Author:
Modell, Alexander D

Title:
Spectral embedding of large graphs and dynamic networks

Spectral embedding of large graphs and dynamic networks

by

Alexander Modell

School of Mathematics
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science.

September 2023

Word count: nineteen thousand, three hundred and thirty-four.

Abstract

The analysis of network data, describing relationships, interactions and dependencies between entities, often begins with embedding: the process of mapping these entities into a low-dimensional vector space in a way which preserves salient information present in the data. Spectral embedding is a family of embedding algorithms in which representations are obtained using the eigenvectors of a specially designed matrix constructed from the network. It has emerged as a simple yet effective approach, which is both highly scalable and interpretable.

In this thesis, we provide a statistical lens into spectral embedding, shining light on how various network structures manifest themselves as geometric patterns in the vector space, and how certain algorithmic choices influence the information which is extracted from the network. We present new methodology which exploits these insights and provide statistical theory, as well as simulated and real data studies to support them.

Chapter 2 introduces some statistical models for network data and reviews existing estimation theory for spectral embedding; Chapter 3 studies spectral embedding using the random walk Laplacian matrix, developing estimation theory which illuminates a key inferential difference between it and other popular matrix constructions; Chapter 4 elucidates the geometric structure which emerges in the spectral embeddings of multipartite networks and develops bespoke statistical methodology which exploits it for dimension reduction; and Chapter 5 presents an algorithmic framework for spectral embedding of dynamic networks which produces representations that evolve in continuous time and reflect the changing structural roles of the nodes in the network.

Acknowledgements and Dedication

First of all, I owe an enormous debt of gratitude to my PhD supervisor, Patrick Rubin-Delanchy. Patrick's enthusiasm towards research and life more generally is infectious, and his kindness and selflessness inspire me daily. He is an excellent teacher, a remarkable thinker and I cannot envision anyone better to have guided me through the adventure that has been the past four years.

Another person I owe immeasurably is Carey Priebe. It was him, with his flamboyant charisma and cryptic wisdom, who, alongside Patrick, first got me hooked on statistics research during my undergraduate degree, and inspired me to pursue this PhD.

I have had the pleasure of working with and learning from a wonderful research group in Bristol — Nick Whiteley, Ian Gallagher, Ed Davis, Annie Gray, Hannah Sansford and Emma Ceccherini — together, you make a supportive, creative and energised team which I have been proud to be a part of.

I was fortunate enough to undertake a summer internship at Microsoft Research, and I thank Anna Bertiger, Jonathan Larson and Melissa Turcotte for making it such an enriching experience.

I must also thank Joshua Cape for his efforts on the work we have done together and Joshua Agterberg, Hayden Helm, Avanti Athreya, Keith Levin, Jesús Arroyo and Vince Lysinski for stimulating discussions about network analysis, spectral methods and statistics more broadly which have fuelled my enduring enthusiasm for these subjects.

Next, I would like to express my gratitude to my personal advisor, Oliver Johnson, all the staff on the Compass PhD programme, in particular to Liz, Crina, Harriet and Helen, and to all of my peers in the Fry Building.

My journey to where I am today began long before I started post-graduate study, and I owe a particular thanks to my A-level maths teachers Josh and Tom for inspiring me to pursue mathematics in the first place. Without their direction, I would have likely ended up doing something entirely different.

I am deeply grateful for the support and encouragement of my friends who have made these four years so profoundly enjoyable. There are too many of you to list but you know who you are. I owe a particular thank you to Rachel for all her love and support in everything I do.

Finally, I owe the most to my family. It is their unconditional love and support that makes all of my successes possible.

This thesis is dedicated to my late nan, Christine, who would have been so proud to call me a doctor.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .... DATE: ....

Table of Contents

List of Tables

List of Figures

Figure **Page**

Chapter 1

Introduction

In the ever-expanding landscape of modern data science, the study of network data describing relationships, interactions, and dependencies between entities has emerged as a pivotal field, with applications spanning a diverse set of domains.

Information about personal relationships, message exchanges, and physical and virtual interactions are collected on mass scales online, and making sense of these relationships is vital for a diverse range of societal applications, from understanding the spread of disease and misinformation to detecting organised crime such as human trafficking and corruption. In the biological sciences, the study of protein-protein interactions, gene regulatory networks, and metabolic pathways provides a rich web of relationships that govern the fundamental processes of life. Understanding these connections is essential for advancements in personalised medicine, drug discovery, and disease modelling. Computer networks produce a deluge of relational data describing packet transfers and user authentication, the analysis of which is imperative for protecting against intrusions and cyber-attacks.

Network data can be broadly categorised as being either static or dynamic. In a static network, the relationship between a pair of entities, known as nodes, can be described using a single number such as a one or a zero, indicating the presence, or not, of a connection. In this thesis, we will also use the word *graph* to describe a static network. In a dynamic network, one instead observes a process between each pair of nodes which takes place over time, such as a point process or a time series.

Exploratory analysis of network data often begins with embedding: the process of mapping the nodes into a low-dimensional vector space in a way which preserves salient information present in the data. These representations provide a holistic view of the underlying relationships, allowing for visual exploration of patterns and latent structures, such as communities, which may be masked by the complexity of the raw data. Embedding is also used as a precursor to many forms of inference, such as clustering, regression, classification and neighbour recommendation, which require Euclidean data as inputs.

Spectral embedding is a family of embedding algorithms for static networks in which representations are obtained using the eigenvectors of a specially designed matrix constructed from the network. It has emerged as a simple yet effective approach, which is both highly scalable and interpretable. The geometric patterns which emerge in spectral embeddings fundamentally depend on two things: the underlying structures present in the network, and the construction of the matrix used for the embedding. Different matrix constructions extract different information from the network, and studying the precise nature of this interplay is an active area of research.

The first goal of this thesis is to build upon this research by elucidating the fundamental inferential differences between some popular matrix constructions. In particular, we focus on the random walk Laplacian matrix and explore some of the methodological implications of our insights. The second goal of this thesis is to demonstrate how special geometric structure emerges in the spectral embeddings of a class of networks known as multipartite networks, and to develop an algorithmic extension to spectral embedding which exploits it for further dimension reduction.

One limitation of the standard spectral embedding algorithm is that, in order to analyse a dynamic network, one must first summarise it as a static network, for example via counting or averaging. This makes it unsuitable for distilling information about the temporal aspects of the data. The final goal of this thesis is to develop a framework for spectral embedding of dynamic networks, which produces representations that evolve in continuous time and reflect the changing structural roles of the nodes in the network, allowing inference in the temporal domain.

1.1 Overview of thesis

We now give a brief chapter-by-chapter overview of this thesis.

In Chapter 2, we review some existing statistical estimation theory for spectral embedding using the eigenvectors of the adjacency and symmetric normalised Laplacian matrices, which lays the foundation for the statistical estimation theory we develop in Chapters 3 and 4. This theory is based on the generalised random dot product graph, a generic model for low-rank random graphs with independent edges, and provides a model-based interpretation of some the geometric structures often observed in these embeddings, such as clusters, rays, and simplexes.

In Chapter 3, we develop statistical estimation theory for spectral embedding using the eigenvectors of the random walk Laplacian matrix which illuminates the fundamental inferential differences between it and the embeddings discussed in this previous chapter. We then explore the methodological implications of our theory for clustering. To illustrate our theory, we present an exploratory analysis of a network describing the enmity relationships between the characters of J.K. Rowling's renowned Harry Potter series. This chapter is based on joint work with Patrick Rubin-Delanchy, and benefitted from conversations with Sean Dewar and Kevin Hughes. An earlier version of this work was posted on ArXiv in May 2021 [1].

In Chapter 4, we elucidate the geometric structure which emerges in the spectral embeddings of multipartite networks and develop bespoke statistical methodology which exploits it for dimension reduction. We develop statistical estimation theory for our new algorithm and demonstrate its effectiveness via an exploratory analysis of a large multipartite network derived from data repositories supporting biomedical research, linking groups of entities such as drugs, diseases, targets, pathways, variant locations and haplotypes. This chapter is adapted from joint work with Ian Gallagher, Joshua Cape and Patrick Rubin-Delanchy, an earlier version of which was posted on ArXiv in February 2022 [2]. The data analysed in this chapter was kindly provided by Nansu Zong.

In Chapter 5, we consider dynamic networks in which data is in the form of a collection of instantaneous interaction events which occur between nodes in continuous time. We develop a new family of spectral embedding algorithms, which we name “Intensity Profile Projection”, which produce low-dimensional trajectories for the nodes in the network: vectors which evolve in continuous time, encoding their changing structural roles in the network. We demonstrate that the learned embeddings possess two key properties known as temporal coherence and structure preservation, which allow them to be meaningfully compared to different points in time. We support our algorithm with statistical estimation theory and demonstrate it via an exploratory analysis of a dataset describing face-to-face interactions between pupils at a primary school in Lyon. This chapter is adapted from joint work with Ian Gallagher, Emma Ceccherini, Nick Whiteley and Patrick Rubin-Delanchy, which has been accepted for publication at NeurIPS 2023. This work was posted on ArXiv in June 2023 [3].

All authors acknowledged and I have made substantive contributions to the development of the ideas presented in Chapters 3, 4 and 5. All of the proofs in this thesis, and the simulated and real data analysis in Chapters 3 and 4 are my sole contribution. The simulated and real data analyses in Chapter 5 were supported by Ian Gallagher and Emma Ceccherini.

1.2 Notation

We pause here to define some notation and conventions which we will use throughout this thesis. The symbols $:=$ and $\equiv$ are used to assign definitions and denote formal equivalence, for two scalars a, b, we write $a \vee b := \max\{a, b\}$ and for any positive integer n, we use the shorthand $[n] := \{1, \dots, n\}$. We use bold letters to denote matrices and regular letters to denote vectors and scalars. We write m_{ij} and M_i to denote the ijth entry and ith row (viewed as a column vector) of a matrix $\mathbf{M}$, respectively. We use the notation $\text{diag}(x_1, \dots, x_d)$ to denote the diagonal matrix with entires $x_1, \dots, x_d$. For any vector x, we use $\|x\|_p$ to denote its ℓ_p-norm and when $p = \infty$, $\|x\|_\infty := \max_i |x_i|$. For any matrix $\mathbf{M}$, $\|\mathbf{M}\|_p$ denotes its corresponding matrix norm and $\|\mathbf{M}\|_{2,\infty} := \max_i \|M_i\|_2$ denotes its $\ell_{2,\infty}$ norm [4]. Additionally, we write $\sigma_i(\mathbf{M})$ to denote its ith largest singular value.

We use the phrase *"for sufficiently large n"* to mean *"there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$"* and we say an event E_n, which depends on n, occurs *with overwhelming probability* if for any constant $c > 0$, there exists a constant $C > 0$, which may depend on c, such that $\mathbb{P}(E_n) \geq 1 - Cn^{-c}$. We use $\lesssim$ to denote the inequality $\leq$ which hides a multiplicative universal constant and, when qualified with the prior probabilistic statement, the constant c. Additionally, we write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. For any two quantities a_n, b_n, depending on n, we write $a_n \ll b_n$ to mean that $a_n/b_n \to 0$ as $n \to \infty$.

Chapter 2

Background

This chapter is devoted to reviewing some statistical estimation theory for spectral embedding, which provides a model-based explanation for many of the geometric patterns observed in the embeddings of real-world networks. The insights presented here are the result of a large body of literature that stems from seminal work on spectral embedding under the stochastic block model [5, 6] and the random dot product graph model [7, 8].

The geometry of a spectral embedding of a graph fundamentally depends on two things: the matrix construction employed in the spectral decomposition and the graph itself. In this chapter, we will consider two matrix constructions: the adjacency matrix, and the symmetric, normalised Laplacian matrix. We treat the random walk Laplacian matrix separately in Chapter 3.

To motivate our discussion, we consider a graph whose nodes are the characters of the Harry Potter novels by J.K. Rowling [9], and with edges between characters who are enemies in the story. This is a publicly available data set [10] that has previously been studied in [11]. Figure 2.1 shows graph embeddings obtained using the first two eigenvectors of the adjacency, and symmetric Laplacian matrices, respectively.

On inspection of the embeddings, one observes that, approximately speaking, the “good” and “evil” characters in the story are concentrated around two distinct rays, and the magnitude of a node’s position broadly reflects the character’s importance in the story. For example, Harry Potter and Lord Voldemort, the protagonist and antagonist of the books, respectively, are positioned at the ends of their respective rays. This section will provide a model-based explanation for the observed patterns in these embeddings.

Many random graph models have been proposed [8, 12–14], and here we restrict our attention to those in which edges occur randomly, and independently of one another, and for which the matrix containing these edge probabilities, which we denote $\mathbf{P}$, has low rank. A major advantage of these assumptions is simplicity: edge independence permits the use of classical concentration inequalities and the low-rank assumption allows convenient decompositions and the direct application of tools from matrix perturbation theory. In practice, these assumptions are unlikely to hold exactly, however, in their full generality, they often provide a reasonable

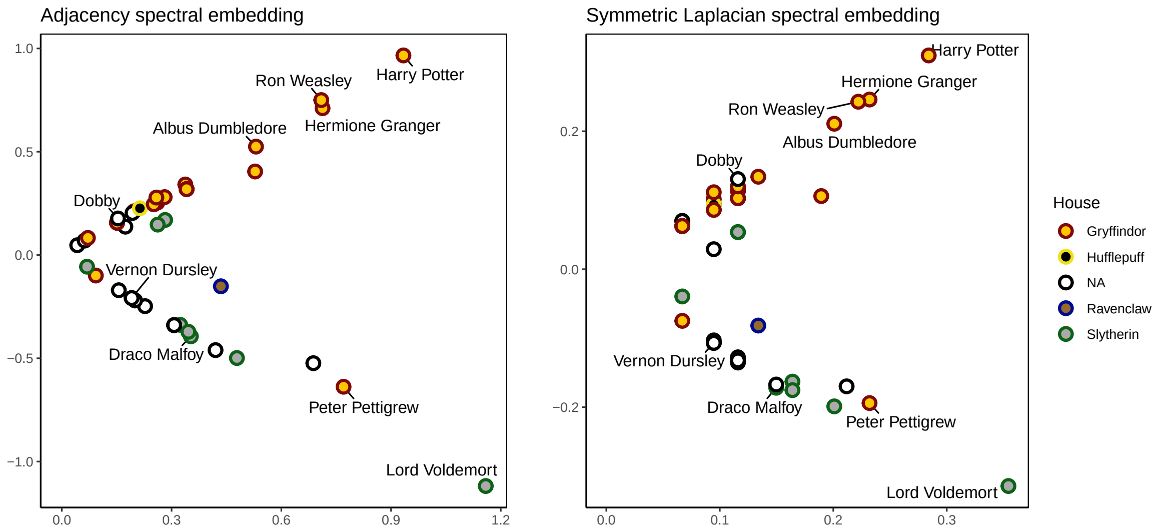


Figure 2.1: The adjacency spectral embedding (left) and symmetric Laplacian spectral embedding (right) of a graph of enmities between characters in the Harry Potter book series. Colour indicates the “house” to which the character belongs in the Hogwarts School, and some points are labelled with the character to which they correspond.

approximation to the truth.

2.1 The generalised random dot product graph model

A useful parametrisation of the independent-edge, low-rank random graph model is as a latent position model [12]. Let r denote the rank of $\mathbf{P}$ and let (p, q) denote its signature (that is, its number of positive and negative eigenvalues, respectively). Define the indefinite inner product with signature (p, q), $\langle \cdot, \cdot \rangle_{p,q}$ by

$$\langle x, y \rangle_{p,q} := \sum_{i=1}^{p} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i \equiv x^\top \mathbf{I}_{p,q} y, \tag{2.1}$$

for $x, y \in \mathbb{R}^{p+q}$, where $\mathbf{I}_{p,q}$ is the diagonal matrix of p ones followed by q minus-ones. It is always possible to choose vectors $X_1, \ldots, X_n \in \mathbb{R}^r$ such that $p_{ij} = \langle X_i, X_j \rangle_{p,q}$. This parametrisation is known in the literature as the *generalised random dot product graph model* [15], and we highlight that any low-rank, independent-edge random graph can be parametrised in this way. We give a formal definition below:

Definition 2.1 (Generalised random dot product graph model [15])**.** A graph is said to follow a generalised random dot product graph model with signature (p, q) and latent positions $X_1, \ldots, X_n \in \mathbb{R}^r$, if $\{a_{ij}\}_{i<j}$ are independent Bernoulli random variables with success probabilities

$$p_{ij} = \langle X_i, X_j \rangle_{p,q} \in [0, 1], \qquad 1 \leq i < j \leq n.$$

It should be noted that this parametrisation is not unique. There are two distinct sources of non-identifiability in the latent positions of a generalized random dot product graph. First, one can increase the dimension of the latent positions, for example by padding them with zeroes, without changing $\mathbf{P}$. We preclude such parametrisations by requiring that $r = \text{rank}(\mathbf{P})$, or equivalently that the latent position span $\mathbb{R}^r$. Second, let $\mathbb{O}(p, q) = \{\mathbf{Q} : \mathbf{Q}^\top \mathbf{I}_{p,q}\mathbf{Q} = \mathbf{I}_{p,q}\}$ denote the indefinite orthogonal group of signature (p, q), i.e. the group of transformations which leave the indefinite inner product invariant. Then, replacing $\{X_i\}_{i=1}^n$ with $\{\mathbf{Q}X_i\}_{i=1}^n$ for any matrix $\mathbf{Q} \in \mathbb{O}(p, q)$ does not change $\mathbf{P}$. If $\mathbf{P}$ has distinct eigenvalues, the model can be made identifiable [16], however, we prefer to consider $\{\mathbf{Q}X_i : \mathbf{Q} \in \mathbb{O}(p, q)\}_{i=1}^n$ as an equivalence class of latent positions, and the theory we present reflects this.

The following lemma, which is a generalisation of Lemma 1 of Rubin-Delanchy et al. [15] and gives some control on the extent to which this non-identifiability can distort the latent positions. A proof is given in Section A.1.

Lemma 2.1. *Let $\mathbf{P} \in [0, 1]^{n\times n}$ be a rank-r probability matrix with signature (p, q) and reduced condition number $\kappa = \sigma_1(\mathbf{P})/\sigma_r(\mathbf{P})$. Let $\{X_i\}_{i=1}^n$, $\{X_i'\}_{i=1}^n$ be two sets of r-dimensional latent positions such that $p_{ij} = \langle X_i, X_j\rangle_{p,q} = \langle X_i', X_j'\rangle_{p,q}$. Then the indefinite orthogonal matrix $\mathbf{Q} \in \mathbb{O}(p, q)$ such that $X_i = \mathbf{Q}X_i'$ for all $i \in [n]$ satisfies $\|\mathbf{Q}\|_2, \|\mathbf{Q}^{-1}\|_2 \leq \kappa$.*

2.2 Adjacency spectral embedding

In this section, we define adjacency spectral embedding and present estimation theory which makes formal the sense in which it performs statistical inference on the generalised random dot product graph model.

Definition 2.2 (Adjacency spectral embedding)**.** Suppose $\mathbf{A}$ has the eigendecomposition $\mathbf{A} = \sum_{i=1}^n \hat\lambda_i \hat u_i \hat u_i^\top$ with $|\hat\lambda_1| \geq \cdots \geq |\hat\lambda_n|$. The adjacency spectral embedding of the graph into $\mathbb{R}^r$, denoted $\hat X_1, \ldots, \hat X_n \in \mathbb{R}^r$, is given by the rows of the matrix

$$\hat{\mathbf{X}} = \begin{pmatrix} \hat X_1^\top \\ \vdots \\ \hat X_n^\top \end{pmatrix} := \begin{pmatrix} |\hat\lambda_1|^{1/2}\hat u_1 & \cdots & |\hat\lambda_r|^{1/2}\hat u_r \end{pmatrix}$$

obtained by stacking the scaled eigenvectors $|\hat\lambda_1|^{1/2}\hat u_1, \ldots, |\hat\lambda_r|^{1/2}\hat u_r$ in columns.

We will review estimation theory developed in Rubin-Delanchy et al. [15] and Xie [17] which makes formal the following statement:

> Under a generalised random dot product graph model, the adjacency spectral embedding, $\hat X_1, \ldots, \hat X_n$, is a uniformly consistent estimate of the latent positions $X_1, \ldots, X_n$, with asymptotically Gaussian error.

The precise nature of the mathematical statements which can be made depends on whether the latent positions are treated as fixed, or random quantities. Under the fixed setup, the latent positions are treated as deterministic quantities, and under the random setup, the latent positions are treated as realisations of independent and identically distributed random variables. While the distinction is philosophical from a practical standpoint, the mathematical tools required for their analysis are very different.

The casual reader who is content with the informal statement above may skip to Section 2.3. For the technical reader, the following subsections are dedicated to reviewing the results by Rubin-Delanchy et al. [15] and Xie [17] which formalise it.

2.2.1 Estimation theory with random latent positions

In this subsection, we review the estimation theory developed in Rubin-Delanchy et al. [15] for adjacency spectral embedding under the generalised random dot product graph model with random latent positions.

In this setup, a sequence of graphs $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$ is considered, where for each $n \in \mathbb{N}$, a set of n r-dimensional independent and identically-distributed random vectors $X_1^{(n)}, \ldots, X_n^{(n)}$ are drawn from a probability distribution, conditional upon which, the graph $\mathbf{A}^{(n)}$ follows a generalised random dot product graph model with signature (p, q) and latent positions $X_1^{(n)}, \ldots, X_n^{(n)}$.

We denote the expected node degrees by $t_i^{(n)} = \sum_{j=1}^n \langle X_i, X_j \rangle_{p,q}$ for $i \in [n]$. Their distribution is either assumed to be fixed or, in order to permit asymptotic regimes in which node degrees grow less than linearly in n, it is allowed to shrink. To achieve this, a sequence $\{\rho_n\}_{n\in\mathbb{N}}$ is introduced which is either fixed or shrinks towards zero, and for each $i \in [n]$, we set $X_i^{(n)} = \rho_n^{1/2} \xi_i^{(n)}$, where $\xi_1^{(n)}, \ldots, \xi_n^{(n)}$ are i.i.d. draws from a distribution F.

It is necessary to make two common-sense assumptions on the support of F, which we denote by $\mathcal{X}$: first, that inner products between any points in $\mathcal{X}$ give valid probabilities, i.e. for any $x, y \in \mathcal{X}$, $\langle x, y \rangle_{p,q} \in [0, 1]$; and second, that $\mathcal{X}$ spans $\mathbb{R}^r$, which ensures that r is not larger than necessary.

Additionally, it is assumed that $n\rho_n$, which scales with the expected node degrees, grows at least polylogarithmically in n, which is stated precisely in the theorem statements. This assumption could likely be relaxed to logarithmic degree growth using cutting-edge proof techniques which have emerged since these theorems first appeared (e.g. those employed in Xie [17]).

The first result is a consistency result, which states that subject to an indefinite orthogonal transformation, the maximum error between a node's position in the adjacency spectral embedding and its latent positions vanishes for large graphs.

Theorem 2.1 (Uniform consistency [15]). *Suppose that $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$ is a sequence of graphs generated as described in Section 2.2.1 and the sparsity factor satisfies $n\rho_n \gg \log^{4c} n$ where $c > 0$ is a universal constant. Then, there exists a sequence of indefinite orthogonal transformations*

$\{\mathbf{Q}^{(n)} \in \mathbb{O}(p,q)\}_{n\in\mathbb{N}}$ such that, for sufficiently large n, the adjacency spectral embedding $\hat{X}_1^{(n)}, \ldots, \hat{X}_n^{(n)}$, satisfies

$$\max_{i\in\{1,\ldots,n\}} \left\| \mathbf{Q}^{(n)}\hat{X}_i^{(n)} - X_i^{(n)} \right\|_2 \lesssim \frac{\log^c n}{n^{1/2}}$$

with overwhelming probability.

The second result is a central limit theorem. It states that for a fixed, finite subset of nodes, indexed without loss of generality as $1, \ldots, m$, their error distributions, scaled by $n^{1/2}$, are asymptotically Gaussian.

Theorem 2.2 (Asymptotic normality [15])**.** *Assume the setting of Theorem 2.1. Conditional on $\xi_i^{(n)} = x_i$, for $i = 1, \ldots, m$, $n \geq m$, the random vectors $n^{1/2}(\mathbf{Q}^{(n)}\hat{X}_i^{(n)} - X_i^{(n)})$ converge in distribution to independent mean-zero normal random vectors with covariance matrices $\mathbf{\Sigma}(x_i)$ respectively, where*

$$\mathbf{\Sigma}(x) = \mathbf{I}_{p,q}\mathbf{\Delta}^{-1}\mathbf{\Gamma}_\rho(x)\mathbf{\Delta}^{-1}\mathbf{I}_{p,q},$$

with

$$\mathbf{\Gamma}_\rho(x) = \begin{cases} \mathbb{E}\left\{\langle x,\xi\rangle_{p,q}(1-\langle x,\xi\rangle_{p,q})\xi\xi^\top\right\} & \text{if } \rho_n \equiv 1, \\ \mathbb{E}\left\{\langle x,\xi\rangle_{p,q}\xi\xi^\top\right\} & \text{if } \rho_n \to 0, \end{cases}$$

where $\mu = \mathbb{E}(\xi)$, $\mathbf{\Delta} = \mathbb{E}\left(\xi\xi^\top\right)$, and where expectations are taken with respect to $\xi \sim F$.

2.2.2 Estimation theory with fixed latent positions

In this subsection, we review the estimation theory developed in Xie [17] for adjacency spectral embedding under the generalised random dot product graph model with fixed latent positions. The statements we give here are corollaries of Corollary 4.1 and Theorem 4.4 of Xie [17], where the reader can find more general statements of these results.

In this setup, a graph $\mathbf{A}$ is considered, which follows a generalised random dot product graph model with signature (p,q) and latent positions $X_1, \ldots, X_n$ which are assumed to be deterministic r-dimensional vectors. Define $\mathbf{\Delta} := n^{-1}\sum_{i=1}^n X_iX_i^\top$, denote its eigenvalues by $\lambda_1 \geq \cdots \geq \lambda_r$, and define the condition number $\kappa := \lambda_1/\lambda_r$. It is assumed that indefinite inner products between the latent positions give valid probabilities, i.e. $\langle X_i, X_j\rangle_{p,q} \in [0,1]$ for all $i,j \in [n]$, and that they span $\mathbb{R}^r$.

We additionally assume that the latent positions are relatively homogeneous. We assume that their sizes are of the same order, by which we mean there exists some $\rho \leq 1$ such that $\|X_i\|_2 \asymp \rho$ for all $i \in [n]$, and that $\mathbf{\Delta}$ is well conditioned, by which we mean $\kappa \asymp 1$. The results in Xie [17] are stated with weaker assumptions, though at the expense of more complicated theorems.

The first result is an analogue of Theorem 2.1.

Theorem 2.3. *Suppose that* $\mathbf{A}$ *is a graph generated as described in Section 2.2.2 and the sparsity factor satisfies* $n\rho_n \gtrsim \log n$*. Then, there exists an indefinite orthogonal transformation* $\mathbf{Q} \in \mathbb{O}(p,q)$ *such that, for sufficiently large* n*, the adjacency spectral embedding* $\hat{X}_1, \ldots, \hat{X}_n$*, satisfies*

$$\max_{i \in \{1,\ldots,n\}} \left\| \mathbf{Q}\hat{X}_i - X_i \right\|_2 \lesssim \left(\frac{\log n}{n} \right)^{1/2}$$

with overwhelming probability.

The second theorem is an analogue of Theorem 2.2.

Theorem 2.4 (Asymptotic normality)**.** *Assume the setting of Theorem 2.3. For each fixed index* $i \in [n]$*, and for any sufficiently large* n*,*

$$\sup_{\omega \in \Omega} \left| \mathbb{P}\left\{ n^{1/2} \mathbf{\Sigma}(X_i)^{-1/2} \left(\mathbf{Q}\hat{X}_i - X_i \right) \in \omega \right\} - \mathbb{P}(z \in \omega) \right| \lesssim \frac{\log(n\rho) \left\| \mathbf{\Sigma}(X_i)^{-1/2} \right\|_2}{(n\rho)^{1/2}}.$$

where Ω *is the collection of all convex measurable sets in* $\mathbb{R}^r$*, and*

$$\mathbf{\Sigma}(x) = \mathbf{I}_{p,q} \mathbf{\Delta}^{-1} \mathbf{\Gamma}_\rho(x) \mathbf{\Delta}^{-1} \mathbf{I}_{p,q},$$

with

$$\mathbf{\Gamma}_\rho(x) = n^{-1} \sum_{i=1}^{n} \langle x, X_i \rangle_{p,q} (1 - \rho \langle x, X_i \rangle_{p,q}) X_i X_i^\top.$$

2.3 Symmetric Laplacian spectral embedding

In this section, we define symmetric Laplacian spectral embedding and present analogous estimation theory to Section 2.2, which makes formal the sense in which it performs statistical inference on the generalised random dot product graph.

We begin by defining the symmetric Laplacian matrix $\mathbf{L}_{\text{sym}}$. Let $d_i = \sum_{j=1}^n a_{ij}$ denote the degree of node i for $i \in [n]$ and define the diagonal degree matrix $\mathbf{D} := \operatorname{diag}(d_1, \ldots, d_n)$. Then the symmetric (normalised) Laplacian matrix $\mathbf{L}_{\text{sym}}$ is defined as

$$\mathbf{L}_{\text{sym}} := \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$$

where $\mathbf{D}^{-1/2} := \operatorname{diag}(1/\sqrt{d_1}, \ldots, 1/\sqrt{d_n})$. Clearly, in order for this to be defined we must assume that the graph has no isolated nodes, i.e. no nodes of degree zero.

We now give a definition of symmetric Laplacian spectral embedding.

Definition 2.3 (Symmetric Laplacian spectral embedding)**.** Suppose $\mathbf{L}_{\text{sym}}$ has the eigendecomposition $\mathbf{L}_{\text{sym}} = \sum_{i=1}^n \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top$ with $|\hat{\lambda}_1| \geq \cdots \geq |\hat{\lambda}_n|$. The symmetric Laplacian spectral embedding of the graph into $\mathbb{R}^r$, denoted $\hat{X}_1, \ldots, \hat{X}_n \in \mathbb{R}^r$, is given by the rows of the matrix

$$\hat{\mathbf{X}} = \begin{pmatrix} \hat{X}_1^\top \\ \vdots \\ \hat{X}_n^\top \end{pmatrix} := \left(|\hat{\lambda}_1|^{1/2} \hat{u}_1 \cdots |\hat{\lambda}_r|^{1/2} \hat{u}_r \right)$$

obtained by stacking the scaled eigenvectors $|\hat{\lambda}_1|^{1/2}\hat{u}_1, \ldots, |\hat{\lambda}_r|^{1/2}\hat{u}_r$ in columns.

We will review estimation theory developed in Rubin-Delanchy et al. [15] which makes formal the following statement:

> Under a generalised random dot product graph model, the symmetric Laplacian spectral embedding, $\hat{X}_1, \ldots, \hat{X}_n$, is a uniformly consistent estimate of the scaled latent positions $\frac{X_1}{\sqrt{t_1}}, \ldots, \frac{X_n}{\sqrt{t_n}}$, with asymptotically Gaussian error.

The symmetric Laplacian spectral embedding does not directly estimate the latent positions, but rather versions of them which have been scaled down by the square root of their expected degrees. Except for some special cases [18], estimation theory for the symmetric Laplacian matrix is only available under the random latent positions setup.

As in the previous section, the casual reader who is content with the informal statement above may skip to Section 2.4. For the technical reader, the following subsection is dedicated to reviewing the results by Rubin-Delanchy et al. [15] which formalise it.

2.3.1 Estimation theory with random latent positions

In this subsection, we review the estimation theory developed in Rubin-Delanchy et al. [15] and Tang and Priebe [19] for symmetric Laplacian spectral embedding under the generalised random dot product graph model with random latent positions.

We consider the asymptotic setup described Section 2.2.1 with the additional assumption that $\langle x, \mu \rangle_{p,q}$ is bounded away from zero for all $x \in \mathcal{X}$, where $\mathcal{X}$ is the support of F and $\mu := \mathbb{E}_{\xi \sim F}(\xi)$ is its mean. This additional assumption is necessary to ensure that all the expected node degree grow with $n\rho_n$, which is a requirement for many components of the proofs. For example, this includes the concentration inequality of Theorem 2 of Lu and Peng [20] (see also Theorem 1.1 of Oliveira [21]). The following lemma shows that this additional condition is sufficient to ensure that all the expected degrees of the graph grow with $n\rho_n$.

Lemma 2.2. *Suppose that $\{\mathbf{A}^{(n)}\}_{n \in \mathbb{N}}$ is a sequence of graphs generated as described in Section 2.2.1, with the additional assumption that $\langle x, \mu \rangle_{p,q}$, where $\mu := \mathbb{E}_{\xi \sim F}(\xi)$, is bounded away from zero for all $x \in \mathcal{X}$. Then, for sufficiently large n, with overwhelming probability*

$$t_i^{(n)} \asymp n\rho_n, \qquad i = 1, \ldots, n.$$

A proof of Lemma 2.2 is given in Section A.2 of the appendix. To see why this additional assumption is necessary, consider the following setup for which the additional assumption is not satisfied.

Proposition 2.1. *Suppose that $\rho_n = 1$ and F is the uniform distribution on $\mathcal{X} := [0, 1]$, then for all $n \geq 1$, with probability greater than $1 - 1/e$, $\min_{i \in [n]} t_i^{(n)} \leq 1$.*

Proposition 2.1 is proved by observing that $\min_{i\in[n]} t_i^{(n)} > 1$ implies that $X_1^{(n)}, \ldots, X_n^{(n)} > n^{-1}$, and since $X_1^{(n)}, \ldots, X_n^{(n)}$ are independent,

$$\mathbb{P}\left(\min_{i\in[n]} t_i^{(n)} > 1\right) \leq \mathbb{P}\left(\bigcap_{i=1}^{n} \left\{X_i^{(n)} > \frac{1}{n}\right\}\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i^{(n)} > \frac{1}{n}\right) = \left(\frac{n-1}{n}\right)^n < \frac{1}{e}.$$

Notably, this assumption is missing from both Theorem 1 and 2 of Rubin-Delanchy et al. [15] and Theorems 3.1 and 3.2 of Tang and Priebe [19], however their proofs remain valid if it is made, and so we state Theorems 1 and 2 of Rubin-Delanchy et al. [15] below as such.

The first result is a consistency result, which states that subject to an indefinite orthogonal transformation, the maximum error between a node's position in the symmetric Laplacian spectral embedding and its scaled latent positions vanishes for large graphs.

Theorem 2.5 (Uniform consistency)**.** *Suppose that $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$ is a sequence of graphs generated as described in Section 2.2.1, with the additional assumption that $\langle x, \mu\rangle_{p,q}$, where $\mu := \mathbb{E}_{\xi\sim F}(\xi)$, is bounded away from zero for all $x \in \mathcal{X}$, and the sparsity factor satisfies $n\rho_n \gg \log^{4c} n$ where $c > 0$ is a universal constant. Then, there exists a sequence of indefinite orthogonal transformations $\{\mathbf{Q}^{(n)} \in \mathbb{O}(p,q)\}_{n\in\mathbb{N}}$ such that, for sufficiently large n, the symmetric Laplacian spectral embedding $\hat{X}_1^{(n)}, \ldots, \hat{X}_n^{(n)}$ satisfies*

$$\max_{i\in\{1,\ldots,n\}} \left\| \mathbf{Q}^{(n)}\hat{X}_i^{(n)} - \frac{X_i^{(n)}}{\sqrt{t_i^{(n)}}} \right\|_2 \lesssim \frac{\log^c n}{n\rho_n^{1/2}}$$

with overwhelming probability.

The second result is a central limit theorem. It states that for a fixed, finite subset of nodes, indexed without loss of generality as $1, \ldots, m$, their error distributions, scaled by $n\rho_n^{1/2}$, are asymptotically Gaussian.

Theorem 2.6 (Asymptotic normality)**.** *Assume the setting of Theorem 2.5. Conditional on $\xi_i^{(n)} = x_i$, for $i = 1, \ldots, m$, $n \geq m$, the random vectors*

$$n\rho_n^{1/2}\left(\mathbf{Q}^{(n)}\hat{X}_i^{(n)} - \frac{X_i^{(n)}}{\sqrt{t_i^{(n)}}}\right)$$

converge in distribution to independent mean-zero normal random vectors with covariance matrices $\mathbf{\Sigma}(x_i)$ respectively,

$$\mathbf{\Sigma}(x) = \frac{\mathbf{I}_{p,q}\tilde{\mathbf{\Delta}}^{-1}\mathbf{\Gamma}_\rho(x)\tilde{\mathbf{\Delta}}^{-1}\mathbf{I}_{p,q}}{\langle x, \mu\rangle_{p,q}}$$

with

$$\mathbf{\Gamma}_\rho(x) = \begin{cases} \mathbb{E}\left\{\langle x,\xi\rangle_{p,q}\left(1-\langle x,\xi\rangle_{p,q}\right)\left(\frac{\xi}{\langle \xi,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}x}{2\langle x,\mu\rangle_{p,q}}\right)\left(\frac{\xi}{\langle \xi,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}x}{2\langle x,\mu\rangle_{p,q}}\right)^\top\right\} & \text{if } \rho_n \equiv 1, \\ \mathbb{E}\left\{\langle x,\xi\rangle_{p,q}\left(\frac{\xi}{\langle \xi,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}x}{2\langle x,\mu\rangle_{p,q}}\right)\left(\frac{\xi}{\langle \xi,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}x}{2\langle x,\mu\rangle_{p,q}}\right)^\top\right\} & \text{if } \rho_n \to 0, \end{cases}$$

where $\mu = \mathbb{E}(\xi)$, $\tilde{\boldsymbol{\Delta}} = \mathbb{E}\left(\frac{\xi\xi^\top}{\langle \xi, \mu \rangle_{p,q}}\right)$, *and where expectations are taken with respect to* $\xi \sim F$.

2.4 Random graph models and the latent geometry of spectral embedding

In this section, we introduce some random graph models which are special cases of low-rank, independent-edge random graphs, and show how they are parameterised as the latent positions of a generalised random dot product graph model. The estimation theory in Section 2.2 suggests that the geometric patterns observed in the adjacency spectral embedding of a graph generated from one of these models will approximately resemble the geometry of the model's underlying latent positions. The estimation theory in Section 2.3 suggests that the geometric patterns observed in the symmetric Laplacian spectral embedding of a graph generated from one of these models will approximately resemble the geometry of the model's underlying latent positions, scaled down by the square root of the node's expected degree. We point out that this scaling is a one-to-one mapping which preserves many global geometric properties. For example, it maps points to points, rays to rays and simplexes to simplexes, albeit distorting inter-point distances.

2.4.1 Foundational random graph models

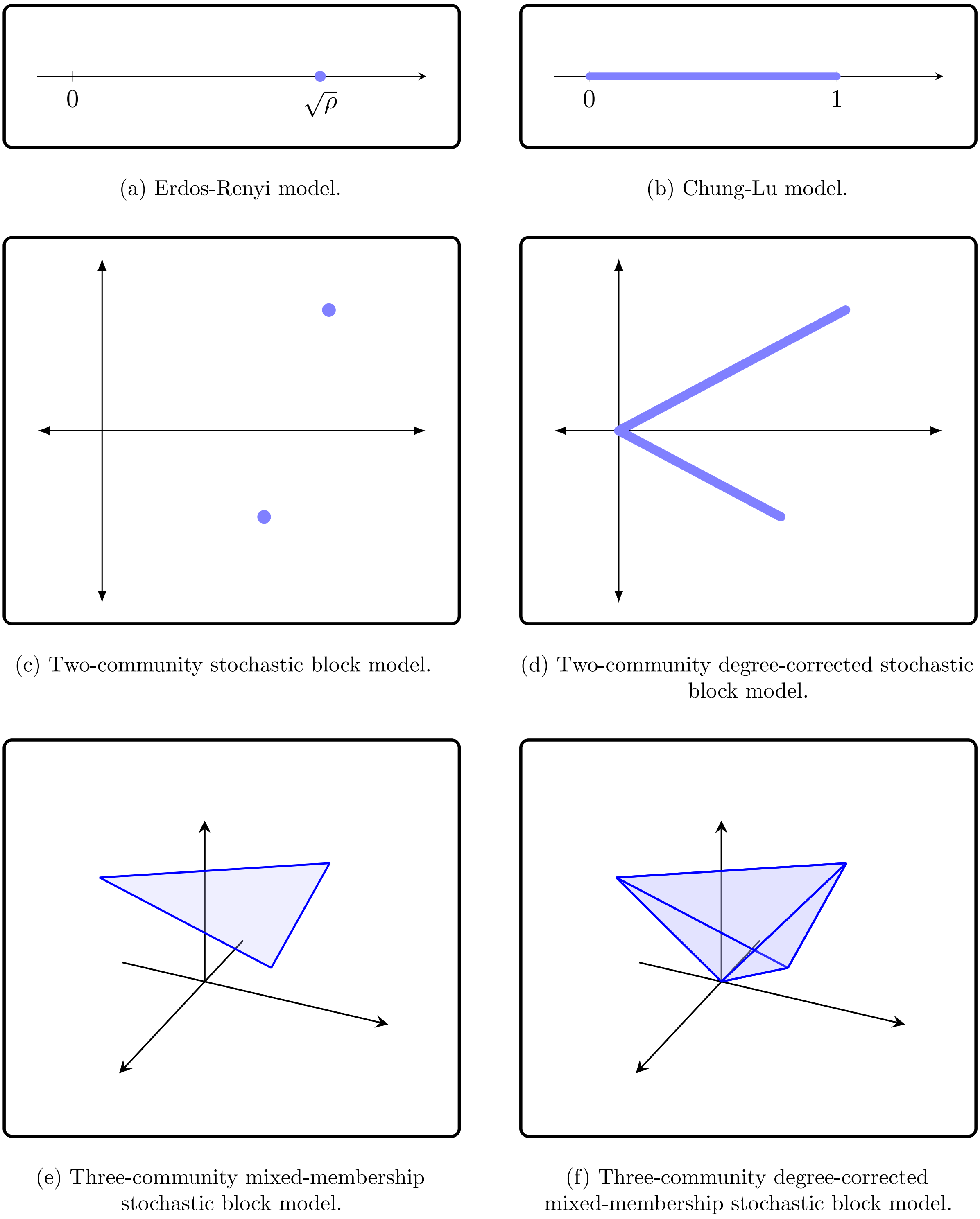
We begin with two simple models which laid the foundations for combinatorial graph theory. Neither of these models are sufficiently flexible to model networks encountered in the real-world today, however, their analysis has led to extraordinary insights into the fundamental properties of random graphs, and they serve as building blocks for more realistic model to come.

Erdös–Rényi model. The origins of random graph theory can be traced back to a hugely influential series of eight papers, co-authored by the great mathematicians Paul Erdös and Alfred Rényi between 1959 and 1968 [22–29], which examined the properties of independent-edge random graphs whose edges occur with some common probability

$$p_{ij} = \rho, \qquad 1 \leq i, j \leq n.$$

When parameterised as a generalised random dot product graph, the latent positions are one-dimensional and all lie at a single point: $X_i = \sqrt{\rho}$ for all $i \in [n]$. Figure 2.2a illustrates this position.

Despite the simplicity of this model, its emergent structure is rich, surprising and abundant in phase transitions. A phase transition which will recur in this thesis is on the connectivity of the graph: they prove that the graph is connected asymptotically almost surely if $n\rho \gg \log n$, and is disconnected asymptotically almost surely if $n\rho \ll \log n$. They additionally prove a myriad results on phase transitions in the sizes of the connected components in the latter



(a) Erdos-Renyi model.

(b) Chung-Lu model.

(c) Two-community stochastic block model.

(d) Two-community degree-corrected stochastic block model.

(e) Three-community mixed-membership stochastic block model.

(f) Three-community degree-corrected mixed-membership stochastic block model.

Figure 2.2: Illustrations of the latent positions of the six random graph models described in Section 2.4, parametrised as generalised random dot product graphs.

regime, on the emergence of subgraphs, on planarity and chromatic number, and on perfect matchings.

These results by Erdös and Rényi have left a lasting legacy, and have challenged the natural intuitions of many about the properties of complex interconnected systems: randomness can induce structure rather than chaos, and simple local rules can lead to complex global behaviours.

Chung-Lu model. The emergence of the internet at the turn of the millennium saw a resurgence of interest in random graph theory. However, while the random graphs studied by Erdös and Rényi had approximately Poisson-distributed degree sequences, the degree sequences of networks emerging from the World Wide Web were vastly more heterogeneous.

The Chung-Lu model [30] is a parsimonious extension of Erdös–Rényi model which permits the generation of graphs with arbitrary expected degree sequences. Given an expected degree sequence $t_1, \ldots, t_n$ satisfying $\max_i t_i^2 \leq \sum_{k=1}^n t_k$, edge probabilities are given by

$$p_{ij} = \frac{t_i t_j}{\sum_{k=1}^n t_k}, \qquad 1 \leq i, j \leq n.$$

The underlying probability matrix has rank one and its associated latent positions lie on the unit interval. Specifically $X_i = t_i / (\sum_j t_j)^{1/2}$ for all $i \in [n]$. Figure 2.2b shows the set of admissible positions.

A series of papers by Fan Chung, Linyuan Lu and others shine light on phase transitions in the global characteristics of these graphs, such as the sizes of connected components [31–34], the average distances between nodes [35] and their spectra [36].

2.4.2 Community-structured random graph models

The following random graph models build on the simple random graph models introduced in the previous subsection by modelling community structure.

Stochastic block model. Arguably the simplest model for a community-structured network is the stochastic block model [37]: each node is assigned a community $z_1, \ldots, z_n \in [K]$ and the probability of an edge between two nodes depends only on their community membership. If we let $\mathbf{B} \in [0,1]^{K \times K}$ be the matrix whose $k\ell$th entry contains the probability of an edge between two nodes in communities k and ℓ, respectively, then the edge probabilities can be written as

$$p_{ij} = b_{z_i z_j}, \qquad 1 \leq i, j \leq n.$$

In this way, nodes in the same community are stochastically equivalent and the associated latent positions lie in one of K places, each corresponding to a community. Specifically, if $\mathbf{B}$ has rank r and signature (p, q), and $v_1, \ldots, v_K$ are r-dimensional positions such that $b_{k\ell} = \langle v_k, v_\ell \rangle_{p,q}$, then $X_i = v_{z_i}$ for all $i \in [n]$. Figure 2.2c shows an example of the latent positions of a two-community stochastic block model. The Erdös–Rényi model is an example of a stochastic block model with one community.

The stochastic block model has been established in the network science community as a canonical model for community detection in networks, and the literature on inference, selection and its fundamental limits is vast (see Abbe [14] and the references therein). While it is not necessarily a realistic model, it can be insightful and it admits many possible refinements that improve its fit to real data.

Degree-corrected stochastic block model. A common criticism of the stochastic block model is that it forces nodes in the same community to have the same expected degree, which makes it too inflexible to model many real-world networks. A proposed solution is the degree-corrected stochastic block model [38], which introduces node-specific scalar parameters $w_1, \dots, w_n \in \mathcal{W} \subset \mathbb{R}_+$ and models edge probabilities as

$$p_{ij} = w_i w_j B_{z_i,z_j}, \qquad 1 \leq i, j \leq n.$$

These parameters allow a node's degree to be independent of its community membership. In practice, it is typically a node's community membership which is of inferential interest, and the weights are viewed as nuisance parameters. Note that $w_1, \dots, w_n$ and $\mathbf{B}$ are only identified up to scale: mapping $w_i \mapsto cw_i$ for each $i \in [n]$ and $\mathbf{B} \mapsto c^{-1}\mathbf{B}$, for some $c > 0$ does not change the edge probabilities.

The latent positions associated with this model lie in the union of one-dimensional subspaces. The angle of a position encodes its community and the magnitude encodes its weight. Specifically, $X_i = w_i v_{z_i}$ for all $i \in [n]$, where $v_1, \dots, v_K$ are r-dimensional vectors such that $b_{k\ell} = \langle v_k, v_\ell \rangle_{p,q}$. An example of the set of admissible latent positions of a degree-corrected stochastic block model is given in Figure 2.2d. The Chung-Lu model is an example of a degree-corrected stochastic block model with one community. Visual inspection of the adjacency and symmetric Laplacian spectral embeddings of the Harry Potter enmity graph introduced at the start of the chapter suggests that the degree-corrected stochastic block model is a good fit for this graph.

Mixed-membership stochastic block model. Another criticism of the stochastic block model is that each node is only allowed to belong to one community, and thus only plays one latent role in the network. In many real-world networks, nodes often play multiple latent roles. The mixed-membership stochastic block model [39] replaces the hard community memberships $z_1, \dots, z_n$ with *soft* community memberships $\tau_1, \dots, \tau_n$: probability-like vectors which indicate the proportion of the time a node acts according to the preferences of each community. The edge probabilities are then given by

$$p_{ij} = \tau_i^\top \mathbf{B} \tau_j, \qquad 1 \leq i, j \leq n.$$

To ensure the community memberships are identifiable, the model is constrained so that each community, k, must have at least one “pure” member: that is a node whose community membership vector has a one in the position k and zero elsewhere.

The latent positions associated with this model lie on a simplex, the corners of which represent “pure” nodes, and the interior of which represent nodes with mixed memberships. Let $v_1, \dots, v_K$ be r-dimensional vectors such that $b_{k\ell} = \langle v_k, v_\ell \rangle_{p,q}$, then the latent positions are $X_i = \sum_{k=1}^K \tau_{ik} v_k$ for all $1 \in [n]$. Figure 2.2e illustrates an admissible set of latent positions of a mixed-membership stochastic block model.

Degree-corrected mixed-membership stochastic block model. The degree-corrected mixed-membership stochastic block model [40] combines the previous two extensions to the stochastic-block model. Each node has a soft community membership and a weight, and edge probabilities are given by

$$p_{ij} = w_i w_j \tau_i^\top \mathbf{B} \tau_j, \qquad 1 \leq i, j \leq n.$$

As with the mixed-membership stochastic block model, the model must be constrained so that each community has at least one “pure” member, and we must additionally impose that $\mathbf{B}$ has full rank. However, this is not enough to ensure the identifiability of the community membership vectors. For any $\alpha \in \mathbb{R}_+$, mapping

$$\mathbf{B} \mapsto \operatorname{diag}(\alpha)^{-1} \mathbf{B} \operatorname{diag}(\alpha)^{-1}, \qquad w_i \mapsto \|\tau_i \circ \alpha\|_1 w_i, \qquad \tau_i \mapsto \frac{\tau_i \circ \alpha}{\|\tau_i \circ \alpha\|_1}$$

for $i = 1 \dots, n$, where $\circ$ denotes the Hadamard product (the entry-wise product), does not change the edge probabilities. For this reason, one must take care when interpreting the soft community memberships in this model.

The latent positions associated with this model lie on a simplicial cone. The angle of a position encodes its community mixture and the magnitude encodes its weight. Specifically, $X_i = w_i \sum_{k=1}^K \tau_{ik} v_k$ for all $1 \in [n]$, where $v_1, \dots, v_K$ are r-dimensional vectors such that $b_{k\ell} = \langle v_k, v_\ell \rangle_{p,q}$. An illustration of the set of admissible latent positions of a degree-corrected mixed-membership stochastic block model is given in Figure 2.2f.

Chapter 3

Spectral embedding with the random walk Laplacian

While a wealth of literature has emerged on the statistical properties of graph embeddings obtained from the adjacency and symmetric Laplacian matrices [6, 8, 14, 15, 19, 40–47], the statistical properties of graph embeddings obtained from the random walk Laplacian matrix are relatively understudied. Despite this, heuristic arguments have made this matrix an incredibly popular choice [48–51]: for example, the Normalised Cuts algorithm [48] which performs graph clustering by applying the k-means algorithm to the principal eigenvectors of the random walk Laplacian matrix has over 19,000 citations according to Google Scholar. In this chapter, we fill this gap and provide a principled statistical interpretation of these embeddings.

To motivate our discussion, we draw the reader's attention back to the real-world graph introduced in Chapter 2, in which nodes represent the characters of the Harry Potter novels by J.K. Rowling, and which has edges between characters who are enemies in the story. Recall that Figure 2.1 shows graph embeddings obtained using the first two eigenvectors of the adjacency and symmetric Laplacian matrices and that the theory reviewed in the previous chapter suggests the following statistical interpretation of these embeddings: magnitude should be interpreted as encoding the *number* of connections made by a node, and angle should be interpreted as encoding the *kinds* of connections made by a node.

The left panel of Figure 3.1 shows the graph embedding obtained using the first two eigenvectors of the random walk Laplacian matrix, and the geometric structure present is visibly different: the points lie on a hyperplane and broadly clustered around two distinct points. While these two clusters appear to reflect the two social groups in the story, the intuition that magnitude encodes degree does not appear to hold.

The purpose of this chapter is to provide a statistically principled interpretation of this geometric phenomenon, which, to our knowledge, is unknown to the statistics community. The crux of our argument is that, under a generalised random dot product graph model, embeddings obtained from the eigenvectors of the random walk Laplacian should not be treated as direct

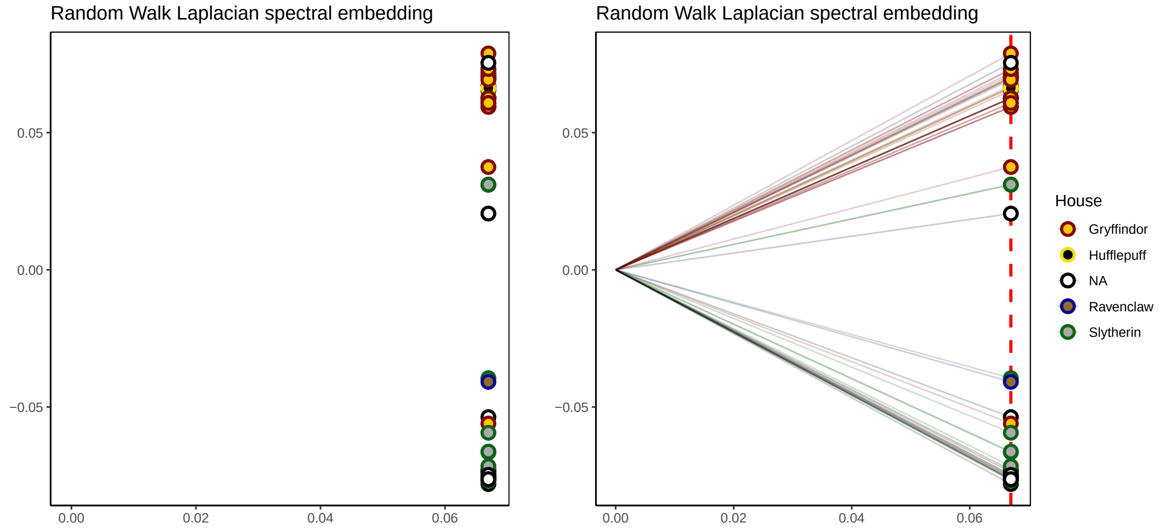


Figure 3.1: The random walk Laplacian spectral embedding of a graph of enmities between characters in the Harry Potter book series. Colour indicates the “house” to which the character belongs in the Hogwarts School. Both panels show the embedding, and the right panel additionally shows the hyperplane on which the embedding lies (dashed line), and the rays which the points of the embedding represent (solid lines).

estimates of latent positions, but rather as estimates of the one-dimensional subspaces passing through each point, which we will herein refer to as “rays”. The right panel of Figure 3.1 shows the hyperplane on which the embedding of the Harry Potter enmity graph lies as a dashed line and the rays which the points of the embedding represent as solid lines.

In other words, each point should be viewed as encoding the equivalence class of points which share its direction, irrespective of magnitude. For this reason, unlike the embeddings studied in Chapter 2, the embeddings obtained from the random walk Laplacian *do not* encode degree.

A pertinent implication of our theory is that under the degree-corrected stochastic block model, graph embeddings obtained using the eigenvectors of the random walk Laplacian matrix will concentrate around distinct points on a hyperplane, and community memberships can be consistently estimated using a standard clustering algorithm such as k-means. Additionally, there are stochastic block models for which spectral clustering using the random walk Laplacian is *inconsistent.*

We provide a uniform consistency result — a high probability bound on the maximum error between a node’s embedded position, and the projection of its latent position onto a specified hyperplane — and a central limit theorem, which states that the errors are asymptotically Gaussian, and quantifies the asymptotic covariance matrices.

We find that the errors are not necessarily spherical, and for sparse graphs, the scale of the error is inversely proportional to the node’s expected degree. Therefore, under a degree-corrected stochastic block model, this suggests that using the k-means algorithm, or Gaussian mixture modelling, will be sub-optimal for clustering and that instead, one should fit a *weighted-data* Gaussian mixture model, which assigns more weight to higher degree nodes, whose positions are more accurately estimated. Through simulation, we provide empirical evidence for this claim an demonstrate that the Bayesian Information Criterion for this fitted model can accurately estimate the number of communities in settings where the same criterion for a fitted standard Gaussian mixture model cannot.

3.1 Random walk Laplacian spectral embedding

In this section, we define the random walk Laplacian matrix and define the spectral embedding obtained from it.

Given a simple, undirected, graph with (symmetric) adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$, with a one in position i, j if there is an edge between nodes i and j and a zero otherwise, the random walk Laplacian matrix, $\mathbf{L}_{\mathrm{rw}}$, is defined as

$$\mathbf{L}_{\mathrm{rw}} := \mathbf{D}^{-1}\mathbf{A},$$

where $\mathbf{D} := \operatorname{diag}(d_1, \ldots, d_n)$ is the diagonal degree matrix where $d_i = \sum_j \mathbf{A}_{ij}$. The reader may be more familiar with the definition $\mathbf{I} - \mathbf{L}_{\mathrm{rw}}$ [51]: both definitions share the same eigenvectors, however in our embedding, we use the eigenvectors with the largest-*in-magnitude* eigenvalues. These are not necessarily the same when $\mathbf{L}_{\mathrm{rw}}$ has large negative eigenvalues, and this will be important when embedding graphs with heterophilic connectivity structure (see [5] and [15] for additional discussion).

We now give a formal definition of the random walk Laplacian spectral embedding.

Definition 3.1 (Random walk Laplacian spectral embedding)**.** Let $u_1, \ldots, u_r \in \mathbb{R}^n$ be right eigenvectors corresponding to the r eigenvalues of $\mathbf{L}_{\mathrm{rw}}$ with the largest magnitude. The random walk Laplacian spectral embedding of the graph $\mathbf{A}$ into $\mathbb{R}^r$, denoted $\hat{X}_1, \ldots, \hat{X}_n \in \mathbb{R}^r$, is given by the rows of the matrix

$$\hat{\mathbf{X}} = \begin{pmatrix} \hat{X}_1^\top \\ \vdots \\ \hat{X}_n^\top \end{pmatrix} := (u_1 \ \cdots \ u_r)$$

obtained by stacking the eigenvectors $u_1, \ldots, u_r$ in columns.

Firstly, it should be noted that the eigenvectors are not unique, and, since $\mathbf{L}_{\mathrm{rw}}$ is not a symmetric matrix, the usual choice of an orthonormal system is not available. In general, this means that the embeddings are only defined up to invertible linear transformations, and our theoretical results will reflect this. Rather than enforce any particular choice, we prefer to state our theoretical results "up to an invertible linear transformation". We highlight that

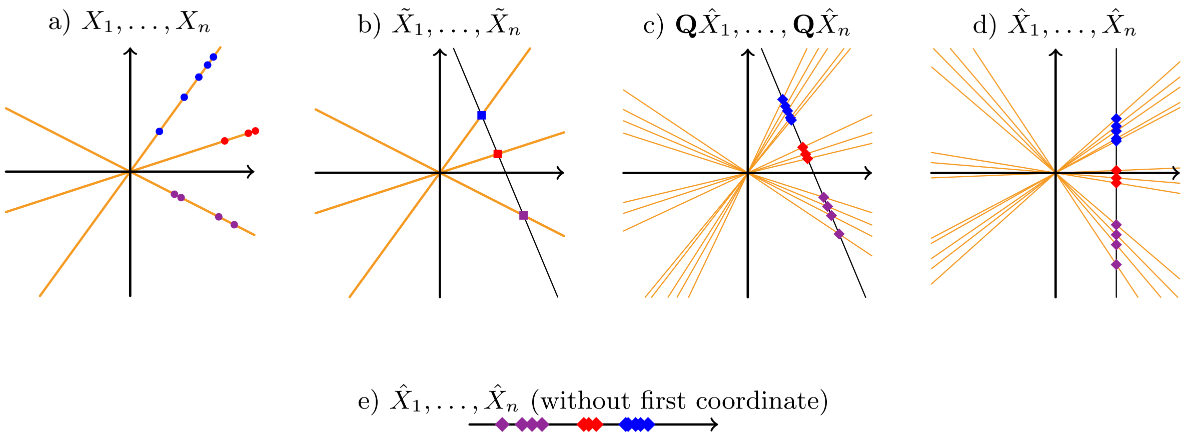

a) $X_1, \ldots, X_n$
b) $\tilde{X}_1, \ldots, \tilde{X}_n$
c) $\mathbf{Q}\hat{X}_1, \ldots, \mathbf{Q}\hat{X}_n$
d) $\hat{X}_1, \ldots, \hat{X}_n$
e) $\hat{X}_1, \ldots, \hat{X}_n$ (without first coordinate)

Figure 3.2: Theory pipeline. a) Latent positions in $\mathbb{R}^2$, corresponding to a degree-corrected stochastic block model with three communities; b) Degree-corrected latent positions, which lie on a one-dimensional hyperplane. Latent positions corresponding to the same community have the same degree-corrected latent position; c-d) r-dimensional random walk spectral embedding which, in c), is aligned to match the true degree-corrected latent positions; e) $r-1$-dimensional random walk spectral embedding (Definition 3.1), for input to subsequent clustering step.

when the graph is connected, it is easily verified that its first eigenvector is proportional to the all-ones vector and the embeddings lie on a hyperplane with constant first coordinate. While this coordinate contains no information, we retain it in our definition as it is useful for our theory. In practice, it is common to omit it.

3.2 Estimation theory

In this section, we make precise the sense in which, under a generalised random dot product graph model, spectral embedding using the eigenvectors of the random walk Laplacian performs statistical inference on the rays which pass through the latent positions.

3.2.1 Projective representations of rays

In geometry, a standard way to represent these rays is by projecting them onto points on a hyperplane which doesn't pass through the origin. In three dimensions, one might imagine rays of light emanating from the origin, shining onto an infinitely large sheet of paper. Technically, we should include points at infinity to represent rays parallel to the hyperplane, although we choose it in such a way that this isn't required. We choose the hyperplane $\mathcal{H} := \{x : n \langle x, \mu \rangle_{p,q} = 1\}$, where $\mu := n^{-1} \sum_{i=1}^{n} X_i$, and let $\tilde{X}_1, \ldots, \tilde{X}_n$ denote the projection of the latent positions $X_1, \ldots, X_n$ onto it. By construction, we can write the projected latent positions as

$$\tilde{X}_i = \frac{X_i}{t_i}, \qquad 1 \leq i \leq n, \tag{3.1}$$

where $t_i = \sum_{j=1}^{n} \langle X_i, X_j \rangle_{p,q}$ denotes the expected degree of node i. Figure 3.2a shows an illustration of some latent positions in $\mathbb{R}^2$ (shown as coloured dots) lying on one of three rays (shown as orange lines). Figure 3.2b displays their associated projected latent positions (shown as coloured squares) and the one-dimensional hyperplane on which they lie (shown as a black line).

3.2.2 Asymptotics under the generalised random dot product graph model

In this section, we formalise the following statement:

> Under a generalised random dot product graph model, the random walk Laplacian spectral embedding, $\hat{X}_1, \ldots, \hat{X}_n$, is a uniformly consistent estimate of $\tilde{X}_1, \ldots, \tilde{X}_n$, the projection of the latent positions $X_1, \ldots, X_n$ onto the hyperplane $\mathcal{H}$, with asymptotically Gaussian error.

For our theory, we consider the random latent position setup detailed in Section 2.2.1, with the additional assumption, described in Section 2.3.1, that $\mathcal{X}$, the support of F, is bounded away from the origin.

Our first results state that $\hat{X}_1^{(n)}, \ldots, \hat{X}_n^{(n)}$, subject to a linear transformation, converge uniformly to $\tilde{X}_1^{(n)}, \ldots, \tilde{X}_n^{(n)}$, in the sense that the maximum error over the whole node set vanishes as n gets large.

Theorem 3.1 (Uniform consistency). *Suppose that $\{\mathbf{A}^{(n)}\}_{n \in \mathbb{N}}$ is a sequence of graphs generated as described in Section 2.2.1, with the additional assumption that $\langle x, \mu \rangle_{p,q}$, where $\mu := \mathbb{E}_{\xi \sim F}(\xi)$, is bounded away from zero for all $x \in \mathcal{X}$. Then there exists a universal constant $c > 0$ and and a sequence of invertible linear transformations $\{\mathbf{Q}^{(n)}\}_{n \in \mathbb{N}}$ such that, providing the sparsity factor satisfies $n\rho_n \gg \log^{4c} n$, then for sufficiently large n,*

$$\max_{i \in \{1,\ldots,n\}} \left\| \mathbf{Q}^{(n)} \hat{X}_i^{(n)} - \tilde{X}_i^{(n)} \right\|_2 \lesssim \frac{\log^c n}{n^{3/2} \rho_n}$$

with overwhelming probability.

We note that $\|\tilde{X}_i\|_2 \asymp 1/(n\rho_n^{1/2})$ for all $i \in [n]$ and since $n\rho_n \gg \log^{4c} n$, the bound is not vacuous. Our second result is a central limit theorem. It states that for a fixed, finite subset of nodes, indexed without loss of generality as $1, \ldots, m$, their error distributions, scaled by $n^{3/2}\rho_n$, are asymptotically Gaussian.

Theorem 3.2 (Central limit theorem). *Assume the setting of Theorem 3.1. Conditional on $\xi_i^{(n)} = x_i$, for $i = 1, \ldots, m$, $n \geq m$, the random vectors $n^{3/2}\rho_n(\mathbf{Q}^{(n)} \hat{X}_i^{(n)} - \tilde{X}_i^{(n)})$ converge in distribution to independent mean-zero normal random vectors with covariance matrices $\mathbf{\Sigma}(x_i)$ respectively, where*

$$\mathbf{\Sigma}(x) = \frac{\mathbf{I}_{p,q} \tilde{\mathbf{\Delta}}^{-1} \mathbf{\Gamma}_\rho(x) \tilde{\mathbf{\Delta}}^{-1} \mathbf{I}_{p,q}}{\langle x, \mu \rangle_{p,q}^2}$$

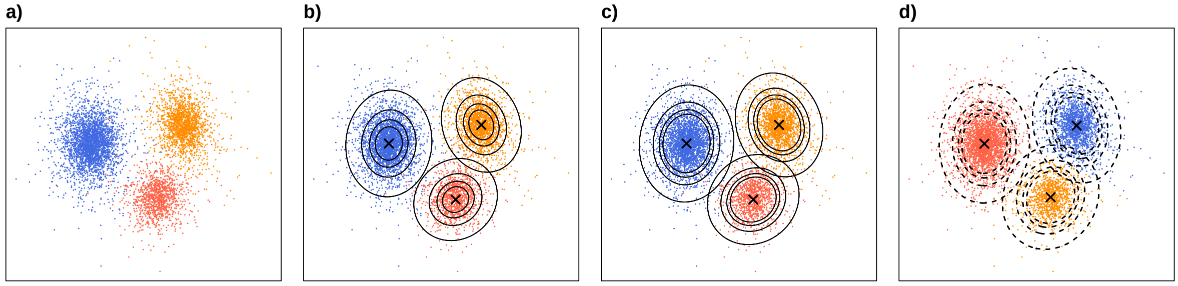


Figure 3.3: Spectral clustering under a degree-corrected stochastic block model using random walk spectral embedding. a) Spectral embedding of a graph on $n = 6000$ nodes, simulated from the degree-corrected stochastic block model described in (3.3), coloured according to community membership. b,c) Theoretical means and 95% level sets of the error distributions, for weights $w_i = 0.25, 0.5, 0.75, 1$, for b) dense ($\rho_n = 1$) and c) sparse ($\rho_n \to 0$) regimes (after re-alignment and neglecting the first coordinate, see main text for details). d) 95% level sets of the weighted-data Gaussian mixture model estimated using the expectation-maximisation algorithm described in Section B.1 of the appendix.

with

$$\mathbf{\Gamma}_\rho(x) = \begin{cases} \mathbb{E}\left\{ \langle x, \xi \rangle_{p,q} \left(1 - \langle x, \xi \rangle_{p,q}\right) \left(\frac{\xi}{\langle \xi, \mu \rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} x}{\langle x, \mu \rangle_{p,q}} \right) \left(\frac{\xi}{\langle \xi, \mu \rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} x}{\langle x, \mu \rangle_{p,q}} \right)^\top \right\} & \textit{if } \rho_n \equiv 1, \\ \mathbb{E}\left\{ \langle x, \xi \rangle_{p,q} \left(\frac{\xi}{\langle \xi, \mu \rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} x}{\langle x, \mu \rangle_{p,q}} \right) \left(\frac{\xi}{\langle \xi, \mu \rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} x}{\langle x, \mu \rangle_{p,q}} \right)^\top \right\} & \textit{if } \rho_n \to 0, \end{cases}$$

where $\mu = \mathbb{E}(\xi)$, $\tilde{\mathbf{\Delta}} = \mathbb{E}\left(\frac{\xi\xi^\top}{\langle \xi, \mu \rangle_{p,q}} \right)$, *and where expectations are taken with respect to* $\xi \sim F$.

To be clear, Theorem 3.2 is a central limit theorem for a set of r-dimensional vectors which, with probability one, live together on a $r-1$-dimensional hyperplane. Accordingly, the derived covariance matrices have rank $r-1$.

The details of the proofs of Theorems 3.1 and 3.2 are given in Section B.2 of the appendix.

3.2.3 Testing for equality of projected latent positions

In this subsection, we give an overview of how the asymptotic covariance in Theorem 3.2 may be estimated from the random walk Laplacian spectral embedding, $\hat{X}_1, \ldots, \hat{X}_n$ and the node degrees $d_1, \ldots, d_n$, and how this may be used to test for the equality of the projected latent positions of two nodes. This approaches follows that of Du and Tang [52] and we refer the reader to that paper for further details.

Suppose $\rho_n = 1$ and let

$$\hat{\mathbf{\Delta}} := \sum_{i=1}^n d_i \hat{X}_i \hat{X}_i^\top,$$

$$\hat{\zeta}_{ik} := \hat{X}_k - \hat{\mathbf{\Delta}} \mathbf{I}_{\hat{p},\hat{q}} \hat{X}_i,$$

$$\hat{\mathbf{\Gamma}}(\hat{X}_i) := n^{-1} \sum_{k=1}^{n} d_i d_k \langle \hat{X}_i, \hat{X}_k \rangle_{p,q} \left(1 - d_i d_k \langle \hat{X}_i, \hat{X}_k \rangle_{p,q}\right) \hat{\zeta}_{ik} \hat{\zeta}_{ik}^{\top},$$

and consider the plug-in estimator

$$\hat{\mathbf{\Sigma}}(\hat{X}_i) := \frac{n^2}{d_i^2} \mathbf{I}_{p,q} \hat{\mathbf{\Delta}}^{-1} \hat{\mathbf{\Gamma}}(\hat{X}_i) \hat{\mathbf{\Delta}}^{-1} \mathbf{I}_{p,q}$$

where $\hat{X}_i$ is plugged-in for $\tilde{X}_i \equiv X_i/t_i$ and d_i is plugged-in for t_i and $n\langle X_i, \mu \rangle$. Define the test statistic

$$T(\hat{X}_i, \hat{X}_j) = n^3 \left(\hat{X}_i - \hat{X}_j\right)^{\top} \left(\hat{\mathbf{\Sigma}}(\hat{X}_i) + \hat{\mathbf{\Sigma}}(\hat{X}_j)\right)^{-1} \left(\hat{X}_i - \hat{X}_j\right).$$

Then, under the null hypothesis $\mathbb{H}_0 : \tilde{X}_i = \tilde{X}_j$, and for $n \to \infty$,

$$T(\hat{X}_i, \hat{X}_j) \overset{\mathrm{d}}{\to} \chi_r^2 \tag{3.2}$$

where $\overset{\mathrm{d}}{\to}$ denotes convergence in distribution. The proof of (3.2) follows by identical arguments to the proof of Theorem 4.1 in Du and Tang [52].

Therefore, the null hypothesis $\mathbb{H}_0$ may be rejected in favour of the alternative hypothesis $\mathbb{H}_1 : \tilde{X}_i \neq \tilde{X}_j$ at the significance level α if $T(\hat{X}_i, \hat{X}_j) > z_{r,1-\alpha}$, where $z_{r,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ_r^2 distribution.

3.2.4 Identifiability

We now briefly pause to discuss the role of the invertible linear transformation $\mathbf{Q}^{(n)}$ which appears in our theorems, which stems from two distinct sources: the non-identifiability of the eigenvectors, and the non-identifiability of the model. One might ask whether by imposing additional constraints, this could be replaced by an orthogonal transformation.

One can employ a relationship between the eigenvectors of the symmetric Laplacian and random walk Laplacian matrices to obtain a canonical set of eigenvectors in Definition 3.1, which are defined up to orthogonal, rather than invertible linear transformations. It is easy to verify that if (λ, u) is an eigenvalue, right-eigenvector pair of $\mathbf{L}_{\mathrm{rw}}$, then $(\lambda, \mathbf{D}^{1/2}u)$ is an eigenvalue, eigenvector pair of $\mathbf{L}_{\mathrm{sym}}$. Then, we let $v_1, \ldots, v_n$ denote orthonormal eigenvectors of $\mathbf{L}_{\mathrm{sym}}$, and let $u_i = |\lambda_i|^{1/2}\mathbf{D}^{1/2}v_i$, $1 \leq i \leq n$ be canonically defined eigenvectors of $\mathbf{L}_{\mathrm{rw}}$. Defined this way, the embedding is identifiable up to coordinate-wise sign flipping and orthogonal transformation in the eigenspaces of repeated eigenvalues. This is the construction we use in our proofs, however in order to keep our work aligned with the literature, we prefer to allow any construction and maintain the invertible linear transformation.

Even using this construction, we are left with an indefinite orthogonal transformation which stems from the non-identifiability of latent positions of a generalised random dot product graph. Under a canonical construction of the distribution F, it has been shown that this converges to an orthogonal transformation [16], however this convergence is not fast enough that the transformation can be replaced in the central limit theorem.

3.2.5 Asymptotics under the degree-corrected stochastic block model

A pertinent implication of our estimation theory is that spectral clustering using the eigenvectors of the random walk Laplacian performs statistical inference under the degree-corrected stochastic block model. We remind the reader of its definition.

Definition 3.2. A graph is said to follow a degree-corrected stochastic block model with community memberships $z_1, \dots, z_n \in [K]$ and weights $w_1, \dots, w_n \in \mathbb{R}_+$ if there exists a matrix $\mathbf{B} \in [0,1]$ such that its edges $\{a_{ij}\}_{i<j}$ are independent Bernoulli random variables with success probabilities

$$p_{ij} = w_i w_j b_{z_i z_j}, \qquad 1 \le i < j \le n.$$

In particular, our central limit theorem (Theorem 3.2) implies that under a degree-corrected stochastic block model, each node's embedded position is distributed around a point which depends only on its community, and the scale of its asymptotic covariance additionally depends on its weight parameter. To make this relationship between a node's weight parameter and asymptotic covariance explicit, we state the central limit theorem for a degree-corrected stochastic block model as a special case.

We consider the following asymptotic regime: Suppose that for each $n \in \mathbb{N}$, the community memberships $z_1^{(n)}, \dots, z_n^{(n)}$ are drawn at random with probabilities $\pi_1, \dots, \pi_K$, conditional upon which, for each $i \in [n]$, $w_i^{(n)}$ is drawn from a distribution $H_{z_i^{(n)}}$, where $H_1, \dots, H_K$ are probability distributions on $\mathbb{R}_+$. We let $\mathbf{B}^{(n)} = \rho_n \mathbf{B}$ be the inter-community probability matrix, where $\mathbf{B} \in [0,1]^{K \times K}$ is a fixed matrix. We let r and (p,q) be respectively the rank and signature of $\mathbf{B}$, and assume that the supports of $H_1, \dots, H_K$, which we denote $\mathcal{W}_1, \dots, \mathcal{W}_K$ satisfy $vwb_{k,\ell} \in [0,1]$ for all $v \in \mathcal{W}_k, w \in \mathcal{W}_\ell$ and $k, \ell \in [K]$. In addition, we define $v_1, \dots, v_K \in \mathbb{R}^r$ to be vectors satisfying $b_{k,\ell} = \langle v_k, v_\ell \rangle_{p,q}$ for all $k, \ell \in [K]$ and set $\tilde{v}_k^{(n)} = v_k / \rho_n^{1/2} \sum_{i=1}^n w_i^{(n)} b_{k, z_i^{(n)}}$.

Corollary 3.1. *Suppose $\{\mathbf{A}^{(n)}\}_{n \in \mathbb{N}}$ is a sequence of graphs following degree-corrected stochastic block models, generated as described in Section 3.2.5. There exists a sequence of linear transformations $\{\mathbf{Q}^{(n)}\}_{n \in \mathbb{N}}$ such that, providing the sparsity factor satisfies $n\rho_n \gg \log^{4c} n$, where $c > 0$ is the same universal constant as in Theorem 3.1, conditional on $z_i^{(n)} = z_i$ and $w_i^{(n)} = w_i$, for $i = 1, \dots, m$, $n \ge m$, the random vectors $n^{3/2}\rho_n(\mathbf{Q}^{(n)} \hat{X}_i^{(n)} - \tilde{v}_{z_i}^{(n)})$ converge in distribution to independent mean-zero normal random vectors with covariance matrices $\mathbf{\Sigma}(z_i, w_i)$, respectively, where*

$$\mathbf{\Sigma}(k, w) = \frac{\sum_{\ell=1}^K \pi_\ell \mathbf{I}_{p,q} \tilde{\mathbf{\Delta}}^{-1} \mathbf{\Gamma}_\ell(k, w) \tilde{\mathbf{\Delta}}^{-1} \mathbf{I}_{p,q}}{w \omega_{z_i}^2},$$

with

$$\mathbf{\Gamma}_\ell(k,w) = \begin{cases} \mathbb{E}\left(\theta_\ell \mathbf{B}_{k\ell}(1 - w\theta_\ell \mathbf{B}_{k\ell})\right) \left(\frac{v_\ell}{\omega_\ell} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} v_k}{\omega_k}\right) \left(\frac{v_\ell}{\omega_\ell} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} v_k}{\omega_k}\right)^\top & \text{if } \rho_n \equiv 1, \\ \mathbb{E}\left(\theta_\ell \mathbf{B}_{k\ell}\right) \left(\frac{v_\ell}{\omega_\ell} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} v_k}{\omega_k}\right) \left(\frac{v_\ell}{\omega_\ell} - \frac{\tilde{\mathbf{\Delta}} \mathbf{I}_{p,q} v_k}{\omega_k}\right)^\top & \text{if } \rho_n \to 0, \end{cases}$$

where $\omega_\ell = \sum_{m=1}^K \pi_m \mathbb{E}(\theta_m)\mathbf{B}_{\ell m}$*,* $\tilde{\boldsymbol{\Delta}} = \sum_{m=1}^K \frac{\pi_m \mathbb{E}(\theta_m) v_m v_m^\top}{\omega_m}$ *and expectations are taken with respect to* $\theta_\ell \sim H_\ell$ *for* $\ell \in [K]$*.*

We highlight that when node degrees grow sublinearly in n, i.e. $\rho_n \to 0$, the scale of the covariance matrix is inversely proportional to the weight parameter, and its shape depends only on the community. This is an observation we will exploit in Section 3.3.

Figure 3.3a shows the second and third dimensions of the spectral embedding $\hat{X}_1, \ldots, \hat{X}_n$ of a graph generated from a degree-corrected stochastic block model with $n = 6000$ nodes and parameters

$$\mathbf{B} = \begin{pmatrix} 0.3 & 0.2 & 0.2 \\ 0.2 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.3 \end{pmatrix}, \quad w_1, \ldots, w_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0.1, 1), \quad \pi = (0.5, 0.3, 0.2), \tag{3.3}$$

coloured according to community membership. To obtain Figures 3.3b,c, we first compute $\mathbf{Q}^{-1}$ to align the projected latent positions $\tilde{v}_1, \ldots, \tilde{v}_3$ with $\hat{X}_1, \ldots, \hat{X}_n$. After this transformation, the induced theoretical error distributions have no error in the first coordinate, so we do not display it, showing only what happens in the second and third coordinates. The second and third coordinates of the aligned projected community latent positions $\mathbf{Q}^{-1}\tilde{v}_1, \ldots, \mathbf{Q}^{-1}\tilde{v}_3$ are shown as crosses. Figure 3.3b shows four ellipses for each community describing the 95% level sets of the aligned, theoretical error distributions for weights $w_i = 0.25, 0.5, 0.75, 1$, assuming the sparsity regime $\rho_n \equiv 1$. Figure 3.3c shows the same assuming the sparsity regime $\rho_n \to 0$.

3.3 Implications for clustering

In this section, we focus on the methodological implications of the estimation theory in Section 3.2. The uniform consistency result of Theorem 3.1 ensures that, under the degree-corrected stochastic block model, asymptotically perfect clustering can be achieved by applying any reasonable clustering algorithm to the eigenvectors of the random walk Laplacian. Traditionally, in the spectral clustering literature, the recommendation has been to use the k-means algorithm [51]. However, recently, central limit theorems for adjacency and symmetric Laplacian spectral embedding under the standard stochastic block model have recently motivated fitting a Gaussian mixture model [15, 19, 47], the actual asymptotic distribution of the embeddings, which has been shown to empirically improve clustering performance [19].

In this section, we explain how our theory suggests fitting a *weighted-data* Gaussian mixture model to the eigenvectors of the random walk Laplacian to cluster nodes under the degree-corrected stochastic block model.

3.3.1 Fitting a weighted-data Gaussian mixture model

Under a degree-corrected stochastic block model, and in an asymptotic regime in which node degrees grow sub-linearly (i.e. $\rho_n \to 0$), our central limit theorem (Corollary 3.1) shows that the error distribution of a node's position is asymptotically Gaussian with mean depending only on its community, and covariance whose shape depends only on its community, and whose scale is inversely proportional to its weight w_i. This suggests the data will approximately fit a *weighted-data* Gaussian mixture model with likelihood

$$L\left(\{\hat{X}_i\}_{i=1}^n, \{w_i\}_{i=1}^n; \Theta\right) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \mathcal{N}\left(\hat{X}_i; \mu_k, \mathbf{\Sigma}_k / w_i\right) \tag{3.4}$$

for some $\Theta := \{\pi_k, \mu_k, \mathbf{\Sigma}_k\}_{k=1}^K$, where $\{\pi_k\}_{k=1}^K$ are positive mixing proportions summing to one, $\{\mu_k\}_{k=1}^K$ are r-dimensional means, $\{\mathbf{\Sigma}_k\}_{k=1}^K$ are $r \times r$ covariance matrices, and $\mathcal{N}(x; \mu, \mathbf{\Sigma})$ is the likelihood of a multivariate Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$.

It is straightforward to notice that $\mathcal{N}(x; \mu, \mathbf{\Sigma}/w) \propto \mathcal{N}(x; \mu, \mathbf{\Sigma})^w$, and the expectation-maximisation (EM) algorithm which optimises (3.4) with respect to Θ turns out to coincide with the EM algorithm to fit a standard Gaussian mixture model, with the exception that influence of each data point is reweighted according to the weights $w_1, \ldots, w_n$. It is for this reason that the model with likelihood (3.4) is referred to as the *weighted-data* Gaussian mixture model. We provide an expectation-maximisation algorithm to optimise (3.4), which is stated in the appendix, and refer the reader to Section III of Gebru et al. [53] for a derivation.

The expectation-maximisation algorithm which fits (3.4) assumes that the weights $w_1, \ldots, w_n$ are known, whereas in practice we do not have access to them, and they must be estimated from the data. Since the weight w_i is proportional to the expected degree t_i, we propose to use the observed degree $d_i = \sum_{j=1}^n A_{ij}$ as a proxy for w_i.

We apply a weighted-data Gaussian mixture model to the embedding of a graph simulated from the degree-corrected stochastic block model described in (3.3). For weights $w_i = 0.25, 0.5, 0.75, 1$, Figure 3.3d shows the 95% level sets of the fitted model, which are closely aligned with the theoretical error distribution shown in Figure 3.3c.

3.3.2 Comparison with Gaussian mixture modelling and k-means

To illustrate the advantage of exploiting our asymptotic theory, and fitting a weighted data Gaussian mixture model to the embeddings obtained from Definition 3.1, we simulate sequences of degree-corrected stochastic block model graphs with $n = 3000, 3500, \ldots, 7000$ nodes, and parameters given in (3.3). For each graph, we embed its random walk Laplacian matrix into $\mathbb{R}^3$ (Definition 3.1) and discard the first, constant coordinate. We cluster the embedded nodes by fitting either a weighted-data Gaussian mixture model using the EM algorithm described in Section B.1 of the appendix with weight estimates $d_1, \ldots, d_n$, fitting standard Gaussian

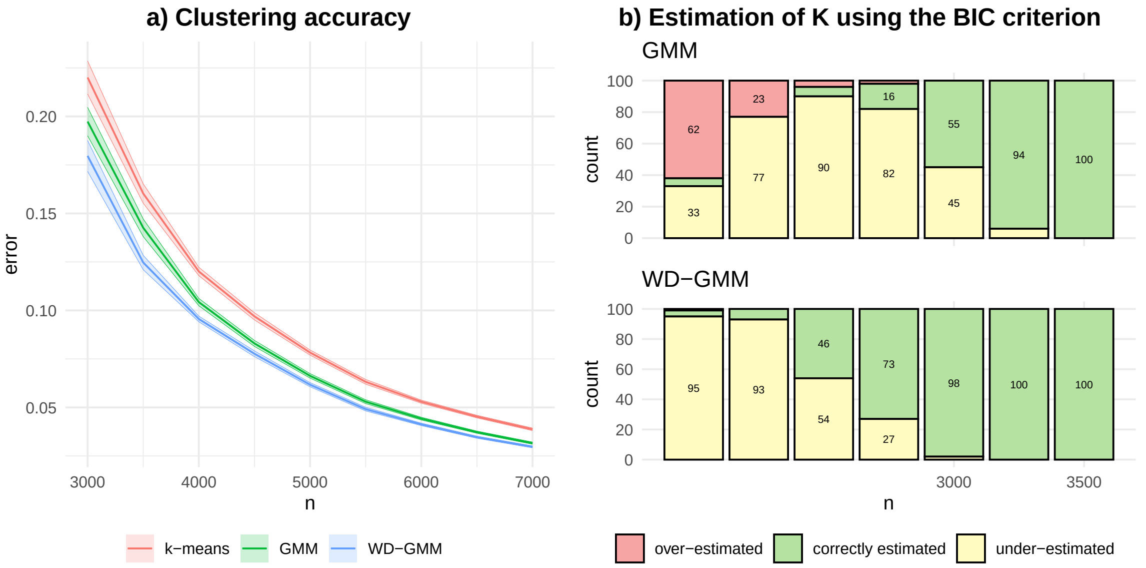


Figure 3.4: Comparison of different clustering algorithms applied to the random walk Laplacian spectral embedding of graphs simulated from a degree-corrected stochastic block model with parameters given in (3.3). The left panel (a) shows the clustering accuracy of k-means, Gaussian mixture modelling (GMM) and weighted-data Gaussian mixture modelling (WD-GMM) for 100 graphs of each size $n = 3000, 3500, \ldots, 7000$. For each algorithm the solid line shows the mean clustering error and the ribbon shows plus and minus two standard errors. The right panel (b) shows the number of times the BIC criterion of a fitted GMM and a fitted WD-GMM over-estimated, correctly estimated and under-estimated the number of communities in the model for 100 graphs each of size $n = 2000, 2250, \ldots, 3500$.

mixture model and applying the k-means algorithm, to obtain community estimates $\hat{z}_1, \ldots, \hat{z}_n$ and measure the clustering error

$$\text{error} := \min_{\sigma \in S_K} \sum_{i=1}^{K} \mathbb{I}\left(\sigma(\hat{z}_i) - z_i\right)$$

where S_k is the permutation group on $\{1, \ldots, K\}$. We simulate 100 such sequences, and for each clustering algorithm and graph size n, we compute the mean clustering error and its standard error. For each clustering algorithm and graph size n, we compute the mean clustering error and plus and minus two standard errors, which are plotted in Figure 3.4a. We use our own implementation of weighted-data and standard Gaussian mixture modelling, and the base-R implementation of k-means [54]. We see that fitting a weighted-data Gaussian mixture model yields a noticeable improvement over fitting a standard Gaussian mixture model, and the k-means algorithm.

3.3.3 Selecting the number of communities using the BIC criterion

In addition to improved clustering performance, a material advantage of fitting the true asymptotic distribution of the embedding is that we can employ standard model selection tools for mixture models to select the number of communities. In particular, we focus on the Bayesian Information Criterion (BIC). We simulate sequences of degree-corrected stochastic block model graphs with $n = 2000, 2250, \ldots, 3500$ nodes, and parameters given in (3.3), and for each graph, we embed its random walk Laplacian matrix into $\mathbb{R}^3$ (Definition 3.1), and to the second and third coordinates, we fit a weighted-data Gaussian mixture model using the EM algorithm described in Section B.1 of the appendix with weight estimates $d_1, \ldots, d_n$, for $K = 1, 2, \ldots, 5$ components. We calculate the BIC criterion of each fitted model and select the number of components which maximises the BIC criterion as an estimate of the number of communities in the underlying degree-corrected stochastic block model. We apply the same approach using a standard Gaussian mixture model. We simulate 100 such sequences, and for each method and graph size n, Figure 3.4b shows the number of graphs for which the true number of communities were over-estimated, under-estimated or corrected estimated, using both a weighted-data and standard Gaussian mixture model.

We see that the BIC criterion for the weighted-data Gaussian mixture model is able to select the correct number of communities in settings where the BIC criterion for the standard Gaussian mixture model cannot.

3.4 Illustration with a fictional character network

In many real-world applications, the degree of a node in a network is a parameter of secondary interest. In social networks, we may wish to model a person's friendship preferences independently of their popularity. In cyber-security, and many other domains, the graph represents a snapshot of a dynamic network describing, for example, packet transfers or other network transactions [55]. The time that a node is present on the network may have a significant bearing on its degree, yet have little to do with its role (e.g. a new laptop connecting to the network). Moreover, the placement of routers and other collection points will result in higher visibility of some nodes' connections compared to others. In this case, node degrees are heavily influenced by the observation process and may not represent an intrinsic property of the nodes themselves.

Stories, real or fictional, often provide network examples to illustrate graph methods, common examples being Zachary's Karate Club [56] and the "Les Miserables" character network [57]. Conversely, graph theory is often used in literature studies [58] to understand character networks and, in this field, degree is often seen as an artefact of the narrative point-of-view: the story spends more time with the protagonist and antagonist, and so we observe more of their connections. As an example, we return to the graph describing the enmity relationships between the characters in the Harry Potter novels of J.K. Rowling [9].

The remainder of this section contains spoilers for the story. Those wishing to read the books should refrain from reading it. Figure ???3.5 shows the second coordinate of the random walk spectral embedding of the graph, coloured, where applicable, according to the characters'

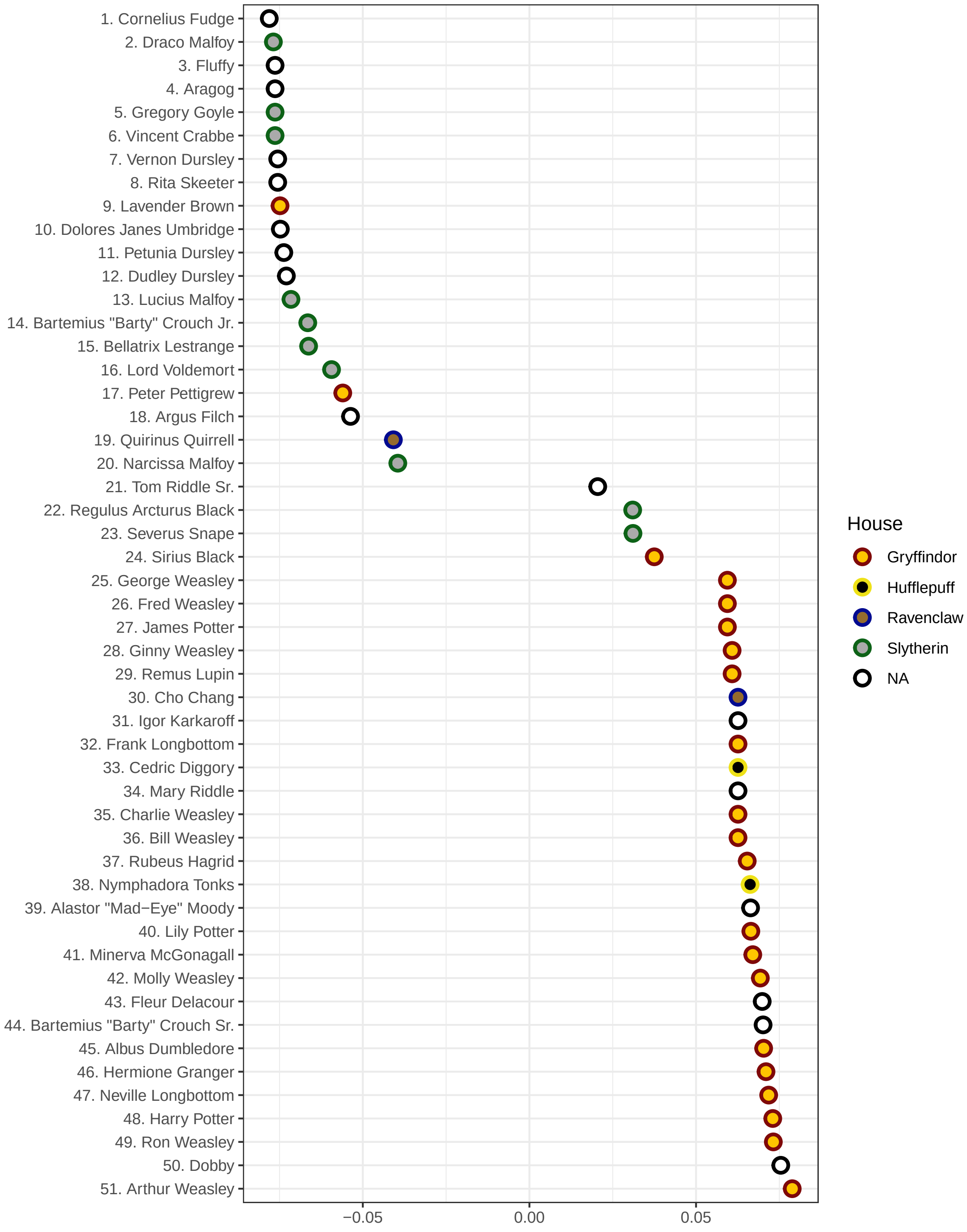


Figure 3.5: The second coordinate of the random walk Laplacian spectral embedding of the Harry Potter enmity network, coloured, where applicable, according to the character's house at the Hogwarts school.

house memberships at the Hogwarts school. The embedding shows a clear separation of the characters into two distinct clusters, broadly reflecting their alignment with the protagonist and antagonist.

However, there are some interesting outliers. Regulus Black, Severus Snape and Sirius Black mix in evil circles throughout the story but their benevolence is revealed in the later books. All three characters are positioned in between the two clusters, reflecting their mixed membership to the two sides.

A surprise is the positioning of Lavender Brown, a “good” character in the story who is positioned alongside the “evil” characters. She has only one enemy, Hermione Granger, and thus only one edge in the graph. Our statistical theory would therefore suggest that her embedding has high variance, which might explain her unexpected position.

Chapter 4

Spectral embedding of multipartite graphs

In this chapter, we develop bespoke statistical methodology for spectral embedding in the case when the graph under study is multipartite.

There are at least two reasons why a dedicated treatment is warranted. First, this special case is ubiquitous across science and technology, encompassing, for example, data linking users and items (bipartite graphs), which support modern recommendation systems, data from various security applications, providing interconnections between groups such as users, computers and processes in cyber-security [59]; or images, phone numbers, locations and names in human-trafficking prevention [60], and large data repositories supporting biomedical research, linking groups such as drugs, diseases, targets, pathways, variant locations and haplotypes [61], a 6-partite example analysed in this paper (Figure 4.4 shows a schematic of this dataset). Second, as we will show, multipartite structure provides an opportunity for dimension reduction.

Under the generalised random dot product graph model, we will show that the spectral embedding of a multipartite network has a special geometric structure, in which the node representations lie in the vicinity of group-specific subspaces, whose dimension may be significantly lower than the population rank of the graph. This motivates a subsequent step to the standard spectral embedding algorithm of estimating these subspaces and projecting onto them, to obtain vector representations of the nodes of each group in these lower, intrinsic dimensions.

Throughout this chapter, we focus on adjacency spectral embedding, however, the ideas presented here extend to symmetric Laplacian spectral embedding and a regularised variant thereof.

4.1 Spectral embedding of bipartite graphs

A graph is said to be bipartite if its vertex set can be partitioned into two disjoint subsets as $V = V_1 \cup V_2$ such that $a_{ij} = 0$ for all i and j in the same group. We write $n_k = |V_k|$, denote a

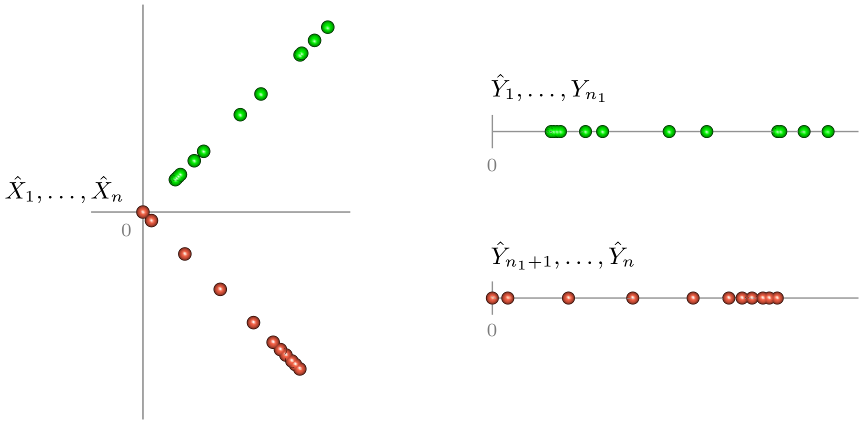


Figure 4.1: The latent geometry of a bipartite graph. The left panel shows the two-dimensional spectral embedding, $\hat{X}_1, \dots, \hat{X}_n$, of a bipartite graph generated from the random graph model (4.1) where $z_i \in \{1, 2\}$, coloured by group. The right panel shows the corresponding one-dimensional biadjacency spectral embedding, $\hat{Y}_1, \dots, \hat{Y}_n$ of the graph.

node's group membership by $z_i \in \{1, 2\}$ and assume without loss of generality that the nodes are indexed such that $z_1 \leq \cdots \leq z_n$. Then, the graph adjacency matrix has the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \boldsymbol{\mathcal{A}} \\ \boldsymbol{\mathcal{A}}^\top & \mathbf{0} \end{pmatrix},$$

where $\boldsymbol{\mathcal{A}} \in \{0, 1\}^{n_1 \times n_2}$ is known as the graph biadjacency matrix. While an embedding of $\mathbf{A}$ can be obtained using Definition 2.2, it is more common to approximately factorize $\boldsymbol{\mathcal{A}}$ using a truncated singular value decomposition, known as biadjacency spectral embedding, as defined in the following.

Definition 4.1 (Biadjacency spectral embedding)**.** Suppose $\boldsymbol{\mathcal{A}}$ has the singular value decomposition $\boldsymbol{\mathcal{A}} = \sum_{i=1}^n \hat{s}_i \hat{u}_i \hat{v}_i^\top$ with $\hat{s}_1 \geq \cdots \geq \hat{s}_n$. The biadjacency spectral embedding of $\boldsymbol{\mathcal{A}}$ into $\mathbb{R}^d$, denoted $\hat{Y}_1, \dots, \hat{Y}_n$, is given by the rows of the matrices

$$\hat{\mathbf{Y}}^{(1)} = \begin{pmatrix} \hat{Y}_1^\top \\ \vdots \\ \hat{Y}_{n_1}^\top \end{pmatrix} := \begin{pmatrix} s_1^{1/2} u_1 & \cdots & s_d^{1/2} u_d \end{pmatrix}, \qquad \hat{\mathbf{Y}}^{(2)} = \begin{pmatrix} \hat{Y}_{n_1+1}^\top \\ \vdots \\ \hat{Y}_n^\top \end{pmatrix} := \begin{pmatrix} s_1^{1/2} v_1 & \cdots & s_d^{1/2} v_d \end{pmatrix},$$

obtained by stacking the scaled left singular vectors $s_1^{1/2} u_1, \dots, s_d^{1/2} u_d$, and scaled right singular vectors $s_1^{1/2} v_1, \dots, s_d^{1/2} v_d$, respectively, in columns.

The matrix $\mathbf{A}$ is known in the literature as the symmetric dilation of $\boldsymbol{\mathcal{A}}$, and the use of the following geometric relationship between the eigenvalues and vectors of $\mathbf{A}$ and the singular values and vectors of $\boldsymbol{\mathcal{A}}$ is widespread in the literature on matrix perturbation theory and dates back to the discovery of the singular value decomposition itself:

Proposition 4.1 (Stewart and Sun [62]). *Suppose s is a singular value of $\boldsymbol{\mathcal{A}}$ and u, v are corresponding left and right singular vectors, then $\pm s$ are eigenvalues of $\mathbf{A}$ and*

$$\frac{1}{\sqrt{2}}\begin{pmatrix} u \\ \pm v \end{pmatrix}$$

are corresponding eigenvectors.

The proof of this statement is a simple computation. This result implies the following geometric relationship between $\{\hat{X}_i\}_{i=1}^n$ and $\{\hat{Y}_i\}_{i=1}^n$:

Lemma 4.1. *Let $\{\hat{X}_i\}_{i=1}^n$ be the adjacency spectral embedding of $\mathbf{A}$ into $\mathbb{R}^{2d}$ (Definition 2.2) and let $\{\hat{Y}_i\}_{i=1}^n$ be the biadjacency spectral embedding of $\boldsymbol{\mathcal{A}}$ into $\mathbb{R}^d$ (Definition 4.1). Then, for compatibly chosen spectral decompositions,*

$$\hat{Y}_i = \frac{1}{\sqrt{2}}\begin{pmatrix} \hat{X}_i \\ \hat{X}_i \end{pmatrix} \quad \text{for } i \in V_1; \qquad \hat{Y}_i = \frac{1}{\sqrt{2}}\begin{pmatrix} \hat{X}_i \\ -\hat{X}_i \end{pmatrix} \quad \text{for } i \in V_2.$$

Figure 4.1 illustrates Lemma 4.1 with a toy example.

4.2 The latent geometry of multipartite networks

A graph is said to be multipartite if its node set can be partitioned into K disjoint group $V = V_1 \cup \cdots \cup V_K$ such that $a_{ij} = 0$ if i and j are in the same group. As before, we write $n_k = |V_k|$, we denote a node's group membership by $z_i \in [K]$ and assume without loss of generality that the nodes are indexed such that $z_1 \leq \cdots \leq z_n$. For the rest of the chapter, we assume $\mathbf{A}$ is multipartite and that the node partitioning is known. In addition, unless otherwise stated, $i, j \in [n]$ and $k, \ell \in [K]$.

4.2.1 A tripartite example

To motivate our discussion of multipartite random graphs, we consider a simple tripartite inhomogeneous random graph, in which each node, i, is assigned a scalar-valued weight, $w_i \in (0, 1]$, and edge probabilities are given by

$$p_{ij} = \begin{cases} w_i w_j & \text{if } z_i \neq z_j, \\ 0 & \text{if } z_i = z_j. \end{cases} \tag{4.1}$$

This model is a multipartite generalisation of the Chung-Lu model [30, 34, 35]. The matrix $\mathbf{P}$ has rank three and is equivalently described by a generalised random dot product graph with signature $(1, 2)$, and latent positions $X_i = w_i \alpha_{z_i}(1, \cos\theta_{z_i}, \sin\theta_{z_i})^\top$ with distinct angles $\theta_1, \theta_2, \theta_3 \in [0, 2\pi)$ and compatibly defined $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$.

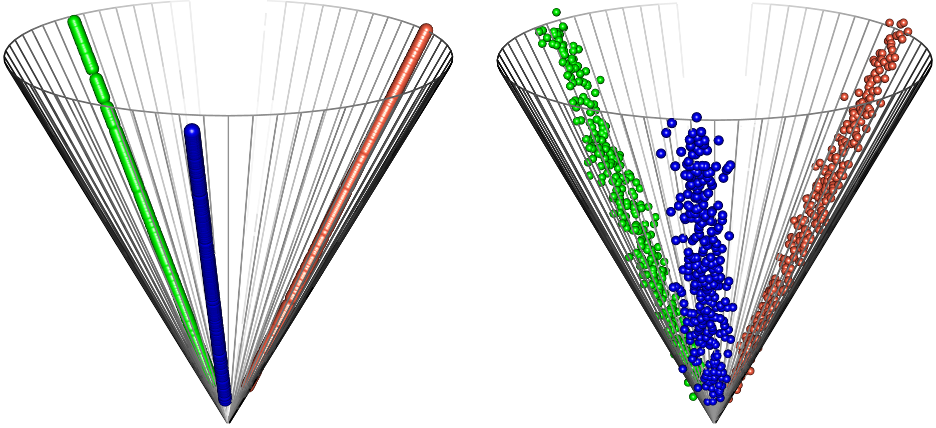


Figure 4.2: The latent geometry of a tripartite graph. The left panel shows three-dimensional latent positions $X_1, \ldots, X_n$ corresponding to the tripartite random graph described in (4.1), coloured by group. The cone represents all the totally isotropic subspaces in $\mathbb{R}^3$ with signature $(1, 2)$. The right panel shows the spectral embedding of a simulated realization of such a graph.

The latent positions lie on three one-dimensional subspaces, corresponding to the three groups. These subspaces are necessarily *totally isotropic* with respect to the indefinite inner product $\langle \cdot, \cdot \rangle_{p,q}$, meaning that the indefinite inner product of any two points on a subspace is zero.

The left panel of Figure 4.2 shows a configuration of these three-dimensional latent positions with 300 nodes in each group and weights drawn uniformly on the interval $[0.1, 1]$. The cone represents all totally isotropic subspaces in $\mathbb{R}^3$ with signature $(1, 2)$. The right panel shows the adjacency spectral embedding of a simulated realization of the graph.

4.2.2 Subspace geometry of multipartite graphs

The subspace geometry observed in the previous example is a special case of a more general phenomenon.

Lemma 4.2. *The latent positions $\{X_i\}_{i \in V_k}$ corresponding to nodes in group k lie on a subspace which is totally isotropic with respect to the indefinite inner product $\langle \cdot, \cdot \rangle_{p,q}$ which has dimension no greater than* $\min\{p, q\}$*.*

Lemma 4.2 follows as a corollary of Witt's theorem of quadratic forms [63], and an elementary proof is given in Section C.1 of the appendix. We denote the dimension of the totally isotropic subspace supporting $\{X_i\}_{i \in V_k}$ by r_k, and we refer to r as the *ambient dimension* and $r_1, \ldots, r_K$ as the *intrinsic dimensions* of the graph.

4.3 Spectral embedding of multipartite graphs

A combination of Lemma 4.2 and Theorem 2.3 says that the adjacency spectral embedding of a low-rank multipartite random graph lies "close" to the union of K lower-dimensional subspaces. This motivates a subsequent step to the standard spectral embedding algorithm of estimating these subspaces and projecting onto them, to obtain vector representations of the nodes of each group in their intrinsic, rather than ambient, dimension. We propose to estimate these subspaces using group-specific, uncentered principal component analysis.

Definition 4.2 (Multipartite adjacency spectral embedding)**.** Given positive integers $r_1, \ldots, r_K \leq r$, let $\hat{X}_1, \ldots, \hat{X}_n$ be the adjacency spectral embedding of $\mathbf{A}$ into $\mathbb{R}^r$ (see Definition 2.2). Then the *multipartite adjacency spectral embedding* of node $i \in V_k$ into $\mathbb{R}^{r_k}$ is

$$\hat{Y}_i = \hat{\mathbf{\Xi}}_k^\top \hat{X}_i,$$

where the columns of $\hat{\mathbf{\Xi}}_k$ are the r_k eigenvectors of $\hat{\mathbf{\Sigma}}_k := n_k^{-1} \sum_{i \in V_k} \hat{X}_i \hat{X}_i^\top$ with the largest eigenvalues.

As indicated by the notation, Definition 4.2 contains biadjacency spectral embedding (Definition 4.1) as a special case when $r_1 = r_2 =: d$ and $r = 2d$, which can be proved using Lemma 4.1.

In modern applications where node degrees are highly heterogeneous, the normalisation step in the following remark is often considered before the spectral decomposition [64–66].

Remark 4.1 (Optional Laplacian regularisation)**.** *Let $d_i = \sum_{j=1}^n a_{ij}$ denote the degree of node i, set a regularisation parameter $\tau \geq 0$ and define the matrix $\mathbf{D}_\tau = \operatorname{diag}(d_1 + \tau, \ldots, d_n + \tau)$. Define the regularised Laplacian matrix $\mathbf{L}_\tau = \mathbf{D}_\tau^{-1/2} \mathbf{A} \mathbf{D}_\tau^{-1/2}$, and replace the adjacency spectral embedding in Definition 4.2 with the embedding obtained from the eigenvectors of the $\mathbf{L}_\tau$ (i.e. replacing $\mathbf{A}$ with $\mathbf{L}_\tau$ in Definitions 2.2).*

When $\tau = 0$, $\mathbf{L}_\tau$ corresponds which the standard symmetric, normalised Laplacian matrix $\mathbf{L}_{\text{sym}}$ introduced in Section 2.2. For a properly chosen τ, $\mathbf{L}_\tau$ has been shown to improve the statistical performance of the spectral embedding for sparse graphs when compared to $\mathbf{A}$ or $\mathbf{L}_{\text{sym}}$ [66–68]. When the Laplacian normalisation step is applied, the estimand in Theorem 2.3 becomes $\{\mathbb{E}(d_i) + \tau\}^{-1/2} X_i$, a rescaling of X_i, and therefore the subspace geometry explored in the preceding section still applies. For simplicity of analysis, we do not consider this step in the theory of this paper, however, we do in the real data example of Section 5.4.

4.3.1 Estimation theory

To facilitate the statistical analysis of point clouds obtained by multipartite spectral embedding, we put down a random graph model parametrised to make the true, intrinsic-dimensional representations of the nodes explicit.

Definition 4.3 (Multipartite bilinear random graph model)**.** Let $\mathbf{\Lambda}_{k,\ell}$ be fixed, $r_k \times r_\ell$ matrices satisfying $\mathbf{\Lambda}_{k,\ell} = \mathbf{\Lambda}_{\ell,k}^\top$ for all $k \neq \ell$ and $\mathbf{\Lambda}_{k,\ell} = 0$ for all $k = \ell$. Let $\{Y_i\}_{i \in V_k}$ be vectors in $\mathbb{R}^{r_k}$ satisfying $Y_i^\top \mathbf{\Lambda}_{z_i,z_j} Y_j \in [0,1]$. The graph $\mathbf{A}$ follows a *multipartite bilinear random graph model* with link matrices $\{\mathbf{\Lambda}_{k,\ell}\}_{k<\ell}$ if $\{a_{ij}\}_{i<j}$ are independent Bernoulli random variables with success probabilities

$$p_{ij} = Y_i^\top \mathbf{\Lambda}_{z_i,z_j} Y_j, \qquad 1 \leq i < j \leq n.$$

While the generalised random dot product graph model requires that edge probabilities be modelled by a common function of the latent positions (i.e. the indefinite inner product), the multipartite inner product graph model allows edges between nodes in different pairs of groups to be modelled using *different* functions. This additional flexibility is what allows nodes to be parametrised in their intrinsic dimension.

Identifiability. There are two distinct sources of non-identifiability in the latent positions and link matrices of a multipartite bilinear random graph model. First, as with the generalised random dot product graph, one can increase the dimension of the latent positions, for example by padding them with zeros without changing the distribution of $\mathbf{A}$. We preclude this by requiring that the matrices $\mathbf{\Lambda}_k := (\mathbf{\Lambda}_{k1} \cdots \mathbf{\Lambda}_{kK})$ and $\mathbf{\Sigma}_k := n_k^{-1} \sum_{i \in V_k} Y_i Y_i^\top$ have full rank. Second, replacing $\{Y_i\}_{i \in V_k}$ with $\{\mathbf{G}_k Y_i\}_{i \in V_k}$ and $\{\mathbf{\Lambda}_{k\ell}\}_{k,\ell=1}^K$ with $\{(\mathbf{G}_k^\top)^{-1} \mathbf{\Lambda}_{k\ell} \mathbf{G}_\ell^{-1}\}_{k,\ell=1}^K$, where $\{\mathbf{G}_k\}_{k=1}^K$ are invertible, linear transformations, does not change the distribution of $\mathbf{A}$. Therefore the latent positions are only identified up to group-wise invertible transformations, a fact reflected in the consistency result to come.

A uniform consistency result. The following theorem asserts that, provided the latent positions $\{Y_i\}_{i=1}^n$ are of the same order of magnitude and the expected degrees of the graph grow at least logarithmically in n, and the model is well-conditioned, then the multipartite adjacency spectral embedding of $\mathbf{A}$, $\{\hat{Y}_i\}_{i=1}^n$, with ambient embedding dimension $r := \operatorname{rank}(\mathbf{\Lambda})$ and intrinsic embedding dimensions equal to the dimension of the corresponding latent positions, provides a uniformly consistent estimate of the latent positions $\{Y_i\}_{i=1}^n$. By “well-conditioned”, we mean that the matrices $\mathbf{\Sigma}_k, \mathbf{\Lambda}_k$ and $\mathbf{\Lambda}$ have (reduced) condition numbers

$$\kappa(\mathbf{\Sigma}_k) = \frac{\sigma_1(\mathbf{\Sigma}_k)}{\sigma_{r_k}(\mathbf{\Sigma}_k)}, \qquad \kappa(\mathbf{\Lambda}_k) = \frac{\sigma_1(\mathbf{\Lambda}_k)}{\sigma_{r_k}(\mathbf{\Lambda}_k)}, \qquad \kappa(\mathbf{\Lambda}) = \frac{\sigma_1(\mathbf{\Lambda})}{\sigma_r(\mathbf{\Lambda})},$$

respectively, which are of constant order for all $k \in [K]$. Here, we assume that the ambient and intrinsic dimensions are fixed and known.

Theorem 4.1. *Suppose* $\mathbf{A}$ *follows a multipartite bilinear random graph model with link matrices* $\{\mathbf{\Lambda}_{k,\ell}\}_{k,\ell=1}^K$ *and latent positions* $\{Y_i\}_{i=1}^n$ *satisfying* $\|Y_i\|_2 \asymp \rho_n^{1/2}$ *for some* $\rho_n \lesssim 1$*. Then, providing* $\kappa(\mathbf{\Sigma}_k) \asymp \kappa(\mathbf{\Lambda}_k) \asymp \kappa(\mathbf{\Lambda}) \asymp 1$*,* $n_1 \asymp \cdots \asymp n_K$ *and* $n\rho_n \gtrsim \log n$*, there exist invertible matrices*

$\{\mathbf{G}_k\}_{k=1}^K$ *such that, for sufficiently large* n,

$$\max_{i\in[n]} \left\|\hat{Y}_i - \mathbf{G}_{z_i} Y_i\right\|_2 \lesssim \sqrt{\frac{\log n}{n}}$$

with overwhelming probability.

4.3.2 Sketch proof of Theorem 4.1.

In this subsection, we present a high-level overview of the main ideas for the proof of Theorem 4.1. The detailed proof is given in Section C.2 of the appendix. We start by constructing linear maps $\{\mathbf{H}_k \in \mathbb{R}^{r\times r_k}\}_{k=1}^K$ satisfying $\mathbf{\Lambda}_{k\ell} = \mathbf{H}_k^\top \mathbf{I}_{p,q}\mathbf{H}_\ell$ so that $Y_i^\top \mathbf{\Lambda}_{z_i,z_j} Y_j = \langle X_i, X_j\rangle_{p,q}$ for all $i,j\in[n]$. Then $\mathbf{A}$ follows a generalised random dot product graph with latent positions $\{X_i\}_{i=1}^n$ and signature (p,q), and verifying the conditions of Theorem 2.3 and applying it, combined with Lemma 2.1, we have that with overwhelming probability, there exists an indefinite orthogonal matrix $\mathbf{Q}^{-1}\in\mathbb{O}(p,q)$ such that, for sufficiently large n,

$$\max_{i\in[n]} \left\|\hat{X}_i - \mathbf{Q}X_i\right\|_2 \lesssim \sqrt{\frac{\log n}{n^{1/2}}}. \tag{4.2}$$

with overwhelming probability. We then show that $\hat{\mathbf{\Sigma}}_k$ concentrates around $\mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^\top$ in spectral norm, where $\mathbf{\Sigma}_k = n_k^{-1}\sum_{i\in V_k} X_i X_i^\top$, in the sense that, for sufficiently large n,

$$\left\|\hat{\mathbf{\Sigma}}_k - \mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^\top\right\|_2 \lesssim \sqrt{\frac{\rho_n \log n}{n^{1/2}}} \tag{4.3}$$

with overwhelming probability.

Recall that $\hat{\mathbf{\Xi}}_k \in \mathbb{R}^{n_k\times r_k}$ is the orthonormal matrix of eigenvectors of $\hat{\mathbf{\Sigma}}_k$ corresponding to its r_k largest eigenvalues, and let $\mathbf{\Xi}_k\in\mathbb{R}^{n_k\times r_k}$ be the orthonormal matrix of eigenvectors of $\mathbf{\Sigma}_k$ corresponding to its r_k non-zero eigenvalues. We have that the smallest non-zero eigenvalue of $\mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^\top$, δ_k, satisfies $\delta_k \asymp \rho_n$, and we then apply the Davis-Kahan theorem, combined with (4.3) to obtain the following eigenvector bound: there exists an orthogonal $\mathbf{W}_k\in\mathbb{O}(r_k)$ such that, for sufficiently large n,

$$\left\|\hat{\mathbf{\Xi}}_k - \mathbf{\Xi}_k\mathbf{W}_k\right\|_2 \le 2^{3/2}\delta_k^{-1}\left\|\hat{\mathbf{\Sigma}}_k - \mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^\top\right\|_2 \lesssim \sqrt{\frac{\log n}{n\rho_n}} \tag{4.4}$$

with overwhelming probability. For each $k\in[K]$, we set $\mathbf{G}_k := \mathbf{W}_k^\top\mathbf{\Xi}_k^\top\mathbf{Q}(\mathbf{H}_k\mathbf{H}_k^\top)^{-1}\mathbf{H}_k$ and derive that

$$\hat{Y}_i - \mathbf{G}_{z_i}Y_i = \hat{\mathbf{\Xi}}_{z_i}^\top\left(\hat{X}_i - \mathbf{Q}X_i\right) + \left(\hat{\Xi}_{z_i} - \mathbf{W}_{z_i}\mathbf{\Xi}_{z_i}\right)^\top \mathbf{Q}X_i,$$

and then we employ the triangle inequality together with (4.2), (4.4) and the facts that $\|\mathbf{Q}\|_2 \lesssim 1$ and $\|X_i\|_2 \asymp \rho_n^{1/2}$ to obtain that for sufficiently large n,

$$\begin{aligned}\max_{i\in[n]}\left\|\hat{Y}_i - \mathbf{G}_{z_i}Y_i\right\|_2 &\le \max_{i\in[n]}\left\|\hat{X}_i - \mathbf{Q}X_i\right\|_2 + \max_{k\in[K]}\left\|\hat{\mathbf{\Xi}}_k - W_k\mathbf{\Xi}_k\right\|_2 \|\mathbf{Q}\|_2 \max_{i\in[n]}\|X_i\|_2 \\ &\lesssim \sqrt{\frac{\log n}{n}}\end{aligned}$$

with overwhelming probability, which establishes the result.

4.3.3 Selecting the embedding dimension

The theory to follow in this paper assumes that the embedding dimensions, both ambient and intrinsic, are known, and correspond to the population ranks of the adjacency matrix and the ambient latent positions corresponding to each group. This represents an “unrealistic ideal” in two respects. First, in practice, these dimensions need to be selected by the practitioner using the data. Second, the finite rank assumption on $\mathbf{P}$ might not be expected to hold exactly. As a result, we prefer to view practical dimension selection as a bias-variance trade-off rather than an estimation problem. Indeed, even if we knew the ambient and intrinsic dimensions, in finite samples they might not be the best to choose. We refer the reader to [69–71] for pragmatic discussions around this topic. Several rank selection methods are available in the literature [69, 72, 73]. We use the elbow method of Zhu and Ghodsi [72] in our real data example but leave the choice open in general. Note that Lemma 4.2 provides the maximal value of the intrinsic dimensions given the ambient dimension. We have found that reasonable rank selection procedures rarely break this inequality in practice (e.g. see Figure 4.5).

4.4 Spectral clustering

A common use of spectral embedding is to uncover community structure in a network. This is achieved via a subsequent clustering procedure such as the k-means algorithm. Algorithm 1 below describes a spectral clustering algorithm for multipartite networks. It takes as input the embedding dimensions and number of communities which in practice must be estimated from the data.

Algorithm 1 Multipartite spectral clustering

```
Input: adjacency matrix $\mathbf{A}$, node partition $V = V_1 \cup \cdots \cup V_K$, embedding dimensions $r_1, \ldots, r_K, r$, number of communities in each group $\kappa_1, \ldots, \kappa_K$.
Compute $\{\hat{Y}_i\}_{i \in V_k} \in \mathbb{R}^{r_k}$, $k \in [K]$, the multipartite spectral embedding of $A$ with ambient dimension $r$, and intrinsic dimensions $r_1, \ldots, r_K$.
(optional) For each $i \in [n]$, set $\hat{Y}_i = \hat{Y}_i / \|\hat{Y}_i\|$.
For each $k \in [K]$, apply $k$-means to $\{Y_i\}_{i \in V_k}$ with $\kappa_k$ clusters.
Output: community partition $\hat{C}_1 \cup \cdots \cup \hat{C}_\kappa = V$.
```

The stochastic block model [37] and degree-corrected stochastic block model [38] are models for community structured networks and are ubiquitous in the community detection literature. We formally define these models in the specific setting of multipartite graphs.

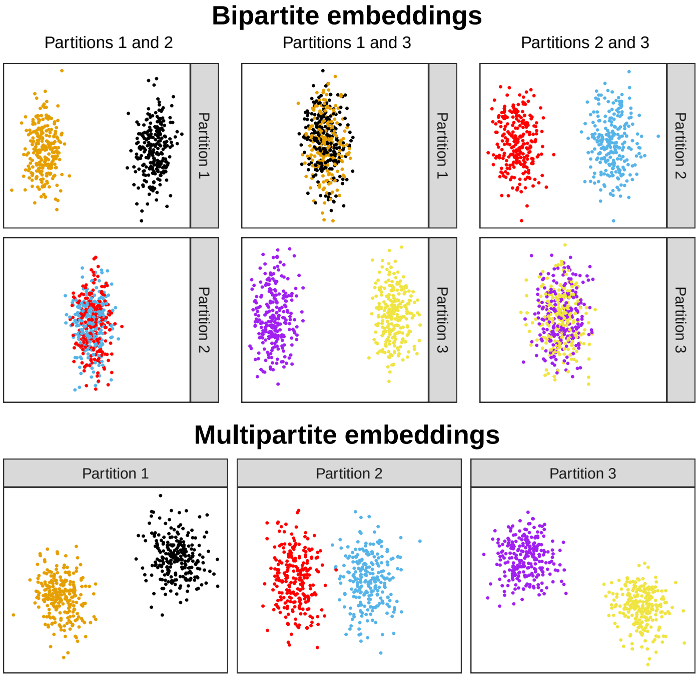


Figure 4.3: For a graph simulated from a multipartite stochastic block model with inter-community probability matrix of the form (4.5), the top panel shows the biadjacency spectral embeddings (Definition 4.1) of the subgraphs corresponding to every pair of groups. For each pair of groups, two of the four relevant communities cannot be distinguished. The bottom panel shows the multipartite spectral embedding (Definition 4.2) of the full tripartite graph, revealing all six communities.

Definition 4.4 (Multipartite degree-corrected stochastic block model)**.** Suppose the node set $V = V_1 \cup \cdots \cup V_K$ of a multipartite graph $\mathbf{A}$ is further sub-partitioned into S disjoint blocks $V = C_1 \cup \cdots \cup C_\kappa$. Let $\mathbf{B} \in [0,1]^{\kappa \times \kappa}$ be a matrix satisfying $b_{uv} = 0$ if $C_u, C_v \subseteq V_k$ for some $k \in [K]$, and let $w_1, \ldots, w_n \in [0,1]$ be a set of weights. The graph A follows a *multipartite degree-corrected stochastic block model* if $\{a_{ij}\}_{i<j}$ are independent Bernoulli random variables with success probabilities

$$p_{ij} = w_i w_j b_{uv}, \qquad i \in C_u, j \in C_v.$$

If $w_1, \ldots, w_n = 1$, then the model is referred to as the multipartite stochastic block model. In the following, we let m_k denote the number of communities in group k.

All multipartite degree-corrected stochastic block models can be parametrised as a multipartite inner product graph model. When $\mathbf{B}$ has full rank, one such parametrisation is to set $\mathbf{\Lambda}_{k,\ell}$ to the $m_k \times m_\ell$ submatrix of $\mathbf{B}$ corresponding to communities in groups k and ℓ, and to set Y_i to be the indicator vector with w_i in the position corresponding to its community. In addition, in this case, the intrinsic dimensions are equal to the number of sub-communities in each group, and therefore intrinsic dimension estimates may also serve as estimates of these quantities.

A corollary of the uniform consistency result of Theorem 4.1 is that, supposing the graph follows a multipartite stochastic block model (and step 2 of Algorithm 1 is not implemented) or a multipartite degree-corrected stochastic block model (and step 2 of Algorithm 1 is implemented), then asymptotically almost surely, Algorithm 1 will perfectly estimate the community memberships, assuming the conditions of the theorem hold and that the number of communities in each group is known. This may be proved using analogous arguments to those employed in Lyzinski et al. [42].

4.4.1 Obscured communities

The following is an example of a multipartite stochastic block model for which multipartite spectral clustering reveals the full community structure of a graph, but spectral clustering using any individual bipartite subgraph fails to reveal a community. This example was inspired by a similar example in Jones and Rubin-Delanchy [74] in the context of multiple graph embedding.

Consider a tripartite stochastic block model, with two sub-communities in each group, where the matrix $\mathbf{B}$ has full rank and the form

$$\mathbf{B} = \left(\begin{array}{cc|cc|cc} 0 & 0 & a & a & c & d \\ 0 & 0 & b & b & c & d \\ \hline a & b & 0 & 0 & e & e \\ a & b & 0 & 0 & f & f \\ \hline c & c & e & f & 0 & 0 \\ d & d & e & f & 0 & 0 \end{array}\right) \tag{4.5}$$

for some $a, b, c, d, e, f \in [0, 1]$. If, for example, we consider only the bipartite subgraph corresponding to groups 1 and 2, the two communities of group 2 are indistinguishable, and in fact, every other bipartite subgraph also obscures a community. No single biadjacency spectral embedding can therefore uncover all of the latent communities. However, they are all revealed through multipartite spectral embedding. This is illustrated by simulation in Figure 4.3.

4.5 Application to a large biomedical knowledge graph

Here we apply Algorithm 1 to a multipartite network representing associations between biomedical entities belonging to six groups: drugs, diseases, targets, pathways, variant locations and haplotypes[1]. Figure 4.4 shows a schematic of the topology of the data. The associations were inferred from several biological databases: Drugbank [75], Kyoto Encyclopedia of Genes and Genomes (KEGG) [76], PharmGKB [77] and the Human Disease network [78]. A superset of the dataset we use was introduced and detailed in Zong et al. [61].

[1] Code to reproduce the analysis in this section is available at `github.com/alexandermodell/multipartite_clustering`.

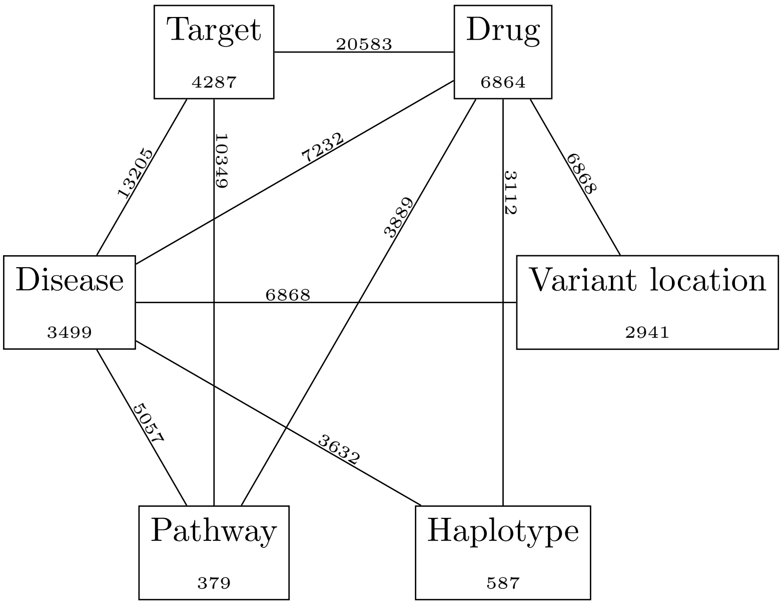


Figure 4.4: Schematic of the biomedical multipartite network analysed in Section 5.4. The number of nodes in each group and the number of edges between groups (zero if unspecified) are indicated.

We apply Algorithm 1 to the graph, implementing Laplacian regularisation as described in Remark 4.1 with τ equal the average degree of the graph as recommended in Qin and Rohe [66]. We use the elbow method of Zhu and Ghodsi [72] for ambient and intrinsic dimension selection, implement the optional spherical projection step, and select the number of communities in each group equal to the intrinsic dimension estimates under the assumption of a full-rank degree-corrected stochastic block model. In addition, we do not include nodes of degree less than five in the clustering steps, as there is not enough signal in their positions for them to be accurately clustered.

The left panel of Figure 4.5 shows the first 1000 singular values of the regularized Laplacian and the dimension ($\hat{r} = 214$) selected by the elbow method of Zhu and Ghodsi [72]. The remaining panels show the singular values of the ambient embedding of each node group. The black lines show the intrinsic dimension selected by the elbow method and the dashed line shows $\min\{\hat{p}, \hat{q}\}$, which is always larger, as predicted by Lemma 4.2.

The intrinsic dimension selected also acts as an estimate of the number of communities in the group. In the Drug and Pathway groups, each item has associated labels, obtained from the DrugBank [75] and KEGG [76] databases, which we use as held-out information to interpret and evaluate the clustering obtained.

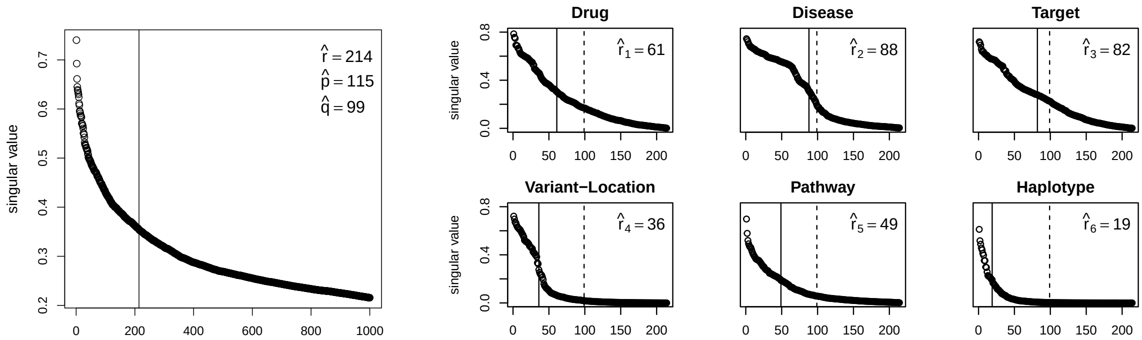


Figure 4.5: Dimension selection for the biomedical multipartite network. The left panel shows the scree plot of the regularized Laplacian matrix and the right panels the scree plots of the ambient embeddings corresponding to each group. The dimension selected — and thus the number of clusters — is shown as a solid line and $\min\{\hat{p}, \hat{q}\} = 99$ is shown as a dashed line.

Table 4.1: Example clusters of pathways.

Cluster 9
Fatty acid degradation **(Li, Me)**
Peroxisome **(Ce, Tr)**
PPAR signaling pathway **(En, Or)**
Fat digestion and absorption **(Di, Or)**
Fatty acid metabolism **(Me, Ov)**
Primary bile acid biosynthesis **(Li, Me)**
alpha-Linolenic acid metabolism **(Li, Me)**
Biosynthesis of unsaturated fatty acids **(Li, Me)**
$\cdots + 3$ more
(Me) 8/80, **(Li)** 5/15, **(Or)** 2/69

Cluster 11
Morphine addiction **(Hu, Su)**
Amphetamine addiction **(Hu, Su)**
Circadian entrainment **(En, Or)**
Amyotrophic lateral sclerosis **(Hu, Ne)**
Nicotine addiction **(Hu, Su)**
Renin secretion **(En, Or)**
Cocaine addiction **(Hu, Su)**
$\cdots + 5$ more
(Hu) 6/71, **(Or)** 5/69, **(Su)** 4/5, **(En)** 2/17, **(Se)** 2/4

Cluster 39
Alzheimer's disease **(Hu, Ne)**
Parkinson's disease **(Hu, Ne)**
Oxidative phosphorylation **(Em, Me)**
Non-alcoholic fatty liver disease **(Em, Hu)**
Huntington's disease **(Hu, Ne)**
(Hu) 4/71, **(Ne)** 3/5

Cluster 61
Prostate cancer **(Ca, Hu)**
Central carbon metabolism in cancer **(Ca, Hu)**
Acute myeloid leukemia **(Ca, Hu)**
Chronic myeloid leukemia **(Ca, Hu)**
Melanoma **(Ca, Hu)**
Non-small cell lung cancer **(Ca, Hu)**
Glioma **(Ca, Hu)**
$\cdots + 4$ more
(Ca) 10/22, **(Hu)** 10/71

"**(Xx)** a/b" means "label **(Xx)** appears a times in the cluster and b times in total".

Labels: **(Ca)** Cancers, **(Ce)** Cellular processes, **(Di)** Digestive system, **(Em)** Energy metabolism, **(En)** Endocrine systems, **(Hu)** Human diseases, **(Li)** Lipid metabolism, **(Me)** Metabolism, **(Ne)** Neurodegenerative diseases, **(Or)** Organismal systems, **(Ov)** Overview, **(Su)** Substance dependence, **(Tr)** Transport and catabolism.

Table 4.2: Example clusters of drugs.

Cluster 3
Pyridoxal Phosphate **(Di, Mi, Su, Vb)**
Citric Acid **(Ac, Ch)**
Alglucosidase alfa **(Ez)**
L-Proline **(Di, Mi, Nea, Su)**
L-Aspartic Acid **(Di, Mi, Nea, Su)**
Pyruvic acid **(Di, Mi, Su)**
Tetrahydrofolic acid **(Di, Mi, Su)**
··· + 22 more
(Di) 12/44, **(Mn)** 12/41, **(Su)** 12/44, **(Aa)** 4/6, **(Nea)** 4/12, **(Vb)** 3/10

Cluster 36
Diazepam **(Ad, Am, Ane, Ax, Cv, Ga, Hy, Mu)**
Midazolam **(Ad, Ane, Ax, Ga, Hy)**
Baclofen **(Mu, Nm)**
Clobazam **(Bz, Cv)**
Propofol **(An, Hy)**
Lorazepam **(Bz, Hy)**
··· + 46 more
(Hy) 23/37, **(Bz)** 16/17, **(Ga)** 14/18, **(Ax)** 12/17, **(Ad)** 6/25, **(Cv)** 4/31

Cluster 45
Caffeine **(Ap, Ce, P1, Ph)**
Theophylline **(Br, Mu, P1, Ph, Va)**
Adenosine monophosphate **(Di, Mi, Su)**
Aminophylline **(Br, Ca, Mu, P1, Ph)**
Oxtriphylline **(Br)**
Flavoxate **(Pa)**
Papaverine
··· + 25 more
(Va) 9/35, **(Br)** 7/22, **(Ph)** 7/8, **(Pl)** 5/15, **(Mu)** 3/19, **(P1)** 3/3

Cluster 61
Methadone **(An, Na, Tu)**
Morphine **(An, Na)**
Heroin **(An, Na)**
Oxycodone **(Ad, An, Na)**
Fentanyl **(Ad, An, Ana, Na)**
Ketamine **(Ag, Ane, Ex)**
Alfentanil **(Ag, An, Na)**
··· + 28 more
(Ag) 20/41, **(Na)** 17/18, **(Naa)** 5/8, **(Ad)** 5/25, **(An)** 5/31, **(Tu)** 5/8

"**(Xx)** *a*/*b*" means "label **(Xx)** appears *a* times in the cluster and *b* times in total".

Labels: **(Aa)** Amino acids, **(Ad)** Adjuvants, **(Ana)** Analgesics, **(Ane)** Anesthetics, **(Ap)** Appetite depressants, **(Ax)** Anti-anxiety agents, **(Br)** Bronchodilator agents, **(Bz)** Benzodiazepines, **(Ca)** Cardiotonic agents, **(Ce)** Central-nervous-system stimulants, **(Cv)** Anticonvulsants, **(Di)** Dietary supplements, **(Ex)** Excitatory amino acid antagonists, **(Ez)** Enzyme replacement agents, **(Ga)** GABA modulators, **(Hy)** Hypnotics and sedatives, **(Mi)** Micronutrients, **(Mu)** Muscle relaxants, **(Na)** Narcotics, **(Naa)** Narcotic antagonists, **(Nea)** Non-essential amino acids, **(Nm)** Neuromuscular agents, **(P1)** Purinergic P1 receptor antagonists, **(Ph)** Phosphodiesterase inhibitors, **(Su)** Supplements, **(Tu)** Antitussive agents, **(Va)** Vasodilator agents, **(Vb)** Vitamin-B complex.

The correspondence between the communities recovered and their labels is strong. Tables 1 and 2 show example clusters in the Drug and Pathway groups. The items with the highest degree are shown, with their labels in brackets. Below, we show the number of occurrences of each label within the cluster and in total, for the most commonly occurring labels in the cluster (if they appear more than once).

In the Drug group, Cluster 3 contains primarily nutrition-related substances, Cluster 36 contains primarily hypnotics and sedatives including all but one of the benzodiazepines, Cluster 45 primarily vasodilators including all but one of the Phosphodiesterase inhibitors and Cluster 61 includes all but one of the narcotics. In the Pathway group, Cluster 9 corresponds to pathways related to metabolism, Cluster 11 to addiction, and Cluster 39 to neuro-degenerative diseases and Cluster 61 to cancer.

Discussion

This chapter elucidates the geometry of low-rank multipartite networks and uses this to motivate a secondary dimension reduction step after spectral embedding. Network communities can then be recovered through k-means clustering, and the estimated intrinsic dimension can serve as an estimate of the number of communities.

Two important assumptions in our statistical model are that the edges are conditionally independent and that the adjacency matrix has some low, ‘true’, rank. These assumptions, standard in statistical graph theory, are the subject of active academic discourse [79–81]. The theory, and the practical scope of our method, is likely to extend to mild edge dependence [82] and fast enough eigenvalue decay [83, 84].

Chapter 5

Spectral embedding of dynamic networks

In this chapter, we divert our attention from static networks to dynamic networks — specifically, to those characterised by a collection of instantaneous interaction events which occur between pairs of nodes in continuous time.

Making sense of patterns of connections occurring over time is a common theme of modern data analysis and is often approached in one of two ways. On the one hand, we may see dynamic network data as a *graph*, in which connections between the same entities over time are somehow treated as one, e.g. through weighting. This view evokes methodological ideas such as community detection [14, 85], topological data analysis [86], or manifold learning [87, 88]. On the other, we may see the data as a set of *point processes* [89], each modelling the event times of connections between two entities. This view evokes temporal notions such as trend, changepoints and periodicity. There are many opportunities for innovation *combining ideas* from these different modelling cultures.

We present a spectral embedding framework for continuous-time dynamic network data which learns continually evolving representations of nodes, in which ideas from both the graph and temporal domains can be combined. Our framework, which we call Intensity Profile Projection, consists of three stages: estimating the intensity functions underlying the interactions between pairs of nodes, e.g. via kernel smoothing; learning a projection which minimises a notion of intensity reconstruction error; and inductively constructing evolving node representations via the learned projection.

Our algorithm has material advantages over existing approaches, broadly relating to statistical precision and interpretability, which open new possibilities for inference. For instance, in Section 5.3, we present a synthetic example involving continuously evolving network topology, which is shown to be hard to infer, not to say impossible, by other methods.

By "precision", above, we mean a uniform error bound: controlling the largest error of any representation over the entire time domain and node-set. By "interpretability", we are referring to a property of our method which to our knowledge is unique among continuous-time methods: two nodes at two points in time exhibiting statistically indistinguishable behaviour are mapped to the same position, up to noise. Properties known in the literature as temporal coherence (or longitudinal stability) and structure preservation (or cross-sectional stability) [90, 91] are established as special cases in which the same node is considered at two distinct points in time, or two distinct nodes are considered at the same point in time. These results assume a generic inhomogeneous Poisson dynamic network model.

Our estimation theory elucidates the role of smoothing as a bias-variance trade-off and shows how we can reduce smoothing as the signal-to-noise ratio increases on account of the algorithm 'borrowing strength' across the network.

Related work. Our proposed framework combines ideas from point process modelling [92] and spectral embedding [15, 47]. The theoretical analysis draws on recent developments in entrywise eigenvector estimation for random matrices [74, 82, 93, 94]. For a specific choice of intensity estimator (the histogram), our method can be viewed as a weighted graph analogue of Unfolded Spectral Embedding [74, 91], but those papers consider different data (multilayer or discrete time networks) and models.

We perform a comprehensive method comparison in Section 5.3. To our knowledge, the only unsupervised representation learning methods for dynamic network data (as defined in the next section) are [95–97], which are based on latent position models and have much weaker theoretical guarantees. There are a number of discrete-time dynamic network representation learning algorithms, which broadly fall under latent position models [98–100], spectral methods [74, 91, 101–103] and word-embedding-based methods [104, 105]. Given how limited the options are for handling continuous time, in our method comparison we also include some discrete-time methods which could reasonably be used as alternatives.

5.1 Intensity Profile Projection

Data. We consider dynamic network data, denoted $\mathcal{G}$, representing instantaneous undirected interactions between nodes over time, which we define formally as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ on a time domain $\mathcal{T} = (0, T]$, containing a node set $\mathcal{V} = [n]$ and a set of triples $\mathcal{E} = \{(i_e, j_e, t_e)\}_{e \geq 1}$, each corresponding to an undirected interaction event, where $i_e < j_e \in \mathcal{V}, t_e \in \mathcal{T}$. We let $\mathcal{E}_{ij} := \{t : (i, j, t) \in \mathcal{E}\}$ denote the interaction events between nodes i and j.

Model. We assume the interaction events $\mathcal{E}_{ij}$ are driven by an independent inhomogeneous Poisson process with intensity $\lambda_{ij}(t)$. Informally:

$$\lambda_{ij}(t)\mathrm{d}t = \mathbb{P}\left\{\text{interaction between nodes } i \text{ and } j \text{ in } (t, t + \mathrm{d}t]\right\}.$$

We represent these intensities in a symmetric time-varying matrix $\mathbf{\Lambda}(\cdot) : \mathcal{T} \to \mathbb{R}_{+}^{n \times n}$.

Procedure. Intensity Profile Projection can be summarised as follows.

1. **Intensity estimation.** Construct intensity estimates $\hat{\lambda}_{ij}(\cdot)$ of $\lambda_{ij}(\cdot)$ from $\mathcal{E}_{ij}$ for all $i < j$.
2. **Subspace learning.** Compute the top d eigenvectors $\hat{\mathbf{U}}_d = (\hat{u}_1, \ldots, \hat{u}_d)$ of

   $$\hat{\mathbf{\Sigma}} := \frac{1}{T}\int_0^T \hat{\mathbf{\Lambda}}^2(t)\mathrm{d}t, \tag{5.1}$$

   where $\hat{\mathbf{\Lambda}}(t)$ has symmetric entries $\hat{\lambda}_{ij}(t)$, and rows denoted $\hat{\Lambda}_i(t)$ called *intensity profiles.*
3. **Projection.** For a query node i at time t, project the intensity profile $\hat{\Lambda}_i(t)$ onto the subspace spanned by $\hat{u}_1, \ldots, \hat{u}_d$, to obtain $\hat{X}_i(t) = \hat{\mathbf{U}}_d^\top \hat{\Lambda}_i(t)$.

While we develop more principled statistical justifications for the procedure in future sections, it is inspired by a simple reconstruction argument. For an arbitrary d-dimensional subspace spanned by the orthonormal columns of a matrix $\mathbf{V}_d \in \mathbb{R}^{n \times d}$, let

$$\hat{r}_i(t; \mathbf{V}_d) := \left\| \mathbf{V}_d \mathbf{V}_d^\top \hat{\Lambda}_i(t) - \hat{\Lambda}_i(t) \right\|_2$$

denote the reconstruction error of node i at time t, and define the *integrated residual sum of squares* as

$$\hat{R}^2(\mathbf{V}_d) := \int_0^T \sum_{i=1}^n \hat{r}_i^2(t; \mathbf{V}_d)\, \mathrm{d}t.$$

Lemma 5.1. *Among all d-dimensional subspaces of $\mathbb{R}^n$, the column span of $\hat{\mathbf{U}}_d$ minimises the integrated residual sum of squares criterion $\hat{R}^2$.*

Lemma 5.1 may be viewed as a dynamic analogue to the classical Eckart-Young theorem on low-rank matrix approximation [106]. A proof is given in Section D.4 of the appendix.

5.1.1 Intensity estimation

The choice of intensity estimator is left fully open, but our theory makes two important recommendations. First, there are computational gains to be made using sparse estimators for subspace learning. Second, the procedure borrows strength across the network and can give precise representations even when the individual intensity estimates are noisy (e.g. inconsistent). In our experiments, we focus on standard non-parametric estimators such as the histogram or kernel smoothers and choose kernels with finite support to induce sparse estimates.

5.1.2 Subspace learning

The subspace learning step of our procedure involves the computation of an integral, and computing the eigendecomposition of the resulting dense matrix $\hat{\mathbf{\Sigma}}$, both of which may be infeasible for large networks. If a sparse intensity estimator is employed in step 1 of the procedure

Algorithm 2 Approximate Intensity Profile Projection

```
Input: Continuous time dynamic graph $\mathcal{G}$, dimension $d$.
1: Construct intensity estimates $\hat{\lambda}_{ij}(\cdot)$ of $\lambda_{ij}(\cdot)$ from $\mathcal{E}_{ij}$ for all $i < j$.
2: Compute the top $d$ left singular vectors $\hat{u}_1, \ldots, \hat{u}_d$ of
       $\left[\hat{\mathbf{\Lambda}}(t_1)\ \hat{\mathbf{\Lambda}}(t_2)\ \cdots\ \hat{\mathbf{\Lambda}}(t_B)\right]$
   where $t_1 < \cdots < t_B$ are equally spaced points on $(0, T]$.
3: Define the trajectory of node $i$ as
       $\hat{X}_i(t) := \hat{\mathbf{U}}_d^\top \hat{\Lambda}_i(t)$
   where $\hat{\mathbf{U}}_d := (\hat{u}_1, \ldots, \hat{u}_d)$.
Output: Node trajectories $\hat{X}_1(t), \ldots, \hat{X}_n(t)$.
```

and we approximate the integral (5.1) using a numerical quadrature scheme, then step 2 can be rephrased as a single sparse, truncated singular value decomposition, which can be computed quickly for very large networks using an efficient solver [107, 108].

Consider the numerical approximation

$$\hat{\mathbf{\Sigma}} \approx \frac{1}{B}\sum_{b=1}^{B} \hat{\mathbf{\Lambda}}^2(t_b) \tag{5.2}$$

where $t_1 < \cdots < t_B$ are equally spaced points on $(0, T]$. The top d eigenvectors of the right-hand-side of (5.2) are then equal[1] to the top d left singular vectors of the matrix

$$\left[\hat{\mathbf{\Lambda}}(t_1)\ \hat{\mathbf{\Lambda}}(t_2)\ \cdots\ \hat{\mathbf{\Lambda}}(t_B)\right],$$

the row concatenation of $\hat{\mathbf{\Lambda}}(t_1), \ldots, \hat{\mathbf{\Lambda}}(t_B)$. This procedure is presented in Algorithm 2.

5.1.3 Projection

The inductive nature of Intensity Profile Projection allows us to obtain representations $\hat{X}_i(t)$ on demand, for example, the full trajectory for a particular node, or the representations of the entire graph at a point in time. It is possible to obtain representations for intensity profiles outside the training sample, corresponding to new nodes or times outside the training domain, allowing online inference. In practice, one will need to retrain occasionally, i.e. return to step 2, although we leave the discussion of this computational and statistical trade-off for future work (see, for example, [109] in the context of static networks).

[1] Up to signs, rotations in the eigenspaces in the case of repeated eigenvalues, and assuming a gap between the dth and $(d+1)$th eigenvalues.

5.2 Estimation theory

In this section, we develop estimation theory showing the sense in which $\hat{X}_i(t)$ is a "good" estimator of $X_i(t) := \mathbf{U}_d^\top \Lambda_i(t)$ where $\mathbf{U}_d = (u_1, \ldots, u_d) \in \mathbb{R}^{n \times d}$ is the matrix containing the top-d orthonormal eigenvectors of

$$\boldsymbol{\Sigma} := \frac{1}{T} \int_0^T \boldsymbol{\Lambda}^2(t) \mathrm{d}t.$$

In this section, we assume, without loss of generality, that $\mathcal{T} = (0, 1]$. We now introduce some quantities which appear in our main theorem. Firstly, we assume that each $\lambda_{ij}(\cdot)$ is Lipschitz with constant L, and is upper bounded by $\lambda_{\max}$. Secondly, we define the (reduced) condition number and the eigengap,

$$\kappa := \frac{\sigma_1}{\sigma_d}, \qquad \text{and} \qquad \delta := \sigma_d - \sigma_{d+1},$$

respectively, where $\sigma_1^2 \geq \cdots \geq \sigma_n^2$ are eigenvalues of $\boldsymbol{\Sigma}$. Thirdle, we introduce the coherence parameter

$$\mu := \sqrt{\frac{n}{d}} \left\| \mathbf{U}_d \right\|_{2,\infty},$$

which is small when, informally, information about a single entry of $\boldsymbol{\Sigma}$ is "spread out" across the matrix [110]. Finally, we define the population residuals

$$r_i(t) := \left\| \mathbf{U}_d \mathbf{U}_d^\top \Lambda_i(t) - \Lambda_i(t) \right\|_2.$$

Rather than attempt to develop a theoretical framework encompassing all intensity estimators, we choose arguably the most rudimentary, the histogram, and we expect more powerful estimators will only improve matters. This choice of estimator is also attractive because it allows us to pinpoint the crucial practical considerations at play.

We now state the assumptions we require for our theorem. Our first assumption is that the intensities are bounded.

Assumption 1 (Bounded intensities)**.** *The intensities are upper bounded by a constant which doesn't depend on the other quantities in the problem; i.e.* $\lambda_{\max} \lesssim 1$.

Our second assumption is on the population integrated residuals. It ensures that the intensity profiles $\Lambda_1(t), \ldots, \Lambda_n(t)$ do not deviate "too much" from a common low-dimensional subspace.

Assumption 2 (Small population residuals)**.** *The population residuals satisfy*

$$r_1(t), \ldots, r_n(t) \lesssim \sqrt{\frac{d}{n}} \mu \delta \log^{5/2} n$$

for all $t \in [0, 1]$.

Our third assumption is a technical condition on the eigengap which, broadly speaking, ensures that there is "enough signal".

Assumption 3 (Enough signal)**.** *The eigengap satisfies* $\delta \log(\delta/\sqrt{n\lambda_{\max}}) \gtrsim \kappa n \lambda_{\max}$.

Our final assumption is on the bin size, and ensures that the bins are not chosen "too small".

Assumption 4 (Large enough bins)**.** *The number of bins satisfies* $M \lesssim n\lambda_{\max}/\log^3 n$.

These assumptions are weaker than those typically required in the literature (e.g. on stochastic block models and random dot product graphs [8, 14]). To emphasise this point, consider the following stronger alternative assumptions which imply Assumptions 1, 2 and 3:

Assumption 1a. *The intensities* $\lambda_{ij}(t)$ *are of comparable order, i.e.* $\lambda_{ij}(t) \asymp \rho$ *for some* $\rho \lesssim 1$ *and all* $i, j \in [n], t \in (0, 1]$

Assumption 2a. *The matrix* $\mathbf{\Sigma}$ *has rank* $d \asymp 1$*; is incoherent, i.e.* $\mu \asymp 1$*; and its non-zero eigenvectors are of comparable order, i.e.* $\sigma_1 \asymp \sigma_d > \sigma_{d+1} = 0$.

It is immediate that Assumption 1a implies Assumption 1 and under Assumption 2a, the population residuals are all exactly zero, $\kappa \asymp 1$ and $\delta \asymp n\rho$, which implies Assumptions 2 and 3.

Assumption 4 requires that the expected number of events involving each node in each bin is at least of the order $\log^3 n$. This is analogous to the $\log n$ degree growth required for perfect clustering under the binary stochastic block model. Since the latter is an information-theoretic bound [111] and the additional logarithmic powers in our work stem from the sub-exponential tails of the Poisson distribution, we do not think this assumption can be weakened.

We now state our main theorem, which under Assumptions 1-4, provides a non-asymptotic bound on the error between the learned representations and their population counterparts, which holds uniformly over the whole node-set and the time domain.

Theorem 5.1. *Suppose that* $\hat{\lambda}_{ij}(t)$ *are histogram estimates with* M *equally-spaced bins and that Assumptions 1-4 hold. Then with overwhelming probability, there exists an orthogonal matrix* $\mathbf{W}$ *such that*

$$\max_{i\in[n]} \sup_{t\in(0,1]} \left\| \mathbf{W}\hat{X}_i(t) - X_i(t) \right\|_2 \lesssim \frac{n^{3/2} L \lambda_{\max}}{M\delta} + \mu\sqrt{M\lambda_{\max} d} \cdot \log^{5/2} n. \tag{5.3}$$

As a corollary to Theorem 5.1, we state a simplified version of this result in which we replace Assumptions 1, 2 and 3 with the stronger Assumptions 1a and 2a. Since the Lipschitz constant L scales with the order of the intensities, and we define the quantity L_0 satisfying $L = \rho L_0$ which is invariant to the rescaling of intensities.

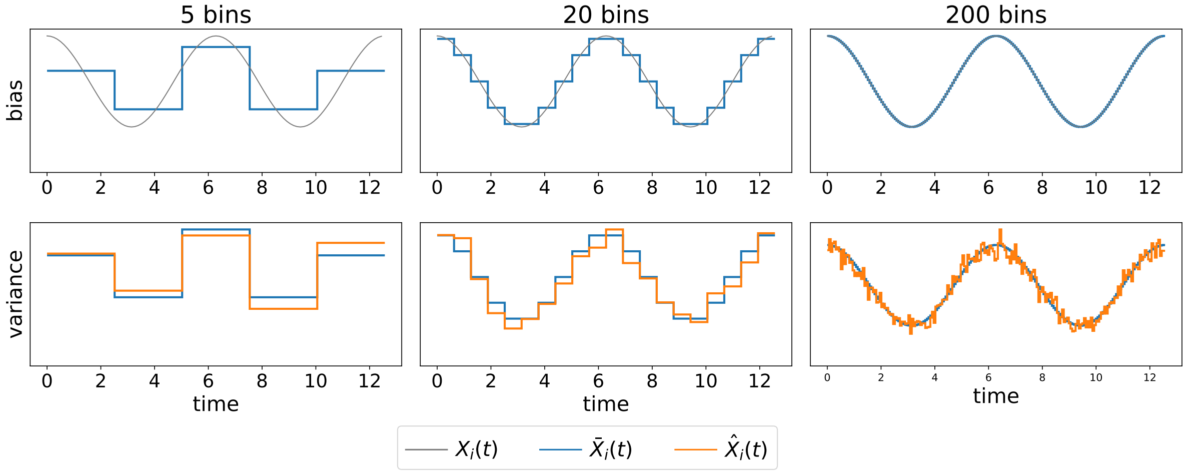


Figure 5.1: A bias-variance trade-off. We simulate a network with common intensities $\lambda_{ij}(t) = 0.7 \times \{2 + \cos(t)\}$ for all i, j, and apply Intensity Profile Projection with a histogram intensity estimator with 5, 20, and 200 bins. In the 'bias' plots, the grey lines show an estimand $X_i(t)$, while blue lines show its histogram approximation. The discrepancy between gray line and the blue line corresponds to the bias of the Intensity Profile Projection estimator. In the 'variance' plots, the blues lines are as in the 'bias' plots and the orange line shows the estimate obtained using Intensity Profile Projection into one dimension. The discrepancy between the blue line and the orange line corresponds to the variance of the Intensity Profile Projection estimator.

Corollary 5.1. *Suppose that $\hat{\lambda}_{ij}(t)$ are histogram estimates with M equally-spaced bins and that Assumptions 1a, 2a and 4 hold. Then with overwhelming probability, there exists an orthogonal matrix $\mathbf{W}$ such that*

$$\max_{i\in[n]} \sup_{t\in(0,1]} \left\| \mathbf{W}\hat{X}_i(t) - X_i(t) \right\|_2 \lesssim \frac{n^{1/2}\rho L_0}{M} + \sqrt{M\rho} \cdot \log^{5/2} n. \tag{5.4}$$

5.2.1 A bias-variance trade-off

The first term in the bound corresponds to the bias between $\bar{X}_i(t)$ and $X_i(t)$, where $\bar{X}_i(t)$ is a histogram approximation to $X_i(t)$ (modulo orthogonal transformation, see Section D.5.3 of the appendix). The second term corresponds to the variance of the estimate.

Theorem 5.1 gives some theoretical guidance on how to select the number of bins in the histogram estimator. For simplicity, we consider the setting of Corollary 5.1. Ignoring logarithmic terms in n, the bound in (5.4) is optimised by choosing

$$M \asymp \left(n\rho L_0^2\right)^{1/3}.$$

Figure 5.1 illustrates this bias-variance trade-off with an example. We simulate a dynamic network with 100 nodes with common intensities $\lambda_{ij}(t) = 0.7 \times \{2 + \cos(t)\}$, for all i, j, on the time domain $(0, 4\pi]$.

The top row shows the population representation $X_i(t)$ of a single node (grey) and its histogram approximation $\bar{X}_i(t)$ (blue) for a variety of bin sizes. The more bins that are chosen, the smaller the bias and the more $\bar{X}_i(t)$ resembles $X_i(t)$. The bottom row shows the histogram approximation $\bar{X}_i(t)$, and the estimate $\hat{X}_i(t)$ (orange) obtained using Intensity Profile Projection. The fewer bins that are chosen, the smaller the variance and the more that $\hat{X}_i(t)$ resembles $\bar{X}_i(t)$.

5.2.2 Sketch proof of Theorem 5.1

In this subsection, we provide a sketch of the main proof techniques used to prove Theorem 5.1. The proof begins by decomposition the error term into bias and variance terms:

$$\max_{i,j\in[n]}\sup_{t\in\mathcal{T}}\left\|\mathbf{W}\hat{X}_i(t)-X_i(t)\right\|_2 = \underbrace{\max_{i,j\in[n]}\sup_{t\in\mathcal{T}}\left\|\mathbf{W}\hat{X}_i(t)-\bar{X}_i(t)\right\|_2}_{\text{variance}} + \underbrace{\max_{i,j\in[n]}\sup_{t\in\mathcal{T}}\left\|\mathbf{W}'\bar{X}_i(t)-X_i(t)\right\|_2}_{\text{bias}},$$

where $\bar{X}_i(\cdot)$ is the histogram approximation of $X_i(\cdot)$ with M equally spaced bins. The bias term is then bounded using standard techniques which make use of the Lipschitz continuity of $\lambda_{ij}(\cdot)$. To bound the variance term, we define the matrices

$$\hat{\mathbf{\Lambda}} = \begin{bmatrix}\hat{\mathbf{\Lambda}}(t_1) & \cdots & \hat{\mathbf{\Lambda}}(t_M)\end{bmatrix}, \qquad \bar{\mathbf{\Lambda}} = \begin{bmatrix}\bar{\mathbf{\Lambda}}(t_1) & \cdots & \bar{\mathbf{\Lambda}}(t_M)\end{bmatrix}$$

where $t_1,\ldots,t_M$ are equally spaces points on $[0,1]$. Let $\hat{\mathbf{S}}$ and $\bar{\mathbf{S}}$, respectively, denote the diagonal matrices containing their top-d singular values and let $\hat{\mathbf{U}},\hat{\mathbf{V}}$ and $\bar{\mathbf{U}},\bar{\mathbf{V}}$, respectively, be the matrices whose columns contain corresponding orthonormal left and right singular vectors.

Using elementary linear algebra, one can show that

$$\max_{i,j\in[n]}\sup_{t\in\mathcal{T}}\left\|\mathbf{W}\hat{X}_i(t)-\bar{X}_i(t)\right\|_2 = \left\|\hat{\mathbf{V}}\hat{\mathbf{S}}-\bar{\mathbf{V}}\bar{\mathbf{S}}\mathbf{W}\right\|_{2,\infty}.$$

Following similar decompositions to those employed in Cape et al. [4], Lyzinski et al. [46], Jones and Rubin-Delanchy [74] and Xie [94], we decompose $\hat{\mathbf{V}}\hat{\mathbf{S}}-\bar{\mathbf{V}}\bar{\mathbf{S}}\mathbf{W}$ as

$$\begin{aligned}\hat{\mathbf{V}}\hat{\mathbf{S}}-\bar{\mathbf{V}}\bar{\mathbf{S}}\mathbf{W} &= \bar{\mathbf{V}}(\bar{\mathbf{V}}^\top\hat{\mathbf{V}}\hat{\mathbf{S}}-\bar{\mathbf{S}}\mathbf{W})\\ &\quad+(\mathbf{I}-\bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top(\hat{\mathbf{U}}-\bar{\mathbf{U}}\mathbf{W})\\ &\quad+(\mathbf{I}-\bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}})^\top\bar{\mathbf{U}}\mathbf{W}\\ &\quad+(\mathbf{I}-\bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}})^\top(\hat{\mathbf{U}}-\bar{\mathbf{U}}\mathbf{W}).\end{aligned}$$

The first and third terms are bounded using classical concentration-of-measure and matrix perturbation tools and the second term is bounded using Assumption 2. However, the final term requires a more delicate treatment.

Part of the challenge of obtaining a good bound on this term is that $(\tilde{\mathbf{A}}-\tilde{\mathbf{\Lambda}})$ and $(\hat{\mathbf{U}}-\bar{\mathbf{U}}\mathbf{W})$ are dependent, and this dependence must be decoupled in order to apply the standard suite of matrix perturbation tools. For this, we employ the *leave-one-out* proof technique pioneered in Bean et al. [112], Javanmard and Montanari [113], Zhong and Boumal [114] and Abbe et al. [93]. We construct the auxiliary matrices $\hat{\boldsymbol{\Lambda}}^{(1)}, \dots, \hat{\boldsymbol{\Lambda}}^{(n)}$ where $\hat{\boldsymbol{\Lambda}}^{(m)}$ is the matrix obtained by replacing the row and columns of $\hat{\boldsymbol{\Lambda}}$ corresponding the mth node with its expectation. In this way, the mth row of $(\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}})$ and $\hat{\boldsymbol{\Lambda}}^{(m)}$ are independent.

Letting $\hat{\mathbf{U}}^{(m)}$ denote the matrix of leading left singular values of $\hat{\boldsymbol{\Lambda}}^{(m)}$, we then decompose the Euclidean norm of $(\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}} - \mathbf{U}\mathbf{W})$ as

$$\begin{aligned}\left\|(\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W})\right\|_2 \leq & \left\|(\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}})^\top_{\cdot,m}\hat{\mathbf{U}}(\mathbf{W} - \hat{\mathbf{U}}^\top\bar{\mathbf{U}})\right\|_2 \\ & + \left\|(\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\bar{\mathbf{U}} - \hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top\bar{\mathbf{U}})\right\|_2 \\ & + \left\|(\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top\bar{\mathbf{U}} - \bar{\mathbf{U}})\right\|_2 .\end{aligned}$$

To bound these terms, we require good bounds on

$$\left\|\hat{\mathbf{U}}\right\|_{2,\infty}, \quad \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} \quad \text{and} \quad \left\|\hat{\mathbf{U}}^{(m)}\mathbf{W}^{(m)} - \mathbf{U}\right\|_{2,\infty}$$

which we obtain directly from a result due to Abbe et al. [93], and good bounds on

$$\left\|\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right\|_2 \quad \text{and} \quad \left\|\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\right\|_2$$

which are obtained by an application of Wedin's inequality, combined with a careful entry-by-entry analysis of $\|(\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}})\hat{\mathbf{U}}^{(m)}\|_{\mathrm{F}}$ and $\|(\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}})\hat{\mathbf{V}}^{(m)}\|_{\mathrm{F}}$. For a more detailed account of the *leave-one-out* proof technique, we refer the reader to Abbe et al. [93].

5.3 Structure preservation and temporal coherence

For many practical inference tasks, it is desirable for a representation learning procedure to possess the following two properties:

- **Structure preserving.** If two nodes exhibit statistically indistinguishable behaviour at a given time, then their representations at that time are similar. That is, if $\Lambda_i(t) = \Lambda_j(t)$, then $\hat{X}_i(t) \approx \hat{X}_j(t)$;

- **Temporally coherent.** If a node exhibits statistically indistinguishable behaviour at two distinct points in time, then its representations at both these times are similar. That is, if $\Lambda_i(s) = \Lambda_i(t)$, then $\hat{X}_i(s) \approx \hat{X}_i(t)$.

It has been observed in a recent survey of [90] that almost all existing dynamic network embedding procedures possess only one of these properties, but not both.

In the following lemma, we formally define $\hat{X}_i(s) \approx \hat{X}_j(t)$ to mean that $X_i(s) = X_j(t)$, referring to equality "up to statistical noise" in the sense of Theorem 5.1.

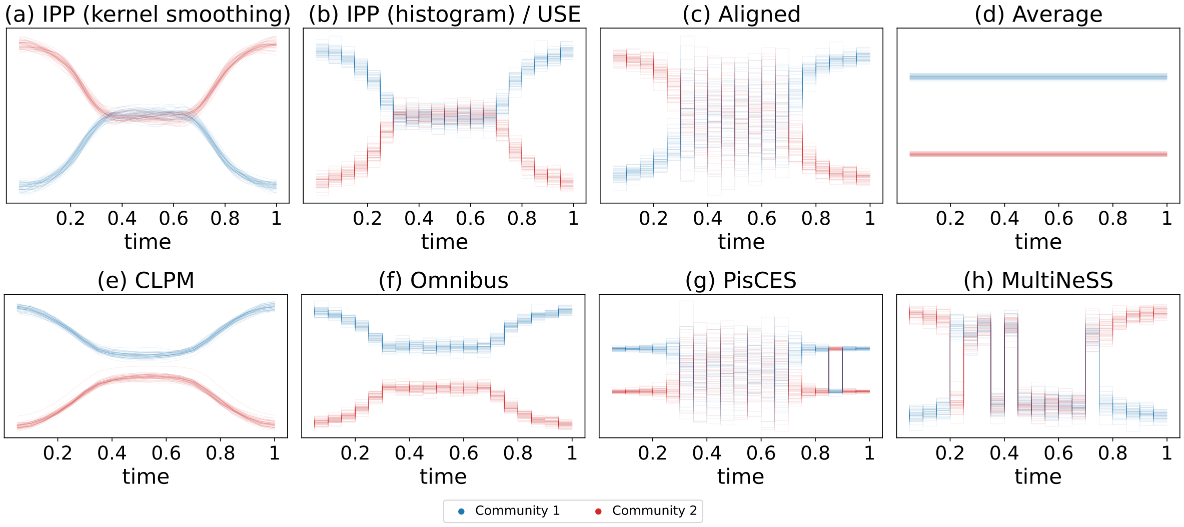


Figure 5.2: One-dimensional PCA visualisation of the two-dimensional node representations obtained using a collection of methods for a network simulated from a bifurcating block model. Colours correspond to the community memberships of the nodes.

Lemma 5.2. *Intensity Profile Projection is both structure preserving and temporally coherent.*

This follows from the simple observation that $X_i(t)$ is a fixed function of $\Lambda_i(t)$ for all i and t. To the best of our knowledge, Intensity Profile Projection is the only existing continuous-time procedure which satisfies these desiderata.

5.3.1 Simulated example: a bifurcating block model

To illustrate these properties, we simulate a two-community dynamic stochastic block model (i.e. where $\mathbf{\Lambda}(t)$ is block structured) in which the intra-community intensities and inter-community intensities are initially distinct, they then gradually merge, remain indistinguishable for some time, and finally diverge. We refer to this model as a *bifurcating block model* and provide full details of the simulation in the supplementary materials.

We apply Intensity Profile Projection to the simulated network, using both a histogram intensity estimator, and a kernel smoother, to produce two-dimensional representations. For visualisation, we reduce the dimension from two to one using a dynamic adaptation of principal component analysis (see Section D.3 of the appendix), and the resulting representations are shown in Figures 5.2a and b.

In both cases, the estimated trajectories mirror the underlying dynamics of the network: the two communities are well separated to begin with, gradually merge, and remain relatively constant before returning to the positions in which they started.

We now illustrate the potential pitfalls of some more naive approaches for embedding dynamic networks. We find that most existing methodology can be viewed as some combination of the following two techniques:

- **Alignment.** Obtain a sequence of static snapshots of the network, embed each of the network snapshots separately and subsequently align the embedding from window $t+1$ with the embedding from window t using orthogonal Procrustes alignment [115].

- **Averaging.** Obtain a static summary of the network by averaging it over time, and embed this to obtain constant node representations.

Alignment preserves the structure of the network at each point in time, however can fail to be temporally coherent. Averaging is temporally coherent but can fail to preserve the structure of the network. To illustrate this point, we apply both approaches, using adjacency spectral embedding into two dimensions, to a network simulated from the bifurcating block model. Figures 5.2c and d show visualisations of the trajectories obtained from each approach.

5.3.2 Method comparison

In this section, we demonstrate how our procedure compares to some existing methods on the simulated data described above. Due to the limited number of continuous-time methods, we include a number of discrete-time methods (Omnibus, PisCSE and MultiNeSS) which we give as input a discrete sequence of snapshots $\mathbf{A}(1), \ldots, \mathbf{A}(M)$ of our simulated continuous-time networks. We compare the following methods:

- **IPP (kernel smoothing)**. Algorithm 2 applied with intensities estimated using kernel smoothing.

- **IPP (histogram) / USE** [74]. Algorithm 2 applied with intensities estimated using a histogram estimator. Equivalent to a weighted extension of the Unfolded Spectral Embedding algorithm of [74].

- **CLPM** [95]. Fits a continuous latent position model $\log \lambda_{ij}(t) = \beta - \|Z_i(t) - Z_j(t)\|^2$ with a penalty on large velocities in the latent space.

- **Omnibus** [101]. Approximately factorises the matrix $\mathbf{A}$ with blocks $\mathbf{A}[k,l] = \frac{1}{2}(\mathbf{A}(k) + \mathbf{A}(l))$, using a spectral decomposition.

- **PisCES** [102]. Minimises the objective function

$$\sum_{k=1}^{M} \|\mathbf{L}(k) - \mathbf{L}^\star(k)\|_F^2 + \alpha \sum_{k=1}^{M-1} \|\mathbf{L}^\star(k) - \mathbf{L}^\star(k+1)\|_{\mathrm{F}}^2 ,$$

  for $\mathbf{L}^\star(1), ..., \mathbf{L}^\star(M)$, where $\alpha \in [0,1]$ and $\mathbf{L}(k)$ are the Laplacian normalisations of $\mathbf{A}(k)$. Then, approximately factorises each $\mathbf{L}^\star(1), ..., \mathbf{L}^\star(M)$ using spectral decompositions.

- **MultiNeSS** [116]. Fits a latent position model $\mathbf{A}_{ij}(k) \sim Q\{\cdot; f(Z_i(k), Z_j(k)), \phi\}$, where $Q(\cdot; \theta, \phi)$ is a parametric distribution.

We use an embedding dimension of $d = 2$ for all methods, and for visualisation, we reduce this to one using PCA. Additional details such as hyperparameter selection, where applicable, are given in the Section D.1 of the appendix.

The CLPM and Omnibus methods produce representations which are temporally coherent, however both fail to capture the complete merging of the communities, shown by Figures 5.2e and f, and are therefore not structure preserving. The PisCES and MultiNeSS methods produce representations which are structure preserving, however, both are unstable when the communities are indistinguishable, shown by Figures 5.2g and h, and are therefore not temporally coherent.

5.4 Illustration with face-to-face interaction data

We demonstrate Intensity Profile Projection on a dataset containing the face-to-face interactions of the pupils of a primary school in Lyon over two days in October 2009 [117]. During the study, discreet radio-frequency identification devices were worn by 232 pupils and 10 teachers which recorded their face-to-face interactions. When two participants were in close proximity over an interval of 20 seconds, the timestamped interaction event was recorded. The school contains five year groups, each divided into two classes, and each class has an assigned room and an assigned teacher. The school day runs from 8:30am to 4:30pm, with a lunch break from 12:00pm to 2:00pm, and no data was gathered on contacts taking place outside the school or during sports activities. For more details about the study and dataset, we refer the reader to [117].

We apply Intensity Profile Projection to the data corresponding to each day of the study using a kernel smoother with an Epanechnikov kernel, choosing a bandwidth of 5 minutes and computing 30-dimensional trajectories.

To visualise the node trajectories, we first rescale them to have unit norm, which has the effect of removing information about the "activeness" of a node from its representation (see, for example, [66]), and apply two dimension reduction techniques. The first is principal component analysis (PCA), which we adapt to our dynamic setting by projecting the (centered) representations onto the direction of maximum average variance over the time domain. This visualisation gives us a temporally coherent view of the trajectories (more details are given in Section D.3 of the appendix). In Figure 5.3, we visualise the trajectories of each pair of classes in each year group using PCA, and for clarity, we just plot the average trajectory for each class, along with one standard deviation above and below.

The second is t-distributed Stochastic Neighbor Embedding (t-SNE), a popular non-linear dimension-reduction tool which provides enough flexibility to visualise the whole set of represen-

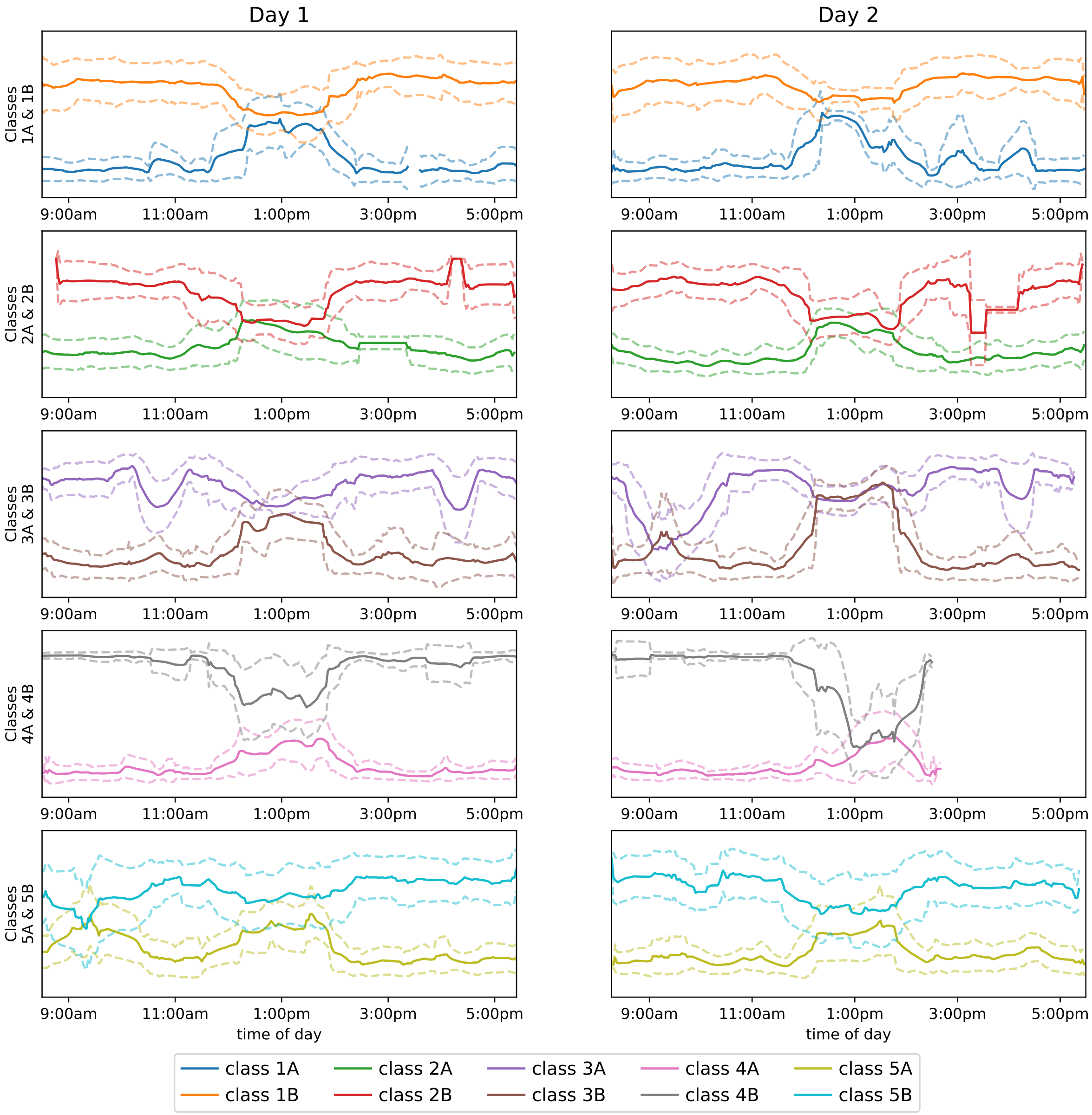


Figure 5.3: One-dimensional PCA visualisation of the 30-dimensional node representations for pairs of classes in the same year group. The solid lines show the average trajectory for each class, and the dashed line shows one standard deviation above and below.

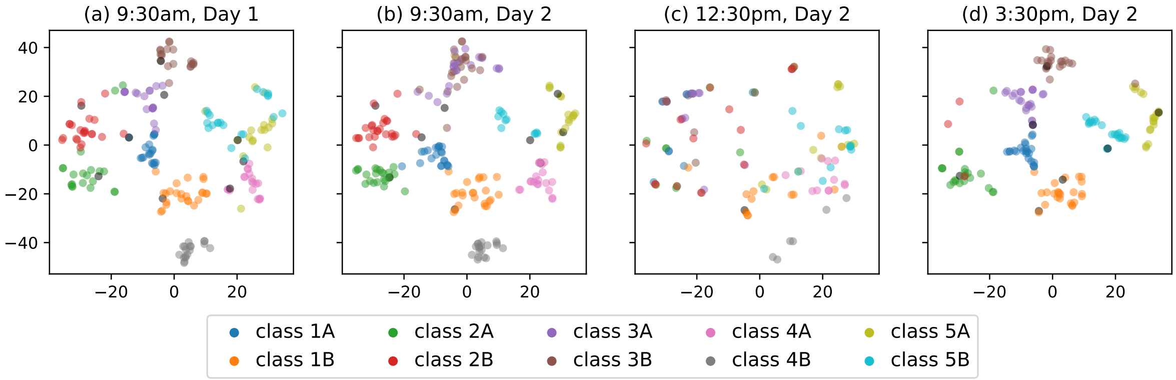


Figure 5.4: Two-dimensional t-SNE visualisation of the 30-dimensional node representations of all pupils and teachers evaluated at 9:30am on Day 1, and 9:30am, 12:30pm and 3:30pm on Day 2.

tations at each point in time. Figure 5.4 shows t-SNE visualisations of the node representations at a collection of times throughout the study.

Figure 5.3 clearly shows the mixing of classes during the lunch hours, and from Figures 5.4, we see that the representations are much more fragmented during the lunch hour (12:30pm, Day 2) than they are during lessons at the other times, where they form tighter clusters corresponding to classes.

While it is reassuring that the geometry of the trajectories reflects the *known* class and timetable structures of the school, it also allows us to uncover structure in the data that *was not known* from the report on the study. For example, classes 5A and 5B (olive and cyan, respectively) merge into a single cluster at approximately 9:30am on Day 1, and classes 3A and 3B (brown and pink, respectively) do the same at approximately 9:30am on Day 2. One might conjecture that this corresponds to a joint lesson, which is taken by the students of both classes in a year group.

5.5 Discussion

We have presented an algorithmic framework to learn continuous-time, low-dimensional trajectories representing the evolving behaviours of nodes in a dynamic network.

A limitation of our framework is the need for bandwidth and dimension selection. These decisions are difficult because they are trade-offs, bias versus variance in the case of bandwidth selection (as seen here), and statistical versus computational in the case of dimension selection (see e.g. [118]). In the presence of a specific supervised downstream task, both decisions could be assisted by cross-validation. In unsupervised settings with reasonably-sized networks, our method is very fast, allowing expedient exploration of different choices.

Our method might be viewed as a dynamic analogue of adjacency spectral embedding for static graphs [6] and, as a result, in future research it could be profitable to find dynamic analogues of other variants of spectral embedding, e.g. applying Laplacian normalisation [19, 51, 119] or regularisation [66, 68].

We view our framework as providing a platform on which novel inference procedures can be developed, particularly combining graph and temporal concepts. For example, in dynamic networks with continuously evolving community structure, it might be interesting to develop procedures for detecting branching points (see bifurcating block model example, Section 5.3), or measures of polarisation and cohesion in the network via the velocities of the trajectories. More generally, we believe there is much left to understand and exploit in the time-evolving topology and geometry of these representations.

Appendix A

Appendix to Chapter 2

A.1 Proof of Lemma 2.1

We begin by defining the matrices

$$\mathbf{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}, \qquad \mathbf{X}' = \begin{pmatrix} (X_1')^\top \\ \vdots \\ (X_n')^\top \end{pmatrix}$$

so that $\mathbf{P} = \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top = \mathbf{X}'\mathbf{I}_{p,q}(\mathbf{X}')^\top$. In addition, let $\mathbf{P} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ be an eigendecomposition of $\mathbf{P}$ and define $\mathbf{X}_\star = \mathbf{U}|\boldsymbol{\Lambda}|^{1/2}$ so that $\mathbf{P} = \mathbf{X}_\star\mathbf{I}_{p,q}\mathbf{X}_\star^\top$, and let $\mathbf{Q}_\star \in \mathbb{O}(p,q)$ be the indefinite orthogonal matrix such that $\mathbf{X} = \mathbf{X}_\star\mathbf{Q}_\star$.

Firstly observe that

$$\mathbf{X}^\top\mathbf{X} = \mathbf{Q}_\star\mathbf{X}_\star^\top\mathbf{X}_\star\mathbf{Q}_\star^\top = \mathbf{Q}_\star|\boldsymbol{\Lambda}|\mathbf{Q}_\star^\top$$

and since for any conformable matrices $\mathbf{M}_1, \mathbf{M}_2$, the matrix products $\mathbf{M}_1\mathbf{M}_2$ and $\mathbf{M}_2\mathbf{M}_1$ share the same non-zero singular values, we have that

$$\sigma_1\left(\mathbf{Q}_\star|\boldsymbol{\Lambda}|\mathbf{Q}_\star^\top\right) = \sigma_1\left(\mathbf{X}^\top\mathbf{X}\right) = \sigma_1\left(\mathbf{X}\mathbf{X}^\top\right) = \sigma_1\left(\mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top\right) = \sigma_1\left(\mathbf{P}\right).$$

Since $\mathbf{Q}_\star|\boldsymbol{\Lambda}|\mathbf{Q}_\star^\top$ is positive definite, by an elementary min-max argument,

$$\sigma_d\left(\mathbf{P}\right)\|\mathbf{Q}_\star\|_2^2 = \sigma_d\left(|\boldsymbol{\Lambda}|\right)\sigma_1^2\left(\mathbf{Q}_\star\right) \le \sigma_1\left(\mathbf{Q}_\star|\boldsymbol{\Lambda}|\mathbf{Q}_\star^\top\right) = \sigma_1\left(\mathbf{P}\right),$$

and therefore

$$\|\mathbf{Q}_\star\|_2 \le \sqrt{\frac{\sigma_1\left(\mathbf{P}\right)}{\sigma_d\left(\mathbf{P}\right)}} = \sqrt{\kappa}.$$

Observe that $\mathbf{Q}_\star^{-1} = \mathbf{I}_{p,q}\mathbf{Q}_\star\mathbf{I}_{p,q}$ and $\|\mathbf{I}_{p,q}\|_2 = 1$, and therefore by the triangle inequality

$$\left\|\mathbf{Q}_\star^{-1}\right\|_2 = \|\mathbf{I}_{p,q}\mathbf{Q}_\star\mathbf{I}_{p,q}\|_2 \le \|\mathbf{Q}_\star\|_2 \le \sqrt{\kappa}.$$

Next, let $\mathbf{Q}'_\star$ be the indefinite orthogonal matrix such that $\mathbf{X}' = \mathbf{X}_\star \mathbf{Q}'_\star$. Then by a similar argument, $\|\mathbf{Q}'_\star\|_2, \|(\mathbf{Q}'_\star)^{-1}\|_2 \leq \sqrt{\kappa}$. Then

$$\mathbf{X} = \mathbf{X}_\star \mathbf{Q}_\star = \mathbf{X}' \left(\mathbf{Q}'_\star\right)^{-1} \mathbf{Q}_\star$$

and therefore, since $\mathbf{X}$ has rank r, $\mathbf{Q} = (\mathbf{Q}'_\star)^{-1}\mathbf{Q}_\star$. By the triangle inequality,

$$\|\mathbf{Q}\|_2 \leq \|(\mathbf{Q}'_\star)^{-1}\|_2 \|\mathbf{Q}_\star\|_2 \leq \kappa,$$

which concludes the proof.

A.2 Proof of Lemma 2.2

It is immediate from the assumption that $\langle x, x' \rangle \in [0,1]$ for all $x, x' \in \mathcal{X}$ that

$$t_i = \rho_n \sum_{i=1}^{n} \langle \xi_i, \xi_j \rangle_{p,q} \leq n\rho_n$$

for all $i = 1, \ldots, n$. To show the reverse inequality, let $C_1 \in (0,1)$ be the constant such that $\langle x, \mu \rangle_{p,q} \geq C_1$ for all $x \in \mathcal{X}$. Then, conditional on $\xi_i = x$, $t_i = \rho_n \sum_{j \neq i} \langle x, \xi_j \rangle$ is the sum of independent and identically distributed, bounded random variables. Therefore, by Hoeffding's inequality, for all $\lambda > 0$,

$$\mathbb{P}\left(t_i < \mathbb{E}t_i - \lambda \mid \xi_i = x\right) \leq \exp\left(-\frac{2\lambda^2}{(n-1)\rho_n^2}\right).$$

Setting $\lambda = (c/2)(n-1)^{1/2}\rho_n \log^{1/2} n$ for any $c > 0$, and recalling that by assumption $\mathbb{E}(t_i \mid \xi_i = x) = (n-1)\rho_n \langle x, \mu \rangle_{p,q} \geq C_1(n-1)\rho_n$, we have that

$$\mathbb{P}\Big(t_i < C_1(n-1)\rho_n + \frac{c}{2}(n-1)^{1/2}\rho_n \log^{1/2} n \mid \xi_i = x\Big) \leq n^{-c} \tag{A.1}$$

For sufficiently large n, the first term on the right hand side of the inequality dominates and there exists a constant $C_2 \in (0,1)$, which depends on c, such that

$$\mathbb{P}\left(t_i < C_2 n\rho_n \mid \xi_i = x\right) \leq n^{-c} \tag{A.2}$$

Then, observing that (A.2) holds for and $x \in \mathcal{X}$, any employing a union bound, we have that for sufficiently large n,

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i=1}^{n} \{t_i \geq C_2 n\rho_n\}\right) &= 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} \{t_i < C_2 n\rho_n\}\right) \\
&\geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(t_i < C_2 n\rho_n\right) \\
&\geq 1 - \sum_{i=1}^{n} \sup_{x \in \mathcal{X}} \mathbb{P}\left(t_i < C_2 n\rho_n \mid \xi_i = x\right) \\
&\geq 1 - n^{-c'}
\end{aligned}$$

where $c' = c - 1$, which concludes the proof.

Appendix B

Appendix to Chapter 3

B.1 Expectation-maximisation algorithm

Given a set of data points $x_1, \ldots, x_n$ and a set of associated weights $w_1, \ldots, w_n$, we will optimise the likelihood (3.4) with the following expectation-maximisation algorithm.

First, we apply k-means to obtain an initial clustering, and initialise $\eta_{ik}^{(0)} = 1$ if x_i is assigned to the kth cluster, and zero otherwise. We then initialise $\Theta^{(0)} := \{\pi_k^{(0)}, \mu_k^{(0)}, \mathbf{\Sigma}_k^{(0)}\}_{k=1}^K$ using the M-step. Then, we alternate between the following steps for $r = 1, 2, \ldots,$ until convergence:
E-step: for $i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}$,

$$\eta_{ik}^{(r+1)} \leftarrow \frac{\pi_k^{(r)} \mathcal{N}(x_i; \mu_k^{(r)}, w_i^{-1} \mathbf{\Sigma}_k^{(r)})}{\sum_{\ell=1}^K \pi_\ell^{(r)} \mathcal{N}(x_i; \mu_\ell^{(r)}, w_i^{-1} \mathbf{\Sigma}_\ell^{(r)})}.$$

M-step: for $k \in \{1, \ldots, K\}$,

$$\begin{aligned}
\pi_k^{(r+1)} &\leftarrow \frac{\sum_i \eta_{ik}^{(r+1)}}{n}; \\
\mu_k^{(r+1)} &\leftarrow \frac{\sum_{i=1}^n w_i \cdot \eta_{ik}^{(r+1)} x_i}{\sum_i w_i \cdot \eta_{ik}^{(r+1)}}; \\
\mathbf{\Sigma}_k^{(r+1)} &\leftarrow \frac{\sum_{i=1}^n w_i \cdot \eta_{ik}^{(r+1)} (x_i - \mu_k^{(r+1)})(x_i - \mu_k^{(r+1)})^\top}{\sum_i \eta_{ik}^{(r+1)}}.
\end{aligned}$$

On convergence, we return the maximum a-posteriori membership for each node, $\hat{z}_i = \text{argmax}_k \eta_{ik}$.

B.2 Proofs of Theorems 3.1 and 3.2

The proofs of Theorems 3.1 and 3.2 make use of results derived in Rubin-Delanchy et al. [15] and Tang and Priebe [19]. Where the techniques employed here are straightforward adjustments of those developed in those papers, we refer the reader to the relevant derivations and omit the details. In this proof, we use the notation $a_n \overset{\mathbb{P}}{\lesssim} b_n$ as shorthand for the statement: *"for sufficiently large n, $a_n \lesssim b_n$ with overwhelming probability"*.

Recall that the symmetric Laplacian and random walk Laplacian matrices of $\mathbf{A}$ are defined as

$$\mathbf{L}_{\text{sym}} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \qquad \mathbf{L}_{\text{rw}} := \mathbf{D}^{-1}\mathbf{A}$$

where $\mathbf{D} \in \mathbb{R}^{d\times d}$ is the diagonal degree matrix with entries $d_i = \sum_j a_{ij}$. Suppose that $\breve{\mathbf{S}}$ is the diagonal matrix containing the d largest-in-magnitude eigenvalues of $\mathbf{L}_{\text{sym}}$ in descending order and $\breve{\mathbf{U}}$ is the matrix containing corresponding orthonormal eigenvectors as columns. Then the symmetric Laplacian spectral embedding (Definition 2.3) is given by the rows of $\breve{\mathbf{X}} := \breve{\mathbf{U}}\breve{\mathbf{S}}^{1/2}$. Recall that both Laplacians share the same eigenvalues, and if u is an eigenvector of $\mathbf{L}_{\text{sym}}$, then $\mathbf{D}^{-1/2}u$ is a right eigenvector of $\mathbf{L}_{\text{rw}}$. We use this spectral relationship to construct a canonical set of eigenvectors for $\mathbf{L}_{\text{rw}}$, namely $|\lambda_1|^{1/2}\mathbf{D}^{-1/2}\breve{u}_1, \dots, |\lambda_r|^{1/2}\mathbf{D}^{-1/2}\breve{u}_r$ which gives a *canonical* random walk Laplacian spectral embedding, which we denote by the rows of $\hat{\mathbf{X}}^{\star} := \mathbf{D}^{-1/2}\breve{\mathbf{U}}\breve{\mathbf{S}}^{1/2}$. Then, we define the invertible linear transformation $\mathbf{Q}_{\hat{\mathbf{X}}}$ satisfying $\hat{\mathbf{X}}\mathbf{Q}_{\hat{\mathbf{X}}} = \breve{\mathbf{X}}$.

Additionally, we let $\bar{\mathbf{X}} = \mathbf{T}^{-1/2}\mathbf{X}$ where $\mathbf{T} \in \mathbb{R}^{n\times n}$ is the diagonal expected degree matrix with entries $t_i = \sum_j p_{ij}$ and let $\tilde{\mathbf{X}} := (\tilde{X}_1, \dots, \tilde{X}_n)^\top \equiv \mathbf{T}^{-1}\mathbf{X}$ (see Section 3.2.1).

By Theorem 7 of Solanki et al. [120], $\mathcal{X}$, the support of F, is a bounded set, and by Lemma 2.2, $t_i \overset{\mathbb{P}}{\asymp} n\rho_n$ for all $i \in [n]$. It therefore follows that $\|\mathbf{X}\|_{2,\infty} \asymp \rho_n^{1/2}$ and $\left\|\bar{\mathbf{X}}\right\|_{2,\infty} \asymp n^{-1/2}$.

A Chernoff bound gives that $|t_i - d_i| \overset{\mathbb{P}}{\lesssim} (n\rho_n)^{1/2}\log n$ and a union bound gives that

$$\|\mathbf{T} - \mathbf{D}\|_\infty \overset{\mathbb{P}}{\lesssim} (n\rho_n)^{1/2}\log n. \tag{B.1}$$

In addition, Lemma 3.1 of [19] states that $\mathbf{D}^{-1/2} - \mathbf{T}^{-1/2}$ admits the decomposition

$$\mathbf{D}^{-1/2} - \mathbf{T}^{-1/2} = \tfrac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T} - \mathbf{D}) + \mathbf{R}_1 \tag{B.2}$$

where $\mathbf{R}_1$ is a diagonal matrix satisfying $\|\mathbf{R}_1\|_\infty \overset{\mathbb{P}}{\lesssim} (n\rho_n)^{-3/2}\log n$.

Theorem 3 from [15], reproduced as Theorem 2.5 in Chapter 1 states that there exists a universal constant $c \geq 0$ and an indefinite-orthogonal matrix $\mathbf{W} \in \mathbb{O}(p,q)$ such that the symmetric Laplacian spectral embedding satisfies

$$\left\|\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}}\right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \frac{\log^c n}{n\rho_n^{1/2}}. \tag{B.3}$$

Recall that $\hat{\mathbf{X}}\mathbf{Q}_{\hat{\mathbf{X}}} = \mathbf{D}^{-1/2}\breve{\mathbf{X}}$ and $\tilde{\mathbf{X}} = \mathbf{T}^{-1/2}\bar{\mathbf{X}}$, and define $\mathbf{Q} = \mathbf{W}\mathbf{Q}_{\hat{\mathbf{X}}}^\top$ we use (B.2) to obtain

$$\begin{aligned}
\hat{\mathbf{X}}\mathbf{Q}^\top - \tilde{\mathbf{X}} &= \mathbf{D}^{-1/2}\breve{\mathbf{X}}\mathbf{W}^\top - \mathbf{T}^{-1/2}\bar{\mathbf{X}} \\
&= (\mathbf{T}^{-1/2} + \tfrac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T} - \mathbf{D}) + \mathbf{R}_1)\breve{\mathbf{X}}\mathbf{W}^\top - \mathbf{T}^{-1/2}\bar{\mathbf{X}} \\
&= \mathbf{T}^{-1/2}(\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}}) + (\tfrac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T} - \mathbf{D}) + \mathbf{R}_1)\breve{\mathbf{X}}\mathbf{W}^\top \\
&= \mathbf{T}^{-1/2}(\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}}) + \tfrac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T} - \mathbf{D})\bar{\mathbf{X}} + \mathbf{R}_2
\end{aligned} \tag{B.4}$$

where $\mathbf{R}_2 = \mathbf{R}_1\bar{\mathbf{X}} + (\frac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T}-\mathbf{D}) + \mathbf{R}_1)(\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}})$. The equations (B.1)–(B.3) give that

$$\begin{aligned}\|\mathbf{R}_2\|_{2,\infty} &\le \|\mathbf{R}_1\|_\infty \left\|\bar{\mathbf{X}}\right\|_{2,\infty} + \left(\tfrac{1}{2}\left\|\mathbf{T}^{-3/2}\right\|_\infty \|\mathbf{T}-\mathbf{D}\|_\infty + \|\mathbf{R}_1\|_\infty\right)\left\|\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}}\right\|_{2,\infty} \\ &\overset{\mathbb{P}}{\lesssim} \frac{\log^c n}{n^2\rho_n^{3/2}},\end{aligned} \tag{B.5}$$

and therefore

$$\begin{aligned}\left\|\hat{\mathbf{X}}\mathbf{Q}^\top - \tilde{\mathbf{X}}\right\|_{2,\infty} &\le \left\|\mathbf{T}^{-1/2}\right\|_\infty \left\|\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}}\right\|_{2,\infty} + \tfrac{1}{2}\left\|\mathbf{T}^{-3/2}\right\|_\infty \|\mathbf{T}-\mathbf{D}\|_\infty \left\|\bar{\mathbf{X}}\right\|_{2,\infty} + \|\mathbf{R}_2\|_{2,\infty} \\ &=\overset{\mathbb{P}}{\lesssim} \frac{\log^c n}{n^{3/2}\rho_n},\end{aligned}$$

establishing Theorem 3.1. To establish Theorem 3.2, we first state an important decomposition derived in [19] for the symmetric Laplacian spectral embedding. We state the decomposition with a minor modification to accommodate both positive and negative leading eigenvalues, where only positive leading eigenvalues are considered in [19] (see [15]). We have

$$\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}} = \mathbf{T}^{-1/2}(\mathbf{A}-\mathbf{P})\mathbf{T}^{-1/2}\bar{\mathbf{X}}(\bar{\mathbf{X}}^\top\bar{\mathbf{X}})^{-1}\mathbf{I}_{p,q} + \tfrac{1}{2}\mathbf{T}^{-1}(\mathbf{T}-\mathbf{D})\bar{\mathbf{X}} + \mathbf{R}_3 \tag{B.6}$$

where each row of $\mathbf{R}_3$, which we denote $\{r_i^{(3)}\}_{i=1}^n$, is such that $n\rho_n^{1/2}r_i^{(3)} \overset{\text{a.s.}}{\to} 0$ for all $i \in [n]$, where $\overset{\text{a.s.}}{\to}$ almost sure convergence as $n \to \infty$. Substituting (B.6) into (B.4) gives

$$\begin{aligned}\hat{\mathbf{X}}\mathbf{Q}^\top - \tilde{\mathbf{X}} &= \mathbf{T}^{-1/2}(\breve{\mathbf{X}}\mathbf{W}^\top - \bar{\mathbf{X}}) + \tfrac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T}-\mathbf{D})\bar{\mathbf{X}} + \mathbf{R}_2 \\ &= \mathbf{T}^{-1/2}\{\mathbf{T}^{-1/2}(\mathbf{A}-\mathbf{P})\mathbf{T}^{-1/2}\bar{\mathbf{X}}(\bar{\mathbf{X}}^\top\bar{\mathbf{X}})^{-1}\mathbf{I}_{p,q} \\ &\qquad\qquad + \tfrac{1}{2}\mathbf{T}^{-1}(\mathbf{T}-\mathbf{D})\bar{\mathbf{X}} + \mathbf{R}_3\} + \tfrac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T}-\mathbf{D})\bar{\mathbf{X}} + \mathbf{R}_2 \\ &= \mathbf{T}^{-1}(\mathbf{A}-\mathbf{P})\mathbf{T}^{-1/2}\bar{\mathbf{X}}(\bar{\mathbf{X}}^\top\bar{\mathbf{X}})^{-1}\mathbf{I}_{p,q} + \mathbf{T}^{-3/2}(\mathbf{T}-\mathbf{D})\bar{\mathbf{X}} + \mathbf{T}^{-1/2}\mathbf{R}_3 + \mathbf{R}_2 \\ &= \mathbf{T}^{-1}(\mathbf{A}-\mathbf{P})\mathbf{T}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{T}^{-1}\mathbf{X})^{-1}\mathbf{I}_{p,q} + \mathbf{T}^{-2}(\mathbf{T}-\mathbf{D})\mathbf{X} + \mathbf{R}\end{aligned}$$

where $\mathbf{R} = \mathbf{T}^{-1/2}\mathbf{R}_3 + \mathbf{R}_2$, the rows of which we denote $\{r_i\}_{i=1}^n$, satisfy $r_i \overset{\text{a.s.}}{\to} 0$ for all $i \in [n]$. We denote by ζ_i, the ith row of $n^{3/2}\rho_n(\hat{\mathbf{X}}\mathbf{Q}^\top - \tilde{\mathbf{X}})$, and from here on, we use r to denote any random vector such that $r \overset{\text{a.s.}}{\to} 0$, which may change from line to line. We have

$$\begin{aligned}\zeta_i &= \mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{T}^{-1}\mathbf{X})^{-1}\frac{n^{3/2}\rho_n}{t_i}\left(\sum_j \frac{a_{ij}-p_{ij}}{t_j}X_j\right) + \frac{n^{3/2}\rho_n}{t_i^2}(t_i - d_i)X_i + r \\ &= \mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{T}^{-1}\mathbf{X})^{-1}\frac{(n\rho_n)^{3/2}}{t_i}\left(\sum_j \frac{a_{ij}-p_{ij}}{t_j}\xi_j\right) + \frac{(n\rho_n)^{3/2}}{t_i^2}\sum_j(a_{ij}-p_{ij})\xi_i + r \\ &= \mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{T}^{-1}\mathbf{X})^{-1}\left(\frac{n\rho_n}{t_i}\right)\left(\sum_j \frac{(n\rho_n)^{1/2}(a_{ij}-p_{ij})}{t_j}\xi_j\right) + \left(\frac{n\rho_n}{t_i}\right)^2\left(\sum_j \frac{(a_{ij}-p_{ij})}{(n\rho_n)^{1/2}}\xi_i\right) + r\end{aligned}$$

Additionally, by an identical argument to that used to obtain Eq. (B.3) and (B.4) in [19],

$$\sum_j \frac{(n\rho_n)^{1/2}(a_{ij}-p_{ij})}{t_j}\xi_j = \sum_j \frac{(a_{ij}-p_{ij})}{(n\rho_n)^{1/2}}\frac{\xi_j}{\langle \xi_j, \mu\rangle_{p,q}} + r$$

and

$$\frac{n\rho_n}{t_i}\sum_j \frac{(a_{ij}-p_{ij})}{(n\rho_n)^{1/2}}\xi_i = \mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{T}^{-1}\mathbf{X})^{-1}\sum_j \frac{(a_{ij}-p_{ij})}{(n\rho_n)^{1/2}}\frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}\xi_i}{\langle\xi_i,\mu\rangle_{p,q}} + r$$

where $\mu = \mathbb{E}_{\xi\sim F}(\xi)$ and $\tilde{\mathbf{\Delta}} = \mathbb{E}_{\xi\sim F}\left(\frac{\xi\xi^\top}{\langle\xi,\mu\rangle_{p,q}}\right)$. Therefore

$$\zeta_i = \frac{n\rho_n}{t_i}\mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{T}^{-1}\mathbf{X})^{-1}\sum_j \frac{(a_{ij}-p_{ij})}{(n\rho_n)^{1/2}}\left(\frac{\xi_j}{\langle\xi_j,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}\xi_i}{\langle\xi_i,\mu\rangle_{p,q}}\right) + r,$$

and for each fixed index $i \in [n]$, conditional on $\xi_i = x_i$,

$$\sum_j \frac{(a_{ij}-p_{ij})}{(n\rho_n)^{1/2}}\left(\frac{\xi_j}{\langle\xi_j,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}\xi_i}{\langle\xi_i,\mu\rangle_{p,q}}\right) \tag{B.7}$$

is $n^{-1/2}$ times the sum of independent and identically-distributed, mean-zero random vectors, ignoring the vanishing contribution of the ith vector.

Therefore, by the multivariate central limit theorem, the sum (B.7) converges in distribution to a multivariate Gaussian random variable with mean zero and the covariance $\mathbf{\Gamma}(x_i)$, where

$$\mathbf{\Gamma}_\rho(x) = \mathbb{E}_{\xi\sim F}\left\{(x^\top\mathbf{I}_{p,q}\xi)(1-\rho_n x^\top\mathbf{I}_{p,q}\xi)\left(\frac{\xi}{\langle\xi,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}x}{\langle x,\mu\rangle_{p,q}}\right)\left(\frac{\xi}{\langle\xi,\mu\rangle_{p,q}} - \frac{\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}x}{\langle x,\mu\rangle_{p,q}}\right)^\top\right\}.$$

where the expectation is taken with respect to $\xi \sim F$.

In [19], it is shown that $(\mathbf{X}^\top\mathbf{T}^{-1}\mathbf{X})^{-1} \overset{\text{a.s.}}{\to} \tilde{\mathbf{\Delta}}^{-1}$ and that $t_i/(n\rho_n) \overset{\text{a.s.}}{\to} \langle\xi_i,\mu\rangle_{p,q}$. By the continuous mapping theorem, $n\rho_n/t_i \overset{\text{a.s.}}{\to} \langle\xi_i,\mu\rangle_{p,q}^{-1}$, and so, by an application of Slutsky's theorem, the vector $\zeta_i = n^{3/2}\rho_n(\mathbf{Q}\hat{X}_i - \tilde{X}_i)$ converges in distribution to a multivariate Gaussian random variable with mean zero and the covariance $\mathbf{\Sigma}(x_i)$, where

$$\mathbf{\Sigma}(x) := \frac{\mathbf{I}_{p,q}\tilde{\mathbf{\Delta}}\mathbf{\Gamma}_\rho(x)\tilde{\mathbf{\Delta}}\mathbf{I}_{p,q}}{\langle x,\mu\rangle_{p,q}^2}.$$

Invoking the Cramér-Wold device establishes Theorem 3.2.

Appendix C

Appendix to Chapter 4

C.1 Proof of Lemma 4.2

Let $\mathcal{V}$ be a totally isotropic subspace with respect to the indefinite inner product $\langle x, y\rangle_{p,q} = x^\top \mathbf{I}_{p,q} y$, so that $\langle x, y\rangle_{p,q} = 0$ for any $x, y \in \mathcal{V}$. Let $\mathcal{W}$ be an arbitrary subspace of dimension $\max\{p, q\}$ which is either positive or negative definite, so that $\langle x, y\rangle_{p,q} = 0$ implies $x = y = 0$, for any $x, y \in \mathcal{W}$. Then $\mathcal{V} \cap \mathcal{W} = \{0\}$, so $\dim(\mathcal{V} + \mathcal{W}) = \dim(\mathcal{V}) + \dim(\mathcal{W})$ and

$$\dim(\mathcal{V}) = \dim(\mathcal{V} + \mathcal{W}) - \dim(\mathcal{W}) \leq p + q - \max\{p, q\} = \min\{p, q\}.$$

Therefore, the maximal dimension of a totally isotropic subspace with respect to the indefinite inner product $\langle \cdot, \cdot\rangle_{p,q}$ is $\min\{p, q\}$.

C.2 Proof of the main theorem

In this proof, we use the notation $a_n \overset{\mathbb{P}}{\lesssim} b_n$ as shorthand for the statement: "for sufficiently large n, $a_n \lesssim b_n$ with overwhelming probability".

The first step of our proof is to define vectors $X_1, \ldots, X_n \in \mathbb{R}^r$ such that $Y_i^\top \mathbf{\Lambda}_{z_i,z_j} Y_j = \langle X_i, X_j\rangle_{p,q}$ for all $i, j \in [n]$, so that $\mathbf{A}$ is described as a generalised random dot product graph, and we can employ existing estimation theory in Rubin-Delanchy et al. [15].

We construct the $(r_1 + \cdots + r_K) \times (r_1 + \cdots + r_K)$ block-matrix $\mathbf{\Lambda}$ whose $k\ell$th block is $\mathbf{\Lambda}_{k\ell}$, and recall the matrices $\mathbf{\Lambda}_k = (\mathbf{\Lambda}_{k1} \cdots \mathbf{\Lambda}_{kK})$, the row concatenation of $\mathbf{\Lambda}_{k1}, \ldots, \mathbf{\Lambda}_{kK}$. Let $r, (p, q)$ denote the rank and signature, respectively, of $\mathbf{\Lambda}$ and let $\mathbf{H}$ be an $r \times (r_1 + \cdots + r_K)$ matrix such that $\mathbf{\Lambda} = \mathbf{H}^\top \mathbf{I}_{p,q} \mathbf{H}$, which can be constructed as $\mathbf{H} = \mathbf{U}_{\mathbf{\Lambda}} |\mathbf{S}_{\mathbf{\Lambda}}|^{1/2}$, where $\mathbf{\Lambda} = \mathbf{U}_{\mathbf{\Lambda}} \mathbf{S}_{\mathbf{\Lambda}} \mathbf{U}_{\mathbf{\Lambda}}^\top$ is the eigendecomposition of $\mathbf{\Lambda}$. In addition, let $\mathbf{H}_k$ denote the kth block of $\mathbf{H}$ containing the rows of $\mathbf{H}$ corresponding to the kth group. Now, by construction, $\mathbf{\Lambda}_{k\ell} = \mathbf{H}_k^\top \mathbf{I}_{p,q} \mathbf{H}_\ell$ and $\mathbf{\Lambda}_k = \mathbf{H}_k^\top \mathbf{I}_{p,q} \mathbf{H}$. Since by assumption, $\mathbf{\Lambda}_k$ has rank d_k, so does $\mathbf{H}_k$. We construct $X_i = \mathbf{H}_{z_i} Y_i$ which satisfies

$$Y_i^\top \mathbf{\Lambda}_{z_i,z_j} Y_j = Y_i^\top \mathbf{H}_{z_i}^\top \mathbf{I}_{p,q} \mathbf{H}_{z_j} Y_j = X_i^\top \mathbf{I}_{p,q} X_j = \langle X_i, X_j\rangle_{p,q},$$

as desired.

Observe that since the transformations $\{\mathbf{H}_k\}_{k=1}^K$ are fixed, $\|X_i\|_2 \asymp \|Y_i\|_2 \asymp \rho_n^{1/2}$ and from the assumptions $\kappa(\mathbf{\Lambda}) \asymp 1$ and $\kappa(\mathbf{\Sigma}_k) \asymp 1$ for all $k \in [K]$, it is straight-forward to derive that $\kappa(\mathbf{\Delta}) := \sigma_1(\mathbf{\Delta})/\sigma_r(\mathbf{\Delta}) \asymp 1$, where $\mathbf{\Delta} := n^{-1}\sum_{i=1}^n X_i X_i^\top$. Therefore the assumptions of Theorem 2.3 are satisfied, and we have that there exists an indefinite orthogonal matrix $\mathbf{Q}^{-1} \in \mathbb{O}(p,q)$ such that

$$\max_{i\in[n]} \left\|\mathbf{Q}^{-1}\hat{X}_i - X_i\right\|_2 \overset{\mathbb{P}}{\lesssim} \sqrt{\frac{\log n}{n}}. \tag{C.1}$$

By Lemma 2.1 we have that $\|\mathbf{Q}\|_2, \left\|\mathbf{Q}^{-1}\right\|_2 \lesssim 1$, and therefore multiplying (C.1) on the left by $\mathbf{Q}$ and applying the triangle inequality we have

$$\max_{i\in[n]} \left\|\hat{X}_i - \mathbf{Q}X_i\right\|_2 \overset{\mathbb{P}}{\lesssim} \sqrt{\frac{\log n}{n}}. \tag{C.2}$$

It therefore follows that

$$\begin{aligned}
\max_{i\in[n]} \left\|\hat{X}_i\right\|_2 &\le \max_{i\in[n]} \left\|\mathbf{Q}X_i + \hat{X}_i - \mathbf{Q}X_i\right\|_2 \\
&\le \|\mathbf{Q}\|_2 \max_{i\in[n]} \|X_i\|_2 + \max_{i\in[n]} \left\|\hat{X}_i - \mathbf{Q}X_i\right\|_2 \\
&\lesssim \rho_n^{1/2} + \sqrt{\frac{\log n}{n}} \\
&\asymp \rho_n^{1/2}.
\end{aligned}$$

where we used that $n\rho_n \gtrsim \log n$. Now recall the matrix $\hat{\mathbf{\Sigma}}_k = n_k^{-1}\sum_{i\in V_k} \hat{X}_i\hat{X}_i^\top$ and that $\hat{\mathbf{\Xi}}_k$ is the matrix whose columns contain the r_k orthonormal eigenvectors of $\hat{\mathbf{\Sigma}}_k$ corresponding to the largest eigenvalues. We define its population counterpart $\mathbf{\Sigma}_k = n_k^{-1}\sum_{i\in V_k} X_i X_i^\top$ and define the matrix $\mathbf{\Xi}_k$ whose columns contain the r_k eigenvectors of $\mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^\top$ corresponding its non-zero eigenvalues. We have that

$$\begin{aligned}
\left\|\hat{\mathbf{\Sigma}} - \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top\right\|_2 &= \left\| n_k^{-1}\sum_{i\in V_k}\left(\hat{X}_i\hat{X}_i^\top - \mathbf{Q}X_iX_i^\top\mathbf{Q}^\top\right)\right\|_2 \\
&= \left\| n_k^{-1}\sum_{i\in V_k}\left\{\hat{X}_i\left(\hat{X}_i - \mathbf{Q}X_i\right)^\top + \left(\hat{X}_i - \mathbf{Q}X_i\right)X_i^\top\mathbf{Q}^\top\right\}\right\|_2 \\
&\le \left(\max_{i\in V_k}\|\hat{X}_i\|_2 + \max_{i\in V_k}\|X_i\|_2\|\mathbf{Q}\|_2\right)\cdot n_k^{-1}\sum_{i\in V_k}\left\|\hat{X}_i - \mathbf{Q}X_i\right\|_2 \\
&\overset{\mathbb{P}}{\lesssim} \sqrt{\frac{\rho_n \log n}{n}}.
\end{aligned}$$

The smallest non-zero eigenvalue of $\mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^\top$, δ_k satisfies $\delta_k \asymp \rho_n$ and by the Davis-Kahan sinΘ theorem, we have that there exists an orthogonal matrix $\mathbf{W}_k \in \mathbb{O}(d_k)$ such that

$$\left\|\hat{\mathbf{\Xi}}_k - \mathbf{\Xi}_k \mathbf{W}_k\right\|_2 \lesssim \delta_k^{-1} \left\|\hat{\mathbf{\Sigma}}_k - \mathbf{Q}\mathbf{\Sigma}_k\mathbf{Q}^\top\right\|_2 \overset{\mathbb{P}}{\lesssim} \sqrt{\frac{\log n}{n\rho_n}}. \tag{C.3}$$

We set $\mathbf{G}_k := \mathbf{W}_k^\top \mathbf{\Xi}_k^\top \mathbf{Q}(\mathbf{H}_k\mathbf{H}_k^\top)^{-1}\mathbf{H}_k$ and then we have

$$\begin{aligned}
\hat{Y}_i - \mathbf{G}_{z_i} Y_i &= \hat{\mathbf{\Xi}}_{z_i}^\top \hat{X}_i - \mathbf{G}_{z_i}\mathbf{H}_{z_i}^\top X_i \\
&= \hat{\mathbf{\Xi}}_{z_i}^\top \hat{X}_i - \mathbf{W}_{z_i}^\top \mathbf{\Xi}_{z_i}^\top \mathbf{Q} X_i \\
&= \hat{\mathbf{\Xi}}_{z_i}^\top \hat{X}_i - \hat{\mathbf{\Xi}}_{z_i}^\top \mathbf{Q} X_i + \hat{\mathbf{\Xi}}_{z_i}^\top \mathbf{Q} X_i - \mathbf{W}_{z_i}^\top \mathbf{\Xi}_{z_i}^\top \mathbf{Q} X_i \\
&= \hat{\mathbf{\Xi}}_{z_i}^\top \left(\hat{X}_i - \mathbf{Q}X_i\right) + \left(\hat{\mathbf{\Xi}}_{z_i} - \mathbf{W}_{z_i}\mathbf{\Xi}_{z_i}\right)^\top \mathbf{Q}X_i
\end{aligned}$$

Therefore it follows from (C.2), (C.3) and the bounded spectral norm of $\mathbf{Q}$, that

$$\begin{aligned}
\max_{i\in[n]} \left\|\hat{Y}_i - \mathbf{G}_{z_i} Y_i\right\|_2 &\leq \max_{i\in[n]} \left\|\hat{X}_i - \mathbf{Q}X_i\right\|_2 + \max_{k\in[K]} \left\|\hat{\mathbf{\Xi}}_k - \mathbf{W}_k\mathbf{\Xi}_k\right\|_2 \|\mathbf{Q}\|_2 \max_{i\in[n]} \|X_i\|_2 \\
&\overset{\mathbb{P}}{\lesssim} \sqrt{\frac{\log n}{n}},
\end{aligned}$$

which establishes the theorem.

Appendix D

Appendix to Chapter 5

D.1 Details of the simulated example and method comparison of Sections 4.1 and 4.2

We simulate data according to the following generative process, which might be viewed as describing as a dynamic, two-community stochastic block model. Assign to each node i a variable $z_i \in \{1, 2\}$ denoting its community (which does not change with time). If nodes i and j are in the same community, i.e., $z_i = z_j$, the point process $\mathcal{E}_{ij}$ follows a homogeneous Poisson process with (fixed) intensity η_0. Otherwise, $\mathcal{E}_{ij}$ follows an inhomogeneous Poisson process with intensity

$$\lambda_{ij}(t) = \begin{cases} \eta_0 \exp\{\eta_1(t - s_1)\} & t < s_1, \\ \eta_0 & s_1 \leq t < s_2, \\ \eta_0 \exp\{-\eta_1(t - s_2)\} & t \geq s_2, \end{cases}$$

where $0 < s_1 < s_2 < T$. This model describes two communities gradually coming together until fully merging by time s_1, before splitting at time s_2 and then gradually drifting apart. We simulate from this model using the parameters $T = 1$, $n = 100$, $z_1, \ldots, z_{50} = 1$, $z_{51}, \ldots, z_{100} = 2$, $\eta_0 = 100$, $\eta_1 = 10$, $s_1 = 0.3$ and $s_2 = 0.7$.

In our method comparison, we used an embedding dimension of $d = 2$ for all methods, unless otherwise stated. For the discrete-time methods, we construct a series of 20 snapshots of the continuous-time network, each a weighted static network whose edge weights are the number of events which occur on the edge in the corresponding time window. The selection of hyperparameters for each method is outlined below:

- **Intensity Profile Projection (histogram)**: We used a bin size of $\frac{1}{M} = \frac{1}{20}$.
- **Intensity Profile Projection (kernel smoothing)**: We used a Epanechnikov kernel with bandwidth 0.1 and applied the approximate Intensity Profile Projection algorithm

with $B = 20$. Different values of bandwidth gave similar results in terms of embedding structure; we chose this bandwidth to achieve the desired smoothness.

- **CLPM** [95]: The dimension is automatically computed by the algorithm as $d = 2$. The hyperparameters are chosen equal to the ones used in "Simulation C" in [95] which is a similar simulated example with two communities. We used 19 changes point which correspond to 20 windows. The implementation was obtained from the Github repository `https://github.com/marcogenni/CLPM`.
- **PisCES** [102]: The dimension is automatically selected by the algorithm as $d = 2$. The smoothing parameter is chosen with cross-validation which results in equivalent log-likelihood values for α from 0.00001 to 0.001. We choose $\alpha = 0.001$ which is the larger value for which the algorithm converges. The implementation was obtained from the Github repository `https://github.com/xuranw/PisCES`.
- **MultiNeSS** [116]: The dimension is automatically selected by the algorithm as $d = 2$ for all windows except windows 5 to 16 for which $d = 1$ is selected. For these windows, we set missing the second dimension to zeros. The implementation is obtained from the `multiness` R package (available on CRAN) and hyperparameters are set to their default values.

D.2 Additional real data analysis

We provide further experiments and details of the Intensity Profile Projection analysis of the Lyon primary school dataset described in Section 5.4. As a comparison to the analysis in the paper, we apply Intensity Profile Projection to the data corresponding to each day of the study with a histogram intensity estimator, choosing a bin size of 10 minutes and computing 30-dimensional trajectories.

Figure D.3 (equivalent to Figure 5.3) visualises the trajectories of each pair of classes in each year group using PCA where we plot the average trajectory for each class, along with one standard deviation above and below. Since every trajectory using the histogram intensity estimator is piece-wise constant, so are the resulting PCA averages. The pairs of trajectories merge and split in a similar way to those obtained using the kernel smoother.

In Figures D.1 and D.2 show the first two spherical coordinates [121] of the trajectories obtained using the histogram intensity estimator and the Epanechnikov kernel smoother, respectively. The six plots correspond to the morning, lunchtime and afternoon across both days.

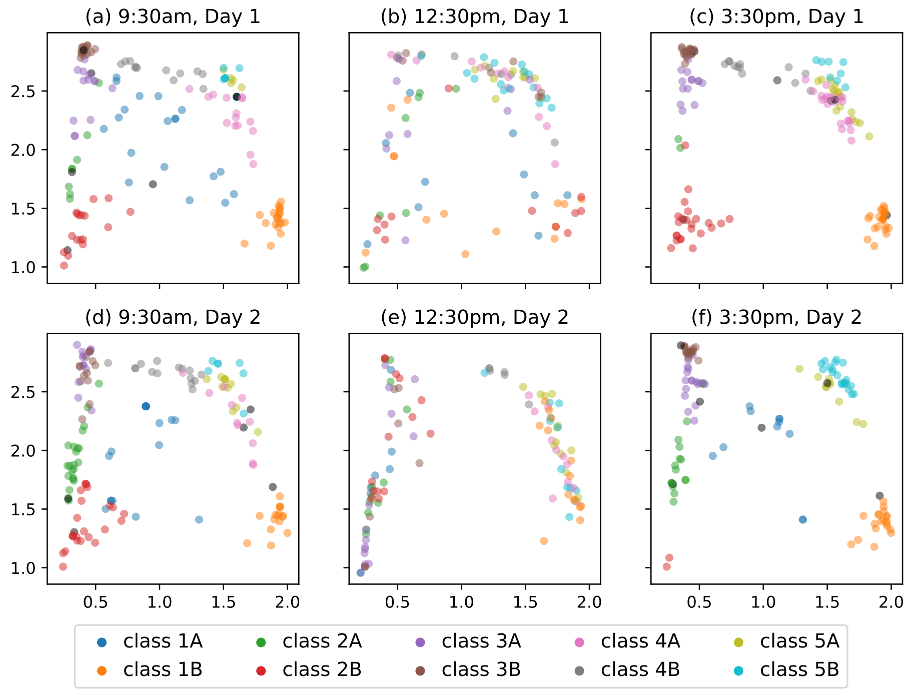


Figure D.1: The first two dimensions of the spherical coordinates of the coordinates $\hat{X}_i(t)$ using the histogram intensity estimator for times corresponding to the morning, lunchtime and afternoon across both days. The colours indicate classes with black points representing teachers.

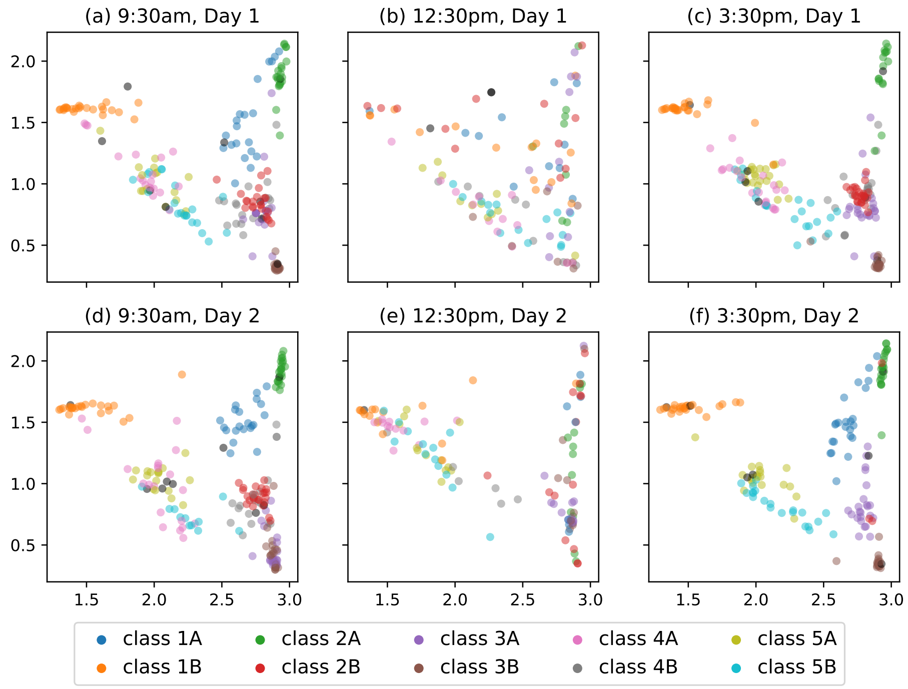


Figure D.2: The first two dimensions of the spherical coordinates of the trajectories $\hat{X}_i(t)$ using the Epanechnikov kernel smoother for times corresponding to the morning, lunchtime and afternoon across both days. The colours indicate classes with black points representing teachers.

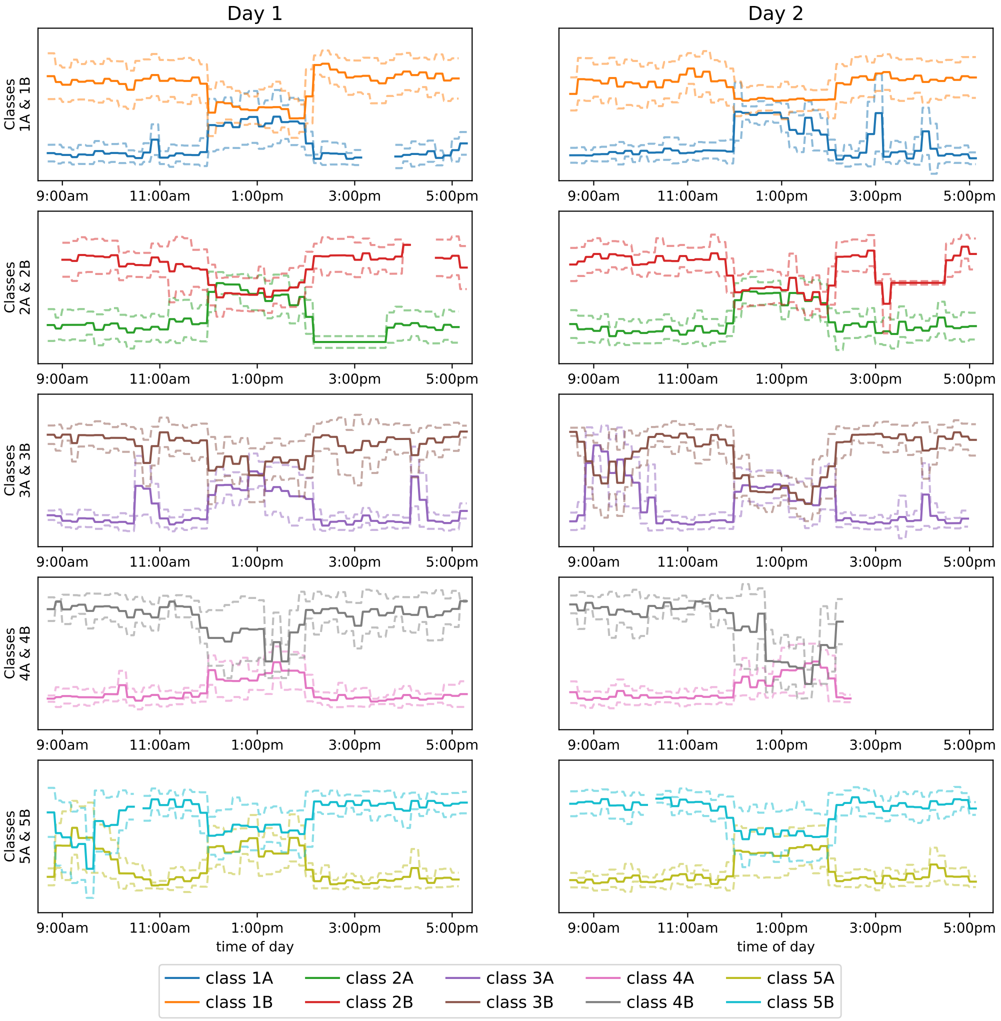


Figure D.3: One-dimensional PCA visualisation of the 30-dimensional node representations for pairs of classes in the same year group. The solid lines show the average trajectory for each class, and the dashed line shows one standard deviation above and below.

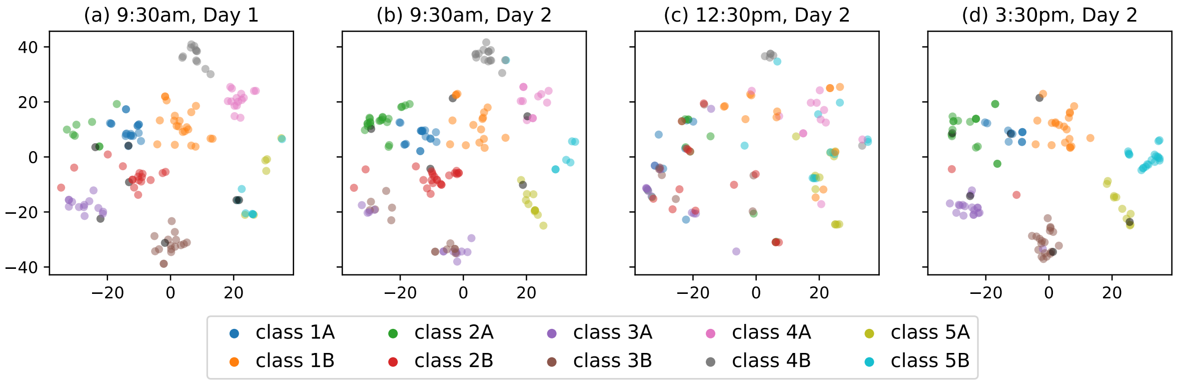


Figure D.4: Two-dimensional t-SNE visualisation of the 30-dimensional node representations of all pupils and teachers evaluated at 9:30am on Day 1, and 9:30am, 12:30pm and 3:30pm on Day 2.

Figure D.4 shows t-SNE visualisations of the node representations at a collection of times throughout the study. The plots are very similar to the equivalent Figure 5.4 for the kernel smoother with almost identical clusters of students before and after lunch, albeit placed differently by the t-SNE algorithm.

D.3 Visualsation

In this section, we give a short overview of the two dimension reduction techniques employed for visualisation in this paper.

For the trajectory visualisation in Figures 5.2 and 5.3, we use a principal component analysis which we extend to the dynamic setting by computing a projection using the leading eigenvectors of the *average* covariance matrix, which we apply to the (globally centered) trajectories. This has a similar flavour to our Intensity Profile Projection algorithm, and since we reduce dimension using a common projection, it gives a temporally coherent view of the trajectories.

The second visualisation technique we apply is t-SNE [122], using the Flt-SNE implementation [123], which we used to obtain Figure 5.4. This visualisation method is not naturally extended to dynamic data, so we initialise the algorithm using the aforementioned dynamic extension PCA, which results in the visualisations at different times being approximately aligned.

D.4 Proof of Lemma 5.1

We begin by writing

$$\begin{aligned}
\operatorname*{arg\,min}_{\mathbf{V}\in\mathbb{O}(n,d)} \hat{R}^2(\mathbf{V}) &= \operatorname*{arg\,min}_{\mathbf{V}\in\mathbb{O}(n,d)} \int_0^T \hat{r}_i^2(t;\mathbf{V})\,\mathrm{d}t \\
&= \operatorname*{arg\,min}_{\mathbf{V}\in\mathbb{O}(n,d)} \int_0^T \sum_{i=1}^n \left\|\mathbf{V}\mathbf{V}^\top\hat{\Lambda}_i(t) - \hat{\Lambda}_i(t)\right\|_2^2 \mathrm{d}t \\
&= \operatorname*{arg\,min}_{\mathbf{V}\in\mathbb{O}(n,d)} \int_0^T \sum_{i=1}^n \left\|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\hat{\Lambda}_i(t)\right\|_2^2 \mathrm{d}t,
\end{aligned}$$

and since $(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)$ is the projection onto the orthogonal complement of the columns space of $\mathbf{V}$, we have that

$$\left\|\hat{\Lambda}_i(t)\right\|_2^2 = \left\|\mathbf{V}\mathbf{V}^\top\hat{\Lambda}_i(t)\right\|_2^2 + \left\|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\hat{\Lambda}_i(t)\right\|_2^2.$$

Therefore, minimising $\hat{R}^2(\mathbf{V})$ is equivalent to maximising

$$\int_0^T \sum_{i=1}^n \left\|\mathbf{V}\mathbf{V}^\top\hat{\Lambda}_i(t)\right\|_2^2 \mathrm{d}t = \int_0^T \sum_{i=1}^n \left\|\mathbf{V}^\top\hat{\Lambda}_i(t)\right\|_2^2$$

where the equality holds due to the invariance of the Euclidean norm under orthogonal transformations. As a result, we have

$$\begin{aligned}
\underset{\mathbf{V}\in\mathbb{O}(n,d)}{\arg\min}\, \hat{R}^2(\mathbf{V}) &= \underset{\mathbf{V}\in\mathbb{O}(n,d)}{\arg\max} \int_0^T \sum_{i=1}^n \left\| \mathbf{V}^\top \hat{\Lambda}_i(t) \right\|_2^2 \\
&= \underset{\mathbf{V}\in\mathbb{O}(n,d)}{\arg\max} \int_0^T \left\| \hat{\boldsymbol{\Lambda}}(t) \mathbf{V} \right\|_{\mathrm{F}}^2 \\
&= \underset{\mathbf{V}\in\mathbb{O}(n,d)}{\arg\max} \int_0^T \operatorname{tr}\left\{ \mathbf{V}^\top \hat{\boldsymbol{\Lambda}}^2(t) \mathbf{V} \right\} \,\mathrm{d}t \\
&= \underset{\mathbf{V}\in\mathbb{O}(n,d)}{\arg\max} \operatorname{tr}\left\{ \mathbf{V}^\top \left(\int_0^T \hat{\boldsymbol{\Lambda}}^2(t)\,\mathrm{d}t \right) \mathbf{V} \right\} \\
&= \underset{\mathbf{V}\in\mathbb{O}(n,d)}{\arg\max} \operatorname{tr}\left\{ \mathbf{V}^\top \hat{\boldsymbol{\Sigma}} \mathbf{V} \right\} \\
&= \hat{\mathbf{U}}
\end{aligned}$$

where the final equality follows from the Courant-Fisher min-max theorem. This concludes the proof.

D.5 Proof of Theorem 5.1

D.5.1 Prerequisites

D.5.1.1 Additional notation

In this proof, we use the notation $a_n \overset{\mathbb{P}}{\lesssim} b_n$ to mean $a_n \lesssim b_n$ with overwhelming probability.

D.5.1.2 Symmetric dilation with change of basis "trick"

Symmetric dilation is a proof technique which allows statements about the eigenvectors of a symmetric matrix to be easily extended to hold for the singular vectors of a (potentially rectangular) asymmetric matrix. Let $\mathbf{M}$ be an $n_1 \times n_2$ matrix with non-zero singular values $\{\sigma_i\}_{i=1}^r$ and corresponding orthonormal left singular vectors $\{u_i\}_{i=1}^r$ and right singular vectors $\{v_i\}_{i=1}^r$. Its *symmetric dilation* is the $n \times n$ matrix (with $n = n_1 + n_2$) constructed as

$$\mathcal{D}(\mathbf{M}) = \begin{pmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M}^\top & \mathbf{0} \end{pmatrix}.$$

One can easily verify that $\mathcal{D}(\mathbf{M})$ has eigenvalues $\{\pm\sigma_i\}_{i=1}^r$ and eigenvectors $\{(u_i^\top, \pm v_i^\top)^\top\}_{i=1}^r$. We stack the first d left and right singular vectors into matrices $\mathbf{U} \in \mathbb{R}^{n_1 \times d}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times d}$, and stack the first $2d$ eigenvectors of $\mathcal{D}(\mathbf{M})$ into a matrix

$$\bar{\mathbf{U}} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{U} & \mathbf{U} \\ \mathbf{V} & -\mathbf{V} \end{pmatrix}.$$

We then have

$$\|\mathbf{U}\|_{2,\infty} \vee \|\mathbf{V}\|_{2,\infty} = \left\|\bar{\mathbf{U}}\right\|_{2,\infty}, \qquad \text{and} \qquad \|\mathbf{M}\|_2 = \|\mathcal{D}(\mathbf{M})\|_2 .$$

While this standard construction is very useful when $n_1 \asymp n_2$, it can lead to suboptimal bounds when $n_2 \gg n_1$, or $n_1 \gg n_2$, due to an issue about incoherence, which was first raised in [124]. The *incoherence* of a subspace $\mathcal{U}_0$ spanned by the orthonormal columns of a matrix $\mathbf{U}_0 \in \mathbb{R}^{n_0 \times d}$ is

$$\mu\left(\mathbf{U}_0\right) = \sqrt{\frac{n_0}{d}}\, \|\mathbf{U}_0\|_{2,\infty} .$$

To obtain a good entrywise eigenvector bound under a signal-plus-noise matrix model it is typically necessary that $\mu(\bar{\mathbf{U}}) \asymp 1$. Observe that

$$\mu(\bar{\mathbf{U}}) = \sqrt{\frac{n_1+n_2}{2d}}\left\|\bar{\mathbf{U}}\right\|_{2,\infty} = \sqrt{\frac{n_1+n_2}{2d}}\left(\|\mathbf{U}\|_{2,\infty} \vee \|\mathbf{V}\|_{2,\infty}\right) = \sqrt{\frac{n_1+n_2}{2n_1}}\mu(\mathbf{U}) + \sqrt{\frac{n_1+n_2}{2n_2}}\mu(\mathbf{V}).$$

If $\mu(\mathbf{U}), \mu(\mathbf{V}) \asymp 1$ and $n_1 \asymp n_2$, then $\mu(\bar{\mathbf{U}}) \asymp 1$ and it is typically possible to obtain good bounds. However, when $n_2 \gg n_1$, we have $\mu(\bar{\mathbf{U}}) \gg 1$, and a good bound can typically not be obtained. The imbalance of n_1 and n_2 can cause similar issues when obtaining spectral norm bounds.

This issue can be overcome by changing to a basis which balances the contribution from its first n_1 and second n_2 elements of each column. Specifically, let $\pi_1 = \sqrt{2n_1/(n_1+n_2)}$ and $\pi_2 = \sqrt{2n_2/(n_1+n_2)}$, and consider the basis $\tilde{e}_1, \dots, \tilde{e}_{n_1+n_2}$, such that

$$e_i = \begin{cases} \pi_1 \tilde{e}_i & \text{if } i \in \{1, \dots, n_1\} \\ \pi_2 \tilde{e}_i & \text{if } i \in \{n_1+1, \dots, n_1+n_2\}, \end{cases}$$

where $\{e_i\}_{i=1}^{n_1+n_2}$ are the standard basis vectors in $\mathbb{R}^{n_0}$. Let $|||\cdot|||_\eta$ denote a norm with respect to the column basis $\{\tilde{e}_i\}_{i=1}^{n_1+n_2}$, then one can verify that

$$|||\mathcal{D}(\mathbf{M})|||_2 = \|\mathbf{M}\|, \qquad \text{and} \qquad \left|\left|\left|\bar{\mathbf{U}}\right|\right|\right|_{2,\infty} = \pi_1 \|\mathbf{U}\|_{2,\infty} \vee \pi_2 \|\mathbf{V}\|_{2,\infty} .$$

As a result, if $\tilde{\mu}(\mathbf{U}_0) = \sqrt{n_0/d}|||\mathbf{U}_0|||_{2,\infty}$, then

$$\tilde{\mu}(\bar{\mathbf{U}}) = \mu(\mathbf{U}) \vee \mu(\mathbf{V}),$$

regardless of the relative sizes of n_1 and n_2. We use this symmetric dilation with change-of-basis "trick" to apply some existing theorems for symmetric matrices to our setting.

D.5.1.3 Concentration inequalities

In this section, we state a collection of lemmas which we will make use of throughout the proof. We begin with a tail bound for a Poisson random variable.

Lemma D.1. *Let* $X \sim \text{Poisson}(\lambda)$*. Then*

$$\mathbb{P}\left(|X-\lambda| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\lambda+t/3)}\right).$$

For $t \geq \lambda$*,*

$$\mathbb{P}(|X-\lambda| \geq t) \leq 2e^{-3t/8}.$$

The bound can be established by approximating the Poisson distribution with mean λ as the sum of k Bernoulli random variables with mean λ/k, applying Bernstein's inequality, and taking $k \to 0$.

Our next result is a concentration bound which adapts Lemma A.1 of [94] and can be proved using a vector version of the Bernstein inequality (Corollary 4.1 in [125]).

Lemma D.2. *Let* $X_i \sim \text{Poisson}(\lambda_i)$ *independently for all* $i = 1, \ldots, n$*, and suppose* $\mathbf{Q} \in \mathbb{R}^{n\times d}$ *is a deterministic matrix whose rows we denote* Q_i*. Let* $\lambda_{\max} := \max_{i\in[n]} \lambda_i$*, then with probability* $1 - 28n^{-3}$

$$\left\|\sum_{i=1}^{n} \left(X_i - \lambda_i\right) Q_i\right\|_2 \leq 3\log^2 n \left\|\mathbf{Q}\right\|_{2,\infty} + \sqrt{6\lambda_{\max}\log n}\left\|\mathbf{Q}\right\|_{\mathrm{F}}.$$

Next, we state a concentration bound for the spectral norm of random matrices with independent entries which appears as Corollary 3.12 in [126]. The original statement of this lemma is for symmetric random matrices, although we general it to arbitrary random matrices using the symmetric dilation with change-of-basis trick described in Section D.5.1.2.

Lemma D.3 (Corollary 3.12 of [126])**.** *Let* $\mathbf{X}$ *be an* $n_1 \times n_2$ *matrix whose entries* x_{ij} *are independent random variables which obey*

$$\mathbb{E}\left(x_{ij}\right) = 0, \qquad \textit{and} \qquad |x_{ij}| \leq B, \qquad i \in [n_1], j \in [n_2].$$

Then there exists a universal constant $c > 0$ *such that for any* $t \geq 0$

$$\mathbb{P}\left\{\|\mathbf{X}\| \geq 4\sqrt{\nu} + t\right\} \leq n\exp\left(-\frac{t^2}{cB^2}\right).$$

where

$$\nu := \max\left\{\pi_1 \max_{i\in[n_1]} \sum_{j=1}^{n_2} \mathbb{E}\left(x_{ij}^2\right), \pi_2 \max_{i\in[n_2]} \sum_{j=1}^{n_1} \mathbb{E}\left(x_{ji}^2\right)\right\}$$

and $\pi_k = 2n_k/(n_1+n_2)$*.*

D.5.1.4 Weyl's inequality and Wedin's sinΘ theorem

The next two lemmas are classical results matrix perturbation theory. Weyl's inequality shows that the singular values of a matrix are stable with respect to small perturbations.

Lemma D.4 (Weyl's inequality). *Let $\mathbf{M}, \mathbf{E}$ be $n_1 \times n_2$ real-valued matrices. Then for every $1 \leq i \leq (n_1 \wedge n_2)$, the ith largest singular value of $\mathbf{M}$ and $\mathbf{M} + \mathbf{E}$ obey*

$$|\sigma_i(\mathbf{M} + \mathbf{E}) - \sigma_i(\mathbf{M})| \leq \|\mathbf{E}\|_2 .$$

One way to measure the distance between two subspaces $\mathcal{U}$ and $\hat{\mathcal{U}}$ is via principal angles. Let $\mathbf{U}, \hat{\mathbf{U}}$ be matrices whose orthonormal columns span $\mathcal{U}$ and $\hat{\mathcal{U}}$ respectively, and let $\{\xi_i\}_{i=1}^d$ denote the singular values of $\mathbf{U}^\top \hat{\mathbf{U}}$. Then the principal angles $\{\theta_i\}_{i=1}^d$ between $\mathcal{U}$ and $\hat{\mathcal{U}}$ are defined by $\xi_i = \cos(\theta_i)$. Let $\sin \Theta(\mathbf{U}, \hat{\mathbf{U}}) := \operatorname{diag}(\sin\theta_1, \ldots, \sin\theta_d)$. Another way to measure the distance between $\mathcal{U}$ and $\hat{\mathcal{U}}$ is via the difference between the projection operators $\mathbf{U}\mathbf{U}^\top$ and $\hat{\mathbf{U}}\hat{\mathbf{U}}^\top$, and in fact, these two characterisations are equivalent. Specifically,

$$\left\|\sin \Theta(\mathbf{U}, \hat{\mathbf{U}})\right\|_2 \equiv \left\|\mathbf{U}\mathbf{U}^\top - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\right\|_2 .$$

We will use this equivalence without mention throughout the proof. Wedin's sinΘ theorem shows that the singular vectors of a matrix are stable with respect to small perturbations.

Lemma D.5. *Let $\mathbf{M}$ and $\hat{\mathbf{M}} = \mathbf{M} + \mathbf{E}$ be two $n_1 \times n_2$ real-valued matrices, and denote by $\mathbf{U}, \hat{\mathbf{U}}$ (respectively $\mathbf{V}, \hat{\mathbf{V}}$) the matrices whose columns contain d orthonormal left (respectively, right) singular vectors, corresponding to the d largest singular values of $\mathbf{M}$ and $\hat{\mathbf{M}}$. Let $\delta = \sigma_d(\mathbf{M}) - \sigma_{d+1}(\mathbf{M})$ and suppose that $\|\mathbf{E}\| < (1 - 1/\sqrt{2})\delta$, then*

$$\left\|\sin \Theta\left(\mathbf{U}, \hat{\mathbf{U}}\right)\right\|_2 \vee \left\|\sin \Theta\left(\mathbf{V}, \hat{\mathbf{V}}\right)\right\|_2 \leq \frac{2\left(\|\mathbf{E}^\top \mathbf{U}\|_2 \vee \|\mathbf{E}\mathbf{V}\|_2\right)}{\delta} \leq \frac{2\|\mathbf{E}\|}{\delta}.$$

See [127] for a proof.

D.5.2 Implications of Assumptions 1-4

We state here some inequalities involving the parameters of our problem which follow from Assumptions 1-4, and elementary linear algebra. We will use these facts throughout the proof without mention.

$$\sqrt{n\lambda_{\max}} \lesssim \delta \leq n\lambda_{\max}; \tag{D.1}$$

$$\delta \log n \gtrsim \kappa n \lambda_{\max}; \tag{D.2}$$

$$\kappa \lesssim \log n. \tag{D.3}$$

The inequality (D.1) holds since $\delta \leq \sigma_1^{1/2}(\mathbf{\Sigma}) \leq \sqrt{n}\|\mathbf{\Sigma}\|_{\max}^{1/2} \leq n\lambda_{\max}$, and

$$\delta \gtrsim \frac{\kappa n \lambda_{\max}}{\log(\delta/\sqrt{n\lambda_{max}})} \gtrsim \frac{n\lambda_{\max}}{\log n} \gtrsim \sqrt{n\lambda_{\max}\log n} \gtrsim \sqrt{n\lambda_{\max}}$$

where we invoked Assumption 4. (D.2) holds by noting that the previous bound implies $\log(\delta/\sqrt{n\lambda_{\max}}) \lesssim \log n$ and invoking Assumption 3. (D.3) follows from (D.1) since $\kappa \lesssim \delta \log n / n\lambda_{\max} \lesssim \log n$.

D.5.3 Setup

We begin by defining M equally spaced bins in $(0, 1]$,

$$B_1 := \left(0, \frac{1}{M}\right], \quad B_2 := \left(\frac{1}{M}, \frac{2}{M}\right], \ldots \quad , B_M := \left(\frac{M-1}{M}, 1\right],$$

and define the piecewise approximation of $\lambda_i(t)$,

$$\bar{\lambda}_i(t) = M \int_{B_m} \lambda_i(t) \mathrm{d}t, \qquad t \in B_m,\ m \in [M].$$

We then define $t_1, \ldots, t_m \in (0, 1]$ such that $\bar{\lambda}_{ij}(t) = \lambda_{ij}(t_m)$ for all $t \in B_m$, which exist by the continuity of $\lambda_{ij}(t)$, and define the piecewise constant approximation of $X_i(t)$ as $\bar{X}_i(t) = \mathbf{U}^\top \bar{\Lambda}_i(t)$. Our strategy to obtain the bound in Theorem 1 is to decompose it into bias and variance terms:

$$\max_{i,j \in [n]} \sup_{t \in \mathcal{T}} \left\| \mathbf{W}_1 \hat{X}_i(t) - X_i(t) \right\|_2 = \underbrace{\max_{i,j \in [n]} \sup_{t \in \mathcal{T}} \left\| \mathbf{W}_1 \hat{X}_i(t) - \bar{X}_i(t) \right\|_2}_{\text{variance}} + \underbrace{\max_{i,j \in [n]} \sup_{t \in \mathcal{T}} \left\| \mathbf{W}_2 \bar{X}_i(t) - X_i(t) \right\|_2}_{\text{bias}}.$$

Section D.5.6 is dedicated to bounding the bias term, and the rest of this section is dedicated to bounding the variance term. Define the unfolding matrices $\hat{\mathbf{\Lambda}}$ and $\mathbf{\Lambda}$ (without arguments) and their (thin) singular value decompositions as

$$\begin{aligned} \hat{\mathbf{\Lambda}} &:= \left(\hat{\mathbf{\Lambda}}(t_1) \ \cdots \ \hat{\mathbf{\Lambda}}(t_M)\right) = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^\top + \hat{\mathbf{U}}_\perp \hat{\mathbf{S}}_\perp \hat{\mathbf{V}}_\perp^\top, \\ \bar{\mathbf{\Lambda}} &:= \left(\mathbf{\Lambda}(t_1) \ \cdots \ \mathbf{\Lambda}(t_M)\right) = \bar{\mathbf{U}}\bar{\mathbf{S}}\bar{\mathbf{V}}^\top + \bar{\mathbf{U}}_\perp \bar{\mathbf{S}}_\perp \bar{\mathbf{V}}_\perp^\top. \end{aligned}$$

Then one has that for $t \in B_m$, $m \in [M]$,

$$\hat{\mathbf{Y}}(t) := \hat{\mathbf{\Lambda}}(t_m)\hat{\mathbf{U}} = \hat{\mathbf{V}}_m \hat{\mathbf{S}}, \qquad \bar{\mathbf{Y}}(t) := \bar{\mathbf{\Lambda}}(t_m)\bar{\mathbf{U}} = \bar{\mathbf{V}}_m \hat{\mathbf{S}}$$

where $\hat{\mathbf{V}}_m, \bar{\mathbf{V}}_m$ denote the mth blocks of $\hat{\mathbf{V}}$ and $\bar{\mathbf{V}}$ respectively. Therefore it follows that,

$$\max_{i,j \in [n]} \sup_{t \in \mathcal{T}} \left\| \mathbf{W}_1 \hat{X}_i(t) - \bar{X}_i(t) \right\|_2 = \left\| \hat{\mathbf{V}}\hat{\mathbf{S}}\mathbf{W}_1^\top - \bar{\mathbf{V}}\bar{\mathbf{S}} \right\|_{2,\infty}.$$

For ease of exposition, we drop the subscript 1 on $\mathbf{W}_1$ in this section. Our bound is based on the following decomposition of $\hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{V}}\bar{\mathbf{S}}\mathbf{W}$.

Proposition D.1. *We have the decomposition*

$$\begin{aligned} \text{(D.4)} \qquad \hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{V}}\bar{\mathbf{S}}\mathbf{W} &= \bar{\mathbf{V}}(\bar{\mathbf{V}}^\top \hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W}) \\ \text{(D.5)} \qquad &+ (\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top(\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W}) \\ \text{(D.6)} \qquad &+ (\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top \bar{\mathbf{U}}\mathbf{W} \\ \text{(D.7)} \qquad &+ (\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top(\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W}). \end{aligned}$$

Proof of Proposition D.1. We begin by adding and subtracting terms to obtain

$$\hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{V}}\bar{\mathbf{S}}\mathbf{W} = \hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\hat{\mathbf{V}}\hat{\mathbf{S}} + \underbrace{\bar{\mathbf{V}}(\bar{\mathbf{V}}^\top\hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W})}_{\text{(D.4)}}.$$

Then, noting that $\hat{\mathbf{V}}\hat{\mathbf{S}} = \hat{\mathbf{\Lambda}}^\top\hat{\mathbf{U}}$ and $(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top\bar{\mathbf{U}} = \mathbf{0}$, we have

$$\begin{aligned}
\hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\hat{\mathbf{V}}\hat{\mathbf{S}} &= \hat{\mathbf{\Lambda}}^\top\hat{\mathbf{U}} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{U}} \\
&= (\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{U}} \\
&= (\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top\hat{\mathbf{U}} - (\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top\hat{\mathbf{U}} \\
&= (\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top\hat{\mathbf{U}} - \underbrace{(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top(\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W})}_{\text{(D.5)}}.
\end{aligned}$$

Next, we decompose $(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top\hat{\mathbf{U}}$ by adding and subtracting terms to obtain

$$(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top\hat{\mathbf{U}} = \underbrace{(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top\bar{\mathbf{U}}\mathbf{W}}_{\text{(D.6)}} + \underbrace{(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top(\hat{\mathbf{U}} - \mathbf{U}\mathbf{W})}_{\text{(D.7)}}.$$

□

D.5.4 Technical propositions

We now outline a series of technical propositions which we require to bound terms (D.4)-(D.7) which we prove in Section D.6.

Our first proposition is a 1-norm and spectral norm bound for $\hat{\mathbf{\Lambda}}$.

Proposition D.2. *The bounds*

$$\left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_1 \lesssim \sqrt{Mn\lambda_{\max}\log n}, \qquad \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_2 \lesssim \sqrt{Mn\lambda_{\max}}$$

hold with overwhelming probability.

The spectral norm bound is obtained using Lemma D.3, and the 1-norm bound is obtained via an application of the classical Bernstein inequality. The next proposition provides control on the singular values of $\hat{\mathbf{\Lambda}}$.

Proposition D.3. *Let $\sigma_i(\cdot)$ denote the ith ordered singular value of a matrix. The singular values of $\hat{\mathbf{\Lambda}}$ satisfy*

$$\sqrt{M\sigma_d(\mathbf{\Sigma})} \lesssim \sigma_d(\hat{\mathbf{\Lambda}}) \leq \sigma_1(\hat{\mathbf{\Lambda}}) \lesssim \sqrt{M\sigma_1(\mathbf{\Sigma})}.$$

The result is obtained using Weyl's inequality. The next proposition provides control of the spectral norm of $\mathbf{Q}^\top(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})\mathbf{R}$, where $\mathbf{Q}, \mathbf{R}$ are conformable, deterministic unit-norm matrices.

Proposition D.4. *For conformable, deterministic unit-norm matrices* $\mathbf{Q}, \mathbf{R}$*, the bound*

$$\left\|\mathbf{Q}^\top(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})\mathbf{R}\right\|_2 \lesssim M \log^{3/2} n \tag{D.8}$$

holds with overwhelming probability.

The proof of Proposition D.4 employs a classical ϵ-net argument to the spectral norm of an appropriately constructed symmetric dilation matrix.

The next proposition states that both the matrices $\bar{\mathbf{U}}^\top\hat{\mathbf{U}}$ and $\bar{\mathbf{V}}^\top\hat{\mathbf{V}}$ are well approximated by a common orthogonal matrix.

Proposition D.5. *There exists an orthogonal matrix* $\mathbf{W}$ *such that*

$$\left\|\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \mathbf{W}\right\|_2 \lesssim \frac{\sqrt{n\lambda_{\max}}}{\delta}, \qquad \left\|\bar{\mathbf{V}}^\top\hat{\mathbf{V}} - \mathbf{W}\right\|_2 \lesssim \frac{\sqrt{n\lambda_{\max}}}{\delta}$$

hold with overwhelming probability.

To prove Proposition D.5, we empoy the Wedin sinΘ theorem to obtain a bound on $\|\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \mathbf{W}\|_2$. We then obtain a bound on $\|\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \bar{\mathbf{V}}^\top\hat{\mathbf{V}}\|_2$, and combine these bounds to establish the proposition.

The next technical tool we require is the ability to "swap" $\mathbf{W}, \bar{\mathbf{S}}$ and $\hat{\mathbf{S}}$.

Proposition D.6. *The bound*

$$\left\|\mathbf{W}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W}\right\|_2 \lesssim M \log^{3/2} n$$

holds with overwhelming probability.

This result follows by applying the previous propositions to an appropriately constructed decomposition.

Part of the challenge of obtaining a good bound on the term (D.7) is that $(\tilde{\mathbf{A}} - \tilde{\mathbf{\Lambda}})$ and $(\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W})$ are dependent, and this dependence must be decoupled in order to apply the standard suite of matrix perturbation tools. For $m = 1, \ldots, n$, let

$$\mathcal{N}_m = \{(i,j) : i = m \text{ or } j \in \{m + (\ell - 1)n, \ell \in [M]\}\}$$

and construct the auxiliary matrices $\hat{\mathbf{\Lambda}}^{(1)}, \ldots, \hat{\mathbf{\Lambda}}^{(n)}$ defined by

$$\hat{\mathbf{\Lambda}}_{ij}^{(m)} = \begin{cases} \hat{\mathbf{\Lambda}}_{ij} & \text{if } (i,j) \notin \mathcal{N}_m, \\ \bar{\mathbf{\Lambda}}_{ij} & \text{if } (i,j) \in \mathcal{N}_m. \end{cases}$$

In words, $\hat{\mathbf{\Lambda}}^{(m)}$ is the matrix obtained by replacing the mth row and columns of each of its blocks with its expectation. In this way, the mth row of $(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})$ and $\hat{\mathbf{\Lambda}}^{(m)}$ are independent. Let $\hat{\mathbf{U}}^{(m)}$ denote the matrix of leading left singular values of $\hat{\mathbf{\Lambda}}^{(m)}$.

We apply a result due to [93], which provides $\ell_{2,\infty}$ control of $\|\hat{\mathbf{U}}\|_{2,\infty}, \|\hat{\mathbf{U}}^{(m)}\|_{2,\infty}$, and $\|\hat{\mathbf{U}}^{(m)}\mathbf{W}^{(m)} - \mathbf{U}\|_{2,\infty}$.

Proposition D.7. *The bounds*

$$\left\|\hat{\mathbf{U}}\right\|_{2,\infty}, \ \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty}, \ \left\|\hat{\mathbf{U}}^{(m)}\mathbf{W}^{(m)} - \mathbf{U}\right\|_{2,\infty} \lesssim \frac{\mu\lambda_{\max}\sqrt{dn}\log n}{\delta}$$

hold with overwhelming probability.

In addition, we require control on the spectral norm difference between the projection matrices $\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top$ and the projection matrices $\bar{\mathbf{U}}\bar{\mathbf{U}}^\top$ and $\hat{\mathbf{U}}\hat{\mathbf{U}}^\top$, which is provided in the following proposition.

Proposition D.8. *The bounds*

$$\left\|\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right\|_2 \lesssim \frac{n\lambda_{\max}}{\delta}, \tag{D.9}$$

$$\left\|\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\right\|_2 \lesssim \frac{\mu\lambda_{\max}^{3/2}\sqrt{d}n\log^{3/2} n}{\delta^2} \tag{D.10}$$

hold with overwhelming probability.

The proof of Proposition D.8 requires a delicate “leave-one-out”–style argument.

D.5.5 Bounding terms (D.4)-(D.7)

Firstly observe that

$$\|\bar{\mathbf{V}}\|_{2,\infty} = \left\|\bar{\mathbf{\Lambda}}^\top\bar{\mathbf{U}}\bar{\mathbf{S}}^{-1}\right\|_{2,\infty} \leq \left\|\bar{\mathbf{\Lambda}}^\top\right\|_\infty \|\bar{\mathbf{U}}\|_{2,\infty}\|\bar{\mathbf{S}}^{-1}\|_2 \leq \frac{\sqrt{nd}\lambda_{\max}\mu}{\sqrt{M\sigma_d(\mathbf{\Sigma})}}$$

and therefore term (D.4) can be bounded as

$$\begin{aligned}
\left\|\bar{\mathbf{V}}(\bar{\mathbf{V}}^\top\hat{\mathbf{V}}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W})\right\|_{2,\infty} &\leq \|\bar{\mathbf{V}}\|_{2,\infty}\left(\left\|\bar{\mathbf{V}}^\top\hat{\mathbf{V}} - \mathbf{W}\right\|_2\left\|\hat{\mathbf{S}}\right\| + \left\|\mathbf{W}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W}\right\|_2\right) \\
&\overset{\mathbb{P}}{\lesssim} \frac{\sqrt{nd}\lambda_{\max}\mu}{\sqrt{M\sigma_d(\mathbf{\Sigma})}}\left(\frac{\sqrt{n\lambda_{\max}}}{\delta}\cdot\sqrt{M\sigma_1(\mathbf{\Sigma})} + M\log^{3/2} n\right) \\
&\lesssim \frac{n\sqrt{Md}\lambda_{\max}^{3/2}\mu\kappa}{\delta} \\
&\lesssim \mu\sqrt{M\lambda_{\max}d}\log n.
\end{aligned}$$

where the third inequality follows from Assumption 4 that $\sqrt{M}\log^{3/2} n \lesssim n\lambda_{\max}$, and the definition $\kappa := \sqrt{\sigma_1(\mathbf{\Sigma})/\sigma_d(\mathbf{\Sigma})}$, and the fourth inequality follows from Assumption 3 that $\delta\log n \geq \delta\log(\delta/\sqrt{n\lambda_{\max}}) \gtrsim \kappa n\lambda_{max}$.

To bound (D.5), we first apply Wedin's $\sin\Theta$ theorem to obtain

$$\left\|\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W}\right\|_2 = \left\|\sin\Theta(\hat{\mathbf{U}}, \bar{\mathbf{U}})\right\|_2 \leq \frac{\left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|}{\sigma_d(\bar{\mathbf{\Lambda}}) - \sigma_{d+1}(\bar{\mathbf{\Lambda}})} \overset{\mathbb{P}}{\lesssim} \frac{\sqrt{Mn\lambda_{\max}}}{\sqrt{M}\delta} = \frac{\sqrt{n\lambda_{\max}}}{\delta}$$

Then, we use Assumption 2 to obtain the bound

$$\left\|(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top\right\|_{2,\infty} = \left\|\bar{\mathbf{\Lambda}}^\top(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}})\right\|_{2,\infty} \lesssim \max_{i\in[n]}\sup_{t\in\mathcal{T}} r_i(t) \lesssim \sqrt{\frac{d}{n}}\mu\delta\log^{5/2} n.$$

Putting these two bounds together, we bound (D.5) as

$$\left\|(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top(\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W})\right\|_{2,\infty} \le \left\|(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)\bar{\mathbf{\Lambda}}^\top\right\|_{2,\infty}\left\|\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W}\right\|_2 \overset{\mathbb{P}}{\lesssim} \mu\sqrt{d\lambda_{\max}}\log^{5/2} n$$

To bound term (D.6), we set $\mathbf{E} = \hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}$ and note that each column of $M^{-1}\mathbf{E}$ contains independent Poisson random variables with means no greater that $M^{-1}\lambda_{\max}$. We will use Lemma D.2 to bound the rows $\mathbf{E}\bar{\mathbf{U}}$ as

$$\begin{aligned}
[\mathbf{E}^\top\bar{\mathbf{U}}]_i = \sum_{j=1}^n e_{ji}\bar{U}_j &\overset{\mathbb{P}}{\lesssim} M\log^2 n\left\|\bar{\mathbf{U}}\right\|_{2,\infty} + \sqrt{M\lambda_{\max}\log n}\left\|\bar{\mathbf{U}}\right\|_{\mathrm{F}} \\
&\lesssim \left(M\log^2 n + \sqrt{Mn\lambda_{\max}\log n}\right)\left\|\bar{\mathbf{U}}\right\|_{2,\infty} \\
&\lesssim \sqrt{M\lambda_{\max} n\log n}\left\|\bar{\mathbf{U}}\right\|_{2,\infty} \\
&\lesssim \mu\sqrt{M\lambda_{\max} d\log n}.
\end{aligned}$$

where the third inequality uses Assumption 4 and a union bound over $i = 1, \dots, n$. Therefore, we have

$$\|(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top\bar{\mathbf{U}}\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \sqrt{M\lambda_{\max} d\log n}. \tag{D.11}$$

Noting that $\left\|\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right\|_\infty \le \left\|\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right\|_2 \lesssim 1$, we bound (D.6) as

$$\left\|(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top\bar{\mathbf{U}}\mathbf{W}\right\|_{2,\infty} \lesssim \left\|\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}\right\|_\infty\left\|\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)^\top\bar{\mathbf{U}}\right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \mu\sqrt{M\lambda_{\max} d\log n}.$$

Let $\hat{\mathbf{\Lambda}}^{(1)}, \dots, \hat{\mathbf{\Lambda}}^{(n)}$ denote the auxiliary matrices described in (D.5.4), and let $\hat{\mathbf{U}}^{(m)}$ denote the matrix of leading left singular values of $\hat{\mathbf{\Lambda}}^{(m)}$. We can then decompose the Euclidean norm of $(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}} - \mathbf{U}\mathbf{W})$ as

$$\begin{aligned}
\left\|(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}} - \bar{\mathbf{U}}\mathbf{W})\right\|_2 \le{}& \left\|(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top_{\cdot,m}\hat{\mathbf{U}}(\mathbf{W} - \hat{\mathbf{U}}^\top\bar{\mathbf{U}})\right\|_2 && \text{(D.12)} \\
&+ \left\|(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}}\hat{\mathbf{U}}^\top\bar{\mathbf{U}} - \hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top\bar{\mathbf{U}})\right\|_2 && \text{(D.13)} \\
&+ \left\|(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top_{\cdot,m}(\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top\bar{\mathbf{U}} - \bar{\mathbf{U}})\right\|_2. && \text{(D.14)}
\end{aligned}$$

The first term (D.12) is bounded as

$$\begin{aligned}
\left\|(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})^\top_{\cdot,m}\hat{\mathbf{U}}(\mathbf{W} - \hat{\mathbf{U}}^\top\bar{\mathbf{U}})\right\|_2 &\le \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_1\left\|\hat{\mathbf{U}}\right\|_{2,\infty}\left\|\mathbf{W} - \hat{\mathbf{U}}^\top\bar{\mathbf{U}}\right\|_2 \\
&\overset{\mathbb{P}}{\lesssim} \sqrt{Mn\lambda_{max}\log n}\cdot\frac{\mu\lambda_{\max}\sqrt{dn}\log n}{\delta}\cdot\frac{\sqrt{n\lambda_{\max}}}{\delta} \\
&= \frac{\sqrt{M}n^{3/2}\lambda_{\max}^2\mu\sqrt{d}\log^{3/2} n}{\delta^2} \\
&\lesssim \mu\sqrt{M\lambda_{\max} d}\log^{5/2} n.
\end{aligned}$$

To bound the second term (D.13), we employ Proposition D.8 to obtain

$$\begin{aligned}\left\|(\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}})_{\cdot,m}^{\top}(\hat{\mathbf{U}}\hat{\mathbf{U}}^{\top}\bar{\mathbf{U}}-\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^{\top}\bar{\mathbf{U}})\right\|_2 &\le \left\|\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}}\right\|_2\left\|\hat{\mathbf{U}}\hat{\mathbf{U}}^{\top}-\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^{\top}\right\|_2\\ &\overset{\mathbb{P}}{\lesssim} \sqrt{Mn\lambda_{\max}}\cdot\frac{\mu\lambda_{\max}^{3/2}\sqrt{d}n\log^{3/2}n}{\delta^2}\\ &= \frac{\sqrt{M}\mu\lambda_{\max}^{2}\sqrt{d}n^{3/2}\log^{3/2}n}{\delta^2}\\ &\lesssim \mu\sqrt{M\lambda_{\max}d}\log^{5/2}n\end{aligned}$$

We now set about bounding the third term (D.14). Let $\mathbf{\Omega}_1\mathbf{\Xi}\mathbf{\Omega}_2^{\top}$ denote a singular value decomposition of $(\hat{\mathbf{U}}^{(m)})^{\top}\bar{\mathbf{U}}$, and set $\mathbf{W}^{(m)} := \mathbf{\Omega}_1\mathbf{\Omega}_2^{\top}$. Let $\theta_i^{(m)}$ denote the principal angles between the column spaces of $\hat{\mathbf{U}}^{(m)}$ and $\bar{\mathbf{U}}$ defined by $\xi_i^{(m)} = \cos(\theta_i^{(m)})$, where $\xi_i^{(m)}$ are the singular values of $(\hat{\mathbf{U}}^{(m)})^{\top}\bar{\mathbf{U}}$. We invoke Wedin's theorem to show that

$$\begin{aligned}&\left\|\mathbf{W}^{(m)}-\left(\hat{\mathbf{U}}^{(m)}\right)^{\top}\bar{\mathbf{U}}\right\|_2 = \|\mathbf{I}-\mathbf{\Xi}\|_2 = \max_{i\in[d]}(1-\xi_i^{(m)}) = \max_{i\in[d]}(1-\cos\theta_i^{(m)})\\ &\le \max_{i\in[d]}(1-\cos^2\theta_i^{(m)}) = \max_{i\in[d]}\sin^2\theta_i^{(m)} \lesssim \frac{\left\|\hat{\mathbf{\Lambda}}^{(m)}-\bar{\mathbf{\Lambda}}\right\|_2^2}{(\sigma_d(\bar{\mathbf{\Lambda}})-\sigma_{d+1}(\hat{\mathbf{\Lambda}}))^2} \overset{\mathbb{P}}{\lesssim} \frac{Mn\lambda_{\max}}{M\delta^2} = \frac{n\lambda_{\max}}{\delta^2} \lesssim 1\end{aligned}$$

We define $\mathbf{H}^{(m)} := \hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^{\top} - \mathbf{U}\mathbf{U}^{\top}$ and note that $\mathbf{H}^{(m)}$ is independent of $(\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}})_{m,\cdot}$ and that

$$\begin{aligned}\left\|\mathbf{H}^{(m)}\right\|_{2,\infty} &\le \left\|\hat{\mathbf{U}}^{(m)}\mathbf{W}^{(m)}-\bar{\mathbf{U}}\right\|_{2,\infty} + \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty}\left\|(\hat{\mathbf{U}}^{(m)})^{\top}\bar{\mathbf{U}}-\mathbf{W}^{(m)}\right\|_2\\ &\lesssim \left\|\hat{\mathbf{U}}^{(m)}\mathbf{W}^{(m)}-\bar{\mathbf{U}}\right\|_{2,\infty} + \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty}\\ &\overset{\mathbb{P}}{\lesssim} \frac{\mu\lambda_{\max}\sqrt{dn}\log n}{\delta}\end{aligned}$$

Then, using Lemma D.2 we have that

$$\begin{aligned}\left\|(\tilde{\mathbf{\Lambda}}-\tilde{\mathbf{\Lambda}})_{\cdot,m}^{\top}\mathbf{H}^{(m)}\right\|_2 &\overset{\mathbb{P}}{\lesssim} M\log^2 n\left\|\mathbf{H}^{(m)}\right\|_{2,\infty} + \sqrt{\lambda_{\max}\log n}\left\|\mathbf{H}^{(m)}\right\|_{\mathrm{F}}\\ &\overset{\mathbb{P}}{\lesssim} \sqrt{M\log n\lambda_{\max}}\left\|\mathbf{H}^{(m)}\right\|_{2,\infty} + \sqrt{\lambda_{\max}d\log n}\left\|\mathbf{H}^{(m)}\right\|_2\\ &\overset{\mathbb{P}}{\lesssim} \sqrt{M\log n\lambda_{\max}}\cdot\frac{\mu\lambda_{\max}\sqrt{dn}\log n}{\delta} + \sqrt{\lambda_{\max}d\log n}\frac{n\lambda_{\max}}{\delta}\\ &\le \frac{\sqrt{M}\mu n\lambda_{\max}^{3/2}\sqrt{d}\log^{3/2}n}{\delta}\\ &\lesssim \mu\sqrt{Md\lambda_{\max}}\log^{5/2}n.\end{aligned}$$

Combining these bounds and taking a union bound over $m\in[n]$, we have

$$\left\|(\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}})^{\top}(\hat{\mathbf{U}}-\bar{\mathbf{U}}\mathbf{W})\right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \mu\sqrt{Md\lambda_{\max}}\log^{5/2}n,$$

and the term (D.7) is bounded as

$$\begin{aligned}\left\|(\mathbf{I}-\bar{\mathbf{V}}\bar{\mathbf{V}}^\top)(\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}})^\top(\hat{\mathbf{U}}-\bar{\mathbf{U}}\mathbf{W})\right\|_{2,\infty} &\leq \left\|\mathbf{I}-\bar{\mathbf{V}}\bar{\mathbf{V}}\right\|_\infty \left\|(\hat{\mathbf{\Lambda}}-\bar{\mathbf{\Lambda}})^\top(\hat{\mathbf{U}}-\bar{\mathbf{U}}\mathbf{W})\right\|_{2,\infty} \\ &\overset{\mathbb{P}}{\lesssim} \mu\sqrt{M\lambda_{\max}d}\log^{5/2} n.\end{aligned}$$

Combining the bounds on (D.4)-(D.7), we have

$$\max_{i,j\in[n]}\sup_{t\in\mathcal{T}}\left\|\mathbf{W}_1\hat{X}_i(t)-\bar{X}_i(t)\right\|_2 = \left\|\hat{\mathbf{V}}\hat{\mathbf{S}}\mathbf{W}_1^\top-\bar{\mathbf{V}}\bar{\mathbf{S}}\right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \mu\sqrt{Md\lambda_{\max}}\log^{5/2} n,$$

which completes the proof.

D.5.6 Controlling the bias term

D.5.6.1 Edge level bias

We begin by studying the edge-level bias of the histogram intensity estimator. Let $\rho_{ij}(t) = \int_0^t \lambda_{ij}(s)\,\mathrm{d}s$ denote the cumulative intensity of edge i, j. Now we have, for $t \in B_\ell$,

$$\begin{aligned}\bar{\lambda}_{ij}(t) &= M\int_{B_\ell}\lambda_{ij}(s)\,\mathrm{d}t \\ &= M\left\{\rho_{ij}\left(\frac{\ell}{M}\right)-\rho_{ij}\left(\frac{\ell-1}{M}\right)\right\} \\ &= \frac{\rho_{ij}\left(\frac{\ell}{M}\right)-\rho_{ij}\left(\frac{\ell-1}{M}\right)}{\frac{\ell}{M}-\frac{\ell-1}{M}} \\ &= \lambda_{ij}(t^\star)\end{aligned}$$

for some $t^\star \in B^\ell$, which follows by an application of the mean value theorem, where $\rho'(t) = \lambda_{ij}(t)$. We then apply the L-Lipschitz continuity of $\lambda_{ij}(t)$ to obtain

$$\begin{aligned}\bar{\lambda}_{ij}(t)-\lambda_{ij}(t)| &= |\lambda_{ij}(t^\star)-\lambda_{ij}(t)| \\ &\leq L\cdot|t^\star-t| \\ &\leq \frac{L}{M}.\end{aligned}$$

D.5.6.2 A subspace perturbation bound

Define the operator $\mathcal{A} : (\mathcal{T} \to \mathbb{R}^n) \to \mathbb{R}^n$ by

$$\mathcal{A}v(\cdot) = \int_{\mathcal{T}}\mathbf{\Lambda}(t)v(t)\,\mathrm{d}t$$

and define the operator $\mathcal{A}^\star : \mathbb{R}^n \to (\mathcal{T} \to \mathbb{R}^n)$ by

$$\mathcal{A}^\star u = \mathbf{\Lambda}(\cdot)u.$$

Then $\boldsymbol{\Sigma} \equiv \mathcal{A}\mathcal{A}^\star$ since

$$\mathcal{A}\mathcal{A}^\star u = \mathcal{A}\left(\boldsymbol{\Lambda}(\cdot)u\right) = \int_{\mathcal{T}} \boldsymbol{\Lambda}^2(t)u \,\mathrm{d}t = \boldsymbol{\Sigma}u.$$

Denote its eigenvalues $\sigma_1^2, \ldots, \sigma_n^2$, and its corresponding orthonormal eigenvectors $u_1, \ldots, u_n$, and define $v_i(\cdot) = \boldsymbol{\Lambda}(\cdot)u_i/\xi_i$ for all $i = 1, \ldots, n$. Then, $\boldsymbol{\Lambda}(\cdot)$ admits the (functional) singular value decomposition

$$\boldsymbol{\Lambda}(\cdot) = \sum_{i=1}^n \sigma_i u_i v_i(\cdot).$$

Define $\bar{\mathcal{A}}$ and its corresponding parameters analogously. By definition,

$$\left\|\bar{\mathcal{A}} - \mathcal{A}\right\|_2 \leq \sup_{t\in\mathcal{T}} \left\|\bar{\boldsymbol{\Lambda}}(t) - \boldsymbol{\Lambda}(t)\right\|_2 \leq \sup_{t\in\mathcal{T}} \max_{i,j\in[n]} \left|\bar{\lambda}_{ij}(t) - \lambda_{ij}(t)\right| \leq \frac{nL}{M}.$$

Therefore, by (a functional version of) Wedin's $\sin\Theta$ theorem

$$\left\|\bar{\mathbf{U}}\mathbf{W}_1 - \mathbf{U}\right\|_2 \lesssim \frac{\left\|\bar{\mathcal{A}} - \mathcal{A}\right\|_2}{\sigma_d - \sigma_{d+1}} \leq \frac{nL}{M\delta}.$$

D.5.6.3 Controlling the bias term

Combining the above bounds, we have that uniformly for all i, j, t,

$$\begin{aligned}
\left\|\bar{X}_i(t)\mathbf{W}_1 - X_i(t)\right\|_2 &= \left\|\bar{\boldsymbol{\Lambda}}(t)\bar{\mathbf{U}}\mathbf{W}_1 - \boldsymbol{\Lambda}(t)\mathbf{U}\right\|_2 \\
&\leq \left\|\bar{\boldsymbol{\Lambda}}(t)\right\|_{2,\infty} \left\|\bar{\mathbf{U}}\mathbf{W}_2 - \mathbf{U}\right\|_2 + \left\|\bar{\boldsymbol{\Lambda}}(t) - \boldsymbol{\Lambda}(t)\right\|_{2,\infty} \left\|\mathbf{U}\right\|_2 \\
&\lesssim \frac{n^{3/2}\lambda_{\max}L}{M\delta} + \frac{\sqrt{n}L}{M} \\
&\leq \frac{n^{3/2}\lambda_{\max}L}{M\delta},
\end{aligned}$$

where the final inequality follows from the fact that $\delta \leq n\lambda_{\max}$.

D.6 Proofs of the technical propositions

D.6.1 Proof of Proposition D.2

We have that M^{-1} times the lower-triangular elements of each block of $\hat{\boldsymbol{\Lambda}}$ are independent Poisson random variables with mean given by M^{-1} times the lower-triangular elements of each block of $\bar{\boldsymbol{\Lambda}}$. Define the matrices $\hat{\boldsymbol{\Lambda}}^{\mathrm{L}}$ and $\hat{\boldsymbol{\Lambda}}^{\mathrm{U}}$ with the upper and lower triangles, respectively, of each block set to zero, and the diagonals of each block halved, and define $\bar{\boldsymbol{\Lambda}}^{\mathrm{L}}$ and $\bar{\boldsymbol{\Lambda}}^{\mathrm{U}}$ similarly, so that $M^{-1}\hat{\boldsymbol{\Lambda}}^{\mathrm{L}}$ (respectively $M^{-1}\hat{\boldsymbol{\Lambda}}^{\mathrm{U}}$) has independent Poisson entries with means $M^{-1}\bar{\boldsymbol{\Lambda}}^{\mathrm{L}}$ (respectively $M^{-1}\bar{\boldsymbol{\Lambda}}^{\mathrm{U}}$), and $\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}} = (\hat{\boldsymbol{\Lambda}}^{\mathrm{L}} - \bar{\boldsymbol{\Lambda}}^{\mathrm{L}}) + (\hat{\boldsymbol{\Lambda}}^{\mathrm{U}} - \bar{\boldsymbol{\Lambda}}^{\mathrm{U}})$.

We condition on the event that $(\hat{\boldsymbol{\Lambda}}^{\mathrm{L}} - \bar{\boldsymbol{\Lambda}}^{\mathrm{L}})_{ij} \lesssim M\log n$ for all i, j, which occurs with overwhelming probability by Lemma D.1 and a union bound. Now, we employ Lemma D.3 with

$B := M \log n$ and $\nu := Mn\lambda_{\max}$ to obtain

$$\mathbb{P}\left(\left\|\hat{\mathbf{\Lambda}}^{\mathrm{L}} - \bar{\mathbf{\Lambda}}^{\mathrm{L}}\right\|_2 \geq 4\sqrt{Mn\lambda_{\max}} + t\right) \leq n \exp\left(-\frac{t^2}{c(M\log n)^2}\right).$$

Setting $t = M \log^{3/2} n$, we have that

$$\left\|\hat{\mathbf{\Lambda}}^{\mathrm{L}} - \bar{\mathbf{\Lambda}}^{\mathrm{L}}\right\|_2 \overset{\mathbb{P}}{\lesssim} \sqrt{Mn\lambda_{\max}} + M\log^{3/2} n \lesssim \sqrt{Mn\lambda_{\max}}$$

where the final inequality follows from Assumption 4. We obtain an analogous bound for $\left\|\hat{\mathbf{\Lambda}}^{\mathrm{U}} - \bar{\mathbf{\Lambda}}^{\mathrm{U}}\right\|_2$ and combine the with the triangle inequality:

$$\left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_2 \leq \left\|\hat{\mathbf{\Lambda}}^{\mathrm{L}} - \bar{\mathbf{\Lambda}}^{\mathrm{L}}\right\|_2 + \left\|\hat{\mathbf{\Lambda}}^{\mathrm{U}} - \bar{\mathbf{\Lambda}}^{\mathrm{U}}\right\|_2 \overset{\mathbb{P}}{\lesssim} \sqrt{Mn\lambda_{\max}}.$$

We now establish a bound on $\left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_1$. We condition on the event $|\hat{\mathbf{\Lambda}}_{ij} - \bar{\mathbf{\Lambda}}_{ij}| \lesssim M \log n$ for all i, j, which occurs with overwhelming probability due to Lemma D.1 and a union bound, and note that we have $\sum_{j=1}^{n} \mathbb{E}(\hat{\mathbf{\Lambda}}_{ji} - \bar{\mathbf{\Lambda}}_{ji})^2 \leq Mn\lambda_{\max}$. Then, by the classical Bernstein inequality, we have for any $t > 0$,

$$\mathbb{P}\left\{\sum_{j=1}^{n}\left|\hat{\mathbf{\Lambda}}_{ji} - \bar{\mathbf{\Lambda}}_{ji}\right| \geq t\right\} \leq 2\exp\left\{\frac{-t^2}{2\left(Mn\lambda_{\max} + tM\log n/3\right)}\right\},$$

and setting $t = \sqrt{nM\lambda_{\max}\log n}$, we obtain

$$\sum_{j=1}^{n}\left|\hat{\mathbf{\Lambda}}_{ji} - \bar{\mathbf{\Lambda}}_{ji}\right| \overset{\mathbb{P}}{\lesssim} \sqrt{nM\lambda_{\max}\log n}.$$

A union bound establishes that

$$\left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_1 \overset{\mathbb{P}}{\lesssim} \sqrt{nM\lambda_{\max}\log n},$$

which establishes Proposition D.2.

D.6.2 Proof of Proposition D.3

Proposition D.3 follows from an application of Weyl's inequality. We have

$$\sigma_1(\hat{\mathbf{\Lambda}}) \leq \sigma_1(\bar{\mathbf{\Lambda}}) + |\sigma_1(\hat{\mathbf{\Lambda}}) - \sigma_1(\bar{\mathbf{\Lambda}})| \leq \sigma_1(\bar{\mathbf{\Lambda}}) + \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_2 \overset{\mathbb{P}}{\lesssim} \sqrt{M\sigma_1(\mathbf{\Sigma})} + \sqrt{Mn\lambda_{\max}} \lesssim \sqrt{M\sigma_1(\mathbf{\Sigma})}$$

since $\sigma_1(\bar{\mathbf{\Lambda}}) = \sqrt{M\sigma_1(\mathbf{\Sigma})} \gtrsim \sqrt{M\delta} \gtrsim \sqrt{Mn\lambda_{\max}}$. Similarly, we have

$$\sigma_d(\hat{\mathbf{\Lambda}}) \geq \sigma_d(\bar{\mathbf{\Lambda}}) - |\sigma_1(\hat{\mathbf{\Lambda}}) - \sigma_1(\bar{\mathbf{\Lambda}})| \geq \sigma_d(\bar{\mathbf{\Lambda}}) - \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_2 \overset{\mathbb{P}}{\gtrsim} \sqrt{M\sigma_d(\mathbf{\Sigma})} - \sqrt{Mn\lambda_{\max}} \gtrsim \sqrt{M\sigma_d(\mathbf{\Sigma})},$$

which establishes the proposition.

D.6.3 Proof of Proposition D.4

We begin by constructing matrices $\bar{\mathbf{Q}}$ and $\bar{\mathbf{E}}$, via a symmetric dilation trick, such that the spectral norms of $\mathbf{Q}^\top(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})\mathbf{R}$ and $\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}$ coincide, and then apply a classical ϵ-net argument to the spectral norm of $\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}$, following the proof of Lemma D.1 in [94].

First, we set $\bar{\mathbf{E}} := \mathcal{D}(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})$, where $\mathcal{D}$ is the dilation operator (see Section D.5.1.2) and $\bar{\mathbf{Q}} = (\mathbf{Q}\ \mathbf{R})$, and observe that

$$\left\|\mathbf{Q}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\mathbf{R}\right\|_2 = \left\|\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}\right\|_2 = \max_{\|v\|_2\le 1}\left|v^\top\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}v\right|$$

where the second equality follows from the Courant-Fischer min-max theorem. Now, let $\mathcal{S}_\epsilon^{d-1}$ be an ϵ-net of the $d-1$–dimensional unit sphere $\mathcal{S}^{d-1} := \{v : \|v\|_2 = 1\}$. By definition, for any $v \in \mathcal{S}^{d-1}$, there exists some $w(v) \in \mathcal{S}_\epsilon^{d-1}$ such that $\|v - w(v)\|_2 < \epsilon$ and

$$\begin{aligned}\left\|\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}\right\|_2 &= \max_{\|v\|_2\le 1}\left|v^\top\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}v\right| \\ &= \max_{\|v\|_2\le 1}\left|\left\{v^\top - w(v) + w(v)\right\}^\top\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}\left\{v - w(v) + w(v)\right\}\right| \\ &\le \left(\epsilon^2 + 2\epsilon\right)\left\|\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}\right\|_2 + \max_{w\in\mathcal{S}_\epsilon^{d-1}}\left|w^\top\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}w\right|.\end{aligned}$$

With $\epsilon = 1/3$, we have

$$\left\|\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}\right\|_2 \le \frac{9}{2}\max_{w\in\mathcal{S}_\epsilon^{d-1}}\left|w^\top\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}w\right|.$$

Now, $\mathcal{S}_{1/3}^{d-1}$ can be selected so that its cardinality can be upper bounded by $|\mathcal{S}_{1/3}^{d-1}| \le 18^d$ (see, for example, Pollard [128]). For a fixed $w \in \mathcal{S}_{1/3}^{d-1}$, we let $z = \bar{\mathbf{Q}}w$ and note that since $\mathcal{S}_{1/3}^{d-1} \subset \mathcal{S}^{d-1}$, that $\|z\|_2 \le 1$, and

$$\left|w^\top\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}w\right| = \left|\sum_{i=1}^{n(M+1)}\sum_{j=1}^{n(M+1)}\bar{e}_{ij}z_iz_j\right| = 2\left|\sum_{i=1}^{n}\sum_{j=1}^{nM}e_{ij}z_iz_{n+j}\right|$$

Now, over the event that entries $e_{ij} \lesssim M\log n$, for all i, j, which occurs which overwhelming probability by Lemma D.1, Hoeffding's inequality and a union bound over $w \in \mathcal{S}_{1/3}^{d-1}$ gives

$$\begin{aligned}\mathbb{P}\left\{\left\|\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}\right\|_2 > t\right\} &\le \sum_{w\in\mathcal{S}_{1/3}^{d-1}}\mathbb{P}\left(\left|w^\top\bar{\mathbf{Q}}^\top\bar{\mathbf{E}}\bar{\mathbf{Q}}w\right| > \frac{2t}{9}\right) \\ &= \sum_{w\in\mathcal{S}_{1/3}^{d-1}}\mathbb{P}\left\{\left|\sum_{i=1}^{n}\sum_{j=1}^{nM}e_{ij}z_iz_{n+j}\right| > \frac{t}{9}\right\} \\ &\le 2\cdot 18^d\exp\left\{-\frac{2t^2}{(9cM\log n)^2}\right\} \\ &= 2\cdot\exp\left\{d\log(18) - \frac{2t^2}{(9cM\log n)^2}\right\}\end{aligned}$$

Setting $t = M \log^{3/2} n$ gives

$$\left\|\mathbf{Q}^\top \left(\hat{\mathbf{\Lambda}} - \tilde{\mathbf{\Lambda}}\right) \mathbf{R}\right\|_2 = \left\|\bar{\mathbf{Q}}^\top \bar{\mathbf{E}} \bar{\mathbf{Q}}\right\|_2 \overset{\mathbb{P}}{\lesssim} M \log^{3/2} n,$$

completing the proof.

D.6.4 Proof of Proposition D.5

Denote the singular value decomposition of $\bar{\mathbf{U}}^\top \hat{\mathbf{U}}$ by $\mathbf{\Omega}_1 \mathbf{\Xi} \mathbf{\Omega}_2^\top$, where $\mathbf{\Xi} = \operatorname{diag}(\xi_1, \ldots, \xi_d)$, and let $\mathbf{W} := \mathbf{\Omega}_1 \mathbf{\Omega}_2^\top$. The principal angles $\{\theta_i\}_{i=1}^d$ between the column spaces of $\bar{\mathbf{U}}$ and $\hat{\mathbf{U}}$ are defined by $\xi_i = \cos(\theta_i)$, and by the Wedin sinΘ theorem, we have

$$\begin{aligned}\left\|\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \mathbf{W}\right\|_2 &= \|\mathbf{\Xi} - \mathbf{I}\|_2 = \max_{i \in [d]} |1 - \xi_i| = \max_{i \in [d]} |1 - \cos\theta_i| \leq \max_{i \in [d]} |1 - \cos^2\theta_i| \\ &= \max_{i \in [d]} \sin^2\theta_i \lesssim \frac{\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\|_2^2}{(\sigma_d(\bar{\mathbf{\Lambda}}) - \sigma_{d+1}(\bar{\mathbf{\Lambda}}))^2} \overset{\mathbb{P}}{\lesssim} \frac{Mn\lambda_{\max}}{M\delta^2} = \frac{n\lambda_{\max}}{\delta^2} \lesssim \frac{\sqrt{n\lambda_{\max}}}{\delta}.\end{aligned} \tag{D.15}$$

We apply the Wedin sinΘ theorem again to obtain a bound which we will require later:

$$\begin{aligned}\left\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right\|_2 \vee \left\|\hat{\mathbf{V}}\hat{\mathbf{V}}^\top - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right\|_2 &= \left\|\sin\Theta\left(\hat{\mathbf{U}}, \bar{\mathbf{U}}\right)\right\|_2 \vee \left\|\sin\Theta\left(\hat{\mathbf{V}}, \bar{\mathbf{V}}\right)\right\|_2 \\ &\lesssim \frac{\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\|_2}{\sigma_d(\bar{\mathbf{\Lambda}}) - \sigma_{d+1}(\bar{\mathbf{\Lambda}})} \\ &\overset{\mathbb{P}}{\lesssim} \frac{\sqrt{n\lambda_{\max}}}{\delta}.\end{aligned}$$

We now establish a bound on $\left\|\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right\|_2$. We start by showing that

$$\begin{aligned}\left\|\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right\|_2 &= \operatorname*{arg\,max}_{x : \|x\|_2 \leq 1} x^\top \left(\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right) x \\ &= \operatorname*{arg\,max}_{x : \|x\|_2 \leq 1} \sum_{i,j=1}^d x_i x_j \left(\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right)_{ij} \\ &\leq \operatorname*{arg\,max}_{x : \|x\|_2 \leq 1} \sum_{i,j=1}^d (1 + \bar{s}_i) x_i (1 + \bar{s}_j^{-1}) x_j \left(\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right)_{ij} \\ &= \operatorname*{arg\,max}_{x : \|x\|_2 \leq 1} \sum_{i,j=1}^d x^\top \left[\left(\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right) + \bar{\mathbf{S}} \left(\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right) \hat{\mathbf{S}}^{-1}\right] x \\ &= \left\|\left(\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right) + \bar{\mathbf{S}} \left(\bar{\mathbf{U}}^\top \hat{\mathbf{U}} - \bar{\mathbf{V}}^\top \hat{\mathbf{V}}\right) \hat{\mathbf{S}}^{-1}\right\|_2,\end{aligned}$$

and then we employ the decomposition

$$\begin{aligned}
\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \bar{\mathbf{V}}^\top\hat{\mathbf{V}} + \bar{\mathbf{S}}\left(\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \bar{\mathbf{V}}^\top\hat{\mathbf{V}}\right)\hat{\mathbf{S}}^{-1}& \\
&= \left[\bar{\mathbf{U}}^\top\hat{\mathbf{U}}\hat{\mathbf{S}} - \bar{\mathbf{S}}\bar{\mathbf{V}}^\top\hat{\mathbf{V}} + \bar{\mathbf{S}}\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \bar{\mathbf{V}}\hat{\mathbf{V}}\right]\hat{\mathbf{S}}^{-1} \\
&= \left[\bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\hat{\mathbf{V}} + \bar{\mathbf{V}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)^\top\hat{\mathbf{U}}\right]\hat{\mathbf{S}}^{-1} \\
&= \bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\left(\hat{\mathbf{V}} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\hat{\mathbf{V}}\right)\hat{\mathbf{S}}^{-1} + \bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\hat{\mathbf{V}}\hat{\mathbf{S}}^{-1} \\
&\quad + \bar{\mathbf{V}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)^\top\left(\hat{\mathbf{U}} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top\hat{\mathbf{U}}\right)\hat{\mathbf{S}}^{-1} + \bar{\mathbf{V}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)^\top\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\hat{\mathbf{U}}\hat{\mathbf{S}}^{-1}.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
\left\|\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \bar{\mathbf{V}}^\top\hat{\mathbf{V}}\right\|_2 &\le \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_2\left(\left\|\hat{\mathbf{V}}\hat{\mathbf{V}}^\top - \bar{\mathbf{V}}\bar{\mathbf{V}}\right\|_2 + \left\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \bar{\mathbf{U}}\bar{\mathbf{U}}\right\|_2\right)\left\|\hat{\mathbf{S}}^{-1}\right\|_2 \\
&\qquad \left\|\bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\bar{\mathbf{V}}\right\|_2\left\|\hat{\mathbf{S}}^{-1}\right\|_2 + \left\|\bar{\mathbf{V}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)^\top\bar{\mathbf{U}}\right\|_2\left\|\hat{\mathbf{S}}^{-1}\right\|_2 \\
&\overset{\mathbb{P}}{\lesssim} \sqrt{Mn\lambda_{\max}}\cdot\frac{\sqrt{n\lambda_{\max}}}{\delta}\cdot\frac{1}{\sigma_d(\bar{\mathbf{\Lambda}})} + \frac{M\log^{3/2}n}{\sigma_d(\bar{\mathbf{\Lambda}})} \\
&= \frac{n\lambda_{\max}}{\delta^2} + \frac{\sqrt{M}\log^{3/2}n}{\delta} \\
&\lesssim \frac{\sqrt{n\lambda_{\max}}}{\delta}.
\end{aligned}$$

Combining this with (D.15), we have

$$\left\|\bar{\mathbf{V}}^\top\hat{\mathbf{V}} - \mathbf{W}\right\|_2 \le \left\|\bar{\mathbf{V}}^\top\hat{\mathbf{V}} - \bar{\mathbf{U}}^\top\hat{\mathbf{U}}\right\|_2 + \left\|\bar{\mathbf{U}}^\top\hat{\mathbf{U}} - \mathbf{W}\right\|_2 \overset{\mathbb{P}}{\lesssim} \frac{\sqrt{n\lambda_{\max}}}{\delta}.$$

D.6.5 Proof of Proposition D.6

We begin by decomposing $\mathbf{W}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W}$ as

$$\begin{aligned}
\mathbf{W}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W} &= \left(\mathbf{W} - \bar{\mathbf{U}}^\top\hat{\mathbf{U}}\right)\hat{\mathbf{S}} + \bar{\mathbf{S}}\left(\mathbf{V}^\top\hat{\mathbf{V}} - \mathbf{W}\right) + \bar{\mathbf{U}}^\top\hat{\mathbf{U}}\hat{\mathbf{S}} - \bar{\mathbf{S}}\bar{\mathbf{V}}^\top\hat{\mathbf{V}} \\
&= \left(\mathbf{W} - \bar{\mathbf{U}}^\top\hat{\mathbf{U}}\right)\hat{\mathbf{S}} + \bar{\mathbf{S}}\left(\mathbf{V}^\top\hat{\mathbf{V}} - \mathbf{W}\right) + \bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\hat{\mathbf{V}} \\
&= \left(\mathbf{W} - \bar{\mathbf{U}}^\top\hat{\mathbf{U}}\right)\hat{\mathbf{S}} + \bar{\mathbf{S}}\left(\mathbf{V}^\top\hat{\mathbf{V}} - \mathbf{W}\right) + \bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\left(\hat{\mathbf{V}}\hat{\mathbf{V}}^\top - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right)\hat{\mathbf{V}} \\
&\quad + \bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\hat{\mathbf{V}},
\end{aligned}$$

and therefore we have that

$$\begin{aligned}
\left\|\mathbf{W}\hat{\mathbf{S}} - \bar{\mathbf{S}}\mathbf{W}\right\|_2 &\le \left\|\mathbf{W} - \bar{\mathbf{U}}^\top\hat{\mathbf{U}}\right\|_2\left\|\hat{\mathbf{\Lambda}}\right\|_2 + \left\|\mathbf{W} - \bar{\mathbf{V}}^\top\hat{\mathbf{V}}\right\|_2\|\bar{\mathbf{\Lambda}}\|_2 \\
&\quad + \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_2\left\|\hat{\mathbf{V}}\hat{\mathbf{V}}^\top - \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right\|_2 + \left\|\bar{\mathbf{U}}^\top\left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right)\bar{\mathbf{V}}\right\|_2 \\
&\overset{\mathbb{P}}{\lesssim} \frac{\sqrt{n\lambda_{\max}}\kappa}{\delta} + \frac{\sqrt{M}n\lambda_{\max}}{\delta} + M\log^{3/2}n \\
&\lesssim M\log^{3/2}n,
\end{aligned}$$

which completes the proof.

D.6.6 Proof of Proposition D.7

A key tool in proving Proposition D.7 is a theorem due [93], providing entrywise eigenvector bounds for random matrices. The original statement is given for the eigenvectors of symmetric random matrices with row and column-wise independence. We state a generalisation for the singular vectors of rectangular matrices with *block-wise* independence structure. The extension to block-wise independence structure has been handled in [82] (see Proposition 2.1(b) of that paper), although the exposition of the results in this paper is more complicated. For this reason, we choose to state the result due to [93] with this generalisation, which can be seen by following through the relevant parts of their proof.

Lemma D.6 (A slight generalisation of Theorem 2.1 of [93])**.** *Let* $\mathbf{M}_0$ *be an* $n_1 \times n_2$ *real-valued random matrix. Define* $n_0 = n_1 + n_2$ *and let* $\pi_1 = \sqrt{2n_1/n_0}$ *and* $\pi_2 = \sqrt{2n_2/n_0}$*. Define* $\kappa_0 := \sigma_1(\mathbb{E}\mathbf{M}_0)/\sigma_d(\mathbb{E}\mathbf{M}_0)$*,* $\delta_0 = \sigma_d(\mathbb{E}\mathbf{M}_0) - \sigma_{d+1}(\mathbb{E}\mathbf{M}_0)$*. Suppose there exists some* $\gamma > 0$ *and a function* $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ *which is continuous and non-decreasing on* $\mathbb{R}_+$*, with* $\varphi(0) = 0$ *and* $\varphi(x)/x$ *non-increasing on* $\mathbb{R}_+$*, such that the following conditions hold:*

B1 (Incoherence)**.** $\|\mathbb{E}\mathbf{M}_0\|_{2,\infty} \vee \left\|\mathbb{E}\mathbf{M}_0^\top\right\|_{2,\infty} \le \gamma\delta_0$.

B2 (Block-wise independence)**.** *Assume that for any* $k \in [n_1], \ell \in [n_2]$*, there exists* $\mathcal{N}_k^1 \subset [n_1]$ *and* $\mathcal{N}_\ell^2 \subset [n_2]$*, such that the* k*th row of* $\mathbf{M}_0$ *is independent of the columns* $\{j : j \notin \mathcal{N}_k^1\}$*, and the* ℓ*th column of* $\mathbf{M}_0$ *is independent of the rows* $\{i : i \notin \mathcal{N}_\ell^2\}$*. Let* $m_0 = \max_{k,\ell}\left\{|\mathcal{N}_k^1| \vee |\mathcal{N}_\ell^2|\right\}$ *and assume* $m_0 \lesssim \delta_0$*.*

B3 (Spectral norm concentration)**.** $\kappa_0 \max\{\gamma, \varphi(\gamma)\} \lesssim 1$ *and* $\mathbb{P}\left(\|\mathbf{M}_0 - \mathbb{E}\mathbf{M}_0\|_2 > \gamma\Delta\right) \le \eta_0$ *for some* $\eta_0 \in (0,1)$*.*

B4. *[Row and column concentration] There exists some* $\eta_1 \in (0,1)$ *such that for any matrices* $\mathbf{Q} \in \mathbb{R}^{n_1 \times d}, \mathbf{R} \in \mathbb{R}^{n_2 \times d}$ *and* $i \in [n_1], j \in [n_2]$*,*

$$\mathbb{P}\left\{\left\|(\mathbf{M}_0 - \mathbb{E}\mathbf{M}_0)_{\cdot,i}\mathbf{R}\right\|_2 \le \delta_0 b_\infty \varphi\left(\frac{b_{\mathrm{F}}}{\sqrt{n_0}b_\infty}\right)\right\} \ge 1 - \frac{\eta_1}{n_0},$$

and

$$\mathbb{P}\left\{\left\|(\mathbf{M}_0 - \mathbb{E}\mathbf{M}_0)_{j,\cdot}\mathbf{Q}\right\|_2 \le \delta_0 b_\infty \varphi\left(\frac{b_{\mathrm{F}}}{\sqrt{n_0}b_\infty}\right)\right\} \ge 1 - \frac{\eta_1}{n_0}$$

where $b_\infty := \pi_1\|\mathbf{Q}\|_{2,\infty} \vee \pi_2\|\mathbf{R}\|_{2,\infty}$*, and* $b_{\mathrm{F}} := \left(\pi_1\|\mathbf{Q}\|_{\mathrm{F}}^2 + \pi_2\|\mathbf{R}\|_{\mathrm{F}}^2\right)^{1/2}$*.*

Let $\hat{\mathbf{U}}_0, \mathbf{U}_0$ *(respectively* $\hat{\mathbf{V}}_0, \mathbf{V}_0$*) be the matrices containing the left (respectively, right) singular vectors corresponding to the* d *leading singular values of* $\mathbf{M}_0$ *and* $\mathbb{E}\mathbf{M}_0$*. Then, with* *probability at least* $1 - \eta_0 - 2\eta_1$*, we have*

$$\begin{aligned}\pi_1\|\hat{\mathbf{U}}_0\|_{2,\infty} \vee \pi_2\|\hat{\mathbf{V}}_0\|_{2,\infty} &\lesssim \{\kappa_0 + \varphi(1)\}\left(\pi_1\|\mathbf{U}_0\|_{2,\infty} \vee \pi_2\|\mathbf{V}_0\|_{2,\infty}\right) \\ &\quad + \gamma(\pi_1\|\mathbb{E}\mathbf{M}_0\|_{2,\infty} \vee \pi_2\|(\mathbb{E}\mathbf{M}_0)^\top\|_{2,\infty})/\delta_0; \\ \pi_1\|\hat{\mathbf{U}}_0\mathbf{O} - \mathbf{U}_0\|_{2,\infty} \vee \pi_2\|\hat{\mathbf{V}}_0\mathbf{O} - \mathbf{V}_0\|_{2,\infty} &\lesssim \left[\kappa_0\{\kappa_0 + \varphi(1)\}\{\gamma + \varphi(\gamma)\} + \varphi(1)\right]\left(\pi_1\|\mathbf{U}_0\|_{2,\infty} \vee \pi_2\|\mathbf{V}_0\|_{2,\infty}\right) \\ &\quad + \gamma(\pi_1\|\mathbb{E}\mathbf{M}_0\|_{2,\infty} \vee \pi_2\|(\mathbb{E}\mathbf{M}_0)^\top\|_{2,\infty})/\delta_0.\end{aligned}$$

The following is an adaptation of Lemma D.2 of [94] (see also Lemma 7 of [93]) who showed an analogous result for Bernoulli random variables.

Lemma D.7. *Let* $Y_i \sim \text{Poisson}(\lambda_i)$ *independently for all* $i = 1, \ldots, n$*, and suppose* $\mathbf{Q}$ *is a deterministic matrix. The* Q_i *denote the ith row of* $\mathbf{Q}$*, and set* $\lambda_{\max} := \max_{i\in[n]} \lambda_i$*. Then for any* $\alpha > 0$*,*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n (Y_i - \lambda_i)\, Q_i\right\| > \frac{(2+\alpha)n\lambda_{\max}\, \|\mathbf{Q}\|_{2,\infty}}{1 \vee \log\left(\sqrt{n}\, \|\mathbf{Q}\|_{2,\infty} / \|\mathbf{Q}\|_{\mathrm{F}}\right)}\right\} \leq 2de^{-\alpha n\lambda_{\max}}.$$

We omit the proof of Lemma D.7, which is identical to the proof of Lemma D.2 of [94] with the Bernoulli moment generating function with the Poisson moment generating function.

With these tools to hand, we begin by obtaining a bound on $\|\hat{\mathbf{U}}\|_{2,\infty}$ using Lemma D.6, with $\mathbf{M}_0 := \hat{\mathbf{\Lambda}}$. We set $\gamma := \sqrt{n\lambda_{\max}}/\delta$ and

$$\varphi(x) := \frac{n\lambda_{\max}}{\delta\{1 \vee \log(1/x)\}}.$$

First observe that $n_0 = n + nM \asymp nM$ and $\pi_1 \asymp M^{-1/2}$ and $\pi_2 \asymp 1$, and that $\kappa_0 = \kappa$ and $\delta_0 = \sqrt{M}\delta$. **B1** holds since $\|\bar{\mathbf{\Lambda}}\|_{2,\infty} \leq \sqrt{n}\lambda_{\max} \lesssim \sqrt{n\lambda_{\max}}$ since $\lambda_{\max} \lesssim 1$ by Assumption 1. Using Assumptions 3 and 4, we have

$$M \lesssim \frac{n\lambda_{\max}}{\log^3 n} \lesssim \frac{\delta\log(\delta/\sqrt{n\lambda_{\max}})}{\kappa\log^3 n} \lesssim \frac{\delta\log n}{\kappa\log^3 n} \lesssim \delta,$$

and therefore **B2** holds. **B3** holds from Proposition D.2, and observing that by Assumption 3, $\kappa_0\max\{\gamma, \phi(\gamma)\} \lesssim 1$.

To see that **B4** holds, note that each row and column of $M^{-1}(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}})$ contains independent Poisson random variables with means not exceeding $n\lambda_{\max}/M$. Then for $\mathbf{Q} \in \mathbb{R}^{n\times d}$, $\mathbf{R} \in \mathbb{R}^{nM\times d}$, setting $\alpha = \log n / n\lambda_{\max}$ in Lemma D.7 implies that

$$\begin{aligned}
\left\| \left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}} \right)_{i,\cdot} \mathbf{R} \right\|_{2,\infty} &= M \left\| \frac{1}{M} \left(\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}} \right)_{i,\cdot} \mathbf{R} \right\|_{2,\infty} \\
&\overset{\mathbb{P}}{\lesssim} M \cdot \frac{\left(n\lambda_{\max}/M + \log n \right) \|\mathbf{R}\|_{2,\infty}}{1 \vee \log \left(\frac{\sqrt{n_0}\|\mathbf{R}\|_{2,\infty}}{\|\mathbf{R}\|_{\mathrm{F}}} \right)} \\
&\lesssim \frac{n\lambda_{\max} \|\mathbf{R}\|_{2,\infty}}{1 \vee \log \left(\frac{\sqrt{n_0}\|\mathbf{R}\|_{2,\infty}}{\|\mathbf{R}\|_{\mathrm{F}}} \right)} \\
&= \delta \|\mathbf{R}\|_{2,\infty} \varphi \left(\frac{\|\mathbf{R}\|_{\mathrm{F}}}{\sqrt{n_0} \|\mathbf{R}\|_{2,\infty}} \right) \\
&\le \delta_0 b_\infty \varphi \left(\frac{b_{\mathrm{F}}}{\sqrt{n_0} b_\infty} \right).
\end{aligned}$$

Similarly, setting $\alpha = M \log n / n\lambda_{\max}$ we have

$$\left\| \left(\tilde{\mathbf{A}} - \tilde{\mathbf{\Lambda}} \right)_{\cdot,i}^{\top} \mathbf{Q} \right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \frac{\sqrt{M} n\lambda_{\max} \|\mathbf{Q}\|_{2,\infty}}{1 \vee \log \left(\sqrt{n_0} \|\mathbf{Q}\|_{2,\infty} / \|\mathbf{Q}\|_{\mathrm{F}} \right)} \le \delta_0 b_\infty \varphi \left(\frac{\sqrt{n} b_\infty}{b_{\mathrm{F}}} \right),$$

which establishes **B4**. Having established **B1-B4**, we are ready to apply Lemma D.6:

$$\left\| \hat{\mathbf{U}} \right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \sqrt{M} \left\{ \kappa_0 + \varphi(1) \right\} \left(\pi_1 \|\mathbf{U}\|_{2,\infty} \vee \pi_2 \|\mathbf{V}\|_{2,\infty} \right) + \sqrt{M} \gamma \left(\pi_1 \left\| \bar{\mathbf{\Lambda}} \right\|_{2,\infty} \vee \pi_2 \left\| \bar{\mathbf{\Lambda}}^{\top} \right\|_{2,\infty} \right) / \delta_0$$

We have

$$\begin{aligned}
\|\bar{\mathbf{V}}\|_{2,\infty} = \|\bar{\mathbf{\Lambda}} \bar{\mathbf{U}} \bar{\mathbf{S}}^{-1}\|_{2,\infty} \le \|\bar{\mathbf{\Lambda}}\|_\infty \|\bar{\mathbf{U}}\|_{2,\infty} \|\bar{\mathbf{S}}\|_2^{-1} \lesssim \frac{n\lambda_{\max} \mu \sqrt{d/n}}{\sigma_d^{1/2}(\mathbf{\Sigma})} \\
\lesssim \frac{n\lambda_{\max} \mu \sqrt{d/n}}{\sqrt{M\delta}} \lesssim \sqrt{\frac{d}{nM}} \mu \log n.
\end{aligned} \tag{D.16}$$

where we used Assumption 3 in the final inequality. Therefore

$$\pi_1 \left\| \bar{\mathbf{U}} \right\|_{2,\infty} \vee \pi_2 \left\| \bar{\mathbf{V}} \right\|_{2,\infty} \le \sqrt{\frac{d}{nM}} \mu \log n,$$

and we have

$$\kappa = \frac{\sigma_1^{1/2}(\mathbf{\Sigma})}{\sigma_d^{1/2}(\mathbf{\Sigma})} \lesssim \frac{n\lambda_{\max}}{\delta} = \varphi(1)$$

and so the first term satisfies

$$\begin{aligned}
\sqrt{M} \left\{ \kappa_0 + \varphi(1) \right\} \left(\pi_1 \|\mathbf{U}\|_{2,\infty} \vee \pi_2 \|\mathbf{V}\|_{2,\infty} \right) &\lesssim \sqrt{M} \varphi(1) \left(\pi_1 \|\mathbf{U}\|_{2,\infty} \vee \pi_2 \|\mathbf{V}\|_{2,\infty} \right) \\
&\lesssim \sqrt{M} \cdot \frac{n\lambda_{\max}}{\delta} \cdot \sqrt{\frac{d}{nM}} \mu \log n \\
&= \frac{\mu \lambda_{\max} \sqrt{nd} \log n}{\delta}.
\end{aligned}$$

To control the second term, we first observe that

$$\begin{aligned}
\pi_1 \left\|\bar{\mathbf{\Lambda}}\right\|_{2,\infty} &\leq M^{-1/2} \left\|\bar{\mathbf{U}}\bar{\mathbf{U}}^\top \bar{\mathbf{\Lambda}}\right\|_{2,\infty} + M^{-1/2} \left\|\left(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}\right) \bar{\mathbf{\Lambda}}\right\|_{2,\infty} \\
&\lesssim M^{-1/2} \left\|\bar{\mathbf{U}}\right\|_{2,\infty} \left\|\bar{\mathbf{\Lambda}}\right\|_2 + M^{-1/2} \max_{i\in[n]} \sup_{t\in(0,1]} r_i(t) \\
&\lesssim \sqrt{\frac{d}{Mn}}\mu \cdot \sqrt{M}\kappa\delta + \mu\sqrt{d\lambda_{\max}} \log^{5/2} n \\
&\lesssim \mu\delta\sqrt{d\lambda_{\max}}
\end{aligned}$$

where the final inequality follows from $\kappa \lesssim \log n \leq \sqrt{n}$ and $\lambda_{\max} \lesssim 1$. Similarly we obtain $\pi_2\|\bar{\mathbf{\Lambda}}^\top\|_{2,\infty} \lesssim \mu\delta\sqrt{d\lambda_{\max}} \log n$ using (D.16), and therefore

$$\pi_1 \left\|\bar{\mathbf{\Lambda}}\right\|_{2,\infty} \vee \pi_2 \left\|\bar{\mathbf{\Lambda}}^\top\right\|_{2,\infty} \lesssim \mu\delta\sqrt{d\lambda_{\max}} \log n.$$

We then have

$$\begin{aligned}
\sqrt{M}\gamma \left(\pi_1 \left\|\bar{\mathbf{\Lambda}}\right\|_{2,\infty} \vee \pi_2 \left\|\bar{\mathbf{\Lambda}}^\top\right\|_{2,\infty}\right) /\delta_0 &\lesssim \sqrt{M} \cdot \frac{\sqrt{n\lambda_{max}}}{\delta} \cdot \mu\delta\sqrt{d\lambda_{\max}} \log n \cdot \frac{1}{\sqrt{M}\delta} \\
&\lesssim \frac{\mu\sqrt{nd}\lambda_{\max} \log n}{\delta}.
\end{aligned}$$

Combining these bounds, we obtain

$$\left\|\hat{\mathbf{U}}\right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \frac{\mu\sqrt{nd}\lambda_{\max} \log n}{\delta}.$$

We now apply Lemma D.6 with $\mathbf{M}_0 = \hat{\mathbf{\Lambda}}^{(m)}$. We set γ and $\varphi(x)$ as before and verify Assumptions **B1-B4** in the same way. By analogous calculations to the above, we obtain the bound

$$\left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} \overset{\mathbb{P}}{\lesssim} \frac{\mu\sqrt{nd}\lambda_{\max} \log n}{\delta}.$$

The final bound is shown in the same way, requiring the additional observation that $\kappa \left\{\gamma \vee \varphi(\gamma)\right\} \lesssim 1$ which follows from Assumption 3, and we obtain

$$\begin{aligned}
&\pi_1\|\hat{\mathbf{U}}_0\mathbf{O} - \mathbf{U}_0\|_{2,\infty} \vee \pi_2\|\hat{\mathbf{V}}_0\mathbf{O} - \mathbf{V}_0\|_{2,\infty} \\
&\quad \overset{\mathbb{P}}{\lesssim} \left[\kappa_0 \left\{\kappa_0 + \varphi(1)\right\} \left\{\gamma + \varphi(\gamma)\right\} + \varphi(1)\right] \left(\pi_1\|\mathbf{U}_0\|_{2,\infty} \vee \pi_2\|\mathbf{V}_0\|_{2,\infty}\right) + \gamma \left(\pi_1 \left\|\bar{\mathbf{\Lambda}}\right\|_{2,\infty} \vee \pi_2 \left\|\bar{\mathbf{\Lambda}}^\top\right\|_{2,\infty}\right) /\delta_0 \\
&\quad \lesssim \varphi(1) \left(\pi_1\|\mathbf{U}_0\|_{2,\infty} \vee \pi_2\|\mathbf{V}_0\|_{2,\infty}\right) + \gamma \left(\pi_1 \left\|\bar{\mathbf{\Lambda}}\right\|_{2,\infty} \vee \pi_2 \left\|\bar{\mathbf{\Lambda}}^\top\right\|_{2,\infty}\right) /\delta_0 \\
&\quad \lesssim \frac{\mu\sqrt{nd}\lambda_{\max} \log n}{\delta}
\end{aligned}$$

D.6.7 Proof of Proposition D.8

We show (D.9) using a simple application of Wedin's inequality:

$$\begin{aligned}\left\|\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right\|_2 &= \left\|\sin\Theta\left(\hat{\mathbf{U}}^{(m)}, \bar{\mathbf{U}}\right)\right\|_2 \lesssim \frac{\left\|\hat{\boldsymbol{\Lambda}}^{(m)} - \bar{\boldsymbol{\Lambda}}\right\|_2}{\sigma_d\left(\bar{\boldsymbol{\Lambda}}\right) - \sigma_{d+1}\left(\bar{\boldsymbol{\Lambda}}\right)} \leq \frac{\left\|\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}}\right\|_2}{\sigma_d\left(\bar{\boldsymbol{\Lambda}}\right) - \sigma_{d+1}\left(\bar{\boldsymbol{\Lambda}}\right)}\\ &\overset{\mathbb{P}}{\lesssim} \frac{\sqrt{Mn\lambda_{\max}}}{\sqrt{M}\delta} = \frac{\sqrt{n\lambda_{\max}}}{\delta}.\end{aligned}$$

The proof of (D.10) requires a more delicate argument. We apply Wedin's theorem to obtain

$$\left\|\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^\top - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\right\|_2 = \left\|\sin\Theta\left(\hat{\mathbf{U}}^{(m)}, \hat{\mathbf{U}}\right)\right\|_2 \lesssim \frac{\left\|\left(\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}}\right)^\top \hat{\mathbf{U}}^{(m)}\right\|_2 \vee \left\|\left(\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}}\right)\hat{\mathbf{V}}^{(m)}\right\|_2}{\sigma_d\left(\hat{\boldsymbol{\Lambda}}\right) - \sigma_{d+1}\left(\hat{\boldsymbol{\Lambda}}\right)}. \tag{D.17}$$

By Weyl's inequality

$$\sigma_d\left(\hat{\boldsymbol{\Lambda}}\right) \geq \sigma_d\left(\bar{\boldsymbol{\Lambda}}\right) + \left\|\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}}\right\|_2 \overset{\mathbb{P}}{\gtrsim} \sigma_d\left(\bar{\boldsymbol{\Lambda}}\right).$$

and

$$\sigma_{d+1}\left(\hat{\boldsymbol{\Lambda}}\right) \leq \sigma_{d+1}\left(\bar{\boldsymbol{\Lambda}}\right) - \left\|\hat{\boldsymbol{\Lambda}} - \bar{\boldsymbol{\Lambda}}\right\|_2 \overset{\mathbb{P}}{\lesssim} \sigma_{d+1}\left(\bar{\boldsymbol{\Lambda}}\right).$$

and therefore

$$\sigma_d\left(\hat{\boldsymbol{\Lambda}}\right) - \sigma_{d+1}\left(\hat{\boldsymbol{\Lambda}}\right) \overset{\mathbb{P}}{\gtrsim} \sigma_d\left(\bar{\boldsymbol{\Lambda}}\right) - \sigma_{d+1}\left(\bar{\boldsymbol{\Lambda}}\right) = \sqrt{M}\left(\sigma_d^{1/2}(\boldsymbol{\Sigma}) - \sigma_{d+1}^{1/2}(\boldsymbol{\Sigma})\right) = \sqrt{M}\delta. \tag{D.18}$$

We now focus our attention on obtaining a bound for $\|(\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}})\hat{\mathbf{U}}^{(m)}\|_{\mathrm{F}}$. Let

$$\mathcal{N}_m = \{m + (\ell - 1)n, \ell \in [M]\}.$$

The ijth entry of $\hat{\boldsymbol{\Lambda}} - \hat{\boldsymbol{\Lambda}}^{(m)}$ is

$$\left(\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}}\right)_{ij} = \left(\hat{\boldsymbol{\Lambda}}^{(m)} - \bar{\boldsymbol{\Lambda}}\right)_{ij} \mathbb{I}\left(i = m, j \in \mathcal{N}_m\right),$$

and so $\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}}$ is independent of $\hat{\boldsymbol{\Lambda}}^{(m)}$ and hence $\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}}$ is independent of $\hat{\mathbf{U}}^{(m)}$. We can then write

$$\begin{aligned}\left\|\left(\hat{\boldsymbol{\Lambda}}^{(m)} - \hat{\boldsymbol{\Lambda}}\right)^\top \hat{\mathbf{U}}^{(m)}\right\|_{\mathrm{F}}^2 &= \sum_{\ell \notin \mathcal{N}_m} \left(\hat{\boldsymbol{\Lambda}}_{m,\ell} - \bar{\boldsymbol{\Lambda}}_{m,\ell}\right)^2 \left\|\hat{\mathbf{U}}_{\ell,\cdot}^{(m)}\right\|_2^2\\ &\quad + \left\|\sum_{\ell \in \mathcal{N}_m} \sum_{i=1}^n \left(\hat{\boldsymbol{\Lambda}}_{i\ell} - \bar{\boldsymbol{\Lambda}}_{i\ell}\right)\hat{\mathbf{U}}_{\ell,\cdot}^{(m)}\right\|_2^2\\ &=: \zeta_1 + \zeta_2.\end{aligned}$$

The (square root of the) first term is easily bounded as

$$\begin{aligned}\zeta_1^{1/2} &\leq \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_{2,\infty} \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} \\ &\leq \left\|\hat{\mathbf{\Lambda}} - \bar{\mathbf{\Lambda}}\right\|_{2} \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} \\ &\overset{\mathbb{P}}{\lesssim} \sqrt{Mn\lambda_{\max}} \cdot \frac{\mu\sqrt{nd}\lambda_{\max}\log n}{\delta} \\ &= \frac{\sqrt{Md}n\lambda_{\max}^{3/2}\mu\log n}{\delta}\end{aligned}$$

and to bound the second term, we employ Lemma D.2 to obtain

$$\begin{aligned}\zeta_2^{1/2} &\overset{\mathbb{P}}{\lesssim} M\log^2 n \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} + \sqrt{M\lambda_{\max}\log n}\left\|\hat{\mathbf{U}}^{(m)}\right\|_{\mathrm{F}} \\ &\leq M\log^2 n \left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} + \sqrt{M\lambda_{\max}n\log n}\left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} \\ &\lesssim \sqrt{M\lambda_{\max}n\log n}\left\|\hat{\mathbf{U}}^{(m)}\right\|_{2,\infty} \\ &\lesssim \sqrt{M\lambda_{\max}n\log n} \cdot \frac{\mu\sqrt{nd}\lambda_{\max}\log n}{\delta} \\ &= \frac{\sqrt{Md}n\lambda_{\max}^{3/2}\mu\log^{3/2} n}{\delta}\end{aligned}$$

where we used Assumption 4 in the third inequality. Therefore

$$\left\|\left(\hat{\mathbf{\Lambda}}^{(m)} - \hat{\mathbf{\Lambda}}\right)^{\top}\hat{\mathbf{U}}^{(m)}\right\|_2 \leq \left\|\left(\hat{\mathbf{\Lambda}}^{(m)} - \hat{\mathbf{\Lambda}}\right)^{\top}\hat{\mathbf{U}}^{(m)}\right\|_{\mathrm{F}} \leq \zeta_1^{1/2} + \zeta_2^{1/2} \overset{\mathbb{P}}{\lesssim} \frac{\sqrt{Md}n\lambda_{\max}^{3/2}\mu\log^{3/2} n}{\delta}.$$

Similar analysis yields an analogous bound for $\|(\hat{\mathbf{\Lambda}}^{(m)} - \hat{\mathbf{\Lambda}})\hat{\mathbf{V}}^{(m)}\|_2$, and combining this, with (D.17) and (D.18) we have

$$\begin{aligned}\left\|\hat{\mathbf{U}}^{(m)}(\hat{\mathbf{U}}^{(m)})^{\top} - \hat{\mathbf{U}}\hat{\mathbf{U}}^{\top}\right\|_2 &\lesssim \frac{\left\|\left(\hat{\mathbf{\Lambda}}^{(m)} - \hat{\mathbf{\Lambda}}\right)^{\top}\hat{\mathbf{U}}^{(m)}\right\|_2 \vee \left\|\left(\hat{\mathbf{\Lambda}}^{(m)} - \hat{\mathbf{\Lambda}}\right)\hat{\mathbf{V}}^{(m)}\right\|_2}{\sigma_d\left(\hat{\mathbf{\Lambda}}\right) - \sigma_{d+1}\left(\hat{\mathbf{\Lambda}}\right)} \\ &\overset{\mathbb{P}}{\lesssim} \frac{\sqrt{d}n\lambda_{\max}^{3/2}\mu\log^{3/2} n}{\delta^2}\end{aligned}$$

which establishes the proposition.

Bibliography

[1] Alexander Modell and Patrick Rubin-Delanchy.
Spectral clustering under degree heterogeneity: a case for the random walk laplacian.
arXiv preprint arXiv:2105.00987, 2021.

[2] Alexander Modell, Ian Gallagher, Joshua Cape, and Patrick Rubin-Delanchy.
Spectral embedding and the latent geometry of multipartite networks.
arXiv preprint arXiv:2202.03945, 2022.

[3] Alexander Modell, Ian Gallagher, Emma Ceccherini, Nick Whiteley, and Patrick Rubin-Delanchy.
Intensity Profile Projection: A framework for continuous-time representation learning for dynamic networks.
arXiv preprint arXiv:2306.06155, 2023.

[4] Joshua Cape, Minh Tang, and Carey E Priebe.
The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics.
The Annals of Statistics, 47(5):2405–2439, 2019.

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

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

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

[8] Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin.

Statistical inference on random dot product graphs: a survey.
The Journal of Machine Learning Research, 18(1):8393–8484, 2017.

[9] J.K. Rowling.
Harry Potter.
Bloomsbury Publishing, 1997–2007.

[10] Efe Karakus, Jatin Pandey, Craig Evans, and Josh Friedman.
potter-network.
`https://github.com/efekarakus/potter-network`, 2014.

[11] Alexandru Mara, Yoosof Mashayekhi, Jefrey Lijffijt, and Tijl De Bie.
CSNE: Conditional signed network embedding.
In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1105–1114, 2020.

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

[13] Dean Lusher, Johan Koskinen, and Garry Robins.
Exponential random graph models for social networks: Theory, methods, and applications.
Cambridge University Press, 2013.

[14] Emmanuel Abbe.
Community detection and stochastic block models: recent developments.
The Journal of Machine Learning Research, 18(1):6446–6531, 2017.

[15] Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E Priebe.
A statistical interpretation of spectral embedding: The generalised random dot product graph.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(4):1446–1473, 2022.

[16] Joshua Agterberg, Minh Tang, and Carey E Priebe.
On two distinct sources of nonidentifiability in latent position random graph models.
arXiv preprint arXiv:2003.14250, 2020.

[17] Fangzheng Xie.
Entrywise limit theorems of eigenvectors for signal-plus-noise matrix models with weak signals.
arXiv preprint arXiv:2106.09840, 2021.

[18] Shaofeng Deng, Shuyang Ling, and Thomas Strohmer.
Strong consistency, graph laplacians, and the stochastic block model.
The Journal of Machine Learning Research, 22(1):5210–5253, 2021.

[19] Minh Tang and Carey E Priebe.
Limit theorems for eigenvectors of the normalized laplacian for random graphs.
The Annals of Statistics, 46(5):2360–2415, 2018.

[20] Linyuan Lu and Xing Peng.
Spectra of edge-independent random graphs.
arXiv preprint arXiv:1204.6207, 2012.

[21] Roberto Imbuzeiro Oliveira.
Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges.
arXiv preprint arXiv:0911.0600, 2009.

[22] P Erdös and A Rényi.
On random graphs I.
Publicationes Mathematicae Debrecen, 6:290–297, 1959.

[23] Erdös and Rényi.
On the evolution of random graphs.
Publication of Mathematics Institute of Hungian Academy of Sciences, 5:17–61, 1960.

[24] Erdös and Rényi.
On the evolution of random graphs.
Bull. Inst. Internat. Statist., 38:343–347, 1961.

[25] Paul Erdös and Alfred Rényi.
On the strength of connectedness of a random graph.
Acta Mathematica Hungarica, 12(1):261–267, 1961.

[26] Paul Erdös and Alfred Rényi.
Asymmetric graphs.
Acta Math. Acad. Sci. Hungar, 14(295-315):15, 1963.

[27] Paul Erdös and Alfred Rényi.
On random matrices.
Publ. Math. lnst. Hung. Acad. Sci., 8(455-461), 1964.

[28] Pal Erdös and Alfred Rényi.
On the existence of a factor of degree one of a connected random graph.
Acta Math. Acad. Sci. Hungar, 17:359–368, 1966.

[29] Paul Erdös and Alfred Rényi.
On random matrices II.
Studia Sci. Math. Hungar, 3:459–464, 1968.

[30] William Aiello, Fan Chung, and Linyuan Lu.
A random graph model for power law graphs.
Experimental Mathematics, 10(1):53–66, 2001.

[31] Nicholas C Wormald.
The asymptotic connectivity of labelled regular graphs.
Journal of Combinatorial Theory, Series B, 31(2):156–167, 1981.

[32] Michael Molloy and Bruce Reed.
A critical point for random graphs with a given degree sequence.
Random structures & algorithms, 6(2-3):161–180, 1995.

[33] Michael Molloy and Bruce Reed.
The size of the giant component of a random graph with a given degree sequence.
Combinatorics, probability and computing, 7(3):295–305, 1998.

[34] Fan Chung and Linyuan Lu.
Connected components in random graphs with given expected degree sequences.
Annals of combinatorics, 6(2):125–145, 2002.

[35] Fan Chung and Linyuan Lu.
The average distances in random graphs with given expected degrees.
Proceedings of the National Academy of Sciences, 99(25):15879–15882, 2002.

[36] Fan Chung, Linyuan Lu, and Van Vu.
Spectra of random graphs with given expected degrees.
Proceedings of the National Academy of Sciences, 100(11):6313–6318, 2003.

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

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

[39] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing.
Mixed membership stochastic blockmodels.
Journal of Machine Learning Research, 9(65):1981–2014, 2008.

[40] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo.
Mixed membership estimation for social networks.
Journal of Econometrics, page 105369, 2023.
ISSN 0304-4076.

[41] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe.
Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown.
SIAM Journal on Matrix Analysis and Applications, 34(1):23–39, 2013.

[42] Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe.
Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding.
Electronic Journal of Statistics, 8(2):2905 – 2922, 2014.

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

[44] Jing Lei and Alessandro Rinaldo.
Consistency of spectral clustering in stochastic block models.
The Annals of Statistics, 43(1):215 – 237, 2015.

[45] Jiashun Jin.
Fast community detection by score.
The Annals of Statistics, 43(1):57–89, 2015.
ISSN 00905364.
URL `http://www.jstor.org/stable/43556508`.

[46] Vince Lyzinski, Minh Tang, Avanti Athreya, Youngser Park, and Carey E Priebe.
Community detection and classification in hierarchical stochastic blockmodels.
IEEE Transactions on Network Science and Engineering, 4(1):13–26, 2016.

[47] Avanti Athreya, Carey E Priebe, Minh Tang, Vince Lyzinski, David J Marchette, and Daniel L Sussman.
A limit theorem for scaled eigenvectors of random dot product graphs.
Sankhya A, 78(1):1–18, 2016.

[48] Jianbo Shi.
Normalized cuts and image segmentation.
IEEE Transactions on pattern analysis and machine intelligence, 22(8):888–905, 2000.

[49] Marina Meila and Jianbo Shi.
Learning segmentation by random walks.
Advances in Neural Information Processing Systems, 13, 2000.

[50] Marina Meilă and Jianbo Shi.
A random walks view of spectral segmentation.
In *International Workshop on Artificial Intelligence and Statistics*, pages 203–208. PMLR, 2001.

[51] Ulrike Von Luxburg.
A tutorial on spectral clustering.
Statistics and computing, 17(4):395–416, 2007.

[52] Xinjie Du and Minh Tang.
Hypothesis testing for equality of latent positions in random graphs.
Bernoulli, 29(4):3221–3254, 2023.

[53] Israel Dejene Gebru, Xavier Alameda-Pineda, Florence Forbes, and Radu Horaud.
Em algorithms for weighted-data clustering with application to audio-visual scene analysis.
IEEE transactions on pattern analysis and machine intelligence, 38(12):2402–2415, 2016.

[54] R Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2021.

[55] Niall Adams and Nick Heard.
Dynamic Networks and Cyber-security, volume 1.
World Scientific, 2016.

[56] Wayne W Zachary.
An information flow model for conflict and fission in small groups.
J. Anthropol. Res., 33(4):452–473, 1977.

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

[58] Vincent Labatut and Xavier Bost.
Extraction and analysis of fictional character networks: A survey.
ACM Comput. Surv., 52(5):1–40, 2019.

[59] Alexander D. Kent.

Cybersecurity Data Sources for Dynamic Network Research.
In *Dynamic Networks in Cybersecurity.* Imperial College Press, June 2015.

[60] Pedro Szekely, Craig A Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, et al.
Building and using a knowledge graph to combat human trafficking.
In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II 14*, pages 205–221. Springer, 2015.

[61] Nansu Zong, Rachael Sze Nga Wong, Yue Yu, Andrew Wen, Ming Huang, and Ning Li.
Drug–target prediction utilizing heterogeneous bio-linked network embeddings.
Briefings in bioinformatics, 22(1):568–580, 2021.

[62] G.W. Stewart and J.G. Sun.
Matrix Perturbation Theory.
Computer Science and Scientifi. ACADEMIC PressINC, 1990.
ISBN 9781493301997.

[63] Tsit-Yuen Lam.
Introduction to quadratic forms over fields, volume 67.
American Mathematical Soc., 2005.

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

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

[66] Tai Qin and Karl Rohe.
Regularized spectral clustering under the degree-corrected stochastic blockmodel.
Advances in neural information processing systems, 26, 2013.

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

[68] Yilin Zhang and Karl Rohe.
Understanding regularized spectral clustering via graph conductance.

Advances in Neural Information Processing Systems, 31, 2018.

[69] Fan Chen, Sebastien Roch, Karl Rohe, and Shuqi Yu.
Estimating graph dimension with cross-validated eigenvalues.
arXiv preprint arXiv:2108.03336, 2021.

[70] Carey E Priebe, Youngser Park, Joshua T Vogelstein, John M Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford.
On a two-truths phenomenon in spectral graph clustering.
Proceedings of the National Academy of Sciences, 116(13):5995–6000, 2019.

[71] Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy.
Discovering latent topology and geometry in data: a law of large dimension.
arXiv preprint arXiv:2208.11665, 2022.

[72] Mu Zhu and Ali Ghodsi.
Automatic dimensionality selection from the scree plot via the use of profile likelihood.
Computational Statistics & Data Analysis, 51(2):918–930, 2006.

[73] Wei Luo and Bing Li.
Combining eigenvalues and variation of eigenvectors for order determination.
Biometrika, 103(4):875–887, 2016.

[74] Andrew Jones and Patrick Rubin-Delanchy.
The multilayer random dot product graph.
arXiv preprint arXiv:2007.10455, 2020.

[75] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey.
DrugBank: a comprehensive resource for in silico drug discovery and exploration.
Nucleic acids research, 34:D668–D672, 2006.

[76] Minoru Kanehisa and Susumu Goto.
KEGG: Kyoto encyclopedia of genes and genomes.
Nucleic acids research, 28(1):27–30, 2000.

[77] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein.
PharmGKB: the pharmacogenetics knowledge base.
Nucleic acids research, 30(1):163–165, 2002.

[78] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási.

The human disease network.
Proceedings of the National Academy of Sciences, 104(21):8685–8690, 2007.

[79] Madeleine Udell and Alex Townsend.
Why are big data matrices approximately low rank?
SIAM Journal on Mathematics of Data Science, 1(1):144–160, 2019.

[80] C Seshadhri, Aneesh Sharma, Andrew Stolman, and Ashish Goel.
The impossibility of low-rank representations for triangle-rich complex networks.
Proceedings of the National Academy of Sciences, 117(11):5631–5637, 2020.

[81] Sudhanshu Chanpuriya, Cameron Musco, Konstantinos Sotiropoulos, and Charalampos Tsourakakis.
On the power of edge independent graph models.
Advances in Neural Information Processing Systems, 34, 2021.

[82] Lihua Lei.
Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices.
arXiv preprint arXiv:1909.04798, 2019.

[83] Minh Tang, Daniel L Sussman, and Carey E Priebe.
Universally consistent vertex classification for latent positions graphs.
The Annals of Statistics, 41(3):1406–1430, 2013.

[84] Mikhail Belkin.
Approximation beats concentration? an approximation view on inference with smooth radial kernels.
In *Conference On Learning Theory*, pages 1348–1361. PMLR, 2018.

[85] Mark EJ Newman.
Modularity and community structure in networks.
Proceedings of the National Academy of Sciences, 103(23):8577–8582, 2006.

[86] Larry Wasserman.
Topological data analysis.
Annual Review of Statistics and Its Application, 5:501–532, 2018.

[87] Patrick Rubin-Delanchy.
Manifold structure in graph embeddings.
Advances in Neural Information Processing Systems, 33, 2020.

[88] Avanti Athreya, Minh Tang, Youngser Park, and Carey E. Priebe.
On estimation and inference in latent structure random graphs.
Statistical Science, 36(1):68 – 88, 2021.

[89] Patrick O Perry and Patrick J Wolfe.
Point process modelling for directed interaction networks.
Journal of the Royal Statistical Society: Series B: Statistical Methodology, pages 821–849, 2013.

[90] Guotong Xue, Ming Zhong, Jianxin Li, Jia Chen, Chengshuai Zhai, and Ruochen Kong.
Dynamic network embedding survey.
Neurocomputing, 472:212–223, 2022.

[91] Ian Gallagher, Andrew Jones, and Patrick Rubin-Delanchy.
Spectral embedding for dynamic networks with stability guarantees.
Advances in Neural Information Processing Systems, 34:10158–10170, 2021.

[92] Peter Diggle.
A kernel method for smoothing point process data.
Journal of the Royal Statistical Society: Series C (Applied Statistics), 34(2):138–147, 1985.

[93] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong.
Entrywise eigenvector analysis of random matrices with low expected rank.
Annals of statistics, 48(3):1452, 2020.

[94] Fangzheng Xie.
Entrywise limit theorems of eigenvectors and their one-step refinement for sparse random graphs.
arXiv preprint arXiv:2106.09840, 2021.

[95] Riccardo Rastelli and Marco Corneli.
Continuous latent position models for instantaneous interactions.
arXiv preprint arXiv:2103.17146, 2021.

[96] Abdulkadir Çelikkanat, Nikolaos Nakis, and Morten Mørup.
Piecewise-velocity model for learning continuous-time dynamic node representations.
arXiv preprint arXiv:2212.12345, 2022.

[97] Igor Artico and Ernst Wit.
Fast inference of latent space dynamics in huge relational event networks.
arXiv preprint arXiv:2303.17460, 2023.

[98] Purnamrita Sarkar and Andrew W Moore.
Dynamic social network analysis using latent space models.
Advances in neural information processing systems, 18:1145, 2006.

[99] Daniel K Sewell and Yuguo Chen.

Latent space models for dynamic networks.
Journal of the American Statistical Association, 110(512):1646–1657, 2015.

[100] Nial Friel, Riccardo Rastelli, Jason Wyse, and Adrian E Raftery.
Interlocking directorates in Irish companies using a latent space model for bipartite networks.
Proceedings of the National Academy of Sciences, 113(24):6629–6634, 2016.

[101] Keith Levin, Avanti Athreya, Minh Tang, Vince Lyzinski, Youngser Park, and Carey E Priebe.
A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference.
arXiv preprint arXiv:1705.09355, 2017.

[102] Fuchen Liu, David Choi, Lu Xie, and Kathryn Roeder.
Global spectral clustering in dynamic networks.
Proceedings of the National Academy of Sciences, 115(5):927–932, 2018.

[103] Joshua Cape.
Spectral analysis of networks with latent space dynamics and signs.
Stat, 10(1):e381, 2021.

[104] Lun Du, Yun Wang, Guojie Song, Zhicong Lu, and Junshan Wang.
Dynamic network embedding: An extended approach for skip-gram based network embedding.
In *IJCAI*, volume 2018, pages 2086–2092, 2018.

[105] Sedigheh Mahdavi, Shima Khoshraftar, and Aijun An.
dynnode2vec: Scalable dynamic network embedding.
In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3762–3765. IEEE, 2018.

[106] Carl Eckart and Gale Young.
The approximation of one matrix by another of lower rank.
Psychometrika, 1(3):211–218, 1936.

[107] Richard Bruno Lehoucq.
Analysis and implementation of an implicitly restarted Arnoldi iteration.
Rice University, 1995.

[108] James Baglama and Lothar Reichel.
Augmented implicitly restarted Lanczos bidiagonalization methods.
SIAM Journal on Scientific Computing, 27(1):19–42, 2005.

[109] Keith Levin, Fred Roosta, Michael Mahoney, and Carey Priebe.
Out-of-sample extension of graph adjacency spectral embedding.
In *International Conference on Machine Learning*, pages 2975–2984. PMLR, 2018.

[110] Emmanuel Candes and Benjamin Recht.
Exact matrix completion via convex optimization.
Communications of the ACM, 55(6):111–119, 2012.

[111] Emmanuel Abbe and Colin Sandon.
Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery.
In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015.

[112] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu.
Optimal m-estimation in high-dimensional regression.
Proceedings of the National Academy of Sciences, 110(36):14563–14568, 2013.

[113] Adel Javanmard and Andrea Montanari.
Debiasing the lasso: Optimal sample size for gaussian designs.
2018.

[114] Yiqiao Zhong and Nicolas Boumal.
Near-optimal bounds for phase synchronization.
SIAM Journal on Optimization, 28(2):989–1016, 2018.

[115] Peter H Schönemann.
A generalized solution of the orthogonal procrustes problem.
Psychometrika, 31(1):1–10, 1966.

[116] Peter W MacDonald, Elizaveta Levina, and Ji Zhu.
Latent space models for multiplex networks with shared structure.
Biometrika, 109(3):683–706, 2022.

[117] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al.
High-resolution measurements of face-to-face contact patterns in a primary school.
PloS one, 6(8):e23176, 2011.

[118] Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou.
Optimality of spectral clustering in the gaussian mixture model.
The Annals of Statistics, 49(5):2506–2530, 2021.

[119] Fan RK Chung.
Spectral graph theory, volume 92.
American Mathematical Soc., 1997.

[120] Vinesh Solanki, Patrick Rubin-Delanchy, and Ian Gallagher.
Persistent homology of graph embeddings.
arXiv preprint arXiv:1912.10238, 2019.

[121] Francesco Sanna Passino, Nicholas A Heard, and Patrick Rubin-Delanchy.
Spectral clustering on spherical coordinates under the degree-corrected stochastic blockmodel.
Technometrics, 64(3):346–357, 2022.

[122] Laurens Van der Maaten and Geoffrey Hinton.
Visualizing data using t-SNE.
Journal of Machine Learning Research, 9(11), 2008.

[123] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger.
Fast interpolation-based t-SNE for improved visualization of single-cell rna-seq data.
Nature Mthods, 16(3):243–245, 2019.

[124] Jianqing Fan, Weichen Wang, and Yiqiao Zhong.
An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation.
Journal of Machine Learning Research, 18(207):1–42, 2018.

[125] Stanislav Minsker.
On some extensions of Bernstein's inequality for self-adjoint operators.
Statistics & Probability Letters, 127:111–119, 2017.

[126] Afonso S. Bandeira and Ramon van Handel.
Sharp nonasymptotic bounds on the norm of random matrices with independent entries.
The Annals of Probability, 44(4):2479 – 2506, 2016.

[127] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al.
Spectral methods for data science: A statistical perspective.
Foundations and Trends® in Machine Learning, 14(5):566–806, 2021.

[128] David Pollard.
Empirical processes: Theory and applications.
In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR, 1990.