 205(1):345–377. Publisher: Oxford Academic.

Nagao, H. (1973). On some test criteria for covariance matrix. *The Annals of Statistics*, pages 700–709.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Ng, M. K.-P., Li, X., and Ye, Y. (2011). MultiRank: co-ranking for objects and relations in multi-relational data. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1217–1225, New York, NY, USA. Association for Computing Machinery.

Ou, M., Cui, P., Pei, J., Zhang, Z., and Zhu, W. (2016). Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1105–1114.

Palla, K., Knowles, D., and Ghahramani, Z. (2012). An infinite latent attribute model for network data. *arXiv preprint arXiv:1206.6416*.

Paul, S. and Chen, Y. (2018). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *arXiv:1805.02292 [stat]*.

Péché, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probability Theory and Related Fields*, 143(3-4):481–516.

Peixoto, T. P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807.

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

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport. *arXiv:1803.00567 [stat]*.

Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J. (2018). Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 459–467.

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

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b). Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2):192–215.

Rudin, W. (1987). *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., USA.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261.

Shao, Q.-M. and Zhou, W.-X. (2014). Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *The Annals of Probability*, 42(2):623–648.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *The Bell System Technical Journal*, 41(2):463–501. Conference Name: The Bell System Technical Journal.

Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics*, 108(5-6):1033–1056.

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society*, 35(2):251–272.

Stam, C. J. (2014). Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10):683–695. Number: 10 Publisher: Nature Publishing Group.

Stam, C. J., Jones, B. F., Nolte, G., Breakspear, M., and Scheltens, P. (2007). Small-World Networks and Functional Connectivity in Alzheimer's Disease. *Cerebral Cortex*, 17(1):92–99.

Stanley, N., Shai, S., Taylor, D., and Mucha, P. J. (2016). Clustering Network Layers With the Strata Multilayer Stochastic Block Model. *IEEE Transactions on Network Science and Engineering*, 3(2):95–105.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077.

Teicher, M. H., Samson, J. A., Anderson, C. M., and Ohashi, K. (2016). The effects of childhood maltreatment on brain structure, function and connectivity. *Nature Reviews Neuroscience*, 17(10):652–666. Number: 10 Publisher: Nature Publishing Group.

van den Heuvel, M. P. and Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534.

Vitale, R. A. (2000). Some Comparisons for Gaussian Processes. *Proceedings of the American Mathematical Society*, 128(10):3043–3046. Publisher: American Mathematical Society.

Wang, D., Cui, P., and Zhu, W. (2016). Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Willink, R. (2004). Bounds on the bivariate normal distribution function. *Comm. Statist. Theory Methods*, 33(10):2281–2297.

Wilson, J. D., Palowitch, J., Bhamidi, S., and Nobel, A. B. (2017). Community extraction in multilayer networks with heterogeneous community structure. *arXiv:1610.06511 [physics, stat]*.

Xu, H., Luo, D., Zha, H., and Duke, L. C. (2019). Gromov-Wasserstein Learning for Graph Matching and Node Embedding. In *International Conference on Machine Learning*, pages 6932–6941. PMLR. ISSN: 2640-3498.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zhang, D., Yin, J., Zhu, X., and Zhang, C. (2018). Network representation learning: A survey. *IEEE transactions on Big Data*, 6(1):3–28.

Zhang, X., Moore, C., and Newman, M. E. J. (2017). Random graph models for dynamic networks. *The European Physical Journal B*, 90(10):200.

Zhang, Y. and Tang, M. (2021). Consistency of random-walk based network embedding algorithms. *arXiv preprint arXiv:2101.07354*.

Zheng, S., Bai, Z., Yao, J., et al. (2015). Substitution principle for clt of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *The Annals of Statistics*, 43(2):546–591.