Foundations of Adjacency Spectral Embedding

by

Daniel L. Sussman

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

December, 2013

Abstract

The eigendecomposition of an adjacency matrix provides a way to embed a graph as points in finite dimensional Euclidean space. This embedding allows the full arsenal of statistical and machine learning methodology for multivariate Euclidean data to be deployed for graph inference. Our work analyzes this embedding, a graph version of principal component analysis, in the context of various random graph models with a focus on the impact for subsequent inference. For the stochastic blockmodel, with a finite number of blocks of stochastically equivalent vertices, Sussman et al. [2012], Fishkind et al. [2013], and Lyzinski et al. [2013] show that clustering the embedded points using k-means accurately partitions the vertices into the correct blocks, even when the embedding dimension is misspecified or the number of blocks is unknown. For the more general random dot product graph model, an example of a latent position model, Sussman et al. [2013] shows that the latent positions are consistently estimated by the embedding which then allows for accurate learning in a supervised vertex classification framework. Tang et al. [2013] strengthens these results to more general latent position models. Athreya et al. [2013] provide distributional results,

ABSTRACT

akin to a central limit theorem, for the residuals between the estimated and true latent positions which provides the potential for deeper understanding of these methods. In summary, these papers demonstrate that for a broad class of graph models and inference tasks, adjacency-spectral embedding allows for accurate graph inference via standard multivariate methodology.

Primary Reader: Carey Priebe

Secondary Reader: Avanti Athreya

Acknowledgments

The journey through graduate school is one best travelled in the company of friends and I am delighted to have found so many willing to help me along. The Applied Math and Statistics Department at Johns Hopkins University has provided a nurturing environment at all levels from the graduate students, post-docs, faculty, staff and our chair, Daniel Naiman (who I thank for the rides). My co-authors and collaborators, including Minh Tang, Donniell Fishkind, Joshua Vogelstein, Avanti Athreya, Vince Lyzinski and many more, have all shaped who I am through their invaluable help, criticism, knowledge, and willingness to listen. My adviser, Carey Priebe, is a model for the task; he provides endless encouragement and valueable advice while posing fascinating problems that bridge the diverse talents of our fantastic team.

Of course, we all need family to guide us and remind us of the important things and I have been especially lucky in this regard. I cannot quantify my appreciation for my loving parents, Mark and Ellen, who have supported me in so many ways. Finally, I want thank my girl friend, Allison Bland, who has endlessly tolerated the

long work hours and pushed me through times of frustration. All in all, my time as a graduate student has been wonderful and I credit it most the people who have surrounded me and helped me during this time.

Dedication

This thesis is dedicated to Nana, Poppop and Grandmommy.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A common task for statistical inference and exploratory data analysis is to cluster a collection of objects in groups that share similar properties. If we observe a vector of measurements for each object then the objects can be represented as points in Euclidean space and there are hundreds of techniques for clustering the points. Methods as simple as k-means clustering or hierarchical clustering and complicated methods that invoke manifold learning or dissimilarity representations all may prove effective [Duda et al., 2001, Izenman, 2008, Pekalska and Duin, 2005]. If the data is high dimensional, then first using dimensionality reduction will often balance the bias-variance trade-off and improve the results of the clustering.

One recently popularized method in the manifold learning community involves creating a graph where each vertex represents an object and edges are placed between objects based on the distances between the points. Dimensionality reduction

is achieved by using the eigenvectors of the Laplacian associated with the graph and clustering is then done using k-means. This method, known as spectral clustering Luxburg [2007], is closely related to the ideas considered in this work.

In our setting, rather than observing vectors associated with each object, we observe only a graph. Though the methods that are used for spectral clustering can again be used in this setting, the theory from that setting does not necessarily translate. We will want to consider other models for graphs beyond those based on constructions from a point cloud. Our underlying goals do not necessarily change—clustering the vertices is one goal and tasks such as classification and estimation are others.

Indeed, graphs or networks are playing a more prominent role in the way we think about and analyze the world around us. Focusing on the connections between objects, rather than the properties of individual objects, creates new opportunities for understanding and utilization.

When studying people, relationships and communities embody the unique social nature of humans. When studying science, the references between articles demonstrate the complex landscape connecting fields. When studying the neurons in the brain, a researcher investigates not only the complex geometric structure of each cell, graph like itself, but also how these structures connect and form processing units. Phone calls, emails, Facebook, Twitter, maps, the electric grid, food webs, all can be naturally viewed as networks.

Penetrating the problem of not only overcoming but exploiting the networked nature of data is the goal of this work. Needless to say, we can only scratch the surface of this deep statistical challenge. Our focus is on the question, given a graph representing the connections between objects, what can be learned about properties of the individual objects and what tools from classical statistics can be brought to bear?

In this work, we consider a series of probabilistic models for graphs and a generic methodology that permits the application of a broad range statistical tools. The models we study capture the idea that relationships between objects depend on properties of the objects. Supposing these properties are unobserved, our methodology provides a tool to estimate these properties and our theory will provide finite sample and asymptotic performance guarantees for these estimates. Finally, we will make a return to more classical statistics but with a twist. Having not observed the original properties of the objects, we show that nonetheless many standard statistical techniques can be applied to the estimates and that these techniques retain some of their performance guarantees in this distinctive setting.

1.1 Overview

In mathematics, the simplest way that connections between a set of objects can be represented is as a graph, consisting of vertices and edges. Encoding a graph as

an adjacency matrix endows it with linear algebraic structures and our main tool will be the adjacency spectral embedding (see Definition 3.3). Based on the eigenvectors and eigenvalues of an adjacency matrix, this embedding represents each vertex in a graph as a point in a finite dimensional Euclidean space in a way that can capture the underlying structure of the graph. This method is akin to principle components analysis, as a low-dimensional representation of data is sought that captures the variation of the original data.

The practical utility of the embedding is that it takes a graph, a complex combinatorial object, and approximates it as a point cloud. This representation permits the use of standard multivariate statistics and machine learning techniques. If the embedding captures the key structures of the graph, then questions about the vertices in a graph can be transformed into questions about the embedding in Euclidean space and these techniques can be fruitfully applied.

The purpose of this work is to provide a deep theoretical analysis of the adjacency spectral embedding. In particular, we will consider a series of probabilistic models for a random graphs and study the quality of the embedding as an estimate for the underlying parameters of the models. Our theory provides guarantees that show that the embedding provides accurate estimates of certain model parameters. We break down the theory in terms of whether the edges are all present independently or not. In the non-independent, case we assume the edges are still independent conditioned on certain properties of the vertices.

Following this we investigate the question of whether these accurate estimates lead to accurate subsequent inference. Such inference tasks include vertex clustering and vertex classification. Using the embedding we will show that as the graph gets large, the performance of techniques such as k-means clustering, k-nearest-neighbors classification, linear classification and plug-in estimation is close to what would be expected in a more typical statistical setup.

We now give a brief outline of this work.

Chapter 1 In this introductory chapter we will present relevant background material.

Chapter 2 A series of models for random graphs of varying complexity are introduced. These models are all variations on latent position random graphs which have valuable statistical properties for studying vertex based inference.

Chapter 3 Bounds in the case of the independent edge random graphs are presented. We start by proving a general bound and then consider the improvements possible for a low rank model. Strong concentration inequalities for eigenvalues are also shown.

Chapter 4 The case in which the latent positions are iid is investigated and we show how the bounds from the previous chapter can be used in this case.

Chapter 5 Inference tasks such as clustering, classification and estimation are considered and we demonstrate consistency of various procedures.

Chapter 6 The results are broadly discussed and we propose extensions and some alternatives.

1.2 Background

In this section we will try to give give an overview of important background material. We will start by discussing dimensionality reduction and then provide some background on the main tools in our proofs. We then give an overview of random graphs and motivations from spectral graph theory. Finally, we discuss the problem of community detection in graphs.

1.2.1 Dimensionality Reduction

Spectral methods are widely used in statistics. Their use stems from their relationship to the fundamental principle of minimizing square error. The most common spectral method in statistics is principle components analysis (PCA). PCA is closely related to the adjacency spectral embedding in that they both provide a low-rank embedding that minimizes the square error of the approximation.

If we consider a mean-zero collection of vectors $x_1, \ldots, x_n \in \mathbb{R}^m$ then let $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times m}$. For a specified rank $d \leq m$, PCA solves the optimization

problem

$$\underset{Y:\text{rank}(Y)\leq d}{\arg\min} \|X - Y\|_F.$$

If we let $\mathcal{P} = VV^\top$, where the columns of V are given by the d right singular vectors of X corresponding to its d largest singular values, then Eckart and Young [1936] showed the solution is given by $Y = X\mathcal{P}$. In statistics, the columns of V represent the d directions that explain the largest amount of variance of any choice of d orthogonal directions. Another viewpoint relates this to reduced rank regression, where X is regressed onto itself [Izenman, 2008].

If we suppress the requirement that the collection is mean-zero, then this is known as *uncentered* principle components analysis (UPCA) and is even more closely related to the adjacency spectral embedding. Unlike PCA, UPCA does not capture variance but instead captures the directions that contribute most to the sample second moment of X. UPCA is not typically used in statistics but will be useful in our graph setting where we seek to capture both the mean and variance components of the graph structure.

PCA is such an important part of statistics because it is one of the simplest dimensionality reduction tools. Dimensionality reduction is necessary to find a balance in the bias-variance trade-off when we are confronted with high dimensional data [Duda et al., 2001, Trunk, 1979]. Finding a suitable low-dimensional space on which one's data is well represented is one of paramount problems of modern statistics and is the study of the field of manifold learning [Izenman, 2008]. The low dimensional

structures of graphs is still not well understood and in this work we seek to expand the foundations of dimensionality reduction for graphs by providing an extensive study of the adjacency spectral embedding.

1.2.2 Linear Algebra and Concentration Inequalities

Our main tools for proving our results are linear algebraic and probabilistic. From linear algebra we exploit various perturbation theory results as well as the theory of matrix decompositions for symmetric matrices. From probability we exploit a series of concentration inequalities for various settings. We will briefly overview these main tools.

1.2.2.1 Linear Algebra

From linear algebra, it is well known that a real symmetric matrix H can be decomposed as $H = USU^\top$ where S is a real diagonal matrix with the eigenvalues of H along the diagonal and U is an orthogonal matrix with columns given by the corresponding eigenvectors of H. This result forms the basis of the adjacency spectral embedding where we consider the *eigendecomposition* of an adjacency matrix.

A key question for this work is if we have two symmetric matrices H and H', then to what extent can the eigenvalues of H' be related to those of H if we know

something about $\|H - H'\|$. These questions are studied in the field of perturbation theory. For the eigenvalues, one can establish the relatively simple inequality that for all $i \in [n]$

$$|\lambda_i(H) - \lambda_i(H')| \leq \|H - H'\|_{2\to 2} \leq \|H - H'\|_F \tag{1.1}$$

where $\lambda_i(H)$ denotes the i^{th} largest eigenvalue of H in magnitude.

Results for eigenvectors are more complicated and simple counter examples to naive analogies of Eq. (1.1) can be demonstrated for 2×2 matrices. Indeed, compare the two matrices we consider the matrix

$$H = \begin{pmatrix} 1 & 0 \\ 0 & 1+\epsilon \end{pmatrix} \text{ and } H' = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}$$

with eigenvectors given by

$$v_1(H) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \; v_2(H) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \text{ and } v_1(H') = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}, v_2(H') = \begin{pmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix}$$

regardless of ϵ so there is no way to bound the eigenvector perturbations in terms of $\|H - H'\|_F$ alone.

The necessary assumption is a gap in the eigenvalues and the resulting theory was developed by Davis and Kahan [1970]. This theorem can take a variety of forms since there are various ways to measure the pertubation of eigenvectors and the form we desire is in terms of projection matrices. Given an interval $\mathcal{S} \subset \mathbb{R}$ and a symmetric matrix H, let $\mathcal{P}_H(\mathcal{S})$ denote the projection matrix onto the eigenvectors of H corresponding to eigenvalues in of H in $\mathcal{S}$.

Theorem 1.1 (Davis and Kahan [1970], see also Bhatia [1997]). *Let $H, H' \in \mathbb{R}^{n\times n}_{sym}$, suppose $\mathcal{S} \subset \mathbb{R}$ is an interval. If γ is the minimum distance between any eigenvalue of H in $\mathcal{S}$ and any eigenvalue of H not in $\mathcal{S}$ then*

$$\|\mathcal{P}_H(\mathcal{S}) - \mathcal{P}_{H'}(\mathcal{S})\|\gamma \leq \|H - H'\| \tag{1.2}$$

for any unitarily invariant norm.

This theorem ensures that if there is a large gap between the eigenvalues of a matrix then we can guarantee that the corresponding eigenvectors will be close. In our example above, the gap is only ϵ so this theorem provides very weak guarantees. In Chapter 3 we will assume that such a gap exists while in Chapter 4 we show that under certain model assumptions the gap is likely to be large.

1.2.2.2 Concentration Inequalities

Concentration inequalities are a key aspect of any probabilistic analysis involving a large number of independent random variables. We will consider three different concentration inequalities: Hoeffding's Inequality for the sum of independent random variables, a variant of McDiarmid's inequality, and a concentration inequality for random matrices.

Hoeffding's inequality is a classic result which states that the sum of bounded independent bounded random variables will concentrate around its expectation at an exponential rate related to the bounds of the random variables. This results is key

for our results in section 3.3 where we decompose the difference we seek to bound in terms of one component which is bounded using purely linear algebraic results and another component which has terms which are all sums of independent random variables.

Later in that same section, we seek to bound a somewhat more complicated function of independent random variables. McDiarmid's inequality provides a concentration inequality where the bound is in terms of deviations of maximum deviations of a function as each argument of the function is changed, keeping the others fixed. However, an application of this inequality turns out to provide unsatisfactory bounds but a refinement of this inequality established by Kutin [2002] provides a measured improvement. This improvement is due to the fact that with high probability the argument-wise deviations are small even though the maximum deviations are large.

Finally, one of the key results for this work are relatively new concentration inequalities for random matrices. Oliveira [2009] provided the first form of this inequality which shows that the adjacency matrix for a random graph will concentrate around its expectation in terms of the spectral norm. Tropp [2012] extended and refined these results to other random matrices.

1.2.3 Random Graphs and Spectral Graph Theory

A graph, composed of vertices and edges between vertices, provides mathematical abstraction that has been used to study data based on relationships in fields as var-

ied as sociology, communications, neuroscience and molecular biology. Graphs were originally studied by Leonard Euler to analyze a problem in the "the geometry of positions" but it. In the second half of the twentieth century, mathematicians such as Paul Erdős deepened the field of graph theory connecting it with ideas in combinatorics as well as probability and introduced the idea of a random graph. Though statistical study of graphs can be traced at least as far back as Gilbert [1959], it was not until around 1980 that methods focused on modelling and analysing real world networks were introduced by researchers such as Holland et al. [1983].

In Chapter 2 we introduce a series of models for random graphs, from the very simple Erdős-Rényi random graph to the very general class of exhcangeable random graphs. The graphs all have the property that edges in the graph appear either indpendently or conditionally indpendentlly independently. Suffice it to say, these are far from the full scope of models for random graphs. Models that don't satisfy this requirement are random regular graphs, graphs with a specified degree sequence, and certain examples of exponential random graphs. We do not study the implications of adjacency spectral embedding in these models as the techniques used to not necessarily translate easily.

Spectral graph theory also has a long history Chung [1997]. Most of the work in spectral graph theory relates properties of the spectrum to graph theoretic properties of the graph. Perhaps most famously, the combinatorial the number of zero eigenvalues of the combinatorial Laplacian is equal to the number of connected components

in the graph. This relates to one of the original motivations for spectral methods, a relaxation of the combinatorial *ratio-cut* problem. Given an adjacency matrix, a cut is a partition of the vertices $\mathcal{V} \cup \mathcal{V}^c = [n]$ and the cost of the cut is

$$\text{cut}(\mathcal{V}) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}^c} A_{ij}.$$

The ratio cut problem seeks

$$\mathcal{V}^* = \arg\min_{\mathcal{V} \subset [n]} \frac{\text{cut}(\mathcal{V})}{|\mathcal{V}|(n - |\mathcal{V}|)}.$$

Solving this problem exactly is NP-hard but if we relax the problem then the relaxation can be solved exactly via spectral methods involving the combinatorial graph Laplacian. Let $D \in \mathbb{R}^{n \times n}$ be the diagonal matrix with diagonal given by the degree vector $A\mathbf{1}$. The ratio cut problem is equivalent to minimizing $v^\top (D - A)v$ over all unit vectors $v \in \mathbb{R}^d$ such that

$$v_i = \begin{cases} \frac{1}{|\mathcal{V}|} & \text{if } i \in \mathcal{V} \\ \\ \frac{-1}{n - |\mathcal{V}|} & \text{if } i \in \mathcal{V}^c \end{cases}$$

for some $\mathcal{V} \subset [n]$. Noting that any v constructed this way is orthogonal $\mathbf{1}$, we can relax the ratio cut problem to

$$\min v^\top (D - A)v \text{ such that } \|v\|_2 = 1, v^\top \mathbf{1} = 0. \tag{1.3}$$

Since the combinatorial Laplacian is positive semidefinite and has a zero eigenvalue with associated eigenvector $\mathbf{1}$, the minimizer for Eq. (1.3) is exactly the eigenvector of

$D-A$ associated with its second smallest eigenvector [Fiedler, 1973]. This eigenvector is frequently called the Fiedler vector.

This story motivates the use of the eigenvector associated with the second smallest eigenvector of $D-A$ in specific and eigen-decompositions of matrices associated with graphs in general. The use of the adjacency matrix as compared to the combinatorial or normalized Laplacian is an important debate. We leave this debate to the discussion section and for now merely state that for our purposes, much of the theoretical results are more easily demonstrated for the adjacency matrix and we believe that the results here will translate favorably when considering decompositions of other matrices.

1.2.4 Graph Partitions and Community Detection

A frequent and well studied problem when confronted with graph data is to partition the vertices into parts or communities such that vertices in the same part are similar in some way. As is the case for vector valued observations, there is no universal method to find these parts nor is there necessarily some optimal or true partition. The indefiniteness of this problem means it has received much atttention from fields including physics, computer science and statistics [Fortunato, 2010]. We will provide a brief discussion of the various methods to solve this problem here but note that this discussion is far from complete and the literature is expanding rapidly.

We have already discussed one graph partitioning criterion and an algorithm to

attempt to solve it, namely the ratio cut problem and the use of the Fiedler vector. Criterion related to cuts are popular as certain community structures are represented in graphs as nearly disconnected components and so removing relatively few edges will lead to a disconnected graph where clustering is trivial. Typically, most cut problems, like the ratio cut, are NP-hard to solve exactly. As demonstrated above, one way around this is to solve some relaxed version of the problem where fast methods can be used.

Another alternative is greedy methods that iteratively remove edges or vertices that maximize some criterion such as betweeness or centrality until a disconnected graph is achieved [Girvan and Newman, 2002]. Depending on the criteria chosen, these methods can be very fast and one advantage is that they can be constructed to use only local information in the graph so that small communities can be found in very large graphs [Bagrow and Bollt, 2005]. Another direction is to greedily collect edges until Modularity is another generalization of the idea of a cut and there are a plethora of algorithms to find solutions to the modularity problem [Newman, 2006].

All of these methods originate from some algorithm or from some particular objective function which either implicitly or explicitly defines the possible structures of the communities found. If the community structure is defined only implicitly this could lead to some unexpected results such as one large part and many very small parts. The performance guarantees associated with these methods are frequently in terms of accuracy of the algorithm in finding the solution to the problem or are discussed

in the context of a particular stochastic model.

This leads to the other main starting place for graph partitioning algorithm development is to assume the graph follows some probability distribution. Methods such as this have the advantage that if the graph does follow the specified distribution then there is usually "true" partition that is sought. Most commonly, the stochastic blockmodel (see section 2.3.2) is used where the desired partition is determined by the block memberships of the vertices [Holland et al., 1983].

This model has received extensive attention and the methods of attach are varied. Standard statistical techniques such as maximum likelihood and variational methods have been explored with strong theoretical guarantees [Celisse et al., 2011]. Bayesian methods were also introduced early on in the development [Snijders and Nowicki, 1997]. It has also been shown that subgraph counts provide a way to create method of moment estimators for this model [Bickel et al., 2011].

Beyond the stochastic blockmodel, various generalizations have been considered. Latent position models [Hoff et al., 2002] have been used to develop model based clustering and mixed membership models allow for more flexible partitions [Airoldi et al., 2008]. Stochastic blockmodel approximation of general exchangeable graphs are now gaining interest as a non-parametric approach to clustering Airoldi et al. [2013].

Our work here combines these various approaches. Though our original motivations were algorithmic, we study the adjacency spectral embedding in terms of

specific stochastic models. Our goals are more broad than just graph partitioning and one of the advantages of our methodology and spectral techniques in general is that typically the embedding that is generated is conducive to multiple analyses and exploratory analysis. This is different than many community detection approaches where the output is the the particular partition but if we desire to perform another type of analysis than we must return the graph and start a new procedure.

Chapter 2

Latent Position Graphs

Our main objects of study will be random graphs. Their is now a fantastic diversity of models for random graphs ranging from the simple yet deep Erdős-Rényi model to the equally deep exchangeable models. The most general model can be viewed as a multinomial distribution taking values on the enormous space of the $2^{\binom{n}{2}}$ distinct graphs on n vertices.

Since one of our goals is an asymptotic analysis as the number of vertices get large, we are interested in models that naturally extend to graphs of any order. For latent position models, if we take the latent positions to be iid then asymptotic analysis becomes natural and the situation is akin to a standard iid setup for vector valued random variables. As the iid setting is of such importance in statistics, this case will a main focus as we analyze subsequent inference tasks.

In Section 2.1 we introduce the main graph theoretic concepts. As our goals are

non-graph-theoretic, we only require a limited amount of graph theory jargon and in some cases our jargon deviate from the graph theory jargon, such as the use of the term embedding. Random graphs and simpler random graph models such as the Erdős-Rényi model and the independent edge model will be introduced in Section 2.2 and in Chapter 3 we give a detailed analysis of the adjacency spectral embedding in these cases. For the independent edge model, our focus will be on finite sample performance guarantees as this model doesn't have a natural asymptotic regime.

If the latent positions are random then the edges are non-independent in a latent position graphs which is frequently more realistic. Furthermore, interpretable geometry of the latent positions imposes structure on the random graph that can be useful in some applications. We introduce these models in section 2.3 and the relevant adjacency spectral embedding theory is presented in Chapter 4.

Two examples which we return to later are the stochastic blockmodel (Section 2.3.2) and the random dot product graph model (Section 2.3.1). The stochastic blockmodel is an excellent model for imposing a strong notion of community. The random dot product graph is one of our main focuses because of the accessible nature of the eigen-decomposition of the adjacency matrix. Finally, in section 2.4 we argue that a broad class of latent position models can be reparametrized to be represented as random dot product graphs.

2.1 Graph Theory Concepts

In this section we will introduce the few elements of graph theory necessary for this work. For a thorough review of graph theory for non-random graphs see West [1996].

A *graph* is a pair $G = (V, E)$ where V is the set of vertices and $E \subset \binom{V}{2} = \{\{u, v\} : u, v \in V, u \neq v\}$ is the set of edges of the graphs. This definition corresponds to a simple, labeled, and undirected graph and we will assume that all graphs are of this form. In chapter 6 we will discuss extensions beyond this framework. Let $n = |V|$ be the number of vertices in the graph. For ease of notation and without loss of generality, we assume $V = [n] = \{1, 2, \ldots, n\}$.

Every graph can be represented by an adjacency matrix $A \in \mathcal{A}$, where $\mathcal{A} = \{0, 1\}^{n \times n}_{sym}$ the set of symmetric binary matrices. For each $u, v \in [n]$, the entry $A_{uv} = 1$ if u is adjacent to v and 0 otherwise. We will always work directly with the matrix A and not consider the traditional graph theoretic setup with $G = (V, E)$. This perspective is appropriate given our linear algebraic methods.

Most of the work in this thesis will be concerned with the linear algebraic aspects of graphs but occasionally we will refer to some traditional graph theoretic concepts. We list a only the bare minimum additional concepts here:

order The order of a graph is the number of vertices $n = |V|$.

size The size of the graph is the total number of edges in the graph which is given

by $\frac{1}{2}\mathbf{1}^\top A\mathbf{1}$.

density The density of a graph is number of edges divided by the number of possible edges given by $\frac{1}{n^2-1}\mathbf{1}^\top A\mathbf{1}$.

dense If we consider a sequence of graphs with order $n \to \infty$ then this is a dense sequence if the density is bounded from below by a fixed constant.

sparse A sequence of graphs is sparse if it is not dense and the density tends to 0.

degree The number of edges incident to a given vertex. This is given in terms of the adjacency matrix as $d_i = (A\mathbf{1})_i = \sum_{j=1}^{n} A_{ij}$.

path An alternating sequence of vertices and edges where each edge is incident to the vertices before and after in a sequence. Informally a path between vertices is a way to get from one vertex to another following edges in the graph.

connected component A subset of vertices such that there is a path between any two vertices in the the subset.

connected/disconnected A graph is connected if it has only one maximal connected component. It is disconnected otherwise.

2.2 Random Graphs

A random graph is a graph whose vertex set is fixed and the edge set is distributed according to some distribution over all possibly edge sets. For the remainder of this work we will assume that we have a fixed an all encompassing probability space $(\Omega, \mathscr{F}, \mathbb{P})$ from which all random variables arise. A random graph $A : \Omega \mapsto \mathcal{A}$ is a map from the probability sample space to the space of all adjacency matrices on n vertices. [1]

Example 2.1 (Erdős-Rényi Graphs)**.** Erdős-Rényi (ER) graphs were the first random graphs introduced originally by Gilbert [1959]. In this model, each possible edge is present in the random graph independently with probability $p \in [0, 1]$. In our notation, for all $u, v \in [n]$ with $u < v$, we write $A_{uv} \overset{iid}{\sim} \text{Bern}(p)$. For example, in the case that $p = \frac{1}{2}$, all possible (labeled) graphs on n vertices are equally likely. For a given p and any $A' \in \mathcal{A}$ we have

$$\mathbb{P}[A = A'] = \prod_{i<j} p^{A'_{ij}} (1-p)^{1-A'_{ij}}.$$

Frequently in the literature on random graphs it is assumed that p somehow depends on n. As discussed above, the *dense* case is where p is fixed in n, so that the density of the graph does not decay with n. If p decays with n,we say graph is *sparse*. We will introduce analogous concepts in other random graph models as they come up.

[1] Note, we will not make a notational distinction between a random and non-random adjacency matrix. We justify this abuse by the fact that we will almost always be dealing with *random* adjacency matrices and when an adjacency matrix is *non-random*, it will be made obvious from the context.

The ER model is particularly simple and the results in this work can be stated very concisely in this case. However, do not be fooled, the literature for the ER model is rich and deep, especially the work on the phase transition, originally observed by Erdős and Rényi [1961], in which various graph properties change rapidly as p passes through various values around the level of $\frac{1}{n}$. Bollobás [2001] provides a detailed analysis of much of the early results in the field of random graphs, particularly related to this model. □

A second example is more general but does not come close to encompassing all random graph distributions. This example will be considered in detail in Chapter 3.

Example 2.2 (Independent Edge Graphs)**.** For an independent edge graph, as in the ER model, the presence or absence of each edge is independent of all other edges but here the probability each edge is present is allowed to vary. In our notation we have

$$\text{for all } u, v \in [n] \text{ with } u < v, \quad A_{uv} \overset{ind}{\sim} \text{Bern}(p_{uv}), \tag{2.1}$$

for some collection $\{p_{uv}\}_{u,v\in[n]}$. It will be frequently convenient to consider the matrix of probabilities denoted by $P = (p_{uv})_{u,v=1}^{n} \in [0,1]_{sym}^{n\times n}$. For a given probability matrix P, we write $A \sim \text{IEG}(P)$ denote an IEG graph with the specified edge probabilities.

Depending on the particular values p_{uv}, different connectivity structures may be more or less likely to arise. Note, that accurate estimation in the case that the p_{uv} are allowed to vary freely is impossible in case of observing only one graph (see remark 3.2). In Chapter 3, we will see that accurate estimation of low-rank approxi-

mations of the matrix P are possible by spectral methods. □

Most generally, a random graph A can be thought of as a multinomial taking values in the large but finite space $\mathcal{A}$. Recall that $|\mathcal{A}| = 2^{\binom{n}{2}}$. This viewpoint yields no intuition on the structures of the adjacency matrices where the random graph may concentrate its probability but can be valuable when pondering the extent to which distributions over graphs may misbehave. In the next section, we discuss in detail a class of models in which the edges are either independent or conditionally independent.

2.3 Latent Position Graphs

When modeling graphs for statistical inference, we want to select a model that imposes a level of parsimony appropriate for the problem at hand. The ER model imposes excessive parsimony for most problems. The set of all models is far too complicated and even the independent edge model can have too little parsimony. In this section, we present a middle ground.

One of the key aspects of graph data is that different collections of vertices *behave* differently. Vertices are selective about their adjacencies: vertices in one group may be frequently adjacent to vertices a second group but rarely adjacent to vertices in a third group. Modeling these differences in vertex properties is one of the goals of latent position graph models, where each vertex is associated with a latent positions

that influences the adjacencies for that vertex.

Latent position graphs (LPG) were introduced in Hoff et al. [2002]. In LPG models, the probability an edge is present in a graph is controlled by latent, unobserved properties of the incident vertices. In a social network setting, an interpretation is that individuals form relationships based on characteristics of each individual such as shared interests, complementary talents, or proximity. Certain shared or distinct characteristics are conducive to relationships while others are not. Hoff et al. [2002] called this latent space of individual characteristics the "social space".

Definition 2.3 (Latent Position Graph (LPG))**.** Formally, let $\mathcal{X}$ be some space and let $x_1, x_2, \ldots, x_n \in \mathcal{X}$. Let $\kappa : \mathcal{X}^2 \mapsto [0,1]$ be a symmetric function. The random graph is then realized such that

$$\text{for all } u, v \in [n], u < v, \quad A_{uv} \overset{ind}{\sim} \text{Bern}(\kappa(x_u, x_v)).$$

Hence, for $A \sim \text{LPG}(\{x_i\}_{i=1}^n, \kappa)$ and for any $A' \in \mathcal{A}$ we have

$$\mathbb{P}[A = A'] = \prod_{i<j} \kappa(x_i, x_j)^{A'_{ij}} (1 - \kappa(x_i, x_j))^{1 - A'_{ij}}.$$

In this notation, $\mathcal{X}$ is the *latent space*, $\{x_i\}_{i=1}^n$ is the set of *latent positions* and the function κ is the *link function* (or kernel). The function κ is a similarity function for elements of $\mathcal{X}$ which returns the probability that two elements of $\mathcal{X}$ will be adjacent in the resulting graph.

If we take the latent positions to be non-random then this is simply a constrained IEG model. For a given κ and $\mathcal{X}$, the matrix of probabilities P can be realized

from this LPG model provided there exist $x_1, \ldots, x_n \in \mathcal{X}$ such that $p_{uv} = \kappa(x_u, x_v)$ for all $u, v \in [n]$. The situation becomes more interesting if we allow the latents to be random, in which case the edges are no longer independent but are instead conditionally independent.

Definition 2.4 (LPG with random latent positions)**.** Suppose the latent positions are given by random variables $X_i : \Omega \mapsto \mathcal{X}$ for $i \in [n]$. Then the adjacencies are conditionally independent so

$$\text{for all } u, v \in [n], u < v, \quad A_{uv} | (X_u, X_v) \overset{ind}{\sim} \text{Bern}(\kappa(X_u, X_v)).$$

More formally, for any $A' \in \mathcal{A}$ we have

$$\mathbb{P}[A = A' | X_1, \ldots, X_n] = \prod_{i<j} \kappa(X_i, X_j)^{A'_{ij}} (1 - \kappa(X_i, X_j))^{1 - A'_{ij}}.$$

Typically, we will suppose that $X_1, \ldots, X_n \overset{iid}{\sim} F$ for some distribution F on $\mathcal{X}$. In this situation we abuse notation somewhat and write $A \sim \text{LPG}(\mathcal{X}, F, \kappa)$.

If the latent positions are i.i.d., we will be able to make statements about statistical inference in the LPG setting analogous to the i.i.d. setting in classical statistics (see Chapter 5). Generalizations beyond the i.i.d. setting are also possible but in those cases we focus primarily on generic IEG and particularly on low rank IEG models such as the random dot product graph model.

2.3.1 Random Dot Product Graphs

The random dot product graph (RDPG) [Young and Scheinerman, 2007, Nickel, 2006] will be our canonical representation for IEGs and LPGs (see Section 2.4).

Definition 2.5 (Random dot product graph (RDPG))**.** The RDPG is an LPG with latent space $\mathcal{X} \subset \mathbb{R}^d$ which satisfies $\langle x, y \rangle \in [0, 1]$ for all $x, y \in \mathcal{X}$ and link function given by $\kappa(x, y) = \langle x, y \rangle$. Hence, if the latent positions are $x_1, \ldots, x_n \in \mathcal{X}$, then for $i < j \in [n]$ we have $A_{ij} \overset{ind}{\sim} \text{Bern}(\langle x_i, x_j \rangle)$. Hence $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$ parametrizes the random dot product graph model. In this case we say an adjacency matrix $A \sim \text{RDPG}(X)$.

As with many LPG models, there is an important non-identifiability for RDPG distributions. In particular, for any orthogonal matrix $W \in \mathcal{O}(d) \subset \mathbb{R}^{d \times d}$, it holds that the distributions $\text{RDPG}(X)$ and $\text{RDPG}(XW)$ are equal. Hence, we do not attempt to estimate X explicitly but are satisfied with merely estimating the equivalence class of X, namely $\{XW : W \in \mathcal{O}(d)\}$.

The matrix of edge probabilities is given by the outer product $P = XX^\top$. Importantly, P is positive semidefinite and $\text{rank}(P) = \text{rank}(X)$. In particular if X has full rank then $\text{rank}(P) = d$. The spectral properties of P are easily analyzed because of its representation as an outer product of a lower dimensional matrix with itself. Indeed, from P we can easily recover an element of the equivalence class of X by computing the spectral decomposition of P (see Section 3.1).

Definition 2.6 (RDPG with iid latent positions)**.** In the iid latent position case we will write $A \sim \text{RDPG}(\mathcal{X}, F)$ for some valid latent space $\mathcal{X}$ some distribution F on $\mathcal{X}$. In this case the latent positions are given by $X_1, \ldots, X_n \overset{iid}{\sim} F$ and $X = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n \times d}$.

Remark 2.7. We note that constraint that the matrix P is positive semidefinite can make modeling certain types of graph data difficult. In this case you could still preserve the simple spectral properties of P if you consider LPG with link function $\kappa(x, y) = x^\top M y$ where $M \in \mathbb{R}^{d \times d}$ is diagonal with $M_{ii} = 1$ for $i \leq d'$ and $M_{ii} = -1$ for $i > d'$ for some $d' > 1$. In this case the P matrix will still have rank at most d. Models of this kind are discussed in indirectly in the context of general IEG distributions but we do not go into the details of inference in this model, though some of our results can easily be extended to this setting.

2.3.2 Stochastic Blockmodel Graphs

The stochastic blockmodel (SBM) [Holland et al., 1983] is a model with a strong notion of community. In the LPG framework, a random graph is said to be distributed according to an SBM with $K \in \mathbb{N}$ blocks if there are only finitely many distinct latent positions.

Definition 2.8 (SBM)**.** Suppose $A \sim \text{LPG}(\{x_i\}_{i=1}^n, \kappa)$ and suppose there are $\xi_1, \ldots, \xi_k \in \mathcal{X}$ such that for all $i \in [n]$ there exists $k \in [K]$ with $x_i = \xi_k$. We define the

block membership function $\tau : [n] \mapsto [K]$ so that $x_i = \xi_{\tau(i)}$. In this case we say $A \sim \text{SMB}(\tau, \{\xi_i\}_{i=1}^k, \kappa)$.

Traditionally, the SBM is parametrized by a matrix $B \in [0,1]^{k\times k}_{sym}$ which specifies the probabilities for edges between vertices in different blocks and one would write $A \sim \text{SBM}(\tau, B)$. In the LPG setting we have $B_{kl} = \kappa(\xi_k, \xi_l)$ and in the RDPG setting we have $B_{kl} = \xi_k^\top \xi_l$. Importantly an SBM graph can be represented as an RDPG graph if and only if the matrix B is positive semi-definite.

Definition 2.9 (SBM with iid block memberships)**.** Finally, if the block memberships are taken to be iid then we say $A \sim \text{SBM}(\pi, \{\xi_i\}_{i=1}^K, \kappa)$ where $\pi \in (0,1)^K$ satisfies $\sum_{k=1}^K \pi_k$. The blockmembership function then is given by $\tau(1), \ldots, \tau(n) \overset{iid}{\sim} \text{Multinomial}(\pi)$.

2.3.3 Exchangeable Graphs

The assumption that a set of random variables are exchangeable is a frequently made assumption and is critical for the theoretical justification of many statistical procedures. This assumption is implied by the iid assumption and is quite reasonable for many scenarios.

One of the most important theorems regarding an exchangeable sequence of random variables is de Finetti's Theorem which characterizes these sequences as conditionally iid sequences. In this section we will present the analogous theory for so

called exchangeable random graphs which provides another theoretical justification for studying latent position graphs.

Definition 2.10 (Exchangeable Random Graph)**.** A random graph is said to be *exchangeable* if for all $n \times n$ permutation matrices E, the adjacency matrices $EAE^\top$ and A are identically distributed.

For completeness we will also present the concept of an exchangeable array so as to introduce the key theorem about exchangeable random graphs.

Definition 2.11 (Exchangeable Array)**.** An infinite rectangular array $(A_{ij})_{i,j=1}^{\infty}$ of real-valued random variables A_{ij} is said to be *exchangeable* if for every bijection $\beta : \mathbb{N} \mapsto \mathbb{N}$, the arrays $(A_{ij})_{i,j=1}^{\infty}$ and $(A_{\beta(i)\beta(j)})_{i,j=1}^{\infty}$ are identically distributed.

The key result regarding exchangeable random graphs is the following analog to de Finetti's Theorem due to Aldous [1981] and Hoover [1979].

Theorem 2.12 ([Aldous, 1981, Hoover, 1979])**.** *Let $(A_{ij})_{i,j=1}^{\infty}$ be a symmetric array of real valued random variables so in particular $A_{ij} = A_{ji}$. The array is exchangeable if and only if there exists a function $f : [0,1]^4 \mapsto \mathbb{R}$ such that*

$$A_{ij} = f(\alpha, \xi_i, \xi_j, \lambda_{ij})$$

for $\alpha, \xi_i, \lambda_{ij} \overset{iid}{\sim} \text{Unif}([0,1])$ for $i, j = 1, 2, \ldots$.

This theorem is analogous to the de Finetti Theorem which asserts a similar statement for an infinite exchangeable sequence of random variables. See Diaconis

and Janson [2008] for further details and other formulations of exchangeable random graphs.

In our case, where the A_{ij} are Bernoulli random variables, the function f can be always be chosen to take the form

$$f(\alpha, \xi_i, \xi_j, \lambda_{ij}) = \mathbb{I}\{\lambda_{ij} < \kappa_\alpha(\xi_i, \xi_j)\}.$$

Equivalently, conditioned on α, ξ_i, ξ_j, then $A_{ij} \sim \text{Bern}(\kappa_\alpha(\xi_i, \xi_j))$, where for every $\alpha \in [0, 1]$, the function $\kappa_\alpha : [0, 1]^2 \mapsto [0, 1]$ is a link function.

Succinctly, the Aldous-Hoover Theorem asserts that a random graph is a subgraph of an infinite exchangeable random graph if and only if it is a mixture (over link functions) of latent position graphs with iid latent positions. Hence, as exchangeability is frequently a reasonable assumption for the vertices, latent position graphs are a natural class of models to consider. In the case $\kappa_\alpha = \kappa$ for all α, then it may be natural to denote such a random graph as an iid random graph, again analogous to the notion of an iid sequence.

2.4 The Random Dot Product Graph Representation

Recall that an LPG is parametrized by the latent position space $\mathcal{X}$, the link function κ and either the latent positions $x_1, \ldots, x_n$ or a distribution for the latent

positions F. If the latent positions are taken to be fixed then it is always possible to reparametrize the graph so that the link function is a bilinear form and the latent space is a subset of $\mathbb{R}^d$ for d at most n. In Section 3.1 we detail how this can be done by considering the eigendecomposition of the matrix $P \in [0,1]^{n\times n}_{sym}$ where $P_{ij} = \kappa(x_i, x_j)$ for all $i, j \in [n]$. Furthermore, if the link function is positive semidefinite (see Definition 2.13) then we can take the link function to be the standard inner product on $\mathbb{R}^d$.

If instead the latent positions are taken to be random, then a reparametrization with bilinear form on $\mathbb{R}^n$ is not necessarily possible. In order to address this case, we will briefly introduce a few ideas from operator theory in order to construct an appropriate latent position distribution on the infinite dimensional space ℓ^2 for which an inner product link function will provide the analogous reparametrization.

2.4.1 Relevant Operator Theory

First, we define a positive semidefinite link function.

Definition 2.13 (Positive semidefinite link function)**.** A link function $\kappa : \mathcal{X}^2 \mapsto [0,1]$ is said to be positive semidefinite if for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \kappa(x_i, x_j) \geq 0.$$

The link function is said to be positive definite if the inequality above is strict.

Note that whether the link function is positive semidefinite is related to whether the matrix $P = (\kappa(x_i, x_j))_{i,j=1}^n$ is positive semidefinite.

To address the iid case, we introduce the integral operator $\mathscr{K}$ associated with the link funciton κ.

Definition 2.14. Let $\kappa : \mathcal{X}^2 \mapsto [0, 1]$ be a positive semidefinite link function and let F be a distribution on $\mathcal{X}$. Let $L^2(\mathcal{X}, F)$ be the Hilbert space of square-integrable functions with respect to F. We define the integral operator $\mathscr{K} : L^2(\mathcal{X}, F) \mapsto L^2(\mathcal{X}, F)$ by

$$(\mathscr{K} f)(x) = \int_{\mathcal{X}} \kappa(x, x') f(x') dF(x').$$

The operator $\mathscr{K}$ has a two key properties that are crucial for subsequent analysis.

Proposition 2.15 (Blanchard et al. [2007])**.** *For a positive semidefinite link function κ, the associated integral operator $\mathscr{K}$ is a compact operator and of trace class. Formally,*

$$\sup_{f \in L^2(\mathcal{X}, F) : \|f\|_{L^2} \leq 1} \|\mathscr{K} f\|_{L^2} < \infty$$

and the trace is well defined meaning there is some finite t such that $\operatorname{tr}(\mathscr{K}) = t$ and

$$t = \sum_{i=1}^{\infty} \langle \mathscr{K} e_i, e_i \rangle$$

for any choice of orthonormal basis $(e_i)_{i=1}^n$ of $L^2(\mathcal{X}, F)$.

Furthermore $\operatorname{tr}(\mathscr{K}) = \sum_{i=1}^{\infty} \lambda_i$ where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq 0$ are the non-negative eigenvalues with multiplicities of $\mathscr{K}$. We let $\psi_1, \psi_2, \ldots \in L^2(\mathcal{X}, F)$ be a set

of orthonormal eigenfunctions of $\mathscr{K}$ so that

$$\mathscr{K}\psi_j = \lambda_j\psi_j \text{ and } \langle\psi_i, \psi_j\rangle_{L^2} = \int_{\mathcal{X}} \psi_i(x)\psi_j(x)dF(x) = \delta_{ij}. \tag{2.2}$$

The Mercer representation theorem characterizes the connection between the link function κ and the eigenpairs of $\mathscr{K}$. Before we stating this result, we define the reproducing kernel Hilbert space associated with a positive semidefinite link function κ.

Definition 2.16 (Reproducing Kernel Hilbert Space (RKHS))**.** For a positive semidefinite link function $\kappa : \mathcal{X}^2 \mapsto [0, 1]$ there is a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ for which κ is a reproducing kernel, meaning for every $f \in \mathcal{H}$ and $x \in \mathcal{X}$, the function $\kappa(x, \cdot) \in \mathcal{H}$ and

$$\langle f, \kappa(x, \cdot)\rangle_{\mathcal{H}} = f(x).$$

Theorem 2.17 (Mercer's Representation Theorem)**.** *Let $(\mathcal{X}, d)$ be a compact metric space, $\kappa : \mathcal{X}^2 \mapsto [0, 1]$ a positive semidefinite link function, and $\mathscr{K}$ its associated integral operator. Let $\lambda_1 \geq \lambda_2 \geq \cdots, \geq 0$ and $\psi_1, \psi_2, \ldots \in L^2(\mathcal{X}, F)$ be the eigenvalues and associated eigenvectors of $\mathscr{K}$. The link function κ can be characterized in the following ways.*

1. *The link function can be written as*

$$\kappa(x, x') = \sum_{j=1}^{\infty} \lambda_j\psi_j(x)\psi_j(x'). \tag{2.3}$$

The sum in Eq. (2.3) converges absolutely for each $x, x' \in \text{supp}(F)$ and uniformly in $\text{supp}(F) \times \text{supp}(F)$.

2. *Let $\mathcal{H}$ denote the RKHS of κ then the elements $\eta \in \mathcal{H}$ are of the form*

$$\eta = \sum_{j=1}^{\infty} a_j \sqrt{\lambda_j} \psi_j \text{ with } (a_j)_{j=1}^{\infty} \in \ell_2$$

and the inner product $\mathcal{H}$ is given by

$$\left\langle \sum_{j=1}^{\infty} a_j \sqrt{\lambda_j} \psi_j, \sum_{j=1}^{\infty} b_j \sqrt{\lambda_j} \psi_j \right\rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} a_j b_j.$$

3. *We can define a feature map $\phi : \mathcal{X} \mapsto \ell_2$ by $\phi(x) = (\sqrt{\lambda_j}\psi_j(x))_{j=1}^{\infty}$ so that $\langle \phi(x), \phi(x') \rangle_{\ell_2} = \kappa(x, x')$.*

Though all three parts of this theorem are of interest, the third part provides the key connection between positive semidefinite link functions and RDPG. Indeed, this theorem indicates that the random graph distribution $\text{LPG}(\mathcal{X}, \kappa, F)$ is equivalent to $\text{LPG}(\ell_2, \langle \cdot, \cdot \rangle_{\ell_2}, F_{\ell_2})$ where F_{ℓ_2} is the distribution of $\phi(X_i)$ for $X_i \sim F$. When studying the spectral properties of random graphs, this particular form will become important and properties of the distribution F_{ℓ_2} will mirror properties of the integral operator $\mathscr{K}$. In Chapter 4 we will go into more detail about how this represantation can be used to study the adjacency spectral embedding. For now, we are content to note the link and provide the following example.

2.4.2 An example

Before ending this chapter we want to note that the positive semidefinite assumption is indeed necessary for the theorems in the previous section to hold. Clearly,

if the integral operator is not positive semidefinite then a representation of the link function as an inner product on ℓ_2 is impossible because the inner product is positive semidefinite. One can argue that since the link function is the difference of two positive semidefinite link we can represent the link function as the difference between two inner products: a function on $\kappa' : (\ell_2 \times \ell_2)' \to \mathbb{R}$ such that $\kappa'((x_+, x_-), (y_+, y_-)) = \langle x_+, y_+ \rangle_{\ell_2} - \langle x_-, y_- \rangle_{\ell_2}$. On the other hand the important notion that the integral operator is trace class can fail leading to the situations where the embedding dimension required for accurate inference can tend to infinity very rapidly.

Example 2.18. As a concrete example of a consider a random *threshold graph*. In this case, we have $A \sim LPG([0,1], \kappa, \text{Unif})$ where $\kappa(x,y) = \mathbb{I}\{x+y > 1\}$. Note, as all edge probabilities are either 0 or 1, the adjacency matrix is completely determined by the sampled latent positions. The integral operator $\mathscr{K}$ acts on $L^2([0,1])$ and is given by

$$(\mathscr{K} f)(x) = \int_0^1 \mathbb{I}\{x + y > 1\} f(y) dy = \int_{1-x}^1 f(y) dy.$$

It is not hard to verify that the (unormalized) eigenfunctions of $\mathscr{K}$ are given by $f_k(x) = \sin((2k+1)\pi/2x)$ and the associated eigenvalue is $\lambda_k = \frac{2}{\pi(2k+1)}(-1)^k$. Clearly, the eigenvalues do not converge absolutely and hence $\mathscr{K}$ is not of trace class.

Additionally, since the link function is discontinuous along the line $x + y = 1$ and the eigenfunctions are all continuous, approximations of the link function with low rank approximations of A will be inaccurate along this line. A simulated example of

various low rank approximations of A is provided in Figure 2.1.

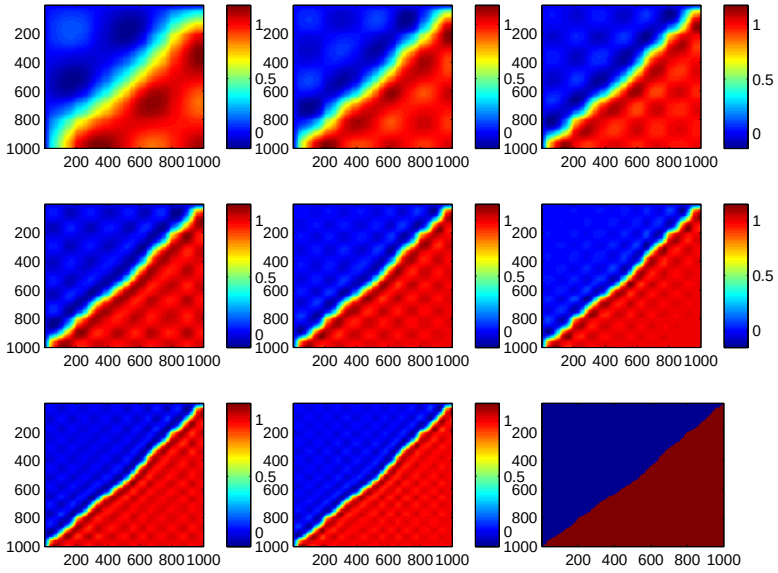


Figure 2.1: The first 8 panels each represent the best rank$(2k)$ approximation for $k = 1, \ldots, 8$ of a adjacency matrix for a random threshold graph on 1000 vertices. Note each low rank embedding uses of k positive and k negative eigenvalue. The perceived "checkerboard" pattern is due to the sinusoidal eigenfunctions. The lower right panel shows the binary adjacency matrix. Though far from the antidiagonal the accuracy of the approximation improves consistently, the discontinuity along the antidiagonal leads to consistently poor accuracy for these low rank approximations.

This model does not fit well into the story for the remainder of this manuscript since the adjacency matrix is exactly equal to the matrix of probabilities, as this matrix itself is binary. However, as all the eigenfunctions of the integral operator can be computed explicitly, it is an interesting and illustrative example.

Chapter 3

Latent Position Estimation

In this chapter we present our main results for the adjacency spectral embedding. Our immediate focus is the independent edge setting, whereas in Chapter 4 we focus on the conditionally independent edge setting where the latent positions are iid. We make an effort to make explicit non-asymptotic performance guarantees where the bounds are in terms of specific properties of the matrix of edge probabilities P. The bounds in this chapter are not necessarily tight—explicit constants are used throughout this chapter and though they can often be improved, we attempt to make the bounds simple and the asymptotic properties of the bounds close to the best possible.

We start by introducing the adjacency spectral embedding and the spectral embedding of P in Section 3.1. The next section is concerned with estimation when no assumptions are made on the matrix of edge probabilities P. In this general setting, we estimate the d-dimensional uncentered principal components of certain

n-dimensional latent positions associated with a spectral decomposition of P. We illustrate the theorem with a simulated example and conclude with a mathematically convenient corollary and a concentration inequality for the d largest eigenvalues of A.

In Section 3.3 we shift our focus to the case where $\text{rank}(P) = d \ll n$. In this case, linear algebraic methods akin to the power methods allow for large improvements in the bounds. We start that section with an overview of the main ideas and a presentation of some key lemmas. We then prove our most applicable result, a bound on the component-wise error in the estimation. This result has powerful implications for subsequent estimation. We end this chapter with a final improvement of the bound for the Frobenius norm in the rank d case.

3.1 Adjacency Spectral Embedding

For the entirety of this chapter we will suppose that $A \sim \text{IEG}(P)$ where $P \in [0,1]^{n\times n}_{sym}$ is a matrix of edge probabilities (see Example 2.2). As P is symmetric and real, it has real eigenvalues denoted $\lambda_1(P), \ldots, \lambda_n(P)$ where, by assumption,

$$|\lambda_1(P)| \geq |\lambda_2(P)| \geq \cdots \geq |\lambda_n(P)|.$$

The associated eigenvectors are $v_1(P), \ldots, v_n(P)$ with $\|v_i(P)\|_2 = 1$ for all $i \in [n]$. In matrix form we have the eigendecomposition $P = \tilde{V}\tilde{S}\tilde{V}^\top$ where $\tilde{S}$ is diagonal with $\tilde{S}_{ii} = \lambda_i(P)$ and $\tilde{V}$ is an orthogonal matrix with column i given by $v_i(P)$.

Remark 3.1. We will occasionally use different orderings of the eigenvalues. The

ordering introduced above is decreasing according to magnitude and denoted $\lambda_i(P)$ with no decorations. We will occasionally use the decreasing ordering denoted $\lambda_i^+(P)$ with $\lambda_1^+(P) \geq \lambda_2^+(P) \geq \cdots \geq \lambda_n^+(P)$ and the increasing ordering denoted $\lambda_i^-(P)$ with $\lambda_1^-(P) \leq \lambda_2^-(P) \leq \cdots \leq \lambda_n^-(P)$.

If we let $\tilde{X} = \tilde{V}|\tilde{S}|^{1/2} \in \mathbb{R}^{n\times n}$, then we have that $P = \tilde{X}\text{sign}(\tilde{S})\tilde{X}^\top$ where

$$\text{sign}(\tilde{S})_{ij} = \begin{cases} 0 & \text{if } \tilde{S}_{ij} = 0, \\ 1 & \text{if } \tilde{S}_{ij} > 0 \text{ and} \\ -1 & \text{if } \tilde{S}_{ij} < 0. \end{cases}$$

If we denote the i^{th} row of $\tilde{X}$ as $\tilde{x}_i^\top$, we can view A as a latent position with non-random latent positions $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ and link function given by

$$\kappa(x, y) = x^\top \text{sign}(\tilde{S})y,$$

a bilinear form. This chapter is concerned with truncated versions of $\tilde{X}$ which are also the uncentered principal components of $\tilde{X}$.

For $d \in [n]$, let $X^{(d)} \in \mathbb{R}^{n\times d}$ be the matrix given by the first d columns of $\tilde{X}$. The dimension d will frequently be obvious from context, in which case we omit the superscript and simply refer to X. Note that if $\text{rank}P = d$ then X is exactly the non-zero columns of $\tilde{X}$. Similar to $\tilde{X}$, let the i^{th} row of X be denoted by $x_i^\top$. We will call X the d-dimensional spectral embedding of P or simply the embedding of P in analogy with Definition 3.3 below.

Remark 3.2. As of now we make no assumptions on P. One may seek an estimate for P or equivelantly $\tilde{X}$ but with no further assumptions, non-trivial guarantees for the accuracy of this estimate will be impossible. Indeed, if we insist that P can be arbitrary then a perfectly reasonable estimate for P is just A itself—A is the maximum likelihood estimate for P under the IEG model with no constraints. On the other hand this can be quite a poor estimate. For $A \sim ER(1/2)$ since $(A_{ij} - P_{ij})^2 = 1/4$ for all $i \neq j$ we have

$$\|A - P\|_F = \sqrt{\frac{1}{4}n(n-1)} \approx n/2.$$

Of course, if we knew A had an Erdos-Renyi distribution then trivial estimates are many orders of magnitude more accurate.

Regardless, the results in the next section show that at least accurate *approximate* estimation in the full IEG model is possible. We accomplish this by restricting our goals to estimating a low rank approximation of P, in particular the d-dimensional spectral embedding X.

Our main object of study is the adjacency spectral embedding which is defined analogously to the spectral embedding of P.

Definition 3.3 (Adjacency Spectral Embedding)**.** Denote the eigendecomposition of A by $A = \hat{\hat{V}}\hat{\hat{S}}\hat{\hat{V}}^\top$.[1] We will again construct the matrix $\hat{\hat{X}} = \hat{\hat{V}}|\hat{\hat{S}}|^{1/2} \in \mathbb{R}^{n\times n}$. The matrix given by the first d columns of $\hat{\hat{X}}$ is denoted by $\hat{X}^{(d)}$ with rows $\hat{x}_i^{(d)}$ (or just

[1] Do not despair, the $\hat{\cdot}$ will disapear soon.

$\hat{X}$ and $\hat{x}_i$ when the dimension d is obvious from context). $\hat{X}$ is the d-dimensional spectral embedding of A or simply the adjacency spectral embedding.

We note that the spectral embedding can be applied to any square matrix and that X is the spectral embedding of P. The definition does not rely on any model assumptions and so can be applied to any graph without making specific assumptions about the generative mechanisms. We use the term embedding to emphasize the fact that this method provides a representation of each vertex as a vector in finite dimensional Euclidean space, despite an alternative use of "embedding" in graph theory literature. As we discuss in more detail in Chapter 5, this representation admits standard multivariate statistical techniques for which performance asymptomatic and non-asymptotic guarantees can be made.

3.2 Estimation for General P

The first main result of this work is the following, which shows that $\hat{X}$ is an accurate estimate of X provided the graph is large and the eigenvalues are well separated.

Theorem 3.4. *Let $A \sim \text{IEG}(P)$ where P be the matrix of edge probabilities and let $d \in [n]$. Let X be the d-dimensional embedding of P and let $\hat{X}$ be the d-dimensional embedding of A. Define*

$$\delta = \|P\mathbf{1}\|_\infty \text{ and } \gamma = \min_{i \leq d, j \geq d} |\lambda_i(P) - \lambda_j(P)|/\delta. \tag{3.1}$$

If $\gamma\sqrt{\delta} \geq 2\sqrt{\log(n/\eta)}$ *then with probability at least* $1-\eta$

$$\min_{W \in \mathcal{O}(d)} \|\hat{X}W - X\|_F \leq \gamma^{-1}\sqrt{32d\log(n/\eta)}, \tag{3.2}$$

where $\mathcal{O}(d)$ *denotes the set of* $d \times d$ *orthogonal matrices.*

All bounds in this chapter will depend only on the constants in Eq. (3.1), the dimension d and the number of vertices n. The constant δ is the maximum expectected degree among the vertices and $\gamma\delta$ gives the smallest gap between the the first d eigenvalues and the remaining eigenvalues.

If we consider γ and d fixed in n, as is natural in the iid latent position RDPG setting of Chapter 4, then the bound in Eq. 3.2 says $\|\hat{X} - X\| = O_P(\sqrt{\log(n)})$. This error rate turns out to be good enough to make certain asymptotic guarantees for subsequent inference such as in Theorem 5.9. However, improvements in the low rank case make many arguments easier.

The proof of this theorem relies on three key results. One of the results is the famous Davis-Kahan Theorem [Davis and Kahan, 1970] which provides a bound on the perturbation of invariant subspaces in terms of the eigengap when a matrix is perturbed. The second result is a powerful concentration inequality for random matrices proved first in Oliveira [2009] and subsequently improved by Tropp [2012] where the bound is in terms of the maximum expected degree δ. The third result is a lemmma that provides a perturbation bound for two spectral embeddings [Tang et al., 2013].

Our theorem builds on and generalizes results in Rohe et al. [2011], Sussman et al. [2012, 2013], and Tang et al. [2013]. In this section we will present the proof

of Theorem 3.4, starting with the proof of the third key result, then proving another simple Lemma and finally completing the proof of the main theorem. In Chapter 1 we gave an overview of the results of Davis and Kahan [1970] and now we state the result of Tropp [2012] for our current setting.

Theorem 3.5 (Oliveira [2009], Tropp [2012]). *Let $P \in [0,1]^{n\times n}$ be a matrix of edge probabilities for an independent edge random graph with adjacency matrix A and let $\delta = \|P\mathbf{1}\|_\infty < n$ be the maximum expected degree of the vertices in A. The following holds for all $\eta > 0$,*

$$\mathbb{P}\left[\|A - P\|_{2\to 2} \leq 2\sqrt{\delta \log(n/\eta)}\right] \geq 1 - \eta. \tag{3.3}$$

This next lemma provides a perturbation bound for the spectral embeddings of two generic symmetric matrices with rank d.

Lemma 3.6 (Tang et al. [2013]). *Let A and B be $n\times n$ positive semidefinite symmetric matrices with $\text{rank}A = \text{rank}B = d$. Let $X, Y \in \mathbb{R}^{n\times d}$ be the d-dimensional spectral embedding of A and B, respectively. Let γ be the smallest non-zero eigenvalue of B. Then there exists an orthogonal matrix $W \in \mathbb{R}^{d\times d}$ such that*

$$\|XW - Y\|_F \leq \frac{1}{\gamma}\|A - B\|_{2\to 2}(\sqrt{d\|A\|_{2\to 2}} + \sqrt{d\|B\|_{2\to 2}}). \tag{3.4}$$

Proof. Note first that $A = XX^\top$ and $B = YY^\top$. Let $R = A - B$. Since Y is of full column rank $Y^\top Y$ is invertible and its smallest eigenvalue is γ. We then have

$$Y = XX^\top Y(Y^\top Y)^{-1} - RY(Y^\top Y)^{-1}.$$

Let $T = X^\top Y(Y^\top Y)^{-1}$ so that

$$T^\top T - I = (Y^\top Y)^{-1}Y^\top XX^\top Y(Y^\top Y)^{-1} = (Y^\top Y)^{-1}Y^\top RY(Y^\top Y)^{-1}.$$

Therefore,

$$-(Y^\top Y)^{-1}Y^\top \|R\|_{2\to 2}Y(Y^\top Y)^{-1} \preceq T^\top T - I \preceq (Y^\top Y)^{-1}Y^\top \|R\|_{2\to 2}Y(Y^\top Y)^{-1},$$

where $\preceq$ denotes the positive semidefinite ordering for matrices. We thus have

$$\|T^\top T - I\|_F \leq \|R\|_{2\to 2}\|(Y^\top Y)^{-1}\|_F \leq \sqrt{d}\|R\|\|(Y^\top Y)^{-1}\|_{2\to 2} \leq \|R\|\sqrt{d}/\gamma.$$

Now, let W be the orthogonal matrix in the polar decomposition $T = W(T^\top T)^{1/2}$. We then have

$$\begin{aligned}
\|XW - Y\|_F &\leq \|XW - XT\|_F + \|XT - Y\|_F \\
&\leq \|X\|_{2\to 2}\|(T^\top T)^{1/2} - I\|_F + \|R\|_{2\to 2}\|Y(Y^\top Y)^{-1}\|_F \\
&\leq \|X\|_{2\to 2}\|(T^\top T)^{1/2} - I\|_F + \|R\|_{2\to 2}\|Y\|_{2\to 2}\|(Y^\top Y)^{-1}\|_F \\
&\leq \frac{\sqrt{d}}{\gamma}(\|X\|_{2\to 2} + \|Y\|_{2\to 2})\|R\|_{2\to 2} \qquad (*) \\
&= \frac{\sqrt{d}}{\gamma}(\sqrt{\|A\|_{2\to 2}} + \sqrt{\|B\|_{2\to 2}})\|A - B\|_{2\to 2},
\end{aligned}$$

where the inequality preceding (*) uses the fact that

$$\|(T^\top T)^{1/2} - I\|_F^2 = \sum_{i=1}^{d}(\lambda_i(T^\top T)^{1/2} - 1)^2 \leq \sum_{i=1}^{d}(\lambda_i(T^\top T) - 1)^2 = \|T^\top T - I\|_F^2.$$

□

This next lemma provides bounds for certain projection matrices and differences between the eigenvectors of A and P. It is a straightforward application of the Davis-Kahan Theorem, Theorem 3.5, and Lemma 3.6.

Lemma 3.7. *Let $A \sim \text{IEG}(P)$ and let $\delta = \|P\mathbf{1}\|_\infty$. Let $V \in \mathbb{R}^{n\times d}$ denote the matrix with orthonormal columns given by the eigenvectors of P corresponding to the d largest* positive *eigenvalues. Let $\hat{V}$ be defined analogously for A. Let $\zeta = \lambda_d^+(P) - \lambda_{d+1}^+(P)$. If $\eta > 0$ satisfies $\zeta > 2\sqrt{\delta \log(n/\eta)}$ then with probability at least $1-\eta$*

$$\|VV^\top - \hat{V}\hat{V}^\top\|_{2\to 2} \leq 2\zeta^{-1}\sqrt{\delta \log(n/\eta)} \tag{3.5}$$

$$\textit{and} \quad \min_{W\in\mathcal{O}(d)} \|V - \hat{V}W\|_F \leq 4\zeta^{-1}\sqrt{d\delta \log(n/\eta)}. \tag{3.6}$$

Proof. We will use the Davis-Kahan theorem with $\mathcal{S} = [\lambda_d^+(P) - \zeta/3, \infty]$ and note that the nearest eigenvalue of P to $\mathcal{S}$ outside of $\mathcal{S}$ is $\lambda_{d+1x}^+(P)$ at distance ζ. Note, by the condition $\zeta > 2\sqrt{\delta \log(n/\eta)}$ and Theorem 3.5, it occurs with probability at least $1-\eta$ that the d largest positive eigenvalues of A will all be in $\mathcal{S}$:

$$\mathbb{P}[\lambda_d^+(P) \in \mathcal{S}, \lambda_{d+1}^+ \notin \mathcal{S}] \geq 1-\eta.$$

Provided this happens, the Theorem 1.1 gives

$$\|VV^\top - \hat{V}\hat{V}^\top\|_{2\to 2} \leq \zeta^{-1}\|A - P\|_{2\to 2} \leq 2\zeta^{-1}\sqrt{\delta \log(n/\eta)}.$$

Lemma 3.6 gives the second part. □

Now we are ready to prove Theorem 3.4.

Proof of Theorem 3.4. Recall, that

$$\gamma\delta = \min_{i\leq d, j\geq d} |\lambda_i(P) - \lambda_j(P)| \leq \zeta,$$

where ζ is as in Lemma 3.7. If we let

$$d^+ = \max\{i : \lambda_i^+(P) \geq |\lambda_d(P)|\} \text{ and } d^- = \max\{i : \lambda_i^-(P) \leq -|\lambda_d(P)|\}$$

so that in particular $d^+ + d^- = d$, then similarly hat $\lambda_{d^+}^+(P) - \lambda_{d^+ +1}^+ \geq \gamma\delta$ and $\lambda_{d^- +1}^-(P) - \lambda_{d^-}^-(P) \geq \gamma\delta$. Also, by the Gershgorin disks theorem $\|P\|_{2\to 2} \leq \delta$ so $\gamma \leq 1$.

Theorem 3.5 ensures the event $\|A - P\|_{2\to 2} \leq 2\sqrt{\delta \log(n/\eta)}$ occurs with probability at least $1 - \eta$. The remainder of the proof is non-probabilistic as we use only deterministic implications of this event.

As before, we will first prove the bound for the columns of X corresponding to positive eigenvalues. Suppose $V, \hat{V} \in \mathbb{R}^{n\times d^+}$ are as in Lemma 3.7. Let $\mathcal{P}_A = \hat{V}\hat{V}^\top$ and let $\mathcal{P}_P = VV^\top$. We have that

$$\begin{aligned}
\|\mathcal{P}_A A - \mathcal{P}_P P\|_{2\to 2} &\leq \|\mathcal{P}_A(A - P)\|_{2\to 2} + \|(\mathcal{P}_A - \mathcal{P}_P)P\|_{2\to 2} \\
&\leq \|\mathcal{P}_A\|_{2\to 2}\|A - P\|_{2\to 2} + \|\mathcal{P}_A - \mathcal{P}_P\|_{2\to 2}\|P\|_{2\to 2} \\
&\leq 2\sqrt{\delta \log(n/\eta)} + 2\frac{\sqrt{\delta \log(n\eta)}}{\gamma\delta}\delta \\
&\leq 4\gamma^{-1}\sqrt{\delta \log(n/\eta)}.
\end{aligned}$$

Now, let X^+ and $\hat{X}^+$ be the submatrices of X and $\hat{X}$, respectively, corresponding to the positive eigenvalues of P. Now using Lemma 3.6 and the fact that $\mathcal{P}_A A$ and $\mathcal{P}_P P$

are rank d^+, we have that there exists an orthogonal matrix $W \in \mathbb{R}^{d\times d}$

$$\begin{aligned}\|\hat{X}^+W - X^+\|_F &\leq \frac{\|A-P\|_{2\to 2}}{\gamma\delta}(\sqrt{d^+\|A\|_{2\to 2}} + \sqrt{d^+\|P\|_{2\to 2}}) \\ &\leq \frac{2\sqrt{\delta\log(n/\eta)}}{\gamma\delta}(\sqrt{2d^+\delta} + \sqrt{d^+\delta}) \\ &\leq 4\gamma^{-1}\sqrt{d^+\log(n/\eta)},\end{aligned}$$

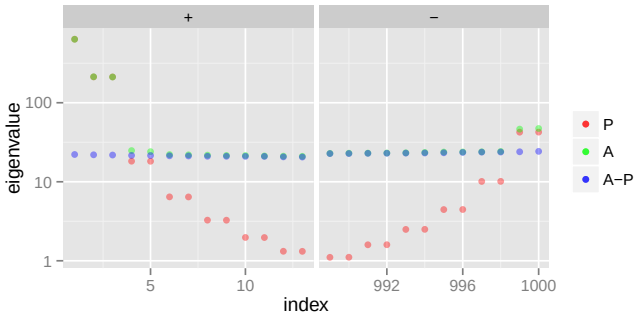
as $\|A\|_{2\to 2} \leq \delta + 2\sqrt{\delta\log(n/\eta)} \leq 2\delta$.

Bounding the coordinates of X and $\hat{X}$ corresponding to negative eigenvalues follows the exactly same proof, noting that Lemma 3.6 and Lemma 3.7 can be easily translated. The result follows since $\sqrt{d^+} + \sqrt{d^-} \leq \sqrt{2d}$. □

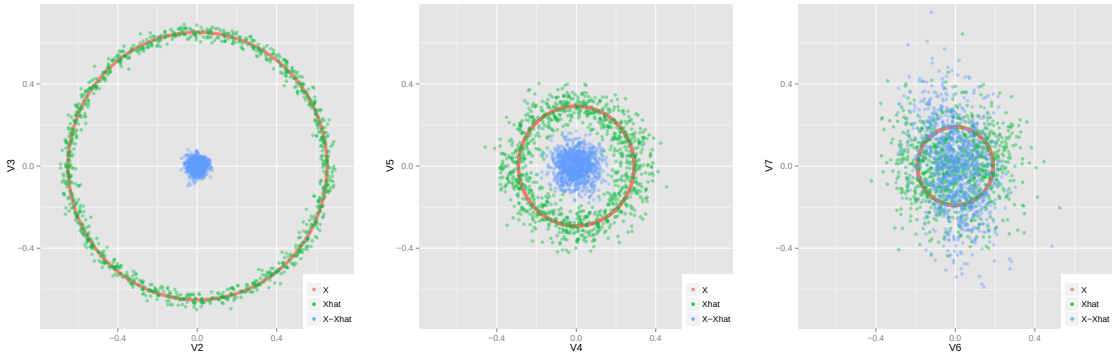
As an example to demonstrate the nature of the bounds we will consider a simple yet illustrative simulated example.

Example 3.8 (Circle Graph)**.** This example has the property that $\text{rank}(P) = n$ and P is indefinite but the eigenvalues decay rather quickly. We simulate a latent position graph where the latent positions $\xi_1, \ldots, \xi_{1000} \in \mathbb{R}^2$ are equispaced around the unit circle and the link function κ is $\kappa(x, y) = \frac{1}{2}\cos(\angle(x, y)/2)$.

Figure 3.1 gives an overview of the eigenvalues and spectral embeddings for this distribution. Figure 3.1(a) shows the largest eigenvalues in magnitude for P, A and $A - P$. For reference, the largest negative eigenvalues of A and P are slightly smaller than $2\sqrt{\delta}$. As Theorem 3.5 provides only an upper bound on $\|A - P\|_{2\to 2}$ we do not necessarily expect that this bound is sharp and indeed the eigenvalues of $A - P$ are much smaller, suggesting that this bound is not particularly tight.



(a) Eigenvalue magnitudes



(b) Dims. 2 & 3 (c) Dims. 4 & 5 (d) Dims. 6 & 7

Figure 3.1: For an instance of $A \sim \text{IEG}(P)$, with P defined in Example 3.8, panel (a) shows the magnitude of the largest eigenvalues for the matrices P, A, and $A - P$ in red, green and blue, respectively. The left panel shows the positive eigenvalues and the right panel shows the negative eigenvalues. We use the log-scale to illustrate the rapid decay of the eigenvalues of P as compared to those of A and $A - P$.
Panels (b), (c) and (d) show the 2^{nd} and 3^{rd}, 4^{th} and 5^{th}, and 6^{th} and 7^{th} dimensions, respectively, of the spectral embeddings. Each point corresponds to a row of either X, in red, $\hat{X}W$ in green, or $\hat{X}W - X$, in blue. Note that these panels are all on the same scale. We see that when the the eigenvalues match closely and are separated from the other eigenvalues, $\hat{X}$ well approximates X whereas by dimensions 6 and 7, the gap in the eigenvalues has diminished and $\hat{X}$ is very noisy.

If we let $W = \arg\min_{W' \mathcal{O}(d)} \|X - \hat{X}W'\|_F$, panels (b), (c), and (d) of Figure 3.1 show the rows of the embeddings X, $\hat{X}W$, and $\hat{X}W - X$. The accuracy of the adja-

cency spectral embedding degrades rapidly as a function of the gap in the eigenvalues. The small gap after eigenvalues 6 and 7 lead to a particularly poor approximation of X by $\hat{X}W$. On the other hand dimensions 2 and 3 show that $\hat{X}W$ is very close to X, which we would predict based on our Theorem because these eigenvalues are large and well separated. The first dimension is not shown because X is constant in its first dimension but again the fit is excellent in this case.

This examples illustrates some of the power of our theorem. Indeed, based on the the eigenvalues of P and the expected max degree, we might predict that accurate estimates are likely for the first three to five dimensions but after that performance would degrade rapidly. In reverse, if we observe A then since the largest five eigenvalues are much more separated than the other eigenvalues we can reasonably conclude that the estimates would be relatively accurate for the top five dimensions but unlikely to be accurate beyond that point.

3.2.1 Estimation without rotation

In the previous section we bounded $\|\hat{X}W - X\|_F$ in terms of the gap γ as defined in Eq. 3.1. In this section, we replace γ with a smaller constant based on gaps among the first d eigenvalues and the smaller eigenvalues of P. This makes the bounds worse but we can directly bound the difference between X and $\hat{X}$ seemingly *without* the need for a rotation. However, this is somewhat misleading as we will exploit any non-uniqueness of the spectral embedding of P to select X to minimize $\|\hat{X} - X\|_F$.

Nonetheless, this result is mostly a convenient and useful trick to deal with the non-identifiability of X by selecting a favorable representative from the equivalence class.

The proof is essentially the same and we do not present the details. The basic idea is that we can apply the method of proof above to each set of eigenvectors corresponding to the unique eigenvalues of P. Since the eigenvalues of A are near the eigenvalues of P we can apply the Davis-Kahan theorem separately to each such group.

Corollary 3.9. *Let d be fixed and let P be the matrix of edge probabilities for the independent edge random graph with adjacency matrix A. Let X be the d-dimensional embedding of P and let $\hat{X}$ be the d-dimensional embedding of A.*

Let $\delta = \|P\mathbf{1}\|_\infty$ and let

$$\gamma = \min \left\{ \min_{\substack{i,j \in [d] \\ \lambda_i(P) \neq \lambda_j(P)}} |\lambda_i(P) - \lambda_j(P)|, \min_{i \in [d], j > d} |\lambda_i(P) - \lambda_j(P)| \right\} / \delta \tag{3.7}$$

and suppose $\lambda_d(P) \neq \lambda_{d+1}(P)$. Provided $\gamma\sqrt{\delta} \geq 2\sqrt{\log(n/\eta)}$, with probability at least $1 - \eta$ we have

$$\|X - \hat{X}\|_F \leq \gamma^{-1}\sqrt{32d\log(n/\eta)} \tag{3.8}$$

$$\text{and } \|V - \hat{V}\|_F \leq 4(\gamma\sqrt{\delta})^{-1}\sqrt{d\log(n/\eta)}, \tag{3.9}$$

where we select the non-unique columns of V to minimize $\|V - \hat{V}\|_F$.

Note, that as $\|\hat{X}W - X\|_F = \|\hat{X} - XW^\top\|_F$ we can work with orthogonal transformations of X. To prove the above corollary, we allow only orthogonal matrices

which commute with S, the matrix of the d largest eigenvalues of P. This is the same as orthogonally transforming the matrix of eigenvectors of V.

Finally, we note that essentially the same bounds can be proved if we consider a spectral embedding in which we do not necessarily use the top d eigenvalues but we select a subset of the eigenvalues. In this situation if we select the dimensions with indices $\mathcal{D} \subset [n]$ then we can let

$$\gamma = \min \left\{ \min_{\substack{i,j \in \mathcal{D} \\ \lambda_i(P) \neq \lambda_j(P)}} |\lambda_i(P) - \lambda_j(P)|, \min_{i \in \mathcal{D}, j \notin \mathcal{D}} |\lambda_i(P) - \lambda_j(P)| \right\} / \delta,$$

and the same bounds hold for the appropriately defined matrices $X_{\mathcal{D}}$ and $\hat{X}_{\mathcal{D}}$.

3.2.2 Eigenvalue Estimation and Concentration

In this section we will show that the largest d eigenvalues of A concentrate around the largest d eigenvalues of P. This concentration is much faster than the concentration for an arbitrary eigenvalue implied by Theorem 3.5 but around the same order as implied by Füredi and Komlós [1981] for ER graphs. Füredi and Komlós [1981] actually prove a central limit theorem for the largest eigenvalues of an ER graph but at present we do not have such a result for these more general models. The eigenvalues plotted in Figure 3.1(a) in Example 3.8 illustrate the next theorem where we see the largest three eigenvalues of A have concentrated very closely around the eigenvalues of P. Note that there is a transition, where the top three to seven eigenvalues have concentrated much more closely then then remaining eigenvalues.

Theorem 3.10. *Let d be fixed and let $A \sim \text{IEG}(P)$. Let S be the diagonal matrix with diagonal given by the d largest eigenvalues in magnitude of P. Let $\hat{S}$ be the analogous matrix for A. Let $\delta = \|P\mathbf{1}\|_\infty$ and let γ be as in Eq. (3.7). For $\eta \in (0, 1/2)$, if $\gamma\sqrt{\delta} \geq 4\sqrt{\log(n/\eta)}$ then with probability greater than $1 - 2\eta$*

$$\|S - \hat{S}\|_F \leq 48d\log(n/\eta)\gamma^{-2} + \sqrt{d\log(2d/\eta)/2} + \sqrt{d}$$

$$\text{and } \|S - \hat{S}\|_{2\to 2} \leq 48d\log(n/\eta)\gamma^{-2} + \sqrt{\log(2d/\eta)/2} + 1.$$

Proof. We will prove the bound for the Frobenius norm. The proof for the $2 \to 2$ norm is nearly identical. First, consider that we can write $S = V^\top PV$ and $\hat{S} = \hat{V}^\top A\hat{V}$, so we can write

$$\|S - \hat{S}\|_F \leq \|V^\top PV - \text{diag}(V^\top AV)\|_F + \|\text{diag}(V^\top AV) - \hat{V}^\top A\hat{V}\|_F, \tag{3.10}$$

where diag denotes the operation of making off-diagonal elements zero. For the second term in Eq. (3.10), we have

$$\begin{aligned}
&\|\text{diag}(V^\top AV) - \hat{V}^\top A\hat{V}\|_F = \|\text{diag}(V^\top AV - \hat{V}^\top A\hat{V})\|_F \\
=&\left\|\text{diag}\left((V-\hat{V})^\top A(V-\hat{V})^\top - 2(\hat{V}-V)^\top A\hat{V}\right)\right\|_F \\
\leq&\left\|\text{diag}\left((V-\hat{V})^\top A(V-\hat{V})^\top\right)\right\|_F + 2\left\|\text{diag}\left((\hat{V}-V)^\top A\hat{V}\right)\right\|_F \\
\leq&\left\|\text{diag}\left((V-\hat{V})^\top A(V-\hat{V})^\top\right)\right\|_F + 2\left\|\text{diag}\left((\hat{V}-V)^\top \hat{V}\hat{S}\right)\right\|_F \\
\leq&\|\hat{S}\|_{2\to 2}\|V-\hat{V}\|_F^2 + 2\|\hat{S}\|\|\text{diag}((\hat{V}-V)^\top\hat{V})\|_F.
\end{aligned} \tag{3.11}$$

To proceed, note that

$$\begin{aligned}\operatorname{diag}((\hat{V}-V)^\top \hat{V}) =& \frac{1}{2}\operatorname{diag}(\hat{V}^\top\hat{V} - V^\top\hat{V} - \hat{V}^\top V + V^\top V)\\ =& \frac{1}{2}\operatorname{diag}((\hat{V}-V)^\top(\hat{V}-V)).\end{aligned}$$

From Theorem 3.5 and our assumptions we have $\|\hat{S}\|_{2\to 2} \leq \|S\|_{2\to 2} + \gamma\delta/2 \leq 3\delta/2$ and Corrollary 3.9 Eq. (3.9) gives $\|V - \hat{V}\|_F^2 \leq \|V-\hat{V}\|_F \leq 16d\log(n/\eta)\gamma^{-2}$. Continuing from Eq. (3.11), we have

$$\begin{aligned}\|\operatorname{diag}(V^\top A V) - \hat{V}^\top A\hat{V}\|_F \leq& 2\|\hat{S}\|_{2\to 2}\|\hat{V}-V\|_F^2\\ \leq& 48d\log(n/\eta)\gamma^{-2}\end{aligned}$$

with probability at least $1-\eta$.

For the first term in Eq. (3.10), we have with probability at least $1-\eta$ that

$$\|\operatorname{diag}(V^\top(A-P)V)\|_F = \sqrt{\sum_{i=1}^{d}\left(V_{i\cdot}^\top(A-P)V_{i\cdot}\right)^2} \leq \sqrt{d\log(2d/\eta)/2} + \sqrt{d}$$

where we use Lemma 3.11 below replacing η with η/d and using a union bound. □

This next Lemma, used above, is a very simple application of Hoeffding's Inequality.

Lemma 3.11. *Let $A \sim \mathrm{IEG}(P)$. For any (non-random) unit vectors $u, u' \in \mathbb{R}^n$ we have*

$$|u^\top(A-P)u'| \leq \sqrt{\log(2/\eta)/2} + 1 \tag{3.12}$$

with probability at least $1-\eta$.

Proof. We have

$$\begin{aligned} |u^\top (A-P)u'| &= |\sum_{i,j}(A_{ij}-P_{ij})u_i u'_j| \\ &\leq 2\left|\sum_{i=1}^{n}\sum_{j=i+1}^{n}(A_{ij}-P_{ij})u_i u'_j\right| + \left|\sum_{i=1}^{n} P_{ii}u_i u'_i\right|. \end{aligned} \tag{3.13}$$

The first term in Eq. (3.13) is the sum of $\binom{n}{2}$ independent random variables with range $u_i u'_j$. Using Hoeffding's inequality, we have

$$\mathbb{P}\left[|\sum_{i<j}(A_{ij}-P_{ij})u_i u'_j| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n} u_i^2 \sum_{j=i+1}^{n} {u'_j}^2}\right) \leq 2\exp(-2t^2).$$

We take $t = \sqrt{\log(1/\eta)/2}$ and use the simple bound for the second term in Eq. (3.13)

$$\left|\sum_{i=1}^{n} P_{ii}u_i u'_i\right| \leq \sum_{i=1}^{n}|u_i u'_i| \leq 1.$$

to get the result. □

We remark that using the fact that u and u' are non-random yields far better bounds than those using a standard spectral norm argument. The result also holds for random vectors provided they are independent of A.

3.3 Estimation for low rank P: Improvements by the power method

Consider again the case of $A \sim \mathrm{ER}(p)$ for some fixed $p \in (0,1)$. The adjacency spectral embedding of P is $X = \sqrt{p}\mathbf{1} \in \mathbb{R}^n$ and the results up to now show that with

probability at least $1 - \eta$ we have

$$\|\hat{X} - X\|_2 \leq \sqrt{32 \log(n/\eta)}.$$

It is natural to ask: Is the order of this bound is sharp? What changes if we consider another norm, such as the ∞-norm? What can we say about $\hat{X} - X$ component-wise? Is there a stable asymptotic distribution for these residuals?

In this section we give a definitive answer to some of these questions. In fact we need not limit ourselves to ER random graphs but can easily demonstrate these results for the IEG model provided we assume $\text{rank}(P) = d$ and the d non-zero eigenvalues are distinct. This is a major restriction compared to the results in Section 3.2, where no assumptions on P are made. The payoff of this restriction is that much stronger theoretical control of the adjacency spectral embedding.

As the title of this section suggests, our technique will be to use the power method from linear algebra. Given an initial vector $u_0 \in \mathbb{R}^n$ and a matrix $M \in \mathbb{R}^{n \times n}_{sym}$ the power methods iterates by setting $u_{k+1} = Mu_k / \|Mu_k\|_2$. Provided $\langle u_o, v_1(M) \rangle \neq 0$ and $\lambda_1(M)$ has multiplicity one then as as k increases u_k converges to $v_1(M)$ and $\|Mu_k\|_2$ converges to $\lambda_1(M)$.

Heuristically, it can be shown that $\|u_{n+1} - v_1(M)\|_2 \approx \frac{|\lambda_2(M)|}{|\lambda_1(M)|} \|u_n - v_1(M)\|_2$. This means that if the ratio between the second and first eigenvalues is small, the convergence of this method will be very fast. This method and its extensions are implemented in many different software packages and provide the basis for finding eigenvalues and eigenvectors numerically.

Our goal in the section is not numerical but instead to use the power method to break down the difference $\hat{X} - X$ into two terms, which can be analyzed separately to give better concentration inequalities. Let $\tilde{X}$ be the result after taking one step in the power method for A starting at X, after appropriate rescaling which we will provide below. Then we can consider the difference $\hat{X} - \tilde{X}$, which we will is negligible compared to $\tilde{X} - X$, which we can tightly control by considering its relation to the matrix $(A - P)X$ which can be tightly controlled.

In section 3.3.1 of this we will lay out this argument in full detail. In section 3.3.2 we will demonstrate a simple L^∞ bound that can be achieved using this method. We will use this bound in Chapter 5 to show that in certain models, vertices can be clustered perfectly asymptotically almost surely. In section 3.3.3, an L^2 bound is shown using an interesting variation on McDiarmid's inequality. Later on in section 4.3, we show that the residuals $\hat{X} - X$ are asymptotically normal.

3.3.1 Overview of the proof method

In this section we will prove some key Lemmas that then make the more applicable results in the next section quite easy to prove. For all the results in this and upcoming section we will make similar assumptions. To keep the statements of the results section simpler and for the convenience of the reader, we summarize these assumptions here.

Assumption 3.12. We suppose that $A \sim \text{IEG}(P)$ and let $\delta = \|P\mathbf{1}\|_\infty$ and $\eta \in (0, 1/2)$. The matrix P satisfies

(i) $\text{rank}(P) = d$,

(ii) the non-zero eigenvalues of P are distinct and

$$\gamma = \min_{i \neq j \in [d+1]} |\lambda_i(P) - \lambda_j(P)|/\delta, \tag{3.14}$$

(iii) $\gamma^2\sqrt{\delta} > 8 \log n/\eta$,

(iv) and $X = VS^{1/2}$ and $\hat{X} = \hat{V}\hat{S}^{1/2}$ are the d-dimensional spectral embedding of P and A, respectively.

Assumption (iii) is stronger than our assumption from before that $2\sqrt{\log(n/\eta)} < \gamma\sqrt{\delta}$. It can be weakened somewhat leaving the order of the bounds largely the same in many contexts but was strengthened to simplify the bounds.

As stated, we will breakdown the difference $\hat{X} - X$ using the power method and so we have

$$\begin{aligned}\hat{X} - X &= \hat{V}\hat{S}^{1/2} - VS^{1/2} = A\hat{V}\hat{S}^{-1/2} - PVS^{-1/2} \\ &= A(\hat{V} - V)\hat{S}^{-1/2} + AV(\hat{S}^{-1/2} - S^{-1/2}) + (A - P)VS^{-1/2}. \end{aligned} \tag{3.15}$$

The first term is analagous to the error after one step of the power method and is bounded in Lemma 3.14. For the second term we will use the strong concentration inequalities for eigenvalues from Theorem 3.10. Finally, for the last term, row i of $(A - P)V$ is

$$\sum_{j=1}^{n} (A_{ij} - P_{ij}) v_j^\top,$$

a sum of *independent* random variables and so standard results such as Hoeffding's inequality, McDiarmid's inequality and the central limit theorem can be applied. Hence, if the first two terms in Eq. (3.15) are small enough than the behavior of $\hat{X} - X$ is determined largely by the behavior of a sum of independent random variables which is well understood.

First, we state and prove a lemma that shows that $V^\top \hat{V}$ is very close to the identity.

Lemma 3.13. *Under assumption 3.12, let V and $\hat{V}$ be the matrices of eigenvectors corresponding to the d largest eigenvalues of P and A, respectively. With probability at least $1 - 2\eta$*

$$\|V^\top \hat{V} - I\|_F \leq \frac{10d \log(n/\eta)}{\gamma^2 \delta}. \tag{3.16}$$

We note that this event occurs provided the events in Corollary 3.9 occur and together with events of the form in Lemma 3.11.

Proof. Working first with the diagonal entries we have that

$$\begin{aligned} \operatorname{diag}(\hat{V}^\top (V - \hat{V})) &= \frac{1}{2} \operatorname{diag}\left(\hat{V}^\top V - \hat{V}^\top \hat{V} - V^\top V + V^\top \hat{V}\right) \\ &= -\frac{1}{2} \operatorname{diag}\left((V - \hat{V})^\top (V - \hat{V})\right) \end{aligned} \tag{3.17}$$

so $\|\operatorname{diag}(\hat{V}^\top \hat{V}) - I\|_F \leq \frac{1}{2}\|V - \hat{V}\|_F^2 \leq 8d \log(n/\eta)(\gamma^2\delta)^{-1}$ with probability at least $1 - \eta$ by Corollary 3.9.

To bound the off-diagonal terms, we adapt an idea from Sarkar and Bickel [2013] to this somewhat different case. First, $V^\top (A - P)\hat{V} = V^\top \hat{V} \hat{S} - S V^\top \hat{V}$ which can be

written entrywise as

$$V_{\cdot i}^\top (A - P)\hat{V}_{\cdot j} = (S_{ii} - \hat{S}_{jj})V_{\cdot i}^\top \hat{V}_{\cdot j}. \tag{3.18}$$

Since the eigenvalues are distinct and $\|A - P\|_{2\to 2} \leq 2\sqrt{\delta \log(n/\eta)}$ we know for $i \neq j$ that

$$|S_{ii} - \hat{S}_{jj}| \geq |S_{ii} - S_{jj}| - \|A - P\|_{2\to 2} \geq \gamma\delta - 2\sqrt{\delta \log(n/\eta)} \geq \gamma\delta/2.$$

We also have

$$V_{\cdot i}^\top (A - P)\hat{V}_{\cdot j} = V_{\cdot i}^\top (A - P)V_{\cdot j} + V_{\cdot i}^\top (A - P)(V_{\cdot j} - \hat{V}_{\cdot j})$$

We can use Lemma 3.11 to get $|V_{\cdot i}^\top (A - P)V_{\cdot j}| \leq \sqrt{\log(2d^2/\eta)/2}$ with probability at least $1 - \eta/d^2$. A simple application of the the Cauchy-Schwarz inequality and Eq. (3.9) yields

$$|V_{\cdot i}^\top (A - P)(V_{\cdot j} - \hat{V}_{\cdot j})| \leq (2\sqrt{\delta \log(n/\eta)})(4\sqrt{\log(n/\eta)}(\gamma^2\delta)^{-1})$$

Dividing through by $S_{ii} - \hat{S}_{jj}$ in Eq. (3.18) and using Assumption 3.12(iii) gives

$$V_{\cdot i}^\top \hat{V}_{\cdot j} = V_{\cdot i}^\top (A - P)\hat{V}_{\cdot j}/(S_{ii} - \hat{S}_{jj}) \tag{3.19}$$

$$\leq \frac{\sqrt{\log(2d^2/\eta)/2} + 8\log(n/\eta)/(\gamma^2\sqrt{\delta})}{\gamma\delta/2} \tag{3.20}$$

$$\leq \frac{2\sqrt{\log(2d^2/\eta)}}{\gamma\delta} \tag{3.21}$$

Combining this with the result for the diagonal and simplifying the expression establishes the result. □

We now bound the first term in Eq. (3.15), which is the term analagous to the error after one step in the power method for A starting at V.

Lemma 3.14. *Under assumption 3.12 the following occurs with probability* $1 - 2\eta$

$$\|A(\hat{V} - V)\hat{S}^{-1/2}\|_F \leq \frac{(30d + 8\sqrt{d})\log(n/\eta)}{\sqrt{\gamma^5\delta}}. \tag{3.22}$$

This event occurs provided the events in Lemma 3.13 and its proof occur.

Proof. Let $E = A - \hat{V}\hat{S}\hat{V}^\top$. We have

$$\begin{aligned}\|A(V - \hat{V})\hat{S}^{-1/2}\|_F &= \|(\hat{V}\hat{S}\hat{V}^\top + E)(V - \hat{V})\hat{S}^{-1/2}\|_F \\ &\leq \|\hat{S}\|_{2\to 2}\|\hat{V}^\top(V - \hat{V})\|_F\|\hat{S}^{-1/2}\|_{2\to 2} \qquad (3.23)\\ &\quad + \|E\|_{2\to 2}\|V - \hat{V}\|_F\|\hat{S}^{-1/2}\|_{2\to 2} \qquad (3.24)\end{aligned}$$

We now bound Eq. (3.23). We also have

$$\|\hat{S}\|_{2\to 2} \leq \|S\|_{2\to 2} + 2\sqrt{\delta\log n/\eta} \leq \delta + \gamma\delta/2 \leq \frac{3}{2}\delta \tag{3.25}$$

$$\text{and } \|\hat{S}^{-1}\|_{2\to 2} \leq (S_{dd} - 2\sqrt{\delta\log n/\eta})^{-1} \leq (\delta\gamma - \delta\gamma/2)^{-1} = \frac{2}{\gamma\delta} \tag{3.26}$$

by Theorem 3.5 and our assumption that $\gamma\sqrt{\delta} \geq 4\sqrt{\log(n/\eta)}$. Together with Lemma 3.13 we have

$$\begin{aligned}&\|\hat{S}\|_{2\to 2}\|\hat{V}^\top(V - \hat{V})\|_F\|\hat{S}^{-1/2}\|_{2\to 2} \\ \leq &\frac{3\delta}{2}\left(\frac{10d\log(n/\eta)}{\gamma\delta}\right)\frac{2}{\sqrt{\gamma\delta}} = \frac{30d\log(n/\eta)}{\sqrt{\gamma^5\delta}}.\end{aligned}$$

Theorem 3.5 also guarantees that $\|E\|_{2\to 2} \leq 2\sqrt{\delta \log(n/\eta)}$. Therefore, we can bound Eq. (3.24)

$$\begin{aligned}\|E\|_{2\to 2}\|V - \hat{V}\|_2\|\hat{S}^{-1/2}\|_{2\to 2} &\leq 2\sqrt{\delta \log(n/\eta)}\sqrt{\frac{4d\log(n/\eta)}{\gamma^2\delta}}\frac{2}{\sqrt{\gamma\delta}} \\ &= \frac{8\sqrt{d}\log(n/\eta)}{\sqrt{\gamma^3\delta}},\end{aligned}$$

from which the desired bound follows. □

Lemma 3.15. *Under Assumption 3.12 the following occurs along with the event in Lemma 3.14*

$$\|AV(\hat{S}^{-1/2} - S^{-1/2})\|_F \leq \frac{100d^{3/2}\log(n/\eta)}{\gamma^6\delta^{1/2}}$$

Proof. First working entrywise and then applying the bound in Theorem 3.10 gives

$$\begin{aligned}\|\hat{S}^{-1/2} - S^{-1/2}\|_{2\to 2} &= \max_{i\leq d}\frac{|S_{ii} - \hat{S}_{ii}|}{(S_{ii}^{1/2} + \hat{S}_{ii}^{1/2})\hat{S}_{ii}^{1/2}S_{ii}^{1/2}} \\ &\leq \frac{\|S - \hat{S}\|_{2\to 2}}{(S_{dd}^{1/2} + \hat{S}_{dd}^{1/2})\hat{S}_{dd}^{1/2}S_{dd}^{1/2}} \\ &\leq \frac{50d\log(n/\eta)\gamma^{-2}}{(\sqrt{\gamma\delta} + \sqrt{\gamma\delta/2}))\sqrt{\gamma\delta/2}\sqrt{\gamma\delta}} \\ &\leq \frac{50d\log(n/\eta)}{\gamma^6\delta^{3/2}}\end{aligned}$$

with probability at least $1 - 2\eta$. Hence

$$\begin{aligned}\|AV(\hat{S}^{-1/2} - S^{-1/2})\|_F &\leq \|A\|_{2\to 2}\|V\|_F\|\hat{S}^{-1/2} - S^{-1/2}\|_{2\to 2} \\ &\leq 2\delta\sqrt{d}\frac{50d\log(n/\eta)}{\gamma^6\delta^{3/2}}\end{aligned}$$

□

For convenience, we combine the previous two Lemmas into the following simplified form which bounds the first two terms in Eq. (3.15).

Lemma 3.16. *Under Assumption 3.12, with probability* $1 - 2\eta$

$$\|\hat{X} - AVS^{-1/2}\|_F \leq \frac{138d^{3/2}\log(n/\eta)}{\sqrt{\gamma^7\delta}}. \tag{3.27}$$

Note that in the iid RDPG case of the next chapter, this bound implies $\|\hat{X} - AVS^{-1/2}\| = \Theta_P(\log(n)n^{-1/2})$, which decays rapidly compared to the bounds from Theorem 3.4.

3.3.2 Concentration of maximum residual error

In this section, we will bound

$$\max_{i\in[n]} \|x_i - \hat{X}_i\|_2,$$

the maximum distance between the estimated and true latent positions. This is the $2 \to \infty$ norm on $n \times d$ matrices which we denote $\|\cdot\|_{2\to\infty}$.

Lemma 3.17. *Under Assumption 3.12 with probability at least* $1 - \eta$

$$\|(A - P)V\|_{2\to\infty} \leq \sqrt{d\log(2dn/\eta)/2}.$$

Proof. First, $(AV - PV)_{ik} = \sum_{j=1}^{n}(A_{ij} - x_i^\top x_j)v_{jk}^\top$ is the sum of n independent random variables with range v_{jk}. By Hoeffding's inequality, we have

$$\mathbb{P}[|(AV - PV)_{ik}| \geq t] \leq 2\exp\left(\frac{-2t^2}{\sum_{j=1}^{n} v_{jk}^2}\right) = 2\exp(-2t^2).$$

To get the reseult let $t = \sqrt{\log(2dn/\eta)/2}$ and use the union bound. □

Putting together the two results above we get the following Theorem.

Theorem 3.18. *Under Assumption 3.12, with probability at least* $1 - 3\eta$

$$\|\hat{X} - X\|_{2\to\infty} \leq \sqrt{\frac{d\log(4dn/\eta)}{2\delta\gamma}} + \frac{138d^{3/2}\log(n/\eta)}{\sqrt{\gamma^7\delta}}.$$

Proof. Recall that $AVS^{-1/2} - X = (A - P)VS^{-1/2}$ Using that

$$\|(A - P)VS^{-1/2}\|_{2\to\infty} \leq \|(A - P)V\|_{2\to\infty}\|S^{-1/2}\|_{2\to 2}.$$

Combining Lemma 3.16, Lemma 3.17 and the fact that $\|S^{-1/2}\|_{2\to 2} = (\gamma\delta)^{-1/2}$. □

Theorem 3.18 will prove to be one of the most useful and powerful results for analyzing various subsequent inference tasks. Unlike our other bounds which focus on the Frobenius norm, the bound of the $2 \to \infty$ norm gives our best simultaneous control over the individual rows of $\hat{X} - X$. If we consider the dense case case where $\delta = \Theta(n)$ then the implied bound of the frobenius norm given by this bound on the $2 \to \infty$ is of essentially the same order as that in Theorem 3.4 and its corollary.

In the iid latent position RDPG case of the next chapter, the bounds in Theorem 3.18 tranlate as $\|X - \hat{X}\| = \Theta_P(\log(n)n^{-1/2})$. In the next section we show that the order for the Frobenius norm can be improved, at least in the case that $\text{rank}(P) = d$ but we cannot use this to improve the $2 \to \infty$-norm bound. In another direction, the distributional result in Section 4.3 provides the most elegant picture for any individual row of $\hat{X} - X$ but cannot be easily be extended to analyze all rows simultaneously.

3.3.3 Concentration of total residual error

In the previous section we bounded $\max_{i\in[n]} \|\hat{X}_i - x_i\|_2$. In this section we will improve our bound bound $\|\hat{X} - X\|_F$ from Corollary 3.9. We will again use Lemma 3.14 and our second lemma for the third term in Eq. (3.15) will require a result similar to McDiarmid's Inequality established by Kutin [2002].

Lemma 3.19. *Under Assumption 3.12 with probability at least* $1 - \eta$

$$\left| \|(A - P)V\|_F^2 - \mathbb{E}[\|(A - P)V\|_F^2] \right| \leq \tau$$

where $\tau = \frac{n}{5}\sqrt{\log(4/\eta)\log(2n/\eta)/(\gamma\delta)}$ *and*

$$\mathbb{E}[\|(A - P)V\|_F^2] = \sum_{i=1}^{n} \|x_i\|_2^4 \|v_i\|_2^2 + \sum_{j\neq i} (1 - x_i^\top x_j) x_i^\top x_j \|v_j\|_2^2.$$

Proof. We use the fact that $\|(A - P)V\|_F$ is a function $\binom{n}{2}$ independent random variables A_{ij} for $i > j$. Let A and A' be two adjacency matrices with the property that $A_{ij} = A'_{ij}$ for all (i, j) except $(i, j) = (k, l)$ and $(i, j) = (l, k)$ and assume without loss of generality that $l > k$ and $A_{kl} = 1 = 1 - A'_{kl}$. Hence, the associated graphs are equal except for one edge is present in A and not in A'. Then for any such pair of

matrices A and A', we have

$$\begin{aligned}
&\left|\|(A-P)V\|_F^2 - \|(A'-P)V\|_F^2\right| \\
&= \left|\sum_{i=1}^n \left\|\sum_{j=1}^n (A_{ij} - x_i^\top x_j)v_j\right\|_2^2 - \sum_{i=1}^n \left\|\sum_{j=1}^n (A'_{ij} - x_i^\top x_j)v_j)\right\|_2^2\right| \\
&= \Bigg|\sum_{i=1}^n \sum_{j=1}^n \sum_{j'=1}^n \big((A_{ij} - x_i^\top x_j)(A_{ij'} - x_i^\top x_{j'})v_j^\top v_{j'} \\
&\qquad\qquad - (A'_{ij} - x_i^\top x_j)(A'_{ij'} - x_i^\top x_{j'})v_j^\top v_{j'}\big)\Bigg| \\
&= \Bigg|2\left(\sum_{j\neq l}(A_{kj} - x_k^\top x_j)v_j^\top v_l\right) + 2\left(\sum_{j\neq k}(A_{lj} - x_l^\top x_j)v_j^\top v_k\right) \\
&\quad + (1 - 2x_k^\top x_l)(v_l^\top v_l + v_k^\top v_k)\Bigg|.
\end{aligned}$$

The third equality follows since $A_{ij} = A'_{ij'}$ as long as $i \notin \{k, l\}$, $A_{kj} = A'_{kj}$ for $j \neq l$, the symmetry across j and j' and the fact that $A_{kl} = 1 = 1 - A'_{kl}$.

For any A, not necessarily with $A_{kl} = 1$, denote the random variable on the last line above as $\rho_{kl} = \rho_{kl}(A)$ for any $k \neq l$. Now note that with probability 1

$$\begin{aligned}
|\sum_{j\neq l}(A_{kj} - x_k x_j)v_j^\top v_l| &\leq \|\sum_{j=1}^n (A_{kj} - x_k x_j)v_j\|\|v_l\| \\
&\leq \sqrt{\sum_{j=1}^n (A_{kj} - x_k x_j)^2}\|V\|_F\|v_l\| \leq \sqrt{\frac{nd}{\gamma\delta}},
\end{aligned}$$

using that $\|v_l\|_2 = \|x_l \hat{S}^{-1/2}\|_2 \leq (\gamma\delta)^{-1/2}$, $\|V\|_F = \sqrt{d}$ and the first term is clearly bounded by $\sqrt{n}$. For the last term of ρ_{kl} we have $(1-2x_k x_l)(v_l^\top v_l + v_k^\top v_k) \leq \frac{4}{\gamma\delta} \leq \sqrt{\frac{nd}{\gamma\delta}}$. Together, we have $\mathbb{P}[\rho_{kl} \leq 4\sqrt{\frac{nd}{\gamma\delta}} + \frac{4}{\gamma\delta}] = \mathbb{P}[\rho_{kl} \leq 5\sqrt{\frac{nd}{\gamma\delta}}] = 1$ for all $k \neq l$ so that

$$\mathbb{P}\left[\max_{k\neq l} \rho_{kl} \leq 5\sqrt{\frac{nd}{\gamma\delta}}\right] = 1.$$

We also have by Hoeffding's Inequality that

$$\mathbb{P}[|\sum_{j\neq l}(A_{kj}-x_kx_j)v_j^\top v_l|\geq t]\leq 2\exp\left\{\frac{-2t^2}{\sum_{j\neq l}(v_j^\top v_l)^2}\right\}\leq 2\exp\left\{-2\gamma\delta t^2\right\},$$

giving that $\mathbb{P}[\rho_{kl}\geq 4\sqrt{\frac{\log(4/\eta)}{2\gamma\delta}}+\frac{4}{\gamma\delta}]\leq\eta$. Replacing η with $2\eta/n^2$ and using a union bound we have that

$$\mathbb{P}\left[\max_{k>l}\rho_{kl}\geq 5\sqrt{\frac{\log(2n/\eta)}{\gamma\delta}}\right]\leq\eta$$

The result of [Kutin, 2002] gives

$$\mathbb{P}\left[|\|(A-P)V\|_F^2-\mathbb{E}[\|(A-P)V\|_F^2]|\geq\tau\right]\leq 4\exp\left(\frac{\tau^2\gamma\delta}{25\binom{n}{2}\log(2n/\eta)}\right)$$

provided

$$\tau\leq 10\sqrt{\binom{n}{2}\log(1/\eta)\frac{\log(4n/\eta)}{\gamma\delta}}\text{ and }\log(1/\eta)\geq 21\log(7n^2/\log(\frac{1}{\eta})).$$

Setting $\eta<n^{-2}$ and then setting $\tau=\frac{n}{5}\sqrt{\log(4/\eta)\log(2n/\eta)/(\gamma\delta)}$ ensures that both conditions hold giving

$$\mathbb{P}\left[|\|(A-P)V\|_F^2-\mathbb{E}[\|(A-P)V\|_F^2]|\geq\frac{n}{5}\sqrt{\log(4/\eta)\log(2n/\eta)/(\gamma\delta)}\right]\leq\eta.$$

Now, we need only establish the form of the expectation,

$$\begin{aligned}\mathbb{E}[\|(A-P)V\|_F^2]&=\sum_{i=1}^n\mathbb{E}[\left\|\sum_{j=1}^n(A_{ij}-x_ix_j)v_j\right\|_2^2]\\&=\sum_{i=1}^n\sum_{j=1}^n\sum_{j'=1}^n\mathbb{E}[(A_{ij}-x_ix_j)(A_{ij'}-x_ix_{j'})v_j^\top v_{j'}].\end{aligned}$$

Now, again A_{ij} and $A_{ij'}$ are independent unless $j = j'$ so the summand is zero if $j \neq j'$. Continuing the computation we have

$$\begin{aligned} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[(A_{ij} - x_i^\top x_j)^2 \|v_j\|_2^2] \\ &= \sum_{i=1}^{n} \left(\|x_i\|_2^4 \|v_i\|_2^2 + \sum_{j\neq i}^{n} x_i^\top x_j (1 - x_i^\top x_j) \|v_j\|_2^2 \right). \end{aligned}$$

□

We can now establish our tighter bound on $\|X - \hat{X}\|_F$.

Theorem 3.20. *Under Assumption 3.12, with probabilitiy at least* $1 - 3\eta$

$$\|\hat{X} - X\|_F \leq \frac{138 d^{3/2} \log(n/\eta)}{\sqrt{\gamma^7 \delta}} + \sqrt{\frac{1}{\gamma\delta}\left(\mathbb{E}[\|(A-P)V\|_F^2] + \tau\right)}, \tag{3.28}$$

with $\tau = \frac{n}{5}\sqrt{\log(4/\eta)\log(2n/\eta)/(\gamma\delta)}$.

Proof. We again use that

$$\|\hat{X} - X\|_F = \|\hat{X} - AVS^{-1/2}\|_F + \|(A-P)VS^{-1/2}\|. \tag{3.29}$$

For the first term in Eq. (3.29), we apply Lemma 3.16 giving the first term in the right hand side of Eq. (3.28).

By assumption $S^{-1/2}\|_F \leq \sqrt{\frac{1}{\delta\gamma}}$ so by Lemma 3.19 we have

$$\mathbb{P}\left[\|(A-P)V\hat{S}^{-1/2}\|_F > \sqrt{\frac{1}{\gamma\delta}\mathbb{E}[\|(A-P)V\|_F^2] + \tau} \right] \leq \eta.$$

A union bound establishes the result. □

The complicated nature of the bound in Eq. 3.28 makes it difficult to interpret the result in the case where P is some fixed rank d matrix. To ease interpretation, consider the case where instead $A \sim \text{RDPG}(\mathcal{X}, F)$ in which case as the number of vertices n becomes large then $\delta = \Theta(n)$ and $\gamma = \Theta(1)$. Additionally, it can be calculated that $\mathbb{E}[\|(A-P)V\|_F^2/(\gamma\delta)] = \Theta(1)$ as well and indeed will coverge to constant. Hence, Eq. 3.28 can be interpreted as $\|\hat{X} - X\|_F = C + O_P(\log(n)/\sqrt{n})$ for some constant C. This concentration of the global error rate will not be used explicitly in the remaining text but has promise in the comparison of two graphs distributed according to two RDPG distributions.

In conclusion, the results in this section have shown that in the case when the rank of P is equal to the dimension of the spectral embedding we establish very strong concentration results for the adjacency spectral embeddings. In particular, we have that $\|X - \hat{X}\|_{2\to\infty} = O_P(\log(n)n^{-1/2})$ and $\|X - \hat{X}\|_F = C + O_P(\log(n)n^{-1/2})$ in the case that $A \sim \text{RDPG}(\mathcal{X}, F)$. In the next chapter, we make these results explicit, generalize to the sparse iid case and end the chapter by proving a central limit theorem for a fixed row of $X - \hat{X}$.

Chapter 4

The iid latent position case

The theorems in Chapter 3 give probabilistic bounds between the d-dimensional adjacency spectral embedding of A and P in terms of the maximum degree and the spectral gap of P in the case that P is some fixed matrix. In order to use this theorem to demonstrate the consistency of certain inference procedures, we must consider how this theorem can be applied to models which are not independent edge models. In particular, we investigate the case where the $A \sim \text{LPG}(\mathcal{X}, F, \kappa)$, so that the latent positions are iid. The iid latent position case is particularly relevant because of its relationship to more classical multivariate techniques. Our methods skirt some of the

i

In this chapter we will first present concentration inequalities for random variables taking values in Hilbert spaces that can be used in a general class of latent position models as well as RDPGs. We then use the results of Chapter 3 to prove analagous

results for finite dimensional RDPG with iid latent positions. We will then extend the results to infinite dimensional random dot product graphs as in Section 2.4. In Chapter 5 we will use these results to prove the consistency of certain clustering, classification and estimation procedures.

4.1 Concentration Inequalities in Hilbert Spaces

Before returning to our results for the adjacency sprectral embedding, we will briefly present two useful concentration inequalities for random variables in Hilbert spaces and empirical second moment operators. The first proposition provides a bound on the sample mean of a collection mean zero iid random variables in a real Hilbert space. The second provides a bound on the spectral norm between the integral operator $\mathscr{K}$ and an empirical version. We adopt our notation and rely upon the theory presented in Section 2.4. For more details about some of these results in the latent position graph setting see Tang et al. [2013].

This first proposition will prove useful in bounding the maximum expected degree of a random graph (see Lemma 4.5).

Proposition 4.1 (Pinelis [1992], see also Rosasco et al. [2010])**.** *Let $\xi_1, \ldots, \xi_n$ be independent mean zero random variables on a real separable Hilbert space $\mathcal{H}$ with*

norm $\|\cdot\|_{\mathcal{H}}$. *If each of the* ξ_i *are uniformly bounded with with probability one, so that* $\|\xi_i\| \leq C$ *for all* i *and some* $C > 0$, *then with probability at least* $1 - 2\eta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\| \leq \frac{C\sqrt{2\log(1/\eta)}}{\sqrt{n}}. \tag{4.1}$$

Now, before providing an analogous lemma for the eigenvalues, we must introduce a few operators associated with the integral operator $\mathscr{K}$ and the matrix P. In the case of $A \sim LPM(\mathcal{X}, \kappa, F)$ we let the latent positions be given by $X_1, \ldots, X_n \overset{iid}{\sim} F$ and the entries of P are $P_{ij} = \kappa(X_i, X_j)$. Again, let $\mathcal{H}$ denote the RKHS associated with κ on $L^2(\mathcal{X}, F)$. Let $\mathscr{K}_{\mathcal{H}} : \mathcal{H} \mapsto \mathcal{H}$ and $\mathscr{K}_{\mathcal{H},n} : \mathcal{H} \mapsto \mathcal{H}$ be the linear operators defined by

$$\begin{aligned} \mathscr{K}_{\mathcal{H}} f &= \int_{\mathcal{X}} \langle f, \kappa(\cdot, x)\rangle_{\mathcal{H}} \kappa(\cdot, x) F(dx) \\ \text{and } \mathscr{K}_{\mathcal{H},n} f &= \frac{1}{n} \sum_{i=1}^{n} \langle f, \kappa(\cdot, X_i)\rangle_{\mathcal{H}} \kappa(\cdot, X_i), \end{aligned}$$

for all $f \in \mathcal{H}$.

Note that these two operators are defined on the same Hilbert space and that $\mathscr{K}_{\mathcal{H},n}$ can be thought of as an empirical version of $\mathscr{K}_{\mathcal{H}}$. On the other hand $\mathscr{K}$ and P are linear operators but they are defined on two different spaces. Being able to directly compare $\mathscr{K}_{\mathcal{H}}$ and $\mathscr{K}_{\mathcal{H},n}$ will be key because we can easily relate the spectra of $\mathscr{K}_{\mathcal{H},n}$ and P, and $\mathscr{K}_{\mathcal{H}}$ and $\mathscr{K}$. The following proposition makes these relationships precise.

Proposition 4.2 (Rosasco et al. [2010])**.** *Suppose* κ *is a positive semidefinite link function. The operators* $\mathscr{K}_{\mathcal{H}}$ *and* $\mathscr{K}_{\mathcal{H},n}$ *are positive, self-adjoint operators and are of*

trace class with $\mathscr{K}_{\mathcal{H},n}$ being of finite rank. The spectrum of $\mathscr{K}$ and $\mathscr{K}_{\mathcal{H}}$ are contained in $[0,1]$ and are the same, possibly up to the zero eigenvalues. Similarly, the spectra of P/n and $\mathscr{K}_{\mathcal{H},n}$ are contained in $[0,1]$ and are the same, possibly up to the zero eigenvalues.

Based on this connection Rosasco et al. [2010] and Blanchard et al. [2007] show the following result.

Proposition 4.3. *In the setting of Proposition 4.2 suppose $\kappa(x,x) \leq C$, for $j = 1,2,\ldots,$ let $\lambda_j(\mathscr{K})$ be a decreasing enumeration of the eigenvalues of $\mathscr{K}_{\mathcal{H}}$ (and hence also $\mathscr{K}$) and let $\lambda_j(\mathscr{K}_{\mathcal{H},n})$ be an extended decreasing enumeration of the eigenvalues of $\mathscr{K}_{\mathcal{H},n}$ (and hence P for $j \leq n$). Then with probability at least $1-2\eta$*

$$\left(\sum_{j=1}^{\infty}(\lambda_j(\mathscr{K}_{\mathcal{H}}) - \lambda_j(\mathscr{K}_{\mathcal{H},n}))^2\right)^{1/2} \leq C\sqrt{\frac{8\log(1/\eta)}{n}} \tag{4.2}$$

Note that since the latent positions are random, $\lambda_j(\mathscr{K}_{\mathcal{H},n})$ is a random variable, while $\lambda_j(\mathscr{K}_{\mathcal{H}})$ is a fixed quantity, depending only on the link function κ and the distribution F.

Finally, we state a bound between the feature mapped latent positions $\phi(X_i)$ and the spectral embedding of P.

Proposition 4.4 (Tang et al. [2013])**.** *Suppose $A \sim \mathrm{LPG}(\mathcal{X}, F, \kappa)$ for a positive semidefinite link function κ. Let $\phi : \mathcal{X} \mapsto \ell_2$ be the feature map for κ, let $\phi_d : \mathcal{X} \mapsto \mathbb{R}^d$ be the truncated feature map, so $\phi_d(x)_i = \phi(x)_i$ for all $i \in [d]$, and define $\Phi_d \in \mathbb{R}^{ntimesd}$ to be the matrix with row i given by $\phi_d(x_i)^\top$. Let X be the d-dimensional spectral*

embedding of the matrix of edge probabilities P. Finally, let $\gamma = \gamma_d = \lambda_d(\mathscr{K}) - \lambda_{d+1}(\mathscr{K})$ and let $C = \max_x \kappa(x,x)$, then with probability at least $1-2\eta$,

$$\min_{W \in \mathcal{O}(d)} \|XW - \Phi_d\|_F \leq \frac{C\sqrt{8\log(1/\eta)}}{\gamma}.$$

4.2 Latent Position Graphs

We will now use the results from the previous section to establish bounds on the two particular parameters of the matrix of edge presence probabilities. Recall that each of the theorems in Chapter 3 are stated in the case that the matrix of edge presence probabilities, P, is fixed (and for the theorems in Section 3.3, $\text{rank}(P) = d$). If A is a latent position graph and the latent positions themselves are random then the matrix P is random. In this case it is still reasonable to try to estimate the latent positions using the adjacency spectral embedding so in this section we seek to re-establish the various error bounds in this setting. Our strategy will be to apply the bounds in the previous two chapters by first conditioning on the latent positions. We will show that for a collection of possible realized latent positions, the two parameters of the matrix P are bounded uniformly and that the probability the latent positions are in this collection is high. These two parameters of P are the largest row sum, which we bound from above, and gap between the d^{th} and $(d+1)^{\text{st}}$ largest eigenvalues of P, which we bound from below.

We change our notation slightly to emphasize the fact that certain quantities

are now random. For $A \sim \text{LPG}(\mathcal{X}, \kappa, F)$ we let the latent positions be denoted by $X_1, \ldots, X_n$ and we will still denote matrix of edge probabilities by P with $P_{ij} = \kappa(X_i, X_j)$. We will now denote the (random) largest row sum of P will be denoted by $\Delta = \|P\mathbf{1}\|_\infty$ and the (random) eigengap as $\Gamma = \Gamma_d = (\lambda_d(P) - \lambda_{d+1}(P))/\Delta$. We will now provide probabilistic upper bounds on Δ and lower bounds on $\Gamma\Delta$, first in the case of that $A \sim \text{RDPG}(\mathcal{X}, F)$ and then in the case that $A \sim \text{LPG}(\mathcal{X}, \kappa, F)$ with κ positive semidefinite. Note that the case κ is not positive semidefinite is also important but we choose to omit it because it would require significant further development in the vein of sections 2.4 and 4.1.

Lemma 4.5. *Suppose that $A \sim \text{RDPG}(\mathcal{X}, F)$ with latent positions $X_1, \ldots, X_n \overset{iid}{\sim} F$ for some distribution F. Suppose $\|X_i\|_2^2 \leq \rho$ with probability 1 and that $\mathbb{E}[X_i] = \mu$ with $\|\mu\|_2^2 = m^2\rho$. Let $\gamma = \lambda_d(\mathbb{E}[X_j X_j^\top])/\rho > 0$. Then with probability at least $1 - 4\eta$,*

$$\Delta \leq \rho\left(nm + \sqrt{2n\log(1/\eta)}\right) \text{ and } \Gamma\Delta \geq \rho(n\gamma - \sqrt{8n\log(1/\eta)}) \tag{4.3}$$

Proof. To bound Δ note that

$$\Delta = \max_i \sum_{j=1}^n P_{ij} = \max_i \langle X_i, \sum_j X_j \rangle \leq \max_{x:\|x\|_2 \leq \rho} \langle x, n\bar{X} \rangle \leq n\sqrt{\rho}\|\bar{X}\|_2.$$

Now, using Proposition 4.1 we have with probability at least $1 - 2\eta$

$$\|\bar{X}\|_2 \leq \|\bar{X} - \mu\|_2 + \|\mu\|_2 \leq \|\mu\|_2 + \sqrt{\frac{2\rho\log(1/\eta)}{n}},$$

which establishes the first inequality in Eq. (4.3). Now, for the RDPG the integral

operator $\mathscr{K}$ has rank d and takes the form

$$(\mathscr{K}f)(x) = \int_{\mathcal{X}} \langle x, x' \rangle f(x') F(dx') = \sum_{j=1}^{d} x_j \int_{\mathcal{X}} x'_j f(x') F(dx')$$

Now, note that $\int_{\mathcal{X}} x'_j f(x') F(dx') = \mathbb{E}[f(X_i) X_{ij}]$ is non-zero only if $f(X_i)$ is not orthogonal to X_{ij} which is the case provided $f(x) = x^\top v$ for some $v \in \mathbb{R}^d$. In this case

$$(\mathscr{K}f)(x) = \sum_{j=1}^{d} x_j \int_{\mathcal{X}} x'_j x'^\top v F(dx') = x^\top \mathbb{E}[X_i X_i^\top] v$$

Hence the eigenfunctions corresponding to non-zero eigenvalues of $\mathscr{K}$ satisfy $f(x) = x^\top v$ for some v satisfying $\mathbb{E}[X_i X_i^\top] v = \lambda v$ so the non-zero eigenvalues of $\mathscr{K}$ are the same as the non-zero eigenvalues of $\mathbb{E}[X_i X_i^\top]$. Using Propositions 4.2 and 4.3 we have with probability at least $1 - 2\eta$

$$\frac{\Gamma \Delta}{n} \geq \gamma \rho - \rho \sqrt{\frac{8 \log(1/\eta)}{n}},$$

establishing the second inequality in Eq. (4.3). □

Lemma 4.5 means we can apply theorems in Chapter 3 in the case that $A \sim \text{RDPG}(\mathcal{X}, F)$. For the sake of asymptotic analysis, we can consider a sequence of graphs with n vertices where $n \to \infty$. If the latent positions are from an iid sequence $X_1, X_2, \ldots \overset{iid}{\sim} F$, then the lemma ensures that $\Delta = \Theta(n)$ and $\Gamma = \Theta(1)$. On the other hand, we may consider that for each graph in the sequence the latent positions are a scaled version of an iid sequence. In other words there is some iid sequence $\xi_1, \xi_2, \ldots \overset{iid}{\sim} F$ and for the n^{th} graph in the sequence, the latent positions are given by

$X_i = \sqrt{\rho}\xi_i$ for each $i \in [n]$ for some ρ depending on n. In this case m does not depend on n and the order of ρ^2 reflects the overall edge density of the graph. Furthermore, γ does not depend on n as the eigenvalues of the second moment are scaled by ρ. This next theorem encapsulates this discussion.

Theorem 4.6. *Suppose $A \sim \text{RDPG}(\mathcal{X}, F)$ with $\mathcal{X} \subset \mathbb{R}^d$. Suppose $\|X_i\|_2^2 \leq \rho$ with probability 1 and that $\mathbb{E}[X_i] = \mu$ with $\|\mu\|_2^2 = m^2\rho$ and $\mathbb{E}[X_i X_i^\top] = \mu^{(2)}$. Suppose the eigenvalues of $\mathbb{E}[X_j X_j^\top]$ are distinct and let*

$$\gamma = (4nm\rho)^{-1} \min\left\{ \min_{i\in[d-1]} \lambda_i(\mu^{(2)}) - \lambda_{i+1}(\mu^{(2)}),\ \lambda_d(\mu^{(2)}) \right\}$$

and $\delta = 2\rho nm$. If $150d^2 \log^2(n/eta) < \gamma^4\sqrt{\delta}$ then the bounds of Theorems 3.4, 3.18, and 3.20 all hold with probability at least $1 - 7\eta$ using γ and δ as defined above and with X replaced by XW where $W = \arg\min_{R\in\mathcal{O}(d)} \|\hat{X} - XR\|_F$. Asymptotically, if $\rho = \omega(\log^2(n)/n)$ then

$$\|\hat{X} - XW\|_{2\to\infty} = O_P(\sqrt{\log(n)/2\rho n}) \text{ and } \|\hat{X} - XW\|_F = O_P(\sqrt{\log(n/\eta)}).$$

and if $\rho = 1$ then

$$\|\hat{X} - XW\|_F = \Theta_P(1)$$

Sketch of proof. The theorem follows by the fact that Lemma 4.5 ensures that Assumption 3.12 holds with probability at least $1-4\eta$. If we condition on the event that these assumptions hold then the bounds in the listed theorems all hold with probability at least $1 - 3\eta$ which proves the result. The asymptotic results hold because

m and γ can be taken to be fixed in n in which case $\delta = \Theta_P(\rho n)$ and $\gamma = \Theta(1)$. Plugging these asymptotics into the bounds gives the result. □

If the distribution of latent positions is allowed to depend more arbitrarily on n then things become more complicated. For example if $m^2 = \Omega(\log(n)/n)$ then Lemma 4.5 ensures that $\Delta = \Omega_p(\sqrt{n \log n}\rho^2)$ while otherwise we have that $\Delta = O_p(\sqrt{n \log(n)}\rho^2)$. We may also consider γ decaying at yet still another rate making our task more complicated still. Note again that the idea of considering a sequence of graphs is typically just a mathematical convenience used to give us a handle on the asymptotic properties of large graphs with different properties.

We now state a similar lemma in the case that that $A \sim \text{LPG}(\mathcal{X}, \kappa, F)$.

Lemma 4.7. *Suppose that $A \sim \text{LPG}(\mathcal{X}, \kappa, F)$ with positive semidefinite link function and latent positions $X_1, \ldots, X_n \overset{iid}{\sim} F$. Let $\mathscr{K} : L^2(\mathcal{X}, F) \mapsto L^2(\mathcal{X}, F)$ and $\phi : \mathcal{X} \mapsto \ell_2$ be the associated integral operator and feature map, respectively. Suppose $\kappa(X_i, X_i) = \|\phi(X_i)\|_{\ell_2}^2 \leq \rho$ with probability 1 and that $\mathbb{E}[\phi(X_i)] = \mu \in \ell_2$ with $\|\mu\|_{\ell_2}^2 = m^2\rho$. Let $\gamma = (\lambda_d(\mathscr{K}) - \lambda_{d+1}(\mathscr{K}))/\rho > 0$. Then with probability at least $1 - 4\eta$,*

$$\Delta \leq \rho\left(nm + \sqrt{2n\log(1/\eta)}\right) \text{ and } \Gamma\Delta \geq \rho(n\gamma - \sqrt{8n\log(1/\eta)})$$

The proof is essentially the same as in the proof of Lemma 4.5. Note the only difference is that the subtracted term in the bound on $\gamma\Delta$ is made a factor of two larger. This is because $\text{rank}(\mathscr{K})$ is not assumed to be exactly d and so we must bound $\lambda_d(P)$ from below and $\lambda_{d+1}(P)$ from above. A similar result to Theorem 4.6 can be

shown to hold either bounding the difference between the spectral embedding of A and P or the difference between the adjacency spectral embedding and the truncated feature by combining the former with Proposition 4.4. In the vertex classification setting, this bound is used to develop a universally consistent classifier for latent position graphs (see Section 5.2.3 and Tang et al. [2013]).

4.3 Asymptotic normality

Before, we proceed to the next chapter we wish to establish the following central limit theorem type result. In this section we will explicitly consider the case where there is a sequence of growing RDPG graphs $A^{(n)}$ and the latent positions are scaled version of an iid sequence. For convenience if $X = \sqrt{\rho}\xi$ for $\xi \sim F$ then we say that $X \sim F_\rho$.

Theorem 4.8. *Let $(\rho_n)_{n=1}^{\infty}$ be a sequence of positive scalars in $(0,1)$ with $\rho_n = \omega(\log^2(n)/n)$. Suppose $A^{(n)} \sim \text{RDPG}(\mathcal{X}, F_{\rho_n})$.Let $X^* \sim F$ and let μ_2 be the second moment matrix of X^* which we assume is diagonal with distinct entries. Let the singular value decomposition of X be $VS^{1/2}W^\top$. Suppressing n in the notation If $\rho_n = o(1)$ then*

$$\sqrt{n}\left(\hat{X}_{i\cdot} - X_{i\cdot}W\right) \xrightarrow{\mathcal{L}} \int_{\mathcal{X}} \mathcal{N}(0, \mu_2^{-1}\mathbb{E}[x^\top X^* X^* X^{*\top}]\mu_2^{-1})F(dx)$$

and if $\rho_n = 1$ *for all* n *then*

$$\sqrt{n}\left(\hat{X}_{i\cdot} - X_{i\cdot}W\right) \xrightarrow{\mathcal{L}} \int_{\mathcal{X}} \mathcal{N}(0, \mu_2^{-1}\mathbb{E}[x^\top X^*(1 - x^\top X^*)X^*X^{*\top}]\mu_2^{-1})F(dx).$$

The integral denotes a F*-weighted mixture over normal distributions each with mean zero and with variance depending on the mixture component.*

We will present the proof in the case of that $\rho = o(1)$; the other case is easier. The proof will be similar to the proofs of the two theorems in Section 3.3.

Lemma 4.9. *Under the conditions of Theorem 4.8 if* $\rho = o(1)$ *then for* $x = x^{(n)} = \sqrt{\rho_n}x^*$ *we have*

$$n\mathbb{E}[x^\top X_j(1 - x^\top X_j)X_jX_j^\top]^{1/2}WS^{-1} \xrightarrow{p} \mathbb{E}[x^{*\top}X^*(X^*X^{*\top})]^{1/2}\mu_2^{-1}. \tag{4.4}$$

Proof. We will show the result by bounding three separate terms. Let $\Sigma(x) = \mathbb{E}[x^\top X_j(1 - x^\top X_j)X_jX_j^\top]$ and let $\Sigma^*(x^*) = \mathbb{E}[x^{*\top}X^*(X^*X^{*\top})]$.

We have

$$n\Sigma(x)^{1/2}WS^{-1} - \Sigma^*(x^*)\mu_2^{-1} = \rho^{-1}\Sigma(x)^{1/2}W(n\rho S^{-1} - \mu_2^{-1}) \tag{4.5}$$

$$+ (\rho^{-1}\Sigma(x)^{1/2} - \Sigma^*(x^*)^{1/2})W\mu_2^{-1} \tag{4.6}$$

$$+ \Sigma^*(x^*)^{1/2}(W - I)\mu_2^{-1} \tag{4.7}$$

Notice that $\Sigma(x) = \rho^2\Sigma^*(x^*) + O_P(\rho^3 I)$ and so $\Sigma(x)^{1/2} = \rho(\Sigma^*(x^*)^{1/2} + o_P(1))$ so Eq. (4.6) tends to 0.

Using Proposition 4.3, the ideas in the proof of Lemma 4.5, and the fact that conditioning on $X_i = x$ can change the eigenvalues of S by at most $\sqrt{\rho/n}$, we have

that

$$\sum_{i=1}^{d}\left(\frac{1}{\rho n}S_{ii}-\mu_{2ii}\right)^2 = O_P(\log(n)n^{-1}).$$

Hence,

$$\left|n\rho S_{ii}^{-1}-\mu_{2ii}^{-1}\right| = \frac{|S_{ii}/(n\rho)-\mu_{2ii}|}{(S_{ii}/(n\rho))\mu_{2ii}} = O_P\left(\sqrt{\frac{\log(n)}{n}}\right). \tag{4.8}$$

This gives for Eq. (4.5) that

$$\rho^{-1}\|\tilde{\Sigma}(x)^{1/2}(\mu_2^{-1}-(S/(n\rho))^{-1})\|_F = O_P\left(\sqrt{\frac{\log n}{n}}\right) = o_P(1). \tag{4.9}$$

Finally, note $\|X^\top X - n\mu^2\|_F = O_P(\sqrt{n})$ and the diagonal entries of μ_2 are distinct so by the Davis-Kahan theorem we have $\|W-I\|_F = O_P(n^{-1/2})$. This establishes that Eq. (4.7) tends to 0 in probability. □

Lemma 4.10. *Under the conditions of Theorem 4.8 if $\rho = o(1)$ and $\rho = \omega(\log^2(n)/n)$ as $n \to \infty$,*

$$\sqrt{n}(AV-PV)_{i\cdot}S^{-1/2} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_2^{-1}\mathbb{E}[x^{*\top}X^*X^*X^{*\top}]\mu_2^{-1})$$

Proof. We can write

$$\begin{aligned}(AV-PV)_{i\cdot}S^{-1/2} &= \sum_{j=1}^{n}(A_{ij}-P_{ij})V_{j\cdot}S^{-1/2} = \sum_{j=1}^{n}(A_{ij}-P_{ij})X_jWS^{-1}\\ &= \left(\sum_{j\neq i}(A_{ij}-P_{ij})X_j^\top\right)WS^{-1} - \|x\|^2x^\top WS^{-1}.\end{aligned} \tag{4.10}$$

Now, note that since we conditioned on $X_i = x = \sqrt{\rho}x^*$, the summands in the last line are independent and identically distributed with mean 0 and variance $\Sigma(x) =$

$\mathbb{E}[x^\top X_j(1-x^\top X_j)X_jX_j^\top]$. By Lyapunov's condition for univariate central limit theorem and the Cramer-Wold theorem for the multivariate normal distribution, we have

$$\frac{1}{\sqrt{n-1}}\left(\sum_{j\neq i}(A_{ij}-P_{ij})X_j^\top\right)\Sigma(x)^{-1/2}\xrightarrow{\mathcal{L}}\mathcal{N}(0,I), \tag{4.11}$$

The ratio between the left hand side of Eq. (4.11) and the first term in Eq. (4.10) after multiplying by $\sqrt{n}$ is

$$\sqrt{n(n-1)}\Sigma(x)^{1/2}S^{-1}\xrightarrow{p}\mu_2^{-1}\mathbb{E}[x^{*\top}X^*X^*X^{*\top}]\mu_2^{-1}$$

as $n\to\infty$ by Lemma 4.9. The multiplicative version of Slutsky's Theorem establishes the result. □

Proof of Theorem 4.8. We use that $\hat{X}-XW=(\hat{X}-AVS^{-1/2})+(A-P)VS^{-1/2}$. Lemma 4.5 ensures that Assumption 3.12 holds with probability tending to one so that we can use Lemma 3.16 to get that

$$\|\hat{X}-AVS^{-1/2}\|_F=O_P\left(\frac{\log(n)}{\sqrt{\rho n}}\right).$$

Now, note that the the rows of $(\hat{X}-AVS^{-1/2})$ are exchangeable so that we have that

$$\mathbb{E}[\|(\hat{X}-AS^{-1/2})_{i\cdot}\|_2^2]=\frac{1}{n}\mathbb{E}[\|\hat{X}-AVS^{-1/2}\|_F^2]=O\left(\frac{\log^2(n)}{\rho n^2}\right).$$

Using Markov's inequality we have that

$$\mathbb{P}\left[n\|(\hat{X}-AVS^{-1/2})_{i\cdot}\|_2^2\geq\epsilon^2\right]=O\left(\frac{\log^2(n)}{n\rho\epsilon^2}\right)=o(1)$$

so that $\sqrt{n}(\hat{X} - AVS^{-1/2})_{i\cdot}$ tends to zero in probability for any fixed row i. Now, we can integrate over all values of x^* in Lemma 4.10 to get that

$$\sqrt{n}(AV - PV)_{i\cdot}S^{-1/2} \xrightarrow{\mathcal{L}} \int_{\mathcal{X}} \mathcal{N}(0, \mu_2^{-1}\mathbb{E}[x^{*\top}X^*X^*X^{*\top}]\mu_2^{-1})F(dx^*)$$

uncondtionally. Applying Slutky's Theorem establishes the result. □

4.3.1 Implications

Theorem 4.8 only provides a limiting distribution for a single row of X at a time. It is not hard to extend this result to show that a fixed collection k rows of $\sqrt{n}(XW - \hat{X})$ converges jointly to a collection of k independent normals. Extending this result to a statement about the distribution of the entire matrix X as $n \to \infty$ is more difficult. At this point no results of this kind are known in the generality of RDPG.

Indeed, the impact of this result is severely weakened by the fact that only a small number of rows can be controlled at once. At this point results such as Theorem 3.18 prove to be much more useful for demonstrating the asymptotic consistency of certain limit procedures. On the other hand, the theorem has some utility and before discussing this we present one more conditional form of Theorem 4.8.

Corollary 4.11. *Under the conditions of Theorem 4.8 suppose that $\mathcal{B} \subset \mathcal{X}$ and let $\beta = \mathbb{P}[\rho_n^{-1/2}X_i \in \mathcal{B}] > 0$, where we note β does not depend on n. If we condition on the event $\rho_n^{-1/2}X_i \in \mathcal{B}$ then if $\rho_n = o(1)$ then*

$$\sqrt{n}\left(\hat{X}_{i\cdot} - X_{i\cdot}W\right) \xrightarrow{\mathcal{L}} \beta^{-1}\int_{\mathcal{B}} \mathcal{N}(0, \mu_2^{-1}\mathbb{E}[x^{\top}X^*X^*X^{*\top}]\mu_2^{-1})F(dx)$$

and if $\rho_n = 1$ *for all* n *then*

$$\sqrt{n}\left(\hat{X}_{i\cdot} - X_i W\right) \xrightarrow{\mathcal{L}} \beta^{-1} \int_{\mathcal{B}} \mathcal{N}(0, \mu_2^{-1}\mathbb{E}[x^\top X^*(1 - x^\top X^*)X^* X^{*\top}]\mu_2^{-1})F(dx).$$

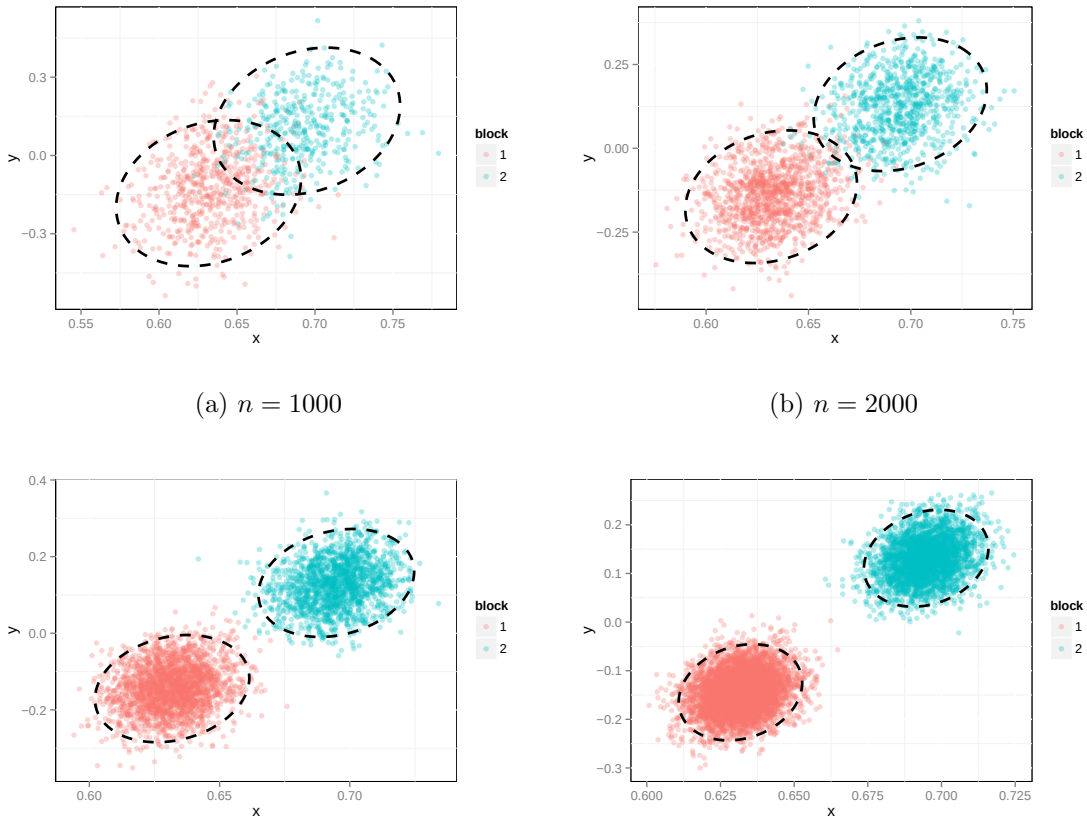
If we consider the case where the support of the distribution F is finite, then we are in the situation of a stochastic blockmodel. The distribution can be written as $F = \sum_{i=1}^{K} \pi_i \delta_{z_i}$ where δ_z denotes point mass at z. Theorem 4.8 then says that as $n\infty$ the distribution of a fixed row of $\hat{X}$ is a mixture of normals with variance tending to zero at the rate $1/n$. Corollary 4.11 ensures that the variance is determined by the the value of X_i.

Example 4.12. In this example, we consider a stochastic blockmodel with latent positions such that the probabilities for between blocks edges and the block membership probabilities are given by

$$B = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix} \text{ and } \pi = \begin{pmatrix} 0.6, 0.4 \end{pmatrix} \tag{4.12}$$

respectively. We simulate graphs of order ranging from 1000 to 16000 and examine the distribution of the latent positions. In Figure 4.1, we see that the covariance structure predicted by Theorem 4.8 and its Corollary accurately fit the observed points of the adjacency spectral embedding. Table 4.1 shows the empirical convariance of $\sqrt{n}(X_i - W^\top \hat{X}_i)$ for various values of n, the number of vertices, as compared to the limiting covariance.

We conclude our discussion of this limit theorem by noting that a Berry-Esseen



(a) $n = 1000$ (b) $n = 2000$

(c) $n = 4000$ (d) $n = 8000$

Figure 4.1: Plot of the rows of $\hat{X}$ for $n \in \{1000, 2000, 4000, 8000\}$ where the graph is a stochastic block models. Dashed ellipses give the 95% level curves for the distributions as specified in Theorem 4.8.

n	$\hat{\Sigma}_1$	$\hat{\Sigma}_2$
2000	$\begin{bmatrix} 0.58 & 0.54 \\ 0.54 & 16.56 \end{bmatrix}$	$\begin{bmatrix} 0.58 & 0.75 \\ 0.75 & 16.28 \end{bmatrix}$
4000	$\begin{bmatrix} 0.58 & 0.63 \\ 0.63 & 14.87 \end{bmatrix}$	$\begin{bmatrix} 0.59 & 0.71 \\ 0.71 & 15.79 \end{bmatrix}$
8000	$\begin{bmatrix} 0.60 & 0.61 \\ 0.61 & 14.20 \end{bmatrix}$	$\begin{bmatrix} 0.58 & 0.54 \\ 0.54 & 14.23 \end{bmatrix}$
16000	$\begin{bmatrix} 0.59 & 0.58 \\ 0.58 & 13.96 \end{bmatrix}$	$\begin{bmatrix} 0.61 & 0.69 \\ 0.69 & 13.92 \end{bmatrix}$
∞	$\begin{bmatrix} 0.59 & 0.55 \\ 0.55 & 13.07 \end{bmatrix}$	$\begin{bmatrix} 0.60 & 0.59 \\ 0.59 & 13.26 \end{bmatrix}$

Table 4.1: For each $n \in \{2000, 4000, 8000, 16000\}$, we show the sample covariance matrix for $\sqrt{n}(W^\top \hat{X}_i - X_i)$ for each block. The last line shows the theoretical covariance for the limiting distribution.

rate is also desirable. Arguments in a similar vein to those used to prove the Theorem can be used to show that a Berry-Esseen rate of at least $O(\log(n)/n^{-1/3})$ are achievable. However, we can reasonable expect that these rates are not optimal and such a result has little value for subsequent theory without being able control the distribution of the entire matrix $\sqrt{n}(XW - \hat{X})$.

Chapter 5

Implications for Inference

Up until now we have been concerned primarily with explicit bounds and distributional results related to the adjacency spectral embedding of a random graph with either independent edges (as in Chapter 3) or conditionally independent edges (as in Chapter 4). The results in these chapters give probabilistic assurances that the adjacency spectral embedding of a random graph with adjacency matrix A will be close in some sense to the spectral embedding of the unit-interval valued matrix of edge probability $P = \mathbb{E}[A]$ in the independent edge case and $P = \mathbb{E}[A|X]$ in the conditionally independent edge case, where X represents the collection of latent positions. If our sole goal were to estimate the latent positions in an RDPG or the feature map of the latent positions for a more general LPG, these results may provide the necessary assurance of accuracy we desire.

However, this is not typically the end goal. One of the values of the adjacency

spectral embedding is that, in the non-standard context of a graph, a variety of exploitation tasks can be performed using standard multivariate statistical techniques. In this chapter we will explore the implications of the results of the previous two chapters on various inference tasks.

The remainder of this chapter will be divided according to the inference task of interest, either estimation, classification, or clustering. Most of the results we consider are focused around the RDPG case with iid latent positions where our strongest bounds are available. For estimation, we will analyze the classical question of estimating a parameter of the distribution of latent position and see that provided the parameter is an appropriately smooth function of the distribution then estimation via the extension principal will given consistent results. For classification, we give the details for two different classifiers and focus on universal consistency. Finally, for clustering we show that under the stochastic blockmodel the asymptotic probability that any vertices are misclustered tends to zero.

Needless to say, these are not the totality of inference tasks that one might wish to perform and we present these select few only to provide a flavor for the kind of results that are possible. Indeed, we hope that the results in Chapters 3 and 4 will have a long life and productive life outside of this current work in a multitude of different statistical applications.

5.1 Estimation

First, for the independent edge setting, suppose that $A \sim \text{RDPG}(X)$ for some fixed latent positions $X \in \mathcal{X}^n \subset \mathbb{R}^{n\times d}$. In this setting, an estimate of a function of the latent positions may be sought. For $h : \mathbb{R}^{n\times d} \mapsto \Theta$, where (Θ, d_Θ) is a metric space, we will $\theta = h(X)$. The plug-in principles of statistics suggests that a natural estimate for θ is $\widehat{\theta} := h(\hat{X})$. Examples of such functions that may be of interest include $h(X) = X_i^\top X_j$ for some i and j, the probability of an edge between two nodes, or $h(X) = \max_i \|X_i^\top \bar{X}\|_2$, the maximum expected degree.

To what extent can we make guarantees about the quality of this estimate? As we might expect no guarantees are possible for arbitrary h. To begin with the function h must invariant under orthogonal transformations because the RDPG distribution is only identifiable up to such transformations. If we make this assumption, namely that $h(X) = h(XW)$ for all $X \in \mathcal{X}^n$ and $W \in \mathcal{O}(d)$, then we can apply all the results in Chapter 3. For example, if we suppose that h is Lipschitz with respect to the $\|\cdot\|_{2\to\infty}$ norm on $\mathbb{R}^{n\times d}$ with Lipschitz constant L, as can be shown for the examples above, then we easily have that

$$d(\theta, \hat{\theta}) \leq L \min_W \|X - \hat{X}\|_{2\to\infty} \leq L\left(\sqrt{\frac{d\log(4dn/\eta)}{2\delta\gamma}} + \sqrt{\frac{d^3}{\gamma^7\delta}}108\log(n/\eta)\right)$$

with probability at least $1-2\eta$. Assumptions of uniform continuity can also yield similar bounds but we will be more interested in more traditional parametric estimation rather than simply functions of the latent positions.

In this case we suppose $A \sim \text{RDPG}(\mathcal{X}, F)$ where $F \in \mathcal{F}$, a model consisting of distributions on $\mathcal{X}$ and let $\mathcal{P}$ consist of all distributions on $\mathbb{R}^d$. Again we have some function $h : \mathcal{P} \mapsto \Theta$ that gives a parameter of the distribution that we wish to estimate.

In the classical statistical setting, we observe $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$ and the plug-in principle suggests that a natural estimate for $\theta = h(F)$ is $\hat{\theta}_X = h(\hat{F}_X)$ where $\hat{F}_X$ is the empirical distribution for the latent positions, namely

$$\hat{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x \preceq X_i\},$$

where $\preceq$ denotes elementwise less-than-or-equal ($x \preceq y \iff x_j \leq y_j$ for all $j \in [d]$). Nearly any graduate text in statistics provides a plethora of theorems developing the consistency, asymptotic normality and efficiency of the estimate $\hat{\theta}_L$ in a variety of settings [Bickel and Doksum, 1976, Lehmann and Casella, 1998, Bickel et al., 1998], provided the the X_is themselves are observed. However, if instead we observe only an adjacency matrix $A \sim \text{RDPG}(\mathcal{X}, F)$, then following the ideas above, we can first estimate the latent positions with $\hat{X} = [\hat{X}_1, \ldots, \hat{X}_n^\top]$ and then estimate θ with $\hat{\theta}_{\hat{X}} = h(\hat{F}_{\hat{X}})$ where now $\hat{F}_{\hat{X}}$ is the empirical distribution for the estimated latent latent positions.

To what extent $\hat{\theta}_{\hat{X}}$ will inherit properties from $\hat{\theta}_X$ will again depend on the specific function h. To start, we must have that h is invariant under rotations in the following sense: for each $F \in \mathcal{P}$ and for each orthogonal matrix W let $F_W \in \mathcal{P}$ be the unique distribution such that if $Z \sim F$ then $WZ \sim F_W$. Then h is said to be orthogonally

invariant if $h(F) = h(F_W)$ for all such F and W. Provided h is orthogonally invariant then we can again hope that $\hat{\theta}_{\hat{X}}$ will have performance guarantees in relation to $\hat{\theta}_X$.

Indeed, we can again appeal to our bound on the $\|\hat{X} - X\|_{2\to\infty}$. To see how this bound can be useful in this setting we must introduce an appropriate metric on $\mathcal{P}$. For our purposes we will consider the Wasserstein metric, also known as the Kantorovich metric or the bounded Lipschitz metric. Of the equivalent definitions for this metric, we will find most useful the following definition:

$$d_W(F_1, F_2) = \inf\{\mathbb{E}[\|Z_1 - Z_2\|_2^2]^{1/2} : \mathcal{L}(Z_i) = F_i, i \in \{1, 2\}\}$$

where the infemum is over all random variables $Z = (Z_1, Z_2)$ taking values in $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginal distribution of Z_i is F_i for $i \in \{1, 2\}$ [Huber and Ronchetti, 2009]. The reason this metric is useful in our setting is because we can bound d_W in terms of the $\|\cdot\|_{2\to\infty}$ norm. For convenience, we will consider the Wasserstein metric on the quotient space induced by orthogonal transformations,

$$d_{WR}(F_1, F_2) = \inf\{\mathbb{E}[\|Z_1 - Z_2\|_2^2]^{1/2} : \mathcal{L}(Z_1) = F_1, \mathcal{L}(WZ_2) = F_2, W \in \mathcal{O}(d)\}.$$

It is now clear that

$$d_{WR}(\hat{F}_X, \hat{F}_{\hat{X}}) \leq \min_{W \in \mathcal{O}(d)} \|X - \hat{X}W\|_{2\to\infty}$$

since we can take Z to be such that for each $i \in [n]$, $\mathbb{P}[Z = (X_i, W\hat{X}_i)] = \frac{1}{n}$ as then

$$\mathbb{E}[\|Z_1 - Z_2\|_2^2] = \frac{1}{n}\sum_{i=1}^{n} \|X_i - W\hat{X}_i\|_2^2 = \frac{1}{n}\|X - \hat{X}W\|_F^2 \leq \|X - \hat{X}W\|_{2\to\infty}^2.$$

The argument above also shows that the Wasserstein metric can be bounded in terms of the Frobenius norm, which will provide better bounds in some cases.

Using these arguments it is straightforward to prove the consistency of estimation provided the function h is sufficiently smooth with respect to the Wasserstein metric. One such theorem is the following, which we state and provide a sketch of the proof.

Theorem 5.1. *Let $A^{(n)} \sim \text{RDPG}(\mathcal{X}, F)$, $\mathcal{X} \subset \mathbb{R}^d$ be a random graph on n vertices. Let $X^{(n)} = [X_1^{(n)}, \ldots, X_n^{(n)}]$ be the matrix of latent positions for $A^{(n)}$ and let $\hat{X}^{(n)}$ be the d-dimensional adjacency spectral embedding of $A^{(n)}$.*

Let $h : \mathcal{P} \mapsto \mathcal{Z}$ be a function from the space $\mathcal{P}$ of all distributions on $\mathbb{R}^d$ to a metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ and suppose that h is rotationally invariant (as defined above) and suppose h is uniformly continuous with respect to the topology induced by the Wasserstein metric in an open neighborhood around F. Then the following hold,

$$d_{\mathcal{Z}}(h(F_X^{(n)}), h(F)) \xrightarrow{p} 0, \quad d_{\mathcal{Z}}(h(F_X^{(n)}), h(F_{\hat{X}^{(n)}})) \xrightarrow{p} 0,$$

$$\text{and} \quad d_{\mathcal{Z}}(h(F_{\hat{X}^{(n)}}), h(F)) \xrightarrow{p} 0,$$

as $n \to \infty$.

Sketch of proof. The first statement is a straightforward consequence of the Glivenko-Cantelli Theorem and the equivalence of the Wasserstein metric topology and the topology induced by the sup-norm on distribution functions.

The second result is due to the argument above that

$$d_{WR}(F_X, F_{\hat{X}}) \leq \min_{W \in \mathcal{O}(d)} \|X - \hat{X}W\|_{2 \to \infty}.$$

In this case, Lemma 4.5 ensures that for n large with probability at least $1 - n^{-2}$, $\Gamma\Delta \geq n\gamma/2$ with $\gamma > 0$ not depending on n. Trivially, $\Delta \leq n$ and d is fixed so Theorem 3.18 gives that $\min_{W^\top W=I} \|\hat{X}W - X\|_{2\to\infty} = O_P(\log(n)/\sqrt{n}) = o_P(1)$. The uniform continuity on an open neighborhood of F proves the result and the rotational invariance of h proves the result. The last result follows from the other two. □

The assumptions and conclusions of this theorem can be strengthened and weakened in various ways depending on the situation at hand. For example, if we make the stronger assumption that h is Lipschitz then we have that $h(F_X, F_{\hat{X}}) = O_P(\log(n)/\sqrt{n})$. Though we do not investigate these results further in this work, there are many potential extensions to consider. Indeed, the results in the next two sections can be viewed as results of the same type as the theorem above, as the key property is that the algorithms used are smooth with respect to the $2 \to \infty$ norm.

Theorem 5.1 shows that to first order, consistency of an estimate in the iid case implies consistency of the analagous estimate using $\hat{X}$ in the RDPG case provided the estimate is sufficiently smooth. Hence, the project of showing the consistency of various estimates can be reduced partially to the project of showing that the estimates are sufficiently smooth. Questions of efficiency of estimates are left open as this theorem gives only first order guarantees, however it is possible that improvements of our distributional result Theorem 4.8 would enable the establishment of thee relative efficiency of our estimates as compared to their counterparts in the iid setting. We leave this project to future work and now focus on the specific tasks of classification

and clustering.

5.2 Classification

In this section we will highlight the implications of the results in Chapters 3 and 4 for the purposes of vertex classification. Before delving into the specifics of vertex classification we will give an overview of classification in classical statistical pattern recognition in Section

5.2.1 Statistical Pattern Recognition

The classical statistical pattern recognition setting involves

$$(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} F_{X,Y},$$

where the $X_i : \Omega \mapsto \mathbb{R}^d$ are observed feature vectors and the $Y_i : \Omega \mapsto \{0, 1\}$ are observed class labels for some probability space Ω. The goal is to predict the class label Y based on the observation X and training data $\{(X_i, Y_i)\}_{i=1}^n$. If the joint distribution $F_{X,Y}$ is known, then it is well known that the Bayes classifier given by $h^*(x) = \arg\max_{y \in \{0,1\}} \mathbb{P}[Y = y | X = x]$ has optimal performance in the sense of minimum probability of error. This probability of error, which we denote $L^* = \mathbb{P}[h^*(X) \neq Y]$, is called the Bayes optimal probability of error.

Given a training set, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, one goal is to learn a classifier $h(\cdot; \mathcal{D}) : \mathbb{R}^d \to \{0, 1\}$ such that the probability of error $L_n = \mathbb{P}[h(X; \mathcal{D}) \neq Y | \mathcal{D}]$ converges in

probability to Bayes optimal as $n \to \infty$ for all distributions $F_{X,Y}$ — this is known as universal consistency [Devroye et al., 1996]. Such a universally consistent classification rule has the property that given $\epsilon > 0$ and a joint distribution $F_{X,Y}$, there exists an N such that if $n > N$ then with high probability $L_n - L^* < \epsilon$.

Many universally consistent classification rules exist but arguably the simplest is the k-nearest-neighbor (NN) classifier which is also one of the first classifiers that was proved to have this property. Another collection of universally consistent classifiers are linear classifiers based on kernels. We will study the k-NN classifier in the RDPG case and linear classifiers in the LPG case.

Before describing the particular classifiers that we will study, we must introduce the vertex classification setting. In this setting, we observe a graph and class labels for a subset of vertices. Hence, we define the *latent position model with class labels*.

Definition 5.2 (LPG with class labels)**.** We say $(A, Y) \sim \text{LPG}_\text{C}(\mathcal{X}, F, \kappa)$ if $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} F$ for the probability distribution F on $\mathcal{X} \times \{0, 1\}$ and $Y = [Y_1, \ldots, Y_n]^\top$ and A is an LPG with latent positions $X_1, \ldots, X_n$ and link function κ. We define $(A, Y) \sim \text{RDPG}_C(\mathcal{X}, F)$ analogously.

For the vertex classification setting we have $(A, Y) \sim \text{LPG}_\text{C}(\mathcal{X}, F, \kappa)$ but we only observe A and $\{Y_i\}_{i \in \mathcal{T}}$ for some training set $\mathcal{T} \subset [n]$. Our goal will be to determine Y_i for $i \in [n] \setminus \mathcal{T}$. In this situation, we will abuse notation slightly and say $(A, \{Y_i\}_{i \in \mathcal{T}}) \sim \text{LPG}_\text{C}(\mathcal{X}, F, \kappa)$ to indicate that we observ the class labels in the training set $\mathcal{T}$ but we assume that the joint distribution of the class labels and latent

positions is the same for all vertices.

5.2.2 k-nearest-neighbor classifier and RDPG

The k-nearest neighbor rule is a classification rule even a child could get behind. It relies on the basic principle that like objects should be classified together. If we are confronted with a new object, we compare it to objects we have seen before and categorize it with ones that are most similar. We measure *likeness* in terms distance and define the k-nearest neighbor rule as follows.

Let $k \in \mathbb{N}$ be odd, denote the training set as $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ and for a given point $x \in \mathbb{R}^d$ let $\mathcal{N}_k(x) = \mathcal{N}_k(x, \mathcal{D}) \subset [n]$ be the collection of indices such that $i \in \mathcal{N}_k(x)$ if and only if X_i is among the k closest elements of $\mathcal{D}$ to x. (For concreteness, assume that ties are broken so that the lowest index is chosen.)

We now define a weight function $W_{ni}(\cdot, \mathcal{D}) : \mathbb{R}^d \mapsto [0, 1]$ as

$$W_{ni}(x) = W_{ni}(x, \mathcal{D}) = \begin{cases} \frac{1}{k} & \text{if } i \in \mathcal{N}_k(x) \\ 0 & \text{otherwise.} \end{cases}$$

The k-nearest neighbor classification rule is defined as

$$h(x, \mathcal{D}) = \mathbb{I}\left\{\sum_{i=1}^n W_{ni}(x, \mathcal{D}) Y_i > \frac{1}{2}\right\}.$$

Informally, $h(X)$ is assigned class one if most of its nearest neighbors are in class one and otherwise it is assigned class 0.

5.2.2.1 Universal Consistency

The universal consistency of the k-nearest neighbors rule was proven by Stone [1977]. In fact he proved the more general result that follows. In this theorem the weight function is allowed to be more general than just the k-nearest neighbor rule.

Theorem 5.3 (Stone [1977])**.** *Assume that for any distribution of X, the weights W_{ni} satisfy the following three conditions:*

(*i*) *There exists a constant c such that for every nonnegative measurable function f satisfying $\mathbb{E}[f(X)] < \infty$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}[f(X)]. \tag{5.1}$$

(*ii*) *For all $a > 0$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{I}\{\|X_i - X\| > a\}\right] = 0 \tag{5.2}$$

(*iii*)

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(X)\right] = 0 \tag{5.3}$$

Then $h_n(x) = \mathbb{I}\{\sum_i W_{ni}(x) > 1/2\}$ is universally consistent.

Condition (i) can be viewed as a smoothness condition and is the most technical of the three conditions. Condition (ii) is a locality constraint ensuring that most of the total weight is on points near the the to-be-classified point X. Condition (iii) ensures that as n grows the weight on any one point must tend to zero so that the

number of points with non-zero weight must tend to infinity. Note from the k-NN standpoint this means we must let $k \to \infty$ so that the number of "votes" goes to infinity. Seen as a voting scheme, the conditions ensure that the votes are not divied up pathologically, lots of votes are cast, and only nearby points are allowed to cast votes.

The following theorem gives that the provided k satisfies some mild constraints, the k-nearest-neighbor rules is universally consistent, hence our extremely simple classifier has powerful asymptotic properties.

Theorem 5.4 (Stone [1977]). *Let* $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} F$ *for some joint distribution* F *on* $\mathbb{R}^d \times \{0, 1\}$. *If* $k \to \infty$ *and* $k/n \to 0$ *as* $n \to \infty$, *then the* $W_{ni}(X)$ *satisfy the conditions of Theorem 5.3 and hence* $\mathbb{E}[\mathbb{P}[h_n(X) \neq Y|\mathcal{D}]] = \mathbb{E}[L_n] \to L^*$ *for all joint distributions* F *on* $\mathbb{R}^d \times \{0, 1\}$.

Note, the assumption that $k \to \infty$ ensures condition (iii) hold and the assumption that $k/n \to 0$ can be used to prove that conditions (i) and (ii) are satisfied.

5.2.2.2 RDPG

Perhaps unsurprisingly, this theorem can be extended to the RDPG case and perhaps less surprising the same asymptotic error rate is achieved as if we had observed the true latent positions. Again, we will apply the k-nearest-neighbors rule to the estimated latent positions, the rows of $\hat{X}$.

In this case we define the neighborhood $\hat{\mathcal{N}}_k(x) \subset \mathcal{T}$ to be a collection of indices where $i \in \hat{\mathcal{N}}_k(X)$ if and only if $\hat{X}_i$ is among the k closest elements of $\{\hat{X}_i\}_{i=\mathcal{T}}$ to x. The weight function is defined only in the context of latent positions for the random graph in the sense that

$$W_{ni}(X_j, A, \mathcal{T}) = W_{ni}(X_j) = \begin{cases} \frac{1}{k} & \text{if } i \in \hat{\mathcal{N}}_k(\hat{X}_j) \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

where $(X_j, \hat{X}_j)$ are the true and estimated latent positions for vertex $j \in [n]$. It might seem like we have to observe the true latent positions to evaluate W_{ni} but in reality it only depends on the estimated latent positions $\hat{X}$.

The classifier is given by

$$h_n(X_j, A, \{Y_i\}_{i\in\mathcal{T}}) = h_n(X_j) = \mathbb{I}\left\{\sum_{i\in\mathcal{T}} W_{ni}(X_j, A, \mathcal{T})Y_i > \frac{1}{2}\right\}.$$

Since we do not actually observe the true latent positions we define $\hat{W}_{ni}(\hat{X}_j) = W_{ni}(X_j)$ as observing X_j is not necessary to evaluate $W_{ni}(X_j)$. Finally, let $\hat{h}_n(\hat{X}_j) = h_n(X_j)$. The somewhat awkward notation is useful to prove the following theorem establishing universal consistency in the RDPG case.

Theorem 5.5 (Sussman et al. [2013]). *Let $(A, \{Y_i\}_{i\in\mathcal{T}}) \sim RDPG(\mathcal{X}, F)$ be a graph on n vertices for some distribution F on $\mathbb{R}^d \times \{0, 1\}$. Let $m = |\mathcal{T}|$.*

If $k \to \infty$, $m \to \infty$ and $k/m \to 0$ as $n \to \infty$, then the $W_{ni}(X_j)$ defined in Eq. (5.4) satisfy the conditions of Theorem 5.3 and hence for $j \notin \mathcal{T}$,

$$\mathbb{E}[\mathbb{P}[\hat{h}_n(\hat{X}_j) \neq Y | A, \{Y_i\}_{i\in\mathcal{T}}]] = \mathbb{E}[L_n] \to L^* = \mathbb{P}[g^*(X) \neq Y]$$

for all joint distributions F *on* $\mathbb{R}^d \times \{0, 1\}$.

We remark that the L^* in the theorem is the Bayes optimal error had we observed the true latent positions so in an asymptotic sense nothing is lost by the fact we observe the graph instead.

Sketch of proof. The proof of the theorem follows *mutatis mutandis* the proof of Theorem 5.4 as proved in Devroye et al. [1996]. Again, condition (iii) is immediate from the definition of the defintion of the W_{ni}. The key fact for the remainder of the proof is that

$$\lim_{n\to\infty} \max_{i\in\mathcal{T}} \mathbb{E}[kW_{ni}\|\hat{X}_j - \hat{X}_i\|_2]] \xrightarrow{p} 0,$$

where we use Theorem 3.18 and the law of large numbers to prove this result. Conditon (ii) is proven by using the triangle inequality and Theorem 3.18 along with this result. Condition (iii) follows from this fact and replicating the proof in Devroye et al. [1996]. □

Like our result for estimation, this theorem establishes that in a supervised learning framework working with the estimated latent positions rather than the true latent positions imposes no cost in the limit, at least to first order. It is not hard to imagine that other simple universally consistent classification rules such as those based on density estimates can also be adopted to this setting. Moving to more general latent position graphs, rather than the RDPG setting, the k-nearest-neighbor rule is no longer as natural and we consider another simple classifications scheme.

5.2.3 Linear classifiers and universal LPG

In some ways even simpler though perhaps less natural than k-nearest-neighbors, linear classifiers are so called because the boundary between objects selected to be hyperplanes. These line-in-the-sand rules at first appears unlikely to be able to achieve universal consistency. This is clear if we consider linear classifiers in the original domain as there is no reason that the Bayes optimal classifier can be in any way reasonably approximated as a linaer classifier. To make these classifiers consistent we must use a trick: first we will map the data *non-linearly* into some high possibly infinite dimensional space and then we will use linear classifiers in this space. As we will see, this leads to another broad class of universally consistent classifiers.

A linear classifier on a Hilbert space $\mathcal{H}$ (either finite or infinite dimensional) is any classifier of the type $h_{w,c}(x) = \mathbb{I}\{\langle w, x\rangle > c\}$ for $w \in \mathcal{H}$ and $c \in \mathbb{R}$. In general it is not sufficient to consider only linear classifiers if universal consistency is the goal since for joint distributions on $\mathcal{H} \times \{0, 1\}$, the Bayes optimal classifier can have arbitarily non-linear properties. However, if we consider a space $\mathcal{X}$ and distributions on $\mathcal{X} \times \{0, 1\}$ then after an appropriate mapping to an infinite dimensional Hilbert space, considering linear classifiers is sufficient.

Again we refer back to the theory of reproducing kernel Hilbert spaces and feature maps (see sections 2.4 and 4.1) and presently introduce the idea of a *universal kernel*. The key idea behind these classifiers is that if the link function is universal then there exists a *linear classifier* on ℓ_2 that achieves Bayes optimal performance on the *feature*

mapped latent positions.

Our strategy will be to use the d-dimensional adjacency spectral embedding for some d depending on n and the observed eigenvalues of the the adjacency matrix. The procedure to choose d will ensure that the embedding provides a sufficiently accurate estimate of the truncated feature map while simultaneously ensuring that d tends to infinity with n. Together, the best linear classifier on the truncated feature maps will approach the best linear classifier on the full feature map.

We will give an overview of the key ideas. First, we will examine the valuable proerties of a universal kernel. We will then overview some theory for empirical risk minimization (ERM) and state a theorem regarding (ERM) using the truncated feature map. Finally, we will state our main theorem for vertex classification for LPG.

5.2.3.1 Universal Kernels

Definition 5.6 (Universal Kernel)**.** Let $\mathcal{X}$ be a compact metric space and let $\kappa : \mathcal{X}^2 \mapsto [0,1]$ be a positive semi-definite kernel (ie link function). Let $\phi : \mathcal{X} \mapsto \ell_2$ be a feature map of κ. The kernel κ is said to be *universal* if the set of functions from $\mathcal{X}$ to $\mathbb{R}$ of the form

$$\mathcal{F}_\phi = \{\langle w, \phi(\cdot)\rangle_{\ell_2} : w \in \ell_2\} \tag{5.5}$$

is dense in the space of continuous functions on $\mathcal{X}$. That is, for any continuous function $g : \mathcal{X} \mapsto \mathbb{R}$ and any $\epsilon > 0$, there exists $g_\phi \in \mathcal{F}_\phi$ such that $\|g - g_\phi\|_\infty < \epsilon$.

Remark 5.7. We note that as stated this definition appears to depend on a particular

feature map ϕ, but it can be shown that this is not the case and that if another feature map is selected, the resulting space of functions will still be dense.

If the link function κ is universal, then this means that we can approximate continuous functions, and hence measurable functions, by functions of the form in Eq. (5.5). In particular, since the Bayes optimal classifier is measurable, then we can show that for any $\epsilon > 0$ there exists a function $g_\phi \in \mathcal{F}_\phi$ such that $\mathbb{P}[g_\phi(X) \neq Y] - L^* < \epsilon$. Hence, if a feature map corresponding to a universal kernel is known then we can remap the training data and then select from among linear classifiers in ℓ_2. Selecting this classifier can be somewhat non-trivial as there will be infinitely many hyperplanes that will separate the data nearly optimally. The next section provides a method to select such a classifier that has the necessary performance guarantees.

5.2.3.2 Empirical Risk Minimization

Another simple principle, given a collection of classifiers, choose the one that performs best on the training data. This basic idea again seems like it will not lead to performance guarantees because we are liable to overfit, but if we appropriately restrict the collection of classifiers, universal consistency can be achieved.

The emprical risk of a classifier h on a training set $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ is given by

$$\hat{L}(h) = \sum_{i=1}^{n} \mathbb{I}\{h(X_i) \neq Y_i\}. \tag{5.6}$$

and the expected risk is given by $L(h) = \mathbb{P}[h(X) \neq Y]$ for $(X, Y) \sim F$. The principle of empirical risk minimization (ERM) states that if we restrict our attention to

classifiers in a set $\mathcal{G}$, then we should select a classifier

$$h^* \in \arg\min_{h \in \mathcal{G}} \hat{L}(h).$$

Given a universal kernel κ, we need only consider classifiers in the collection $\mathcal{G} = \{\mathbb{I}\{\langle w, \cdot \rangle_{\ell_2} > c\} : w \in \ell_2, c \in \mathbb{R}\}$. However, using ERM on all of $\mathcal{G}$ will not give a unique minimizer, so we need another criterion to select the "best" classifier from among these minima. A standard way to deal with this is to use *structural risk minimization* rather than ERM and our strategy is related to this idea. Indeed, in order to have performance guarantees we will need to restrict $\mathcal{G}$ so that it does not necessarily contain a good approximation of the Bayes optimal classifier. We will at first be content to find a classifier that is close to best possible in this restricted class.

Specifically, we restrict our classifier h to be from the collection

$$\mathcal{G}_d = \{\mathbb{I}\{\langle w, \cdot \rangle_{\ell_2} > c\} : w \in \ell_2, w_i = 0, \forall i > d; c \in \mathbb{R}\}$$

for some $d < n$ that depends on n. Importantly, restricting to $\mathcal{G}_d$ ensures that the expected risk is near the best possible for classifiers in that set, which is not true for larger classes like $\mathcal{G}$.

Proposition 5.8. *Let $d \in [n]$ and let h^*_{nd} be the ERM classifier in $\mathcal{G}_d$ trained on $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} F$. Then*

$$\mathbb{E}[L(h^*_d)] - \inf_{h \in \mathcal{G}_d} L(h) \leq 16\sqrt{\frac{(d+1)\log(n) + 4}{n}},$$

regarless of the distribution F

This theorem asserts that the ERM classifier will be nearly as good as the best classifier in $\mathcal{G}_d$. Hence if we let $d \to \infty$ grow so that $d\log(n)/n \to 0$, then $\lim_{n\to\infty} \mathbb{E}[L(h_d^*)] - L^* \to 0$ since $\lim_{d\to\infty} \inf_{h\in\mathcal{H}_d} L(h) \to L^*$ by the universality of the kernel.

Up until now we have considered directly minimizing 0-1 loss. In our situation, where we observe essentially noisy versions of a training set, minimizing 0-1 loss as in Eq. 5.6 is inappropriate because it is disontinous, so slight changes in the training data can drastically change the selected classifier. In these situations, we will minimize a smoother loss function known as a convex surrogate. The convex surrogate is a function $\psi : \mathbb{R} \mapsto [0, \infty)$ which we assume is differentiable with $\psi'(0) < 0$. (This is known as a classification calibrated convex surrogate.) For a function $g : \mathcal{X} \mapsto \mathbb{R}$ the empirical ψ-risk is given by

$$\hat{L}_\psi(g) = \sum_{i=1}^{n} \psi\left(g(X_i)(2Y_i - 1)\right). \tag{5.7}$$

and the expected ψ-risk is given by $L_\psi(h) = \mathbb{E}[\psi(g(X)(2Y-1))]$. The associated classifier is given by $h(x) = \mathbb{I}\{g(x) > 0\}$. Our theorem for LPG will use the classifier associated with empirical ψ-risk minimization given by

$$g_d^* = \arg\min_{g\in\mathcal{F}_d} \hat{L}_\psi(g) \text{ where } \mathcal{F}_d = \{\langle w, \cdot\rangle : w \in \ell_2, w_i = 0, \forall i > d\}. \tag{5.8}$$

A classic and easily computed example of ER$_\psi$M is Fisher's Linear discrimant, the Bayes plug-in rule when the class conditional distributions are multivariate normal with equal covariances.

5.2.3.3 LPG

Putting together the ideas presented so far, we will suppose that we observe $(A, \{Y_i\}_{\mathcal{T}}) \sim \text{LPG}(\mathcal{X}, F, \kappa)$ where κ is a positive semi-definite universal link funciton. If we had observed the true feature vectors, the link function κ, and we were able to compute the truncated feature map then we would be able to select a classifier via ER_ψM. Observing none of this, we will use our knowledge of the adjacency spectral embedding to select an embedding dimension and then use a convex surrogate to select the classifier.

Theorem 5.9 (Tang et al. [2013])**.** *Suppose* $(A, \{Y_i\}_{\mathcal{T}}) \sim \text{LPG}_C(\mathcal{X}, F, \kappa)$. *Let* $\epsilon \in (0, 1/4)$ *be fixed, let* ψ *be a convex surrogate and for a given* d *let* $C_d = \max_{\{} \psi'(-d), \psi'(d)\}$. *Let* $m = |\mathcal{T}|$.

Let d *be given by*

$$d = d_n = \max \left\{ d \leq n : |\lambda_d(A)| - |\lambda_{d+1}(A)| \geq 32n\sqrt{dC_d} \left(\frac{d \log m}{m} \right)^{1/4-\epsilon} \right\}$$

Let h *be the classifier associated with by empirical* ψ*-risk minimizer over* $\mathcal{F}_d = \{\langle w, \cdot \rangle_{\mathbb{R}^d} + c : w \in \mathbb{R}^d, c \in \mathbb{R}\}$ *using* $\{\hat{X}_i, Y_i\}_{\mathcal{T}}$ *as the training set. If* $m \to \infty$ *as* $n \to \infty$ *then the expected 0-1 risk of* h *tends to the Bayes risk for* F*:*

$$\mathbb{E}[\mathbb{P}[h_n(\hat{X}_i) \neq Y_i]] \to L^* \text{ as } n \to \infty.$$

The proof of this theorem requires significant development but the key ideas have been presented so far. We note that as d is defined, we are ensured that $d \to \infty$

as $n \to \infty$ provided $m \to \infty$ since by Lemma 4.7, for any fixed d, the eigengap $\lambda_d(A) - \lambda_{d+1}(A) = \Theta_P(n)$. Finally, we note that the specific choice of embedding dimension here is not claimed to be optimal but merely sufficient to guarantee universal consistence of the learned classifier. Other heuristics and ideas for selecting the dimension d will be considered in the discussion.

5.3 Clustering

Vertex clustering is the one of the most widely studied inference tasks for graphs. As we discussed in section 1.2.4, there are many different approaches for clustering and spectral based methods are some of the most popular. For the remainder of this section we will focus on the adjacency spectral embedding in the context of the stochastic blockmodel (see section 2.3.2).

The basic clustering algorithm we consider is to use the adjacency spectral embedding to to get $\hat{X}$ and then we will use minimum square error (MSE) clustering to cluster the rows of $\hat{X}$. The MSE clustering of the rows of $\hat{X}$ into K clusters is given by

$$\hat{\tau} \in \underset{\tau \in [K]^n}{\arg\min} \min_{C \in \mathcal{C}_\tau} \|C - \hat{X}\|_F$$

where $\mathcal{C}_\tau = \{C \in \mathbb{R}^{n \times d} : C_i = C_j \iff \tau_i = \tau_j\}$. This clustering criterion is the same one that the famous k-means algorithm attempts to minimize. It is also a restriction of Gaussian mixture modeling where all covariance matrices are taken

to be the identity. The following theorem illustrates the power of our bounds from Chapter 3 for minimum square error clustering.

Theorem 5.10 (Adapted from Lyzinski et al. [2013]). *Suppose* $A \sim \text{RDPG}(X)$ *for* $X \in \mathbb{R}^{n\times d}$ *and suppose* X *has exactly* K *distinct rows which we denote* $\xi_1, \ldots, \xi_K \in \mathbb{R}^d$. *Then equivalently* $A \sim \text{SBM}(\tau, \{\xi_1, \ldots, \xi_K\}, \langle\cdot,\cdot\rangle)$ *for an appropriately defined block membership function* τ. *Let* γ *and* δ *be as in Theorem 3.18 denote the bound on* $\|\hat{X} - X\|_{2\to\infty}$ *in Theorem 3.18 as* $\beta = \beta(d, n, \eta, \gamma, \delta)$.

Let $r > 0$ *be such that for all* $i \neq j \in [K]$, $\|\xi_i - \xi_j\|_2 > 4r$. *Let* $\hat{\tau} : [n] \to [K]$ *be the MSE clustering of the rows of* $\hat{X}$ *into* K *clusters. Let* S_K *denote the symmetric group on* K, *and* $\pi \in S_K$ *a permutation of the blocks. Finally, let* $n_{min} = \min_{k\in[K]} |\{i : \tau_i = k\}|$ *be the smallest block size. If* $r > \beta\sqrt{n/n_{min}}$ *and* $\gamma\sqrt{\delta} > 4\sqrt{\log(n/\eta)}$ *then with probability at least* $1 - 2\eta$,

$$\min_{\pi\in S_K} |\{i \in [n] : \tau(i) \neq \pi(\hat{\tau}(v))\}| = 0.$$

Taken into the asymptotic, realm this theorem ensures that if the block memberships are iid and the number of blocks K is kept fixed (or alternatively allowed to grow slowly) then as the number of vertices tends to infinity, the probability that any vertex is misclustered tends to zero. Theorem 5.10 is the first theorem to show that "perfect clustering" of the vertices in a stochastic blockmodel will occur asymptotically using spectral methods. Bickel and Chen [2009] showed that maximizing a likehihood based modualrity also has similar guarantees but their method is more

computationally demanding to perform and is useful only for clustering as opposed to our embedding method which admits a plethora of subsequent inference tasks.

Lyzinski et al. [2013] also showed that asymptotically perfect clustering is achieved in the case of the degree-corrected stochastic blockmodel Karrer and Newman [2011] under similar assumptions. In the degree-corrected stochastic blockmodel, the latent positions for each block are concentrated along a ray and the estimated latent positions must be projected onto the sphere before using the k-means algorithm.

Clustering in other models can also be considered however in this case a natural set of labels or block memberships for the vertices does not necessarily exist. In this case we can appeal to the ideas in Section 5.1 and seek a clustering that corresponds to some population version of the clustering if the distribution of the latent positions were known. Pollard [1981] shows that if iid observations from a distribution are observed then asymptotically the clustering given by the k-means criterion converges almost surely to an appropriately defined population clustering. It can be argued that the k-means criterion satisfies the appropriate smoothness conditions so that the results of Pollard [1981] can be extended to the RDPG setting. Beyond the finite-dimensional RDPG setting, techniques similar to those in Section 5.2.3 may be fruitful but we do not explore those at this time.

Example 5.11 (2-Block SBM, Example 4.12 continued)**.** This example is reproduced from Athreya et al. [2013]. We return to the model considered in Example 4.12 where we have a stochastic blockmodel with latent positions such that the probabilities for

between blocks edges and the block membership probabilities are given by

$$B = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix} \text{ and } \pi = \begin{pmatrix} 0.6, 0.4 \end{pmatrix} \tag{5.9}$$

In that example we demonstrated the approximate normality of the estimated of the estimated latent positions. Here, we consider clustering using two different clustering methodologies, the k-means method discussed above as well as the Gaussian mixture modelling (GMM) method which generalizes k-means to allow for non-spherical and non-equal covariances.

For each $n \in \{1000, 1500, \ldots, 4000\}$ we simulated 200 random graphs according to the distribution in Eq. 5.9. Figure 5.1 shows two curves giving the mean empirical error rates (with standard error) for k-means and GMM clustering. The top cyan curve shows the decay rate of $\log(n)/n$ which were the best bounds for spectral methods established in Rohe et al. [2011], Sussman et al. [2012] before the establishment of Theorem 5.10 in Lyzinski et al. [2013]. The bottom magenta curve shows what the error rate would be if the clustered data were distributed according to the finite sample version of the limiting distribution in Theorem 4.8 and this distribution were known. This is the Bayes optimal classification rate for a mixture of normals with the means and covariances specified by the limit theorem.

It is notable that the error rate for GMM closely mirrors the this theoretical lower bound error rate. Unsurprisingly, GMM outperforms k-means significantly since the distributions do not have spherical covariance and performs nearly as well as the best

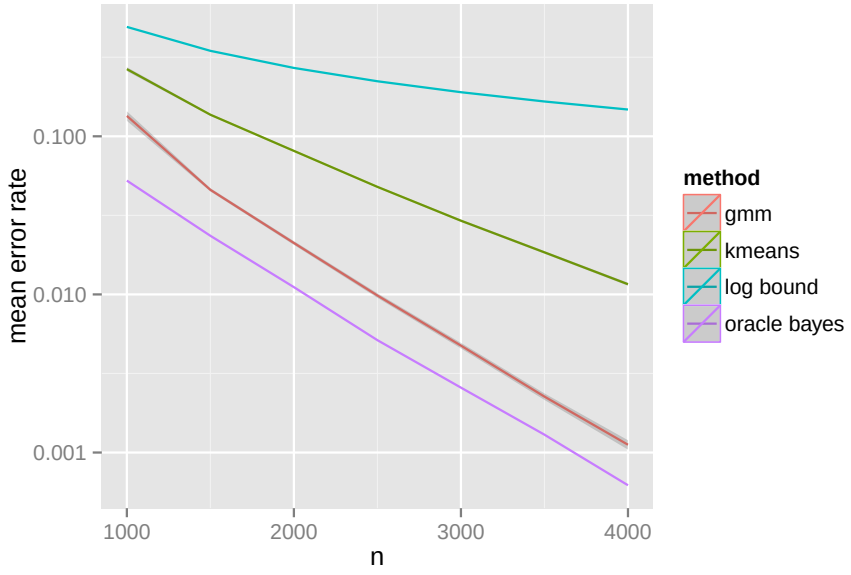


Figure 5.1: The plot shows various theoretical and emprical error rates for the stochastic blockmodel as specified in Eq. (5.9). For n ranging from 1000 to 4000 we simulate a graph from this model 200 times and each time cluster the graph using either k-means (green) or GMM (red). The cyan curve shows the decay rate if the error rate decayed as $\log(n)/n$ which were the rates of the best prior results using spectral methods. The magenta curve gives a theoretical best possible error rate based on the distributional result in Theorem 4.8.

possible error rate given the embedding. Note that though Theorem 5.10 ensures that asymptotically the k-means error rate will be zero it does not give the rate and and it also does not mean that this clustering is optimal. We do not investigate the theory comparing GMM to k-means but it appears clear that for an SBM, GMM will typically outperform standard k-means.

Chapter 6

Discussion

This work has argued the case that for a particular class of models for random graphs, the adjacency spectral embedding provides excellent estimates of underlying properties of the latent positions that allows for powerful results in regard to subsequent inference. In particular, in the RDPG model and the SBM model the value of this embedding procedure and the associated theory translates into the consistency of a broad swath of statistical procedures when used on the embedding.

Our results show how the spectral embedding of the adjacency matrix concentrates around the spectral embedding of the matrix of edge probabilities. We have shown that this holds for both IEG models and for graphs with conditionally independent edges. Furthermore, we showed that for a variety of inference tasks, the adjacency spectral embedding provides excellent performance. In this chapter we wish to discuss various extensions, drawbacks and alternatives for the adjacency spectral embedding.

First, we seek to provide an admittedly brief discussion of to what extent our program is the right one or even a desirable one and we disucss some alternatives. In particular, throughout we considered embeddings using the adjacency matrix and we discuss when this may perform well as compared to embedding of other matrices. In the next section we return to working with the adjacency spectral embedding but consider extensions beyond the setting of a simple undirected graph. These extensions include weighted and directed graphs as well as the case when there are multiple graphs. Finally, we will conclude and discuss some future directions for research.

6.1 Should I embed? If so what?

For the entirety of this work we have worked in the situation where we assume that, no matter what, we will first embed using the adjacency spectral embedding and then use whatever algorithm we desire on the embedded points. Our theory justifies this program in the sense that it provides performance guarantees in certain models but we are never able to claim that this program is optimal. There are many other paths for analysis, some of which involve embedding while others work more directly with the graph. In this section we seek to argue that embedding is at least a reasonable approach but that there are still choices to be made and that optimal strategies, especially those that balance computational time and performance, are still being sought.

Before arguing in favor of any particular embedding approach let us return to more solid statistical footing. For example, consider the case where we observe a graph known to be an ER graph but with unknown p which we seek to estimate. Naturally, one would use the observed density of the graph as an estimate. This is a minimal sufficient statistic, the maximum likelihood estimate and the uniformly minimum variance unbiased estimate. In this situation it may seem that an embedding approach would be inadvisable and as we will show it can be in this case.

Consider the case of embedding to one dimension which would be the natural dimension for an ER graph. In this setting, no function of the embedding will return the in many ways optimal estimate given by the graph density. The reason is that, at least for $n > 4$, the embedding is not a sufficient statistic. Figure 6.1 gives an example of two graphs, one with four edges and one with three edges. However, it is relatively straightforward to see that the largest eigenvalue and its corresponding eigenvectors are the same for the adjacency matrices of both graphs. If the embedding were sufficient then we would be able to recover the density from the embedding, by the minimality of this statistic. However, this example clearly demonstrates that we cannot.

This is quite a disconcerting result as even in the simplest case, the embedding is provably sub-optimal and provably loses information. The question of whether this inconsistency is due purely to the fact that the graph can be disconnected is still open and results in this vain may give us more hope that the embedding is

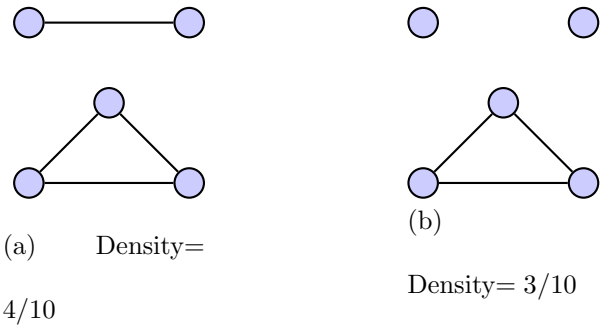


Figure 6.1: An example demonstrating the insufficiency of the embedding for ER graphs. The largest eigenvalue and its associated eigenvectors are the same for both graphs but the densities, a minimial sufficient statistic, are different.

a good strategy provided the graph is disconnected. Indeed, if we go beyond the ER case, disconnected graphs provided difficulties in general as making a sensible estimate about the probability of an edge betweeen the connected components would be difficult.

In general this discussion suggests that other techniques besides embedding may be preferable and indeed a plethora of non-embedding techniques exist. These techniques usually focus on either community detection but there has also been a recent expansion into techniques for non-parametric estimation of the parameters for a general exchangeable random graph. In Section 1.2.4 we discussed the wealth of community detections algorithms and do not expand on that here.

The area of non-paramateric estimation is relatively new and expanding rapidly. We view the work considered as more closely related to these non-parametric methods. Indeed, our method is an embedding variant of the method of singular value

thresholding which has been shown to be able to consistently in a non-parametric exchangeable graph model [Chatterjee, 2012]. Spectral methods, in the spirit of minimizing square error, can be contrasted to other statistically motivated methods. For example, maximum likelihood and variational methods have been shown to consistently estimate the parameters in a stochastic blockmodel [Bickel et al., 2013]a and furthermore it has been shown that blockmodels can be fit to nonparametric models by likelihood methods [Wolfe and Olhede, 2013]. Airoldi et al. [2013] uses similar ideas to show how graphons can be approximated by stochastic blockmodels. Another method, again inspired by classical statistical methods, is the idea of subgraph counts which function as analog to moments and also provide consistent estimates in some cases Bickel et al. [2011].

Overall, the methods above mostly have stronger theoretical guarantees than our spectral methods but nonetheless spectral methods have some advantages. Practically, they are accessible to any research with a linear algebraic package and require only a few lines of code to implement in languages such as R, Matlab, and Python. Furthermore, as we have noted already, spectral methods are conducive to multiple analyses on the same dataset. Whereas the methods above just fit a model, spectral methods allow for any number of techniques to be deployed on the embedding. If we suppose that our only goal is clustering, then it is likely that there are better methods if the goal is optimal performance or alternatively if the goal is speed, but to some extent spectral methods provide a middle ground since they can be computed with

$O(n^2 d)$ complexity. Needless, to say the question of when to embed is an open one but we believe that the theory in this work suggests that it could be useful in the network analysts toolbox.

Given that we are going to embed, many choices still exist. Up to now we have assumed that we will use the adjacency matrix directly however, far from advocating that the adjacency matrix should be the default choice when using spectral methods, we focus on this matrix primarily for mathematical convenience. In general, the adjacency matrix is not best the matrix to consider and in fact it is frequently inferior to other choices. In Section 1.2.4 we explained one of the main motivations for spectral method which used the second smallest eigenvector of the combinatorial Laplacian $D-A$ where D is the diagonal matrix of degrees. Another alternative is the normalized Laplacian given by $I - D^{-1/2}AD^{-1/2}$. Like the combinatorial Laplacian, the second smallest eigenvector of the normalized Laplacian gives a solution to a relaxation of the so-called *Ncut* problem. Luxburg [2007] argued that typically the normalized Laplacian is preferable to the combinatorial Laplacian for general clustering tasks.

Note that the normalized Laplacian has the same eigenvectors as what we call the normalized adjacency matrix $D^{-1/2}AD^{-1/2}$. For this matrix, Rohe et al. [2011] provided the first proof that any spectral clustering algorithm consistently estimates the blocks in a stochastic blockmodel using the normalized adjacency. Some of the foundations of our work are found in the Rohe et al. [2011] and though we do not investigate it here, some of the methods we use can likely be translated to improve

bounds for the normalized adjacency. We note that unlike the adjacency matrix, the eigenvector corresponding to the largest positive eigenvalue of the normalized Laplacian can be computed explicitly and is proportional to the vector with components given by the square root of the degrees of the vertices.

So, which is better, normalized or unnormalized? This question was investigated in the case of a two-block stochastic blockmodel in Sarkar and Bickel [2013]. They argue that for many values of the relevant parameters the normalized adjacency has lower within block variance in its eigenvectors when compared to the adjacency matrix. The picture they provide is far from complete but provides a window into why the normalized version is frequently preferable. This phenomenon is not universal and indeed Sussman et al. [2012] provided an example where that adjacency matrix has better empirical performance for specific model parameters (notably those considered in Examples 4.12 and 5.11). Indeed, in that same paper it was demonstrated that for a graph derived from Wikipedia, the two methods provide two qualitatively different partitions of the data both of which have advantages. We reconsider that example here.

Example 6.1 (Wikipedia Graph Sussman et al. [2012])**.** In this example we consider a graph derived from the Wikipedia online encyclopedia. This graph was collected by considering the article "Algebraic Geometry" and including every article which can be accessed by clicking two hyperlinks starting at "Algebraic Geometry". The graph has $n = 1382$ vertices and each document was labeled based on human inspection as

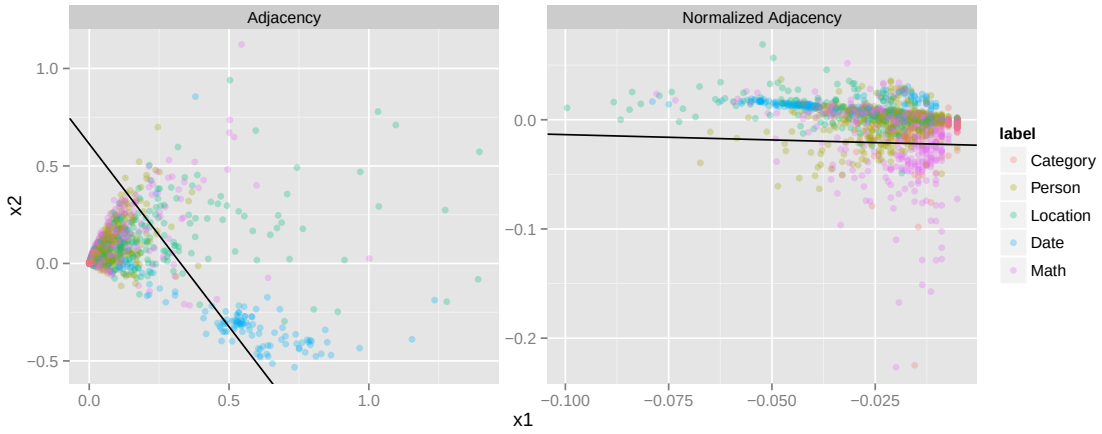
being in one of 5 classes: Category, Person, Location, Date, or Math.

A comparison of using the 2-dimensional spectral embedding of the adjacency matrix and the normalized adjacency matrix is shown in Figure 6.2. Panel (a) shows the two embedding colored according to the 5 classes above. Visually, it is clear that the two embeddings are capturing different aspects of the graph and the 2-means clustering, denoted by the corresponding linear boundary between the two classes, clusters the vertices very differently. Sussman et al. [2012] gives a more detailed comparison of these two clusters in how they compare to the five classes.

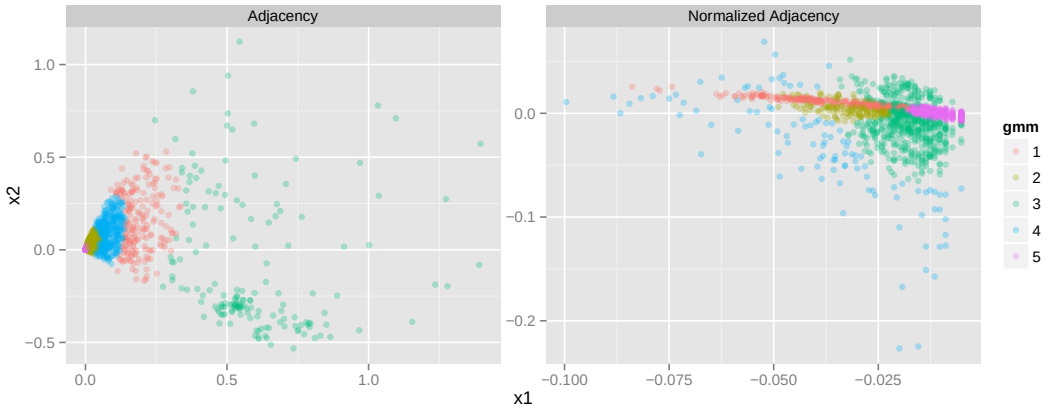
Panel (b) shows the two embeddings colored according to a fitted mixture of five normal distributions. Here we see that by a quick visual inspection, the clusters determined for the normalized adjacency seem to more closely match the manual class labels than those determined using the adjacency matrix. This example provides an affirmation of the idea that both embeddings can provide value and that they can capture very different aspects of the graph, but the normalized adjacency may perform better overall.

We see that in some situations there may be value in investigating both embeddings and that overall neither approach dominates the other. There are yet still other matrices that we may consider. For example, Qin and Rohe [2013] argue that a regularized version of the normalized adjacency matrix improves performance in the case that there are some vertices with very small degrees.

Another possible enhancement is what we call diagonal augmentation. The com-



(a) 2-means clustering



(b) Mixture of 5 normals

Figure 6.2: These figures show the two-dimensional spectral embeddings of the adjacency and the normalized adjacency matrices for the Wikipedia graph. Panel (a) is colored according to the five manually defined classes. The black line shows the division between the clusters as determined by using 2-means clustering. Panel (b) is colored according to a fit using a mixture of five multivariate normals in which the covariances are allowed to vary freely. Overall, we see that the two embeddings provide very different pictures of the graph.

binatorial Laplacian is an extreme form of this, equivalent to putting the negative degrees along the diagonal. Other proposed ideas are to take $A + D/(n-1)$ which can improve performance in some cases Marchette et al. [2011]. In the ER(n, p) distribu-

tion, the work of Füredi and Komlós [1981] suggests that taking $A + (1 - 2p)I$ serves to correct the asymptotic bias in the first eigenvalues. All of these augmentations can improve finite sample performance but have no impact on the asymptotic results provided the added matrix is bounded entrywise.

In practice, given the ability to consider only one spectral decomposition, the normalized adjacency is a reasonable default choice. However, given that the different matrices give different breakdowns of the data, a comparison of different methodologies will likely be illuminating for any given data set. As research on this issue continues, we are likely to see the introduction of more matrices for spectral embeddings as well as a deeper understanding of the differences between the matrices, which we hope will lead to a better answer to this question.

6.2 Beyond one simple undirected graph

Another simplifying assumption we have made throughout the text is that we observe one simple undirected graph. The undirected setting means that the use of spectral theory for symmetric matrices. The simple setting means that the edge have no weights and all have Bernoulli distributions which are bounded. Finally, we assume that we observe single graph, which is not always the case. Some of these simplifying assumptions can relaxed without effecting the results at all—others present more challenges. In this section, we will divide our discussion along the lines of extensions

within the single graph framework and extensions to the multi-graph framework.

Before continuing, we would be remiss if we failed to mention the most important simplifying assumption that we make: the fact that the random graphs have relatively simple distributions. Asymptotic or finite sample analysis of the spectral embedding for some models may within easy reach—for other models there is likely no hope. We do not discuss other models here but believe that investigation in this vein are an exciting future research direction.

6.2.1 Single Graph

There are many ways to generalize a simple undirected graph. For example, we could allow the edges to have real valued weights, considered here, or categorical or vector valued weights, considered in Section 6.2.2. Another extension allows the edges to be directed, which is frequently meaningful for applications. Finally, we consider that the graph may not be fully observed or that we suspect that our recording of whether an edge is present or not is faulty.

6.2.1.1 Weighted Graph

The case that each edge has a non-negative real valued weight is a quite natural extension and requires very few changes to the current theory. This situation could occur if we are counting the number of communications or relationships between objects or if we have some quantification of the strength of the relationships. In this

case, if the weights are real and positive we can then define the weight matrix W where W_{ij} is again zero if there is no edge between vertices i and j and otherwise it is equal to the given edge weight.

In order to apply spectral methods that take edge weights into account we can simply perform the spectral embedding on the weight matrix W. We can again consider latent position models or random dot product models for the weight matrix. For example, if we suppose $\mathbb{E}[W_{ij}|X_i, X_j] = \langle X_i, X_j \rangle$ this is a random dot product weighted graph. If we suppose additionally that $W_{ij} \in [0, 1]$ with probability one than all the theory from Chapters 3 and 4 transfers easily to this case (the limiting covariance for our central limit theorem will change slightly). In general if the weights are bounded then we can apply the same theory by normalizing.

On the other hand if the weights are unbounded, more work is needed. In this case, we cannot use results such as Hoeffding's inequality or the concentration inequalities of Tropp [2012]. Provided the tail of the weight distribution decays rapidly enough, one possible avenue is to simply truncated versions of the distributions. Another possibility is to use concentration inequalities for unbounded random variables. We suspect that for sufficiently well behave distributions the bounds proved will not decay greatly for unbounded weights.

Finally, we note that even if weights are present, it is not necessarily optimal to use them. For example in the Wikipedia graph, we actually have access to the number of hyperlinks between each pair of vertices but we do not use this since

visual inspection suggests that the embedding is more reasonable using a simple graph. Thresholding or performing some other transform of the weights will frequently improve the performance of the embedding, especially if there are large outliers in the weights.

6.2.1.2 Directed Graph

Directed graphs frequently better represent real data then undirected graphs. For example, in communication networks there is frequently a sender and a receiver and hence the edge should follow the direction of the communication. In neuroscience, a synapse is between an axon and a dendrite. In social networks, friendship is (hopefully) undirected but other relationships may be directed, such as the author referee-relationship for peer reviewed publications.

Unlike the weighted case, where the linear algebraic tools do not change, a directed graph can have an asymmetric adjacency matrix and so we must change our tools to accommodate this. One method is to use the singular value decomposition of the adjacency matrix, in which case we would have two embeddings, corresponding to the left and right singular vectors. From a communication graph, the two embeddings would represent the distinct sending and retrieving characteristics of each vertex.

Sussman et al. [2012] used some of the same methods we have used to approach the directed problem in the stochastic blockmodel case and showed that consistent clustering can be achieved. The key idea is to analyze $A^\top A$ or $AA^\top$ which are

symmetric and whose eigenvectors correspond to the right and left singular values of A, respectively. Rohe and Yu [2012] used similar ideas to analyze the normalized adjacency matrix.

Beyond the stochastic blockmodel, the theory for spectral embeddings in directed models have not been investigated greatly. Again, the matrix concentration inequalities in Tropp [2012] do not all extend to the non-symmetric case. Similarly, results like the Davis-Kahan theorem cannot be used directly on A but can be used on matrices such as $AA^\top$. The difficulties with the directed case are more challenging than with the weighted case but it seems reasonable to suspect that the spirit of many of the results for the undirected case would carry over. Finally, like the case of a weighted graph, there are also instances where it is better to work with a symmetrized version of the directed graph.

6.2.1.3 Faulty Observations

There are many ways that an observed graph could be faulty. In terms of social networks, we may find that two individuals are friends when really they don't know each other or visa versa. Alternatively, we may fail to observe whether there is an edge or not in a graph, a form of missing data. These types or deviations from the ideal scenario where the entire graph is observed perfectly are one of the most challenging extensions.

There have been a series of investigations into these scenarios that frequently

highlight the challenges associated with missing data in this setting [Priebe et al., 2012, 2013, Balachandran et al., 2013]. If the entire graph is observed but some of the edges are observed inaccurately, then this induces a new random graph distribution.

Depending on the mechanism for the inaccurate observations, the resulting distribution may share many properties with the original distribution. For example, if the true distribution is an independent edge graph but in the observed graph, each edge is removed with a fixed probability then this results in a scaling of the matrix of edge probabilities P. Hence, our bounds would still apply but our estimates would be further biased. Somewhat more complicated error models can change the rank of the P matrix which could lead to a misspecification of the embedding dimension [Chen et al., 2013]. Clearly, the mechanism for the errors can be arbitrarily nefarious and results in full generality are impossible.

Even more challenging than inaccurate observation for the application of spectral methods are edge observations that are missing completely. Methods to approximate the eigenvalues and eigenvectors of a matrix with missing entries do exist but this is also an active research area.

We will also take this opportunity to note that missing data makes the distinction between the independent edge case and the conditionally independent edge case more important. Indeed, if a latent position graph is observed without errors, then we can ask if from a statistical inference perspective there is a difference between thinking of the latent positions as random or fixed. If we make no assumptions about the

distribution of the latent positions, like a particular parametric form for the distribution, then we would argue that there is no difference based on the conditionality principle. If we make some assumption, then the randomness of the latent positions can be taken into account in terms of some prior information.

But if the whole graph is not observed, then the question becomes very important. Indeed, if the edges are assumed independent then there observing some edges does not inform us about the presence or absence of the unobserved edges. If the edges are not independent, then depending on the setup we may be able to improve our predictions about the unobserved edges based on the observed edges. These differences highlight the extensive challenges in the case that the graph has been inaccurately or only partly observed.

6.2.2 Multiple Graphs

Many problems present us with not just one graph but with many graphs or a time series of graphs. Our methods do not explicitly deal with these situations but can be applied, at least naively. As an example for a time series of graphs, it is typical to observe a sequence of communication events where each event represents a communication between two vertices at a specified time. Combining all the communication events, we can represent this as a single graph but we may want to combine the events into distinct time periods in which case this can be represented by a sequence of graphs. Another way multiple graphs could arise is by considering each edge to

have categorical attributes such as topics or colors associated with it. If there are K categories then the attributed-edge graph can be broken up into K simple graphs Fishkind et al. [2013].

If we have multiple graphs our methodology does not exploit any relationship between them but can still be used to embed each graph separately. For example, if we have a sequence of graphs $A_1, \ldots, A_T$, all on the the same vertex set, then we could embed each graph separately to get a sequence of point clouds $\hat{X}_1, \ldots, \hat{X}_T$. The embeddings can be aligned in various ways including Procrustes analysis, regular or generalized canonical correlation analysis, and other methods [Izenman, 2008]. Once the embeddings are aligned, we can use methods for multivariate time series if the graphs are aligned in time or use general multivariate methods if the graphs have no particular order.

Like in the single graph setting, there are many applications for this sort of analysis. For example, if each graph is associated with a class label, we can use the embeddings as the first step in building a classifier. Alternatively, we could use the embeddings to test whether two random dot product graphs have the same set of latent positions. We note that it that a joint embedding of the graphs could improve subsequent inference for many tasks by borrowing strength between similar graphs. For example, in an anomaly detection task for a time series of graphs, a joint embedding could function to smooth the time series of embeddings so that noise could be better separated from the true anomaly. Our methods do not easily extend to the

joint embedding setting.

The situations described above assumed that the graphs were all on the same vertex set and that the correspondence between the vertices in different graphs is known. Another possibility is that this correspondence is unknown, for example if the graphs represent users of two different social networks and the user names are not matched. This setting again introduces more difficulties as it becomes more difficult to line up the the different embeddings. In the setting where the correspondence is known for a small subset of vertices, Lyzinski et al. [2013] used multiple embeddings to simultaneously cluster two graphs in order to proceed with a divide and conquer strategy for finding a correspondence between the vertices in two large graphs.

Extending our theory to the multiple graph setting will again depend on the particular joint distribution of the graphs. If the graphs are sufficiently independent our theory can be easily translated while if there are dependencies among the graphs then it is like that different results would arise that take into account the dependencies among the edges.

Bibliography

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

Edoardo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *arXiv preprint arXiv:1311.1731*, 2013.

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Avanti Athreya, Vince Lyzinski, David J Marchette, Carey E Priebe, Daniel L Sussman, and Minh Tang. A limit theorem for scaled eigenvectors of random dot product graphs. *arXiv preprint arXiv:1305.7388*, 2013.

James P. Bagrow and Erik M. Bollt. Local method for detecting communities. *Phys. Rev. E*, 72:046108, Oct 2005. doi: 10.1103/PhysRevE.72.046108.

BIBLIOGRAPHY

Prakash Balachandran, Edoardo Airoldi, and Eric Kolaczyk. Inference of network summary statistics through network denoising. *arXiv preprint arXiv:1310.0423*, 2013.

Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997. ISBN 0-387-94846-5. doi: 10.1007/978-1-4612-0653-8.

P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50):21068–73, 2009.

P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Annals of Statistics*, 39(5):38–59, 2011.

Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4):1922–1943, 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1124.

Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics*. Holden-Day Inc., San Francisco, Calif., 1976. Basic ideas and selected topics, Holden-Day Series in Probability and Statistics.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and John A. Wellner. *Efficient*

and adaptive estimation for semiparametric models. Springer-Verlag, New York, 1998. ISBN 0-387-98473-9. Reprint of the 1993 original.

Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294, 2007.

Béla Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001. ISBN 0-521-80920-7; 0-521-79722-5. doi: 10.1017/CBO9780511814068.

Alain Celisse, J-J Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *arXiv preprint arXiv:1105.3288*, 2011.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *arXiv preprint arXiv:1212.1247*, 2012.

Li Chen, Joshua Vogelstein, and Carey Priebe. Robust vertex classification. *arXiv preprint arXiv:1311.5954*, 2013.

F. R. K. Chung. *Spectral Graph Teory*, volume 92. American Mathematical Society, 1997.

C. Davis and W. Kahan. The rotation of eigenvectors by a pertubation. III. *Siam Journal on Numerical Analysis*, 7:1–46, 1970.

BIBLIOGRAPHY

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer Verlag, 1996.

Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, 28(1):33–61, 2008. ISSN 1120-7183.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. Wiley-Interscience, New York, second edition, 2001. ISBN 0-471-05669-3.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

P. Erdős and A. Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist.*, 38:343–347, 1961.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23 (2):298–305, 1973.

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.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010. ISSN 03701573. doi: 10.1016/j.physrep.2009.11.002.

BIBLIOGRAPHY

Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30:1141–1144, 1959. ISSN 0003-4851.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99(12):7821–7826 (electronic), 2002. ISSN 1091-6490. doi: 10.1073/pnas.122653799.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

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

D. N. Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 1979.

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2009. ISBN 978-0-470-12990-6. doi: 10.1002/9780470434697.

Alan Julian Izenman. *Modern multivariate statistical techniques*. Springer Texts

in Statistics. Springer, New York, 2008. ISBN 978-0-387-78188-4. doi: 10.1007/978-0-387-78189-1. Regression, classification, and manifold learning.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E (3)*, 83(1):016107, 10, 2011. ISSN 1539-3755. doi: 10.1103/PhysRevE.83.016107.

Samuel Kutin. Extensions to mcdiarmid's inequality when the differences are bounded with high probability. *Technical Report, University of Chicago Department of Computer Science*, pages 1–24, 2002.

E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98502-6.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

Vince Lyzinski, Daniel Sussman, Minh Tang, Avanti Athreya, and Carey Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *arXiv preprint arXiv:1310.0532*, 2013.

David Marchette, Carey Priebe, and Glen Coppersmith. Vertex nomination via attributed random dot product graphs. In *Proceedings of the 57th ISI World Statistics Congress*, 2011.

BIBLIOGRAPHY

Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

C. L. M. Nickel. *Random dot product graphs: A model for social networks*. PhD thesis, Johns Hopkins University, 2006.

R. I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *Arxiv preprint ArXiv:0911.0600*, 2009.

Elżbieta Pekalska and Robert PW Duin. *The dissimilarity representation for pattern recognition: foundations and applications*. Number 64. World Scientific, 2005.

Iosif Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. In *Probability in Banach spaces, 8 (Brunswick, ME, 1991)*, volume 30 of *Progr. Probab.*, pages 128–134. Birkhäuser Boston, Boston, MA, 1992.

David Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140, 1981. ISSN 0090-5364.

Carey E Priebe, Daniel L Sussman, Minh Tang, and Joshua T Vogelstein. Statistical inference on errorfully observed graphs. *arXiv preprint arXiv:1211.3601*, 2012.

Carey E Priebe, Joshua Vogelstein, and Davi Bock. Optimizing the quantity/quality trade-off in connectome inference. *Communications in Statistics-Theory and Methods*, 42, 2013.

BIBLIOGRAPHY

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. *arXiv preprint arXiv:1309.4111*, 2013.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39(4):1878–1915, 2011.

Karl Rohe and Bin Yu. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *arXiv preprint arXiv:1204.2296*, 2012.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *J. Mach. Learn. Res.*, 11:905–934, 2010. ISSN 1532-4435.

Purnamrita Sarkar and Peter J Bickel. Role of normalization in spectral clustering for stochastic blockmodels. *arXiv preprint arXiv:1310.1495*, 2013.

T. A. B. Snijders and K. Nowicki. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1): 75–100, 1997.

C. J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5(4):595–620, 1977.

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

BIBLIOGRAPHY

D. L. Sussman, M. Tang, and C. E. Priebe. Universally consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (Accepted)*, 2013.

M. Tang, D. L. Sussman, and C. E. Priebe. Universally consistent vertex classification for latent positions graphs. *Annals of Statistics (Accepted)*, 2013.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012. ISSN 1615-3375. doi: 10.1007/s10208-011-9099-z.

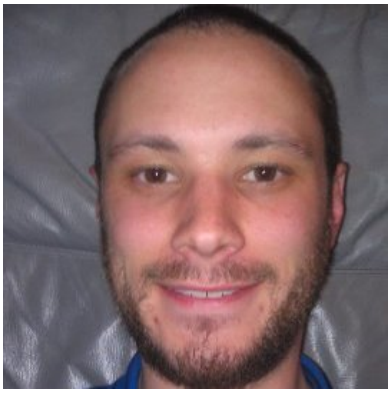
GV Trunk. A problem of dimensionality: A simple example. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (3):306–307, 1979.

Douglas B. West. *Introduction to graph theory*. Prentice Hall Inc., Upper Saddle River, NJ, 1996. ISBN 0-13-227828-6.

Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

S. Young and E. Scheinerman. Random dot product graph models for social networks. *Algorithms and models for the web-graph*, pages 138–149, 2007.

Vita



Daniel L. Sussman received the B. A. degree *magna cum laude* in Mathematics from Cornell University and enrolled in the Applied Math and Statistics Ph.D. program at Johns Hopkins University in 2010. Daniel has been awarded the Charles and Catherine Counselman Endowed Fellowship, the Newman Family Fellowship, and the Whiting School of Engineering Centennial Fellowship in support of his pursuit of a doctoral degree. He has also received the Acheson J. Duncan Fund for the Advancement of Research in Statistics Travel Award. Daniel was recognized as a National Science Foundation Graduate Research Fellowship Program Honorable Mention (2012).

Next, Daniel will continue to work with Dr. Carey Priebe before working in a post-doctoral position with Dr. Edo Airoldi at the Harvard University Department of Statistics.