UNIVERSITÉ CÔTE D'AZUR

Inria INVENTEURS DU MONDE NUMÉRIQUE

**École Doctorale STIC**
**Sciences et Technologies de L'information et de la Communication**
**Unité de recherche: INRIA (équipe NEO)**

# Thèse de Doctorat

Présenté en vue de l'obtention du
grade de Docteur en Sciences

de
l'UNIVERSITE COTE D'AZUR
Mention : INFORMATIQUE

par

Arun KADAVANKANDY

# Spectral analysis of random graphs with application to clustering and sampling

Dirigée par : Konstantin AVRACHENKOV

Laura COTATTELLUCCI

**à soutenir le July 18, 2017**
Devant le Jury Composé de:

| | | | |
|---|---|---|---|
| Konstantin AVRACHENKOV | - | Inria, France | *Directeur* |
| Laura COTTATELLUCCI | - | Eurecom, France | *Directeur* |
| Bruce HAJEK | - | Univeristy of Illinois at Urbana-Champaign, USA | *Rapporteur* |
| Marc LELARGE | - | Inria-ENS, France | *Rapporteur* |
| Pawel PRALAT | - | Ryerson University,Toronto | *Examinateur* |
| Alain JEAN-MARIE | - | Inria, France | *Président* |

# Acknowledgments

I thank my advisors Konstantin Avrachenkov and Laura Cottatellucci for trusting me with this thesis topic and for all the guidance, constant encouragement, the ideas and the many opportunities for collaborations they gave me.

I am extremely grateful to the members of team Neo - Dr. Alain Jean-Marie, Dr. Giovanni Neglia and Dr. Sara Alouf - for their support and the many group discussions, which helped me to deepen my knowlege on various topics related to my thesis and also helped me get acclimatised to life in a new environment. I thank Laurie Vermeersch, our team assistant, who contributed greatly to the smooth running of my PhD.

I thank Andrei Raigorodskii for hosting me at the Moscow Institute of Physics and Technology and his help during my stay and also Liudmila Prokhorenkova for collaborating with me.

I convey my gratitude to Prof. Bruce Hajek and Dr. Marc Lelarge for their thorough review of my thesis and their valuable comments and suggestions, which have helped me enormously to bring my thesis to its current form.

I am thankful for my friend Christophe for making my life in France pleasant and memorable. I also thank my friends from Inria and elsewhere for the company and support they gave me.

Above all I thank my parents and the rest of my family for their support, patience and understanding.

## Spectral analysis of random graphs with application to clustering and sampling


### Abstract:

In this thesis, we study random graphs using tools from Random Matrix Theory and probability to tackle key problems in complex networks and Big Data. First we study graph anomaly detection. Consider an Erdős-Rényi (ER) graph with edge probability q and size n containing a planted subgraph of size $m$ and probability $p$. We derive a statistical test based on the eigenvalue and eigenvector properties of a suitably defined matrix to detect the planted subgraph. We analyze the distribution of the derived test statistic using Random Matrix Theoretic techniques. Next, we consider subgraph recovery in this model in the presence of side-information. We analyse the effect of side-information on the detectability threshold of Belief Propagation (BP) applied to the above problem. We show that BP correctly recovers the subgraph even with noisy side-information for any positive value of an effective SNR parameter. This is in contrast to BP without side-information which requires the SNR to be above a certain threshold. Finally, we study the asymptotic behaviour of PageRank on a class of undirected random graphs called fast expanders, using Random Matrix Theoretic techniques. We show that PageRank can be approximated for large graph sizes as a convex combination of the normalized degree vector and the personalization vector of the PageRank, when the personalization vector is sufficiently delocalized. Subsequently, we characterize asymptotic PageRank on Stochastic Block Model (SBM) graphs, and show that it contains a correction term that is a function of the community structure.



**Keywords:** Random Matrix Analysis, Spectral Graph Theory, Random Graphs, Sampling, Community Detection

# Contents

# List of Symbols

$\Delta$ The absolute ratio between the dominant eigenvalue and the edge of the spectrum.

$\boldsymbol{\mathcal{A}}$ Shifted adjacency matrix in Chapter 4.

$\mathbf{A}^T, \mathbf{A}^H$ represent the transpose and conjugate transpose of $\mathbf{A}$ respectively.

$\mathbf{P}$ Column Stochastic Markov Matrix.

$\overline{\boldsymbol{\mathcal{A}}}$ Mean shifted adjacency matrix in Chapter 4.

$\mathbf{A}$ Adjacency Matrix.

$\mathcal{B}(p)$ Bernouli random variable with success probability $p$.

$\chi(C)$ Indicator function for condition $C$.

$\mathcal{L}(X)$ Denotes the law or distribution of the random variable $X$.

$\bar{\mathbf{u}}$ Dominant eigenvector of the mean shifted adjacency matrix $\overline{\boldsymbol{\mathcal{A}}}$.

$\mathbf{u}$ Dominant eigenvector of the shifted adjacency matrix $\boldsymbol{\mathcal{A}}$.

$\sim$ For random variables: has the distribution, for nodes of a graph: are connected.

$\mathcal{N}(\mu, \sigma^2)$ Normal random variable with mean $\mu$ and variance $\sigma^2$.

$p_b$ Edge probability of background graph.

$p_s$ Edge probability of the embedded subgraph.

# List of Abbreviations

a.a.s asymptotically almost surely.

a.s. almost surely.

API Application Programming Interface.

BP Belief Propagation.

CDF Complimentary Cumulative Density Function.

CDF Cumulative Density Function.

CLT Central Limit Theorem.

ER Erdős-Rényi.

i.i.d. independent and identically distributed.

ID Identity.

MAP Maximum A Posteriori.

MH-MCMC Metropolis-Hastings Markov Chain Monte Carlo.

ML Maximum Likelihood.

OSN Online Social Network.

RDPG Random Dot Product Graph.

RDS Respondent-Driven Sampling.

RL Re-inforcement Learning.

RT Ratio with Tours.

rv random variable.

RW Random Walk.

s.t. such that.

SBM Stochastic Block Model.

t.p.m Transition Probability Matrix.

w.r.t. with respect to.

whp with high probability.

wlog without loss of generality.

Chapter 1

# Introduction and Thesis Organisation

The spread of internet and the ubiquity of mass accessible computational power have led, in recent years, to an explosion of data, often branded Big Data, which test the limits of traditional data processing methods. At the same time, the unprecedented growth of social networks like Facebook™ and other online communities like NetFlix™ has given rise to network sizes orders of magnitude larger than before. Such networks with several key defining characteristics are called complex networks [Newman 2003].

Graphs provide a parsimonious representation of interacting heterogenous entities, and hence are versatile and flexible as a tool for developing data processing algorithms. The advent of complex networks and Big Data has therefore renewed and galvanized an interest in graph processing and learning algorithms in disciplines ranging from Signal Processing [Shuman *et al.* 2013], Computational Biology [Kitano 2002, Hou *et al.* 2016] to Theoretical Physics and Information Theory [Mezard & Montanari 2009]. Graph based data processing has been highly successful and many important problems in machine learning can be formulated and solved efficiently in this framework, for e.g. [Koutra *et al.* 2011]. The analysis of graph algorithms is therefore of great importance.

However, complex networks owing to their large sizes and heterogeniety can often be extremely difficult to study. A remedy to this problem is to model networks using random graphs that capture key network properties of interest. Random graphs are probabilistic models where links are added between nodes according to some probabilistic rule [Bollobás 1998]. Random graph theory was set in motion by the work of Erdős and Rényi, who found out that limiting properties of graphs can be studied by analysing a suitably constructed random graph model [Erdős & Wilson 1977, Erdős & Rényi 1959]. In the following years, several random graph models have been proposed to model important defining characteristics of complex networks such as *clustering*, *small-world property*, *power law degree distributions* and the presence of tighly linked groups of nodes, called *communities* [Newman 2003, Hofstad 2016].

In this thesis we focus especially on the problem of hidden community detection. Community structure has important implications and significance in different domains. For example, in graphs made from datasets of genes or stocks, communities represent correlated datapoints [Firouzi *et al.* 2013], whereas in online communities such as NetFlix or Amazon, communities correspond to users with similar interests in movies, or similar buying habits. Hence, community detection in complex networks has righly garnered significant research attention in recent years [Fortunato 2010, Newman 2006]. However, the heterogeneity of real-world networks and the absence of a universal definition of a community make the design and analysis of community detection algorithms difficult. Random graphs with community structure present a tractable means to compare the performance and detection limits of various community detection algorithms that have been proposed in the literature e.g. [Rohe *et al.* 2011].

A effect graph analysis technique is by way of their matrix representations. Many graph

algorithms can be rephrased in terms of matrices and operations on matrices [Kepner & Gilbert 2011]. The theory of eigenvalues and eigenvectors of these matrices and their relationship to key graph properties, known as **Spectral Graph Theory**, has been a subject of deep research [Chung 1997, Spielman 2007]. In the analysis of random graphs, the matrices encountered are random, and thus the asymptotic spectral theory of random matrices, known as **Random Matrix Theory**, is central to the study of random graphs.

In this thesis, we use techniques from Random Matrix Theory and Random Graph Theory to tackle key problems in complex networks and machine learning. We consider *anomaly detection* and *hidden community detection* on random graphs, both important problems in machine learning on graphs. The anomaly detection algorithm, which we describe in detail in Chapter 4, is *unsupervised* and global and is based on interesting spectral properties of a *shifted adjacency matrix* of the graph considered. To solve the problem of hidden community detection, we propose and analyze a message passing algorithm based on Belief Propagation (BP) that uses prior information about the target community, and is *semi-supervised*. Furthermor, we analyse the behaviour of PageRank, an important algorithm for local community detection [Andersen & Chung 2007] as well as web search and link prediction [Gleich 2015], on a class of large random graphs using Random Matrix Theory. Finally, we propose new local algorithms based on random walks for the problem of estimating the average of an arbitrary function defined on the nodes of a graph.

In this chapter, we describe different matrix representations of graphs and review pertinent results from Spectral Graph Theory. Furthermore, we briefly describe some relevant random graph models. In the following section, we discuss in detail the problem of hidden community detection and the motivation behind studying it. Later, we provide a brief description of the well-known PageRank algorithm, widely used for web ranking as well as for solving important graph problems such as community detection and link prediction. We conclude this chapter by describing in detail the major contributions and the structure of this thesis.

## 1.1 Graph Matrices and Spectral Graph Theory

The study of matrix representations of graphs has a long history [Mohar & Woess 1989, Lovász & Pelikán 1973, Cvetković *et al.* 1980]. Matrices provide a parsimonious representation for graphs, but at the same time the algebraic properties of these matrices can be related to important graph properties. This is the subject of study in Spectral Graph Theory [Chung 1997, Spielman 2007]. Spectral analysis of graphs is a mature field with many applications in varied domains such as Markov chain analysis, Cryptography, and also Quantum Mechanics and other areas of theoretical physics. In the following section, we provide an overview of the role of matrices in the study of graphs.

### 1.1.1 Matrix Graph Representations

Consider a graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of vertices and $E \subset V \times V$ is the set of edges. A simple matrix representation of this graph is in the form of the adjacencies of the nodes. Let us denote the adjacency matrix by $\mathbf{A}$. For a graph with $n$ nodes, $\mathbf{A} \in \mathbb{R}^{n \times n}$ has rows and columns corresponding to the nodes and for any two nodes $i, j \in V$,

$$A_{ij} = \begin{cases} 1 & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases}$$

Here $i \sim j$ denotes the relation that there is an edge between $i$ and $j$. If the graph is directed, i.e., the edges have a source and a destination, then, in general, $\mathbf{A} \neq \mathbf{A}^T$, i.e., $\mathbf{A}$ is asymmetric. For undirected graphs, $\mathbf{A}$ is symmetric, and in our work, we limit ourselves to undirected graphs.

The degree $d_i$ of vertex $i$ is the cardinality of the set $\{j : j \sim i\}$ :

$$d_i = \sum_j A_{ji}.$$

We denote by $\mathbf{D} \in \mathbb{R}^{n \times n}$, the diagonal matrix such that $D_{ii} = d_i$.

A matrix related to the adjacency matrix is the modularity matrix denoted by $\mathbf{B}$ [Newman 2006], given as

$$\mathbf{B} = \mathbf{A} - \frac{\mathbf{d}\mathbf{d}^T}{\mathbf{d}^T\mathbf{1}}, \tag{1.1}$$

where $\mathbf{d} = [d_1, d_2, \ldots, d_n]$.
The modularity matrix has been used to assess the goodness of a community paritioning [Newman 2006, Fortunato & Barthélemy 2007]. Consider the problem of partitioning a graph with degrees $d_i$ into two communities. Let $s = \{s_1, s_2, \ldots s_n\}$ denote a partitioning of the graph such that $s_i = 1$ if node $i$ is mapped to community 1 and $s_i = -1$ otherwise. Then the modularity $Q$ is defined as [Newman 2006]

$$Q := \sum_{ij} s_i B_{ij} s_j = \mathbf{s}^T \mathbf{B} \mathbf{s}. \tag{1.2}$$

A good community assignment $\mathbf{s}$ is then proposed as the one that maximizes $Q$. However, this problem is NP-hard, but a convex relaxation of the problem can be solved exactly and the solution is related to the principal eigenvectors of $\mathbf{B}$ as shown in [Newman 2013].
The combinatorial Laplacian $\mathbf{L}$ is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A},$$

i.e.,

$$L_{ij} = \begin{cases} -A_{ij} & \text{if } i \neq j \\ D_{ii} & \text{if } i = j \end{cases}$$

It can be verified that $\mathbf{L}$ is positive semidefinite. It is the generator matrix for a continuous time Markov chain defined on the graph vertices [Brémaud 2013].

A related matrix is the normalized Laplacian $\boldsymbol{\mathcal{L}}$, which is defined as [Chung 1997]

$$\boldsymbol{\mathcal{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} \tag{1.3}$$

$$= \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{1.4}$$

Markov matrix $\mathbf{P}$ is key in the analysis of random walks on graphs. For an undirected graph $G$, $\mathbf{P}$ is defined as a *column-stochastic matrix* [1] such that

$$P_{ij} = \begin{cases} \frac{1}{d_j} & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

A simple Random Walk (RW) process on a graph is a discrete time process that starts by choosing an initial vertex from $V$ under some distribution at time $t = 0$. At $t = 1$, the process jumps to one of the neighbours of this initial vertex chosen uniformly at random.

[1] The sum across each column is 1

At $t = 2$, the process jumps to a random neighbour of this new vertex, and so on. The transition probability from node $i$ to $j$ is therefore given by $P_{ji}$.

In the next subsection, we review some important results from Spectral Graph Theory that are closely related to the topics studied in this thesis.

### 1.1.2 Spectral Graph Theory

Spectral Graph Theory is the study of the spectra, i.e., the eigenvalues and eigenvectors, of graph matrices and their relationship to important graph properties. Graph properties such as connectivity, bipartitedness, graph diameter, and the evolution of various random processes defined on the graph are closely related to the eigenvalues of a suitable graph matrix [Lovász 1993, Chung 1997, Aldous & Fill 2002, Spielman 2007]. In this section, we provide a brief review of some of key results.

For a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ the eigenvalues $\lambda_i(\mathbf{A})$ are defined as numbers such that there exist vectors $\mathbf{v}_i \in \mathbb{R}^{n\times 1}, \mathbf{v}_i \neq 0$ such that

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i.$$

The pair $(\lambda_i, \mathbf{v}_i)$ is known as the eigenvalue-eigenvector pair [Bhatia 2013]. In general the numbers $\lambda_i$ can be complex, but for symmetric matrices, the eigenvalues are always real and can therefore be ordered [Bhatia 2013].

Let us consider the normalized Laplacian $\boldsymbol{\mathcal{L}}$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of $\boldsymbol{\mathcal{L}}$. The eigenvalue properties of $\boldsymbol{\mathcal{L}}$ have been well studied. It can be shown that $\lambda_n = 0$ and that $0 = \lambda_n \leq \lambda_1 \leq 2$ [Chung 1997]. The second smallest eigenvalue $\lambda_{n-1}$ contains important information about the connectivity of the graph. If the graph is connected, this eigenvalue is strictly positive [Chung 1997]. In addition, the multiplicity of zero eigenvalue is equal to the number of connected components of the graph [Chung 1997, Spielman 2007].

Furthermore, the magnitude of $\lambda_{n-1}$, sometimes referred to as the *spectral gap* is a key property of the graph, related to the dynamics of many processes on the graph such as random walks [Levin *et al.* 2009] and average consensus algorithms [Olshevsky & Tsitsiklis 2009]. It can also related to some intrinsic graph properties. For e.g. the diameter $D$ of the graph, defined as the shortest distance between any two vertices of the graph, maximized over all pairs, is related to $\lambda_{n-1}$ by the following lemma from [Chung 1997]. We have

$$D = \max_{x,y} d(x, y),$$

where $d(x, y)$ is the length of the shortest path between two vertices $x$ and $y$.

**Lemma 1.1.** *[Chung 1997, Lemma 1.9] For a connected graph $G$ with diameter $D$, we have*

$$D \geq \frac{1}{\lambda_{n-1}\mathrm{vol}(G)},$$

*where* $\mathrm{vol}(G) := \sum_{i\in V} d_i$.

Intuitively, the above lemma states that the less connected the graph, i.e., the smaller the spectral gap, the larger is the graph diameter.

The eigenvalues of graph matrices also play a crucial role in the time to *stationarity* of a simple RW defined on the graph.

An interesting property of the simple RW on graphs is that when the graph is connected and non-bipartite, the distribution of the RW after $t$ steps, given by $\mathbf{P}^t\mu$, where $\mu$ is the initial distribution, gets closer and closer to a fixed distribution $\pi$ as $t$ increases [Levin

*et al.* 2009]. The unique distribution $\pi$, known as the stationary distribution, satisfies [Levin *et al.* 2009]

$$\pi = \mathbf{P}\pi.$$

On an undirected random graph, $\pi$ is given as

$$\pi(x) = \frac{d(x)}{\text{vol}(G)}$$

for a vertex $x$. This property, also known as *mixing*, is important in many applications to obtain samples from a desired distribution, or to find averages with respect to the stationary distribution. See [Brémaud 2013, Levin *et al.* 2009] and references therein.

In applications, the time it takes for a RW to reach stationarity, called the *mixing time* $t_{\text{mix}}$ of the RW, is crucial. It is defined in terms of the *total variation* distance between the $t$-step distribution and the stationary distribution as below [Levin *et al.* 2009]. Let us define the distance $d(t)$ as below

$$d(t) = \sup_f \|\mathbf{P}^t f - \pi\|_{\text{TV}},$$

where the supremum is taken over all distributions $f$ on $V$ and

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2}\sum_{i \in V} |\mu_i - \nu_i|,$$

for any two distributions $\mu, \nu$ is the the total variation distance. Then $t_{\text{mix}}$ is defined as [Levin *et al.* 2009]

$$t_{\text{mix}}(\varepsilon) := \min\{t : d(t) \leq \varepsilon\}.$$

Oftentimes $t_{\text{mix}}$ is taken to be $t_{\text{mix}}(1/4)$. It can be related to the eigenvalues of $\mathbf{P}$ as follows.

Let $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_n$ be the eigenvalues of $\mathbf{P}$. Denote by $\beta_*$ the second largest eigenvalue of $\mathbf{P}$ in absolute value. The largest eigenvalue $\beta_1$ of $\mathbf{P}$ is 1, since it is a stochastic matrix. Then the *absolute spectral gap* of the RW is defined as [Levin *et al.* 2009]

$$\gamma_* = 1 - \beta_*.$$

We then have the following important result that bounds the mixing time in terms of the $\gamma_*$ from [Levin *et al.* 2009].

**Theorem 1.1.** *[Levin et al. 2009, Theorem 12.3] Let $P$ be the transition matrix of a simple RW on a graph $G = (V, E)$ with degrees $d_i, \forall i \in V$. Then*

$$t_{mix}(\varepsilon) \leq \log\left(\frac{\sum_i d_i}{\varepsilon \min_i d_i}\right)\frac{1}{\gamma_*}.$$

The above theorem says that the smaller the eigenvalue $\beta_*$, the faster the chain mixes.

To conclude, we look at another important graph property, the conductance, and the associated Cheeger inequality. The conductance of a graph is related to the concept of a cut and is a key metric to study the community structure of a graph [Andersen & Chung 2007]. An (edge) cut is a set of edges whose removal separates the graph into two parts [Chung 1997]. The conductance of a graph cut, which divides the graph vertices into two sets $S, S^c$ is defined as [Chung 1997]

$$h_G(S) = \frac{|E(S, S^c)|}{\min(\text{vol}(S), \text{vol}(S^c))},$$

where $S$ is a set of vertices, and $S^c$ is its compliment;

$$E(S, S^c) = \{(x, y) \in E : x \in S, y \in S^c\}$$

and for any $C \subset V$, $\text{vol}(C) = \sum_{i \in C} d_i$. Then, the Cheeger constant or the conductance of the graph is defined as

$$h_G = \min_S h_G(S). \tag{1.6}$$

A small value of $h_G$ indicates that the graph has weakly connected components or communities, and a set with a small $h_G(S)$ is a good candidate for a community because it has very few outgoing links compared to its volume. The Cheeger constant is bounded on both sides by functions of the spectral gap $\lambda_{n-1}$ as stated in the following result from [Chung 1997], called the Cheeger inequality.

**Theorem 1.2.** *[Chung 1997, Theorem 2.2, Lemma 2.1] For any connected graph $G$,*

$$2h_G \geq \lambda_{n-1} \geq \frac{h_G^2}{2}.$$

A large spectral gap implies a large value of conductance and vice versa, and a small value of $\lambda_{n-1}$ signals the presence of a densely connected community, weakly connected to the rest of the graph [Andersen & Chung 2007].

## 1.2 Random Graph Models

Thus far we dealt with deterministic graphs and their properties. In this section, we give a brief overview of different random graphs. A random graph is a probabilistic object where edges are added between groups of nodes according to some probabilistic rule. Different random graph models have been proposed to model various network properties. One of the earliest random graph models to be studied is the Erdős-Rényi (ER) graph model in [Erdős & Rényi 1959].

An ER graph, denoted by $G(n, p)$, consists of $n$ nodes such that a link exists between any pair of nodes with probability $p$, which can be a function of $n$ [Bollobás 1998]. A related model is $G_{n,m}$, where links are added randomly between nodes such that the total number of edges is $m$ [Bollobás 1998]. In $G(n, p)$, the number of edges is a binomial random variable, and henceforth only this model is considered. This graph model, though simple, has many interesting asymptotic properties. The case when $p_n$ goes to zero as $n$ grows to infinity is an interesting regime to consider, and it has been shown that in this case $G(n, p)$ manifests many important phase transition phenomena.

When $p \geq \frac{\log(n)}{n}$, $G(n, p)$ is connected, but otherwise it has many connected components [Hofstad 2016]. If $p > 1/n$, then $G(n, p)$ has several connected components with one giant component. Otherwise, the graph has no connected components of size larger than $\Theta(\log(n))$. For a comprehensive treatment of asymptotic properties of ER graphs the reader is referred to [Hofstad 2016, Chapter 4,5]. We present a survey of important spectral properties of Erdős-Rényi graphs in Section 2.2.

In $G(n, p)$ all nodes have the same average degree and the degree distribution is asymptotically Poisson, which is a light-tailed distribution. This poses a serious limitation to modeling real-world networks, since most networks have heavy-tailed degree distribution, meaning the tail probability $\mathbb{P}(d_i > \tau)$ for any node $i$ decays slowly as a function of $\tau$ [Hofstad 2016]. Real-world networks also have heterogenous degrees.

A generalization of Erdős-Rényi (ER) graphs that mitigates these drawbacks is the Chung-Lu graph [Chung & Lu 2002b]. In a Chung-Lu graph $G(\mathbf{w})$, the vector $\mathbf{w}$ is such

that $w_i$ is the average degree of node $i$. From the average degrees, the graph is constructed such that for any two nodes $i, j$, an edge appears with probability

$$p_{ij} = \min\left(\frac{w_i w_j}{\sum_k w_k}, 1\right).$$

The Chung-Lu graph is more versatile, in the sense that it can be used to model graphs with different degree distributions, by chosing the vector $\mathbf{w}$ appropriately. We discuss some important spectral properties of Chung-Lu graphs in Chapter 6.

An important feature of many real-world networks is the presence of communities. The Stochastic Block Model (SBM), also known as the Planted Partition Model, is a class of random graphs proposed to model communities [Holland *et al.* 1983]. Consider a SBM with $M$ communities. It is specified by a symmetric matrix $\mathbf{B} \in \mathbb{R}^{M\times M}$ with $B_{ij} < 1$. The entries $B_{ij}, i \neq j$ is the probability that there is an edge between a node in community $i$ and a node in community $j$. Similarly $B_{ii}$ is the edge probability between any two nodes in community $i$. There have been various research efforts to develop and test community detection algorithms on SBM [Rohe *et al.* 2011, Massoulié 2014, Saade *et al.* 2015, Abbe & Sandon 2015a].

A drawback of the standard SBM as described above is that the mean degree of all nodes in a given community is the same. The degree-corrected SBM (DC-SBM) [Karrer & Newman 2011] mitigates this defect, where each node in a given community is allowed to have a different expected degree. In addition, several other important random graph models exist such as the Preferential Attachment model [Albert *et al.* 1999], exchangeable random graphs [Diaconis & Janson 2007] and random geometric graphs [Penrose 2003].

## 1.3 Hidden Community Detection Problem

In Chapter 5 of this thesis, we deal in detail with the hidden community detection problem. The hidden community detection problem is concerned with identifying a subset of graph nodes that are highly connected to one another, but weakly connected to the rest of the graph, i.e., a subset of nodes with a small conductance. This problem is also referred to as *dense subgraph detection* or *dense subgraph discovery*.

The interest in studying this problem is twofold. From a practical point of view, many problems in machine learning on Big Data can be mapped to a problem of detecting a dense subgraph embedded in a sparse graph. For example, detecting a set of highly correlated images in an image dataset [Firouzi *et al.* 2013], detecting fraudulent activity in an auction network [Chau *et al.* 2006], finding a group of friends in a social network, and finding users with similar interests in a website such as Netflix™are all instances of the dense subgraph detection problem.

Secondly, it can be seen as a relaxation of the clique detection problem, where the goal is to detect the largest subset of nodes where every node is connected to all other nodes of the set, and this latter problem is NP-hard [Karp 1972]. Therefore, it is interesting from a computational point of view, since this problem displays a phase transition as the subgraph parameters are changed between an easy regime, where computationally inexpensive algorithms can detect the subgraph and a hard regime, where global exhaustive search has to be employed.

In a general graph, the problem of detecting the nodes of a dense subgraph can be solved by choosing an objective function and relating it to a max-flow instance on the graph [Goldberg 1984]. A commonly used objective function is the edge density defined as $\frac{E(S)}{|S|}$, for any set $S$, where $E(S)$ is the number of edges among nodes in $S$ and $|S|$ is its

cardinality. A survey of other algorithms related to dense subgraph detection can be found in [Lee *et al.* 2010].

In this thesis we look at an instance of this problem on random graphs. We consider $G(n,q)$, an Erdős-Rényi graph of edge probability $q$ and $n$ vertices. A subset of nodes of size $K$ is picked and the edges are inserted in this subset with probability $p$ with $p > q$. Clearly when $n$ and $K$ are large enough, the densest community in this graph corresponds to the planted dense subgraph. One is interested in the minimum detectable subgraph size and further, the minimum detectable subgraph that can be detectable in polynomial time.

By the following result from [Mifflin *et al.* 2004], this problem is characterized by a phase transition.

**Theorem 1.3.** *Let $F$ denote any subgraph on the vertices $V$ of an ER graph $G(n,q)$. Then,*

$$\lim_{n\to\infty} \mathbb{P}(F \subseteq G(n,q)) = \begin{cases} 0 & \text{if } q \ll n^{-\frac{1}{m_F}} \\ 1 & \text{if } q \gg n^{-\frac{1}{m_F}}, \end{cases}$$

*where* $m_F := \max\left\{ \frac{|E(H)|}{|V(H)|} : H \subset F, |V(H)| > 0 \right\}$.

Consequently, a planted subgraph in an ER graph is only distinguishable when $m_F \gg \frac{\log(n)}{\log(1/q)}$.

In the study of planted clique detection in ER graphs, there exist phase transitions between easy, hard and impossible regimes. Consider a $G(n,1/2)$ with a planted clique of size $K$. If $K \leq 2(1-\varepsilon)\log_2(n)$, the clique is impossible to detect; however, an exhaustive search detects the clique nodes when $K \geq 2(1+\varepsilon)\log(n)$. In contrast, the smallest detectable clique size by known polynomial time algorithms is only $\Omega(\sqrt{n})$ [Alon *et al.* 1998, Deshpande & Montanari 2015]. The hidden subgraph detection problem also displays a phase transition phenomenon discussed in detail in Chapter 5.

Many approaches have been proposed in the literature to solve the hidden subgraph problem and the clique detection problem, both global and local. In [Alon *et al.* 1998], the authors use a spectral algorithm to detect the largest clique, i.e., the case when $p = 1$. Similar techniques have been adopted in [Martinsson 2013]. Similarly, there are approaches based on relaxations of Maximum Likelihood detection e.g. [Hajek *et al.* 2016a].

In Chapter 5, we consider a local Belief Propagation based approach. Our approach is semi-supervised, i.e., we assume that some side-information about the community of interest is known to the detector. Semi-supervised learning represents an important class of problems [Avrachenkov *et al.* 2012], but it is so far not well explored in the context of subgraph detection limits on random graphs. Our contribution is to study the impact of side-information on the detectability threshold of local algorithms in hidden community detection.

## 1.4 Personalized PageRank

In Chapter 6, we present an analysis of PageRank on random graphs. PageRank, since its introduction in [Page *et al.* 1997] in the context of web ranking, has found application in many different areas of graph processing such as recommendation systems, link prediction and community partitioning [Gleich 2015, Andersen & Chung 2007].

The Personalized PageRank vector $\boldsymbol{\pi}$ with preference vector $\mathbf{v}$ is defined as the stationary distribution of a modified Markov chain with transition matrix

$$\widetilde{\mathbf{P}} = \alpha\mathbf{P} + (1-\alpha)\mathbf{v}\mathbf{1}^T,$$

where $\alpha \in (0,1)$ is called the damping factor and $\mathbf{v}$, the personalization vector, is any probability distribution on $V$ [Haveliwala 2002].
In other words, $\boldsymbol{\pi}$ satisfies [Langville & Meyer 2004]

$$\boldsymbol{\pi} = \widetilde{\mathbf{P}}\boldsymbol{\pi},$$

or,

$$\boldsymbol{\pi} = (1-\alpha)[\mathbf{I} - \alpha\mathbf{P}]^{-1}\mathbf{v}$$

when $\alpha < 1$.

In [Andersen & Chung 2007, Andersen *et al.* 2006] the authors proposed local paritioning algorithms based on computing the PageRank scores starting from a seed node. The seed node is a node known to belong to the community of interest. If $i$ is a seed node then PageRank is computed by taking $v_i = 1$. In practice, only an approximate computation of PageRank is sufficient, and this is done by means of the power iteration [Langville & Meyer 2004]

$$\boldsymbol{\pi}^{k+1} = \widetilde{\mathbf{P}}\boldsymbol{\pi}^{k}.$$

There have been many other works in this field analyzing the performance of PageRank-based diffusion algorithms for community detection in graphs, e.g. [Gleich & Kloster 2016]. Other diffusion-based algorithms such as the Heat-Kernel have also been proposed [Chung 2009].

The first work in the direction of analyzing PageRank in random graphs for community detection is [Kloumann *et al.* 2016]. They analyze seeded PageRank on a Stochastic Block Model, and show that PageRank arises as a natural weight vector for jump $k$ probabilities as $n \to \infty$. In [Chen *et al.* 2016], the asymptotic distribution of PageRank was derived on heavy-tailed directed configuration models. In Chapter 6, we consider the behaviour of PageRank on undirected random graphs, including the Chung-Lu random graph and the Stochastic Block Model graph and show that as $n \to \infty$, PageRank on these graphs have simple expressions. This constitutes a first step towards comparing PageRank-based community detection with other methods in terms of detection limits, which is absent in [Kloumann *et al.* 2016].

## 1.5 Thesis Organization and Contributions

### 1.5.1 Chapter 2

In the first chapter of this thesis, we provide a survey of fundamental results in Random Matrix Theory. We also review the application of these results to the study of Erdős-Rényi graphs and its matrix representations. In addition, we give a brief background on Belief Propagation and message passing algorithms, and their application to distributed algorithms on graphs.

### 1.5.2 Chapter 3

In this chapter, we derive the limiting form of the empirical spectral distribution of the adjacency and normalized Laplacian matrices of the standard Stochastic Block Model (SBM) with a fixed number of communities. We make use of Girko's stochastic fixed point equations and degree concentration results of SBM to derive the limiting empricial spectral distribution. We also derive a sharp bound for the spectral norm of the centered SBM adjacency matrix using the moment method for bounding the largest eigenvalue of random symmetric matrices. In addition we analyze the limiting eigenvector distribution of

the SBM adjacency matrix by characterizing the asymptotic limit of a modified spectral function that incorporates the eigenvectors. For the Stochastic Block Model with identical communities, we show that this modified empirical distribution also has the same limit as the empirical spectral distribution under certain assumptions.

**Related Publications**

[1] A. Kadavakandy, L. Cottatellucci, and K. Avrachenkov. "Characterization of Random Matrix Eigenvectors for Stochastic Block Model," Asilomar Conference on Signals, Systems, and Computer 2015. IEEE, 2015.

[2] K. Avrachenkov, L. Cottatellucci and A. Kadavankandy. "Spectral properties of random matrices for Stochastic Block Model.", International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt) 2015.

### 1.5.3 Chapter 4

In Chapter 4, we study an important problem in machine learning called Anomaly Detection. We consider a specific anomaly model where the anomaly is a ER subgraph of size $K$ and edge probabiliy $p$ embedded in an ER graph of size $n > K$ and edge probability $q < p$. We analyze an algorithm based on thresholding the $L^1$-norm of the dominant eigenvector of a shifted adjacency matrix defined as $\boldsymbol{\mathcal{A}} = \mathbf{A} - q\mathbf{1}\mathbf{1}^T$, where $\mathbf{A}$ is the adjacency matrix. The main contribution of this thesis is to derive a Central Limit Thorem (CLT) for the suitably scaled dominant eigenvector components under certain assumptions on $K, p$ and $q$. Specifically we consider $q > C\frac{\log^4(n)}{n}$, $\lim_{n\to\infty} \frac{p}{q} = C$, a constant and $K(p-q) = \omega((nq)^{2/3})$. Under these assumptions, we show that when the first $K$ nodes correspond to the anomaly

$$\sqrt{\frac{K\delta_p}{p(1-p)}}\left(x_i - \sqrt{p-q}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

for $1 \le i \le K$, and

$$\sqrt{\frac{K\delta_p}{q(1-q)}}x_i \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

for $1 + K \le i \le n$, with $\mathbf{x} = \sqrt{\lambda}\mathbf{u}$ where $(\lambda, \mathbf{u})$ is the dominant eigenvalue-eigenvector pair of $\boldsymbol{\mathcal{A}}$. In addition, using this fact, we devise an algorithm that **recovers** the subgraph nodes given a graph instance containing the subgraph and we delineate the parameter range where the algorithm succeeds such that a suitably defined error probability goes to zero as $n \to \infty$. Our algorithm works for dense to moderately sparse graphs. We also use the above distribution to derive an approximate distribution of the $L^1$-norm of $\mathbf{u}$ and derive a statistical test to detect the presence of such a subgraph, which only needs the knowledge of $q$ and $n$ and not $p$ or $K$. An algorithm for subgraph detection was proposed in [Hajek *et al.* 2015b] which thresholding the total number of edges; however, this algorithm requires the knowledge of $K$ and $p$.

**Related Publications**

[1] A. Kadavankandy, L. Cottatellucci and K. Avrachenkov. "Characterization of $L^1$-norm Statistic for Anomaly Detection in Erdős Rényi Graphs", IEEE Conference on Decision and Control (CDC), 2016.

### 1.5.4 Chapter 5

In this chapter we consider recovery of planted dense subgraph in a sparse ER graph in the presence of side-information. In recent works [Montanari 2015, Hajek *et al.* 2015a], it

was shown that a local BP is sub-optimal for this problem in that there is a well-defined threshold of the subgraph parameters below which correct recovery of subgraph nodes is not possible. This phase transition is characterized by an effective Signal-to-Noise ratio parameter $\lambda$ defined below:

$$\lambda = \frac{K^2(p-q)^2}{(n-K)q}.$$

In [Montanari 2015, Hajek *et al.* 2015a], the authors show, under certain assumptions on the subgraph parameters that $\lambda > 1/\exp(1)$ is required for BP to achieve weak recovery of a subgraph with sub-linear size in sparse graphs, i.e., $\lim_{n\to\infty} \frac{\mathbb{E}|S\Delta\overline{S}|}{K} = 0$ iff $\lambda > \frac{1}{\exp(1)}$, where $S$ is the hidden subgraph and $\overline{S}$ is the BP output. In this chapter, we study the influence of side-information on this BP threshold. We consider two types of side-information: perfect and imperfect. In the case of perfect side-information, a fraction $\alpha$ of the subgraph nodes are known. In the case of imperfect side-information the cues may be incorrect and the correctness of cues is characterized by a parameter $\beta$. We design a BP-based algorithm that takes advantage of both kinds of side-information. We derive the asymptotic distribution of BP messages and analyse its error performance. We show that BP succeeds in weak-recovery when $K = o(n)$ for any $\lambda, \alpha, \beta > 0$.

**Related Publications**

[1] A. Kadavankandy, K. Avrachenkov, L. Cottatellucci and R. Sundaresan. "Belief Propagation for Subgraph Detection with Imperfect Side-information", to appear in IEEE International Symposium on Information Theory (ISIT) 2017.

[2] A. Kadavankandy, K. Avrachenkov, L. Cottatellucci and R. Sundaresan. "The Power of Side-information in Subgraph Detection", IEEE Transactions on Signal Processing (submitted).

### 1.5.5 Chapter 6

In this chapter, we turn our attenton to the analysis of PageRank on random graphs. Not many analytic studies are available for PageRank in undirected random graph models. We mention the work [Avrachenkov & Lebedev 2006] where PageRank was analysed in preferential attachment models and the more recent works [Chen *et al.* 2014, Chen *et al.* 2016], where PageRank was analysed in directed configuration models.

In our work, we focus on class of graphs with two properties: fast mixing, i.e., the second eigenvalue $\lambda_2(\mathbf{P})$ of Markov matrix $\mathbf{P}$ is such that $\lambda_2(\mathbf{P}) = o(1)$ with high probability (whp) as $n \to \infty$, and restricted degrees, i.e., $\frac{d_{\max}}{d_{\min}} \leq K$ w.h.p. for some $K > 0$. We show that on this class of random graphs $\|\boldsymbol{\pi} - \overline{\boldsymbol{\pi}}\|_1 = o(1)$ whp, where

$$\overline{\pi}_i = \alpha \frac{d_i}{\sum_k d_k} + (1-\alpha)v_i,$$

with $d_i$ being the degree of node $i$. This result substantiates the observation that PageRank is correlated with node degrees on some graph models [Pandurangan *et al.* 2002, Fortunato *et al.* 2006]. The above result is proven thanks to the limiting spectral properties of Markov matrix of undirected random graphs. Next, we show a stronger result that

$$\max_i \frac{|\pi_i - \overline{\pi}_i|}{\pi_i} = o(1)$$

for Chung-Lu graphs with mean degrees $w_i$ such that $\max_i w_i / \min_i w_i \leq K$ for some $K > 0$.

We then consider the Stochastic Block Model (SBM) graphs with two or more communities. On such a graph, we show that under certain conditions, the PageRank satisfies a concentration similar to the above formulations. In particular, we show that for SBM with equi-sized communities, with inter-community edge probability $q$ and intra-community edge probability $p$, the asymptotic PageRank $\overline{\boldsymbol{\pi}}_{\text{SBM}}$ on a SBM is given as follows:

$$\overline{\boldsymbol{\pi}}_{\text{SBM}} = \alpha\frac{1}{n}\mathbf{1} + (1-\alpha)\left(\mathbf{v} + \frac{\alpha\beta}{1-\alpha\beta}(\mathbf{v}^T\mathbf{u})\mathbf{u}\right),$$

where $\beta := \frac{p-q}{p+q}$, and $\mathbf{u} \in \mathbb{R}^n$ is the community partitioning vector such that $u_i = \frac{1}{\sqrt{n}}$, for $i \in C_1$ and $u_i = -\frac{1}{\sqrt{n}}$ for $i \in C_2$, where $C_1, C_2$ represent the set of nodes in community 1 and community 2, respectively. Thus we can see that PageRank on SBM incorporates community partioning information. This preliminary analysis can be used to analyze PageRank performance for community detection. It would be interesting to derive the limits of detectability for PageRank community detection algorithm.

**Related Publications**

[1] K. Avrachenkov, A. Kadavankandy et al "PageRank in Undirected Random Graphs," Workshop on Algorithms and Models for the Web Graph (WAW), 2015.

[2] K. Avrachenkov, A. Kadavankandy et al "PageRank in Undirected Random Graphs," Internet Mathematics, 2016.

### 1.5.6 Chapter 7

In the framework of network sampling, random walk (RW) based estimation techniques provide many pragmatic solutions while uncovering the unknown network as little as possible. Despite several theoretical advances in this area, RW based sampling techniques usually make a strong assumption that the samples are in the stationary regime, and drop the samples collected before the burn-in period. This work proposes two sampling schemes without the burn-in constraint to estimate the average of an arbitrary function defined on the network nodes, for e.g. the average age of users in a social network.

The central idea of the algorithms lies in exploiting regeneration of RWs at revisits to an aggregated super-node or to a set of nodes and in strategies to enhance the frequency of such regenerations either by contracting the graph or by making the hitting set larger. Our first algorithm, which is based on Reinforcement Learning (RL), takes advantage of the regeneration of RWs, and it uses stochastic approximation to derive an estimator. This method can be seen as intermediate between purely stochastic Markov Chain Monte Carlo iterations and deterministic relative value iterations.

We study this method via simulations on real networks and observe that its trajectories are much more stable than those of standard random walk based estimation procedures, and its error performance is comparable to that of respondent driven sampling (RDS) which has a smaller asymptotic variance than many other estimators. The second algorithm, which we call the RT estimator, is a modified form of RDS that accommodates the idea of regeneration. Simulation studies show that the mean squared error of RT estimator decays much faster than that of RDS with time.

**Related Publications**

[1] K. Avrachenkov, V.S. Borkar, A. Kadavankandy and J. K. Sreedharan. *Comparison of Random- walk Based Techniques for Estimating Network Averages*, International Conference on Computa- tional Social Networks (CSoNet) 2016.

[2] K. Avrachenkov, V. Borkar, A. Kadavankandy, J. K. Sreedharan *Revisiting Random Walk based Sampling in Networks: Evasion of Burn-in Period and Frequent Regenerations* Computational Social Networks Journal (submitted).

Chapter 2

# Introduction to Random Matrix Theory and Message Passing Algorithms

In this chapter, we provide a short introduction to Random Matrix Theory and its application to the study of random graphs. In addition, we introduce the concept of Belief Propagation (BP) on graphs, and its application to the solution of hidden community detection problem.

## 2.1 Survey of Random Matrix Theoretic Results

Traditional treatments in statistics and data processing have focused on finite matrices. The analysis of random matrices when their sizes grow to infinity requires new tools. This is the topic under consideration in the field of Random Matrices.

### 2.1.1 Empirical Spectral Distribution and Stieltjes Transform

One of the pioneering results in Random Matrix Theory is *Wigner's Semicircle Law*. While studying the energy levels of a nuclei, Wigner modeled the Hamiltonian as a symmetric matrix with independent entries that are $\pm 1$ with equal probability. He found out that as the matrix size is increased, the histogram of the eigenvalues of the above matrix, when suitably scaled, settles down to a deterministic function that resembles a semicircle [Wigner 1955]. In a later paper, this result was shown to hold for symmetric matrices with independent entries drawn from a general distribution with zero odd order moments and finite even-order moments [Wigner 1958]. For a review of these early connections between physics and Random Matrix Theory, refer to [Wigner 1967]. It later turned out that this property is universal and extends to a larger class of symmetric random matrices with looser conditions on the distribution [Bai 1999]. This property that the spectral properties of a matrix are not too sensitive to the specific entry distribution is called *universality* [Anderson *et al.* 2009].

A *Wigner matrix* can be a real *symmetric* [1] or a complex *Hermitian*[2] matrix with independent upper triangular entries. In this thesis, we define a Wigner matrix as a class of random matrices with zero mean entries, unit variance, and the entries are in addition required to satisfy a higher moment condition. We therefore provide the following definition.

**Definition 1.** *[Anderson et al. 2009, Tao 2012] A Wigner matrix* $\mathbf{X}$ *is a symmetric matrix such that* $X_{ij}, 1 \leq i < j \leq n$ *are i.i.d. random variables such that* $\mathbb{E}(X_{ij}) = 0$ *and* $\mathbb{E}(X_{ij}^2) = 1$ *and* $X_{ii}, 1 \leq i \leq n$ *are i.i.d. with* $\mathbb{E}(X_{ii}) = 0$ *and* $\mathbb{E}(X_{ii}^2) < \infty$.

[1] $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric if $\mathbf{A} = \mathbf{A}^T$

[2] $\mathbf{A} \in \mathbb{C}^{n \times n}$ is Hermitian symmetric or Hermitian if $\mathbf{A} = \mathbf{A}^H$

Commonly known examples of Wigner matrices are the Gaussian Unitary Ensemble (GUE) and Gaussian Orthogonal Ensemble (GOE), which are made up of gaussian entries.

**Definition 2.** *[Anderson et al. 2009] A symmetric random matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ is said to be drawn from a Gaussian Orthogonal Ensemble if its upper diagonal entries are independently drawn from gaussian $\mathcal{N}(0,1)$ and diagonal entries are independently drawn from $\mathcal{N}(0,2)$.*

*A Hermitian random matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ is said to be drawn from a Gaussian Unitary Ensemble if its upper diagonal entries are indepedently drawn from $\mathcal{N}_{\mathbb{C}}(0,1)$ and diagonal entries are independently drawn from $\mathcal{N}(0,1)$, where $\mathcal{N}_{\mathbb{C}}(0,1)$ represents a circular symmetric gaussian random variable with unit variance.*

Explicit expressions for the distribution of eigenvalues of GOEs and GUEs can be derived. For e.g. for a GOE matrix, the joint distribution $p_n(\lambda_1, \lambda_2, \ldots, \lambda_n)$ of ordered eigenvalues is given as follows [Anderson *et al.* 2009].

$$p_n(\lambda_1, \lambda_2, \ldots, \lambda_n) := \begin{cases} \frac{1}{Z} \prod_{1\le i<j\le n} |\lambda_i - \lambda_j| \prod_{i=1}^{n} e^{-\frac{\lambda_i^2}{4}}, & \text{if } \lambda_1 \ge \lambda_2 \ldots \ge \lambda_n, \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $Z$ is a normalization factor.

The importance of gaussian ensembles is that various interesting properties such as separation of eigenvalues, delocalization of eigenvectors etc are easier to study than for a general distribution [Anderson *et al.* 2009]. It is then possible to extend these results to general random matrices by moment matching methods [Tao & Vu 2011, Tao & Vu 2012].

When the entries have a general distribution, it is more tractable to study their distribution by means of what is called the empirical spectral distribution (e.s.d.). Since $\mathbf{X}$ is symmetric its eigenvalues are real, let $\lambda_1 \ge \lambda_2 \ge \lambda_3 \ldots \ge \lambda_n$ be the ordered eigenvalues of $\mathbf{X}$. The e.s.d. of $\mathbf{X}$ is defined as follows.

**Definition 3.** *[Anderson et al. 2009] The empirical distribution function $F^{\mathbf{X}}(x)$ is defined as*

$$F^{\mathbf{X}}(x) = \frac{1}{n} \sum_{i=1}^{n} \chi(\lambda_i \le x),$$

*where $\chi(\cdot)$ is the indicator function, i.e., it is one if the condition in its argument is satisfied and zero otherwise.*
*In future, we will drop the superscript in the notation of the e.s.d.*

Similarly, one can define the derivative of the e.s.d.

$$dF^{\mathbf{X}}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(\lambda_i - x),$$

where $\delta(x)$ is Dirac's delta function at 0. The e.s.d $F(x)$ can also be defined in terms of the integral of a continuous or measurable function $g(x)$ as

$$\int g(x) dF(x) = \frac{1}{n} \sum_{i=1}^{n} g(\lambda_i). \tag{2.2}$$

From (2.2), $\int x^k dF(x) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^k = \frac{1}{n}\mathrm{tr}(\mathbf{X}^k)$. In particular we have,

$$\int x dF(x) = \frac{1}{n}\mathrm{tr}(\mathbf{X}) \qquad \int x^2 dF(x) = \frac{1}{n}\mathrm{tr}(\mathbf{X}^2). \tag{2.3}$$

Therefore, a straightforward way to study the limiting properties of the e.s.d. is to look at the average moments of the eigenvalues. This is known as the moment method, and it was used by Wigner in the proof of the semicircle law [Tao 2012].

A useful functional that is important in the study of random matrices is the Stieltjes transform of a probability distribution, defined as

$$s(z) = \int \frac{1}{x - z} dF(x), \tag{2.4}$$

for $z \in \mathbb{C}, \text{Im}(z) > 0$, where $F(x)$ is any probability distribution.
When $F(x)$ is taken to be the e.s.d. we get

$$s(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i - z} = \frac{1}{n} \text{trace}(\mathbf{X} - z\mathbf{I})^{-1}.$$

The matrix $(\mathbf{X} - z\mathbf{I})^{-1}$ is called the resolvent of $\mathbf{X}$.
We provide here some important properties of the Stieltjes transform.
**Properties of the Stieltjes Transform** [Anderson *et al.* 2009, Tao 2012]

- *Analytic*: The function $s(z)$ is analytic from $\mathbb{C}_+ \to \mathbb{C}_+$.
- *Boundedness*: Since $|x - z| = \sqrt{(\Re(x) - \Re(z))^2 + \Im(z)^2} \geq |\Im(z)|$, for $x \in \mathbb{R}$, $|s(z)| \leq |\text{Im}(z)|^{-1}$, by the ordered property of expectation.
- *Invertibility*: Stieltjes transform $s(z)$ can be inverted to get the empirical spectral distribution.
$$dF(x) = \lim_{y \searrow 0} \frac{1}{\pi} \Im(s(x + iy)). \tag{2.5}$$

Since matrix $\mathbf{X}$ is Hermitian, its eigenvalue distribution is stable, i.e., a small Hermitian perturbation of the entries of $\mathbf{X}$ does not perturb the eigenvalues of $\mathbf{X}$ by much. This is in marked contrast to non-Hermitian matrices, which displays what is called *Pseudospectrum*, i.e., there exist *small* perturbations that can drastically change the spectrum of a non-Hermitian matrix [Rump 2006]. We denote the class of Hermitian matrices of size $n$ by $\mathcal{H}_n$. The continuity of the eigenvalues means that the e.s.d. and also the Stieltjes transform are continuous functions of matrix entries. These results are important in the derivation of the limiting spectral distribution of Wigner matrices.

**Lemma 2.1.** *(Interlacing inequalities) [Horn & Johnson 2012] If two Hermitian matrices* $\mathbf{A}, \mathbf{B} \in \boldsymbol{\mathcal{H}}_n$ *are such that* $\text{rank}(\mathbf{A} - \mathbf{B}) \leq r$, *then,*

$$\lambda_{k-r}(\mathbf{A}) \geq \lambda_k(\mathbf{B}) \geq \lambda_{k+r}(\mathbf{A}),$$

*for* $1 \leq k \leq n$.

From Lemma 2.1, using counting arguments, one can get the following bound on the change in e.s.d. due to a finite rank perturbation.

**Lemma 2.2.** *(E.s.d. of finite rank perturbation) [Bai & Silverstein 2009] If* $\text{rank}(\mathbf{A} - \mathbf{B}) \leq r$, *then*

$$|F^{\mathbf{A}}(x) - F^{\mathbf{B}}(x)| \leq \frac{r}{n},$$

*where* $\mathbf{A}, \mathbf{B} \in \mathcal{H}_n$.

Thus, Lemma 2.2 implies that as long as the change in rank $r = o(n)$, both $\mathbf{A}$ and $\mathbf{B}$ have the same asymptotic e.s.d. This result is useful, for e.g., when analysing the limiting spectra of random Hermitian matrices with non-zero mean.

Another important perturbation is finite norm perturbation. Recall that the Frobenius norm of a matrix $\mathbf{A}$, denoted as $\|\mathbf{A}\|_F$ is given as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |A_{ij}|^2} = \sqrt{\text{trace}(\mathbf{A}\mathbf{A}^H)}.$$

The following lemma gives a bound on the change in the eigenvalues when the matrix $\mathbf{X}$ is subjected to a finite norm perturbation.

**Lemma 2.3.** *(Hoffman Wieldland inequality) [Anderson et al. 2009, Lemma 2.1.19] Let $\mathbf{A}, \mathbf{B}$ be two Hermitian random matrices with eigenvalues $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \ldots \geq \lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{B}) \geq \lambda_2(\mathbf{B}) \ldots \geq \lambda_n(\mathbf{B})$. Then*

$$\sum_{i=1}^{n} |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})|^2 \leq \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Lemma 2.3 is important, because it leads to the important result that any Lipshitz function $g(x)$ defined on the eigenvalues of the matrix gives a Lisphitz continuous function on the matrix entries. Let $g(x) : \mathbb{R} \to \mathbb{R}$ be a Lipshitz continuous function with constant $\|g\|_L$. From Lemma 2.3 and (2.2) we have

$$\begin{aligned}
\left| \int g(x) dF^{\mathbf{A}}(x) - \int g(x) dF^{\mathbf{B}}(x) \right| &= \frac{1}{n} \left| \sum_{i=1}^{n} \left( g\left(\lambda_i(\mathbf{A})\right) - g\left(\lambda_i(\mathbf{B})\right) \right) \right| \\
&\leq \frac{1}{n} \|g\|_L \sum_{i=1}^{n} |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \\
&\leq \frac{1}{\sqrt{n}} \|g\|_L \|\mathbf{A} - \mathbf{B}\|_F,
\end{aligned}$$

where in the first inequality, we used Lipshitz continuity of $g(x)$ and in the last inequality we used Cauchy-Schwartz inequality and Lemma 2.3. Thus, the functional $\int g(x) dF(x)$ is also Lipshitz continuous with constant $\frac{1}{\sqrt{n}} \|g\|_L$. Also see [Anderson *et al.* 2009, Lemma 2.3.1].

When $g(x) = \frac{1}{x-z}$, we get the Stieltjes transform with $\|g\|_L = \frac{1}{|\Im(z)|}$, and hence the latter is also Lipshitz continuous with constant $\frac{1}{\sqrt{n}} \frac{1}{|\Im(z)|}$.

**Lemma 2.4.** *(Lipshitz continuity of Stieltjes transform) The Stieltjes transform $s(z)$ of a matrix $\mathbf{X}$ is Lipshitz continuous with respect to the matrix entries $X_{ij}, 1 \leq i, j \leq n$, and the Frobenius norm with constant $\frac{1}{\sqrt{n}} \frac{1}{|\Im(z)|}$.*

Now we describe *weak convergence of measures*. Consider a sequence of probability measures $\mu_n$ on $\mathbb{R}$. In addition, let $\mathcal{C}_b$ be the class of continuous functions with bounded support defined on $\mathbb{R}$. Then, $\mu_n$ is said to converge weakly to a probability measure $\mu$ if for any $f \in \mathcal{C}_b$,

$$\lim_{n \to \infty} \left| \int f(x) d\mu_n(x) - \int f(x) d\mu(x) \right| = 0.$$

An important property of the Stieltjes transform is that if the Stieltjes transform of a sequence of probability distributions converges, then the sequence converges *weakly*. This is the *weak convergence* property of the Stieltjes transform.

**Lemma 2.5.** *[Anderson et al. 2009, Theorem 2.4.4] Let $\mu$ and $(\mu_n)_{n\geq 1}$ be a sequence of probability measures. Then as $n \to \infty$ $\mu_n$ converges to $\mu$ weakly if and only if $s^{\mu_n}(z) \to s^{\mu}(z)$ for all $z \in \mathbb{C}_+$.*

Using the above result on the limit of probability distribution, one can prove the celebrated Wigner's semicircle law. A key role in the convergence of the e.s.d. of large random Hermitian matrices is played by the following assumption.

**Assumption 2.1.** *[Girko 2001, Girko 1990] The Hermitian matrix $\mathbf{X}$ with zero mean independent upper diagonal entries $X_{ij}$ of variance $\sigma^2$ satisfies the Lindeberg's condition, i.e. for any $\delta > 0$,*

$$\lim_{n\to\infty} \mathbb{E}\left(\chi(|X_{ij}| > \sqrt{n}\delta)X_{ij}^2\right) = 0. \tag{2.6}$$

This assumption essentially implies that the tails of the distributions characterizing the random variables $X_{ij}$ diminish as $n \to \infty$. Under this assumption, it is known that the sequence of the e.s.d. converges weakly to a limiting eigenvalue distribution in the almost sure sense as stated by the following theorem known as *Wigner's Semicircle Law.*

**Theorem 2.1.** *[Girko 2001, Girko 1990, Chapter 1] Let the Wigner matrix $\mathbf{X}$ with zero mean independent random entries $X_{ij}$ satisfy Assumption (2.1) and additionally, all the equal variances satisfy $\sigma_{i,j}^2 = \sigma^2$ with $0 < \sigma^2 < +\infty$. Then, the sequence of the e.s.d. of $\mathbf{X}/\sqrt{n}$ converges weakly to a the Wigner semicircle law in the almost sure sense, i.e. for any bounded continuous function $f$*

$$\int f(x)F^{\mathbf{X}_n}(x)dx \xrightarrow{a.s.} \int f(x)\mu_{sc}(x,\sigma^2)dx$$

*where $F^{\mathbf{X}_n}(x)$ denotes an e.s.d. of the Wigner matrix of size $n$ and $\mu_{sc}(x,\sigma^2)$ is the **Wigner semicircular distribution** with parameter $\sigma^2$ given by*

$$\mu_{sc}(x,\sigma^2) = \frac{1}{2\pi\sigma^2}\sqrt{(4\sigma^2 - x^2)_+},$$

*where $(x)_+ := \max(x, 0)$.*

### 2.1.2 Spectral Norm and Largest Eigenvalues

As noted in Lemma 2.2, the asymptotic e.s.d. is unchanged when a finite number of eigenvalues are changed. This means that, even though as stated by Theorem 2.1, the e.s.d. converges to a function supported on $[-2\sqrt{n}\sigma, 2\sqrt{n}\sigma]$, there might possibly be eigenvalues that fall outside this spectrum.

By definition, the maximum eigenvalue of $\mathbf{X}$ in absolute value is $\|\mathbf{X}\|_2$ [Bhatia 2013]. By Theorem 2.1 we know

$$\|\mathbf{X}\|_2 \geq 2\sqrt{n}\sigma \text{ a.s.}$$

It turns out that the above lower bound is sharp as $n \to \infty$. The first result on an upper-bound on the spectral norm of $\mathbf{X}$ was given by Furedi and Komlos [Füredi & Komlós 1981]. Their approach is based on the observation that

$$\begin{aligned}\mathbb{E}|\mathbf{X}|^{2k} &\leq \mathbb{E}\sum_{i=1}^{n}|\lambda_i|^{2k} \\ &= \mathbb{E}\text{trace}(\mathbf{X}^{2k}).\end{aligned}$$

The above expected trace can be expanded as a sum of products of the matrix elements, where due to independence and zero mean property, many of the terms are trivially zero.

The proof then proceeds by combinatorially bounding the total number of non-zero terms that contribute to the above expected expected trace.
A sharper version of this result was derived in [Vu 2007], which we state below.

**Theorem 2.2.** *[Vu 2007] Let* $\mathbf{X}$ *be a Wigner matrix with independent random elements* $X_{ij}$, $i, j = 1, \dots n$ *having zero mean and variance at most* $\sigma^2(n)$. *If the entries are bounded by* $K(n)$ *and there exist a constant* $C'$ *such that* $\sigma(n) \geq C' n^{-1/2} K(n) \log^2(n)$, *then there exists a constant* $C$ *such that almost surely*

$$\|\mathbf{X}\|_2 \leq 2\sigma(n)\sqrt{n} + C(K(n)\sigma(n))^{1/2} n^{1/4} \log(n). \tag{2.7}$$

Similar results are known for other matrix models such as Wishart matrices [Bai & Silverstein 1998]. The combinatorial method used in the proof can become quite tedious, especially with more complicated matrix models. There are alternative ways to bound the spectral norm of a random matrix, especially when it can be represented as the sum of many simpler independent random matrices. This methodology extends many known concentration bounds of sums of independent random variables to the norm of sums of independent random matrices. We provide a brief overview of these methods in the literature in the subsection.

#### 2.1.2.1 Matrix Concentration Inequalities

There have been significant research in the direction of obtaining concentration results for the norm of matrix sums similar to the concentration results for sums of scalar random variables. These results extend well known scalar concentration inequalities such as the Bernstein's, Azuma-Hoeffding and others to sums of independent random matrices. These results are based on bounds on the trace of the MGF (moment generating function) $M_{\mathbf{X}}(\theta)$ or the CGF ( Cumulant Generating Function) $\Xi_{\mathbf{X}}(\theta)$ defined for any $\mathbf{X}$ as follows [Tropp 2012a]

$$M_{\mathbf{X}}(\theta) = \mathbb{E}(e^{\theta \mathbf{X}}),$$

$$\Xi_{\mathbf{X}}(\theta) = \log \mathbb{E} e^{\theta \mathbf{X}}.$$

Unlike in the case of random variables, the matrix exponential cannot be factored, i.e., the exponential of the sum of two matrices is not generally equal to the product of the exponentials of the two matrices. This is the major impediment in the analysis of the concentration phenomenon in random matrices. However this problem can be mitigated to an extend by using the Golden Thomson inequality [Bhatia 2013, Sec. IX.3]:

$$\text{trace}(e^{\mathbf{A}+\mathbf{B}}) \leq \text{trace}(e^{\mathbf{A}} e^{\mathbf{B}}).$$

This inequality was used by Ahlswede and Winter [Ahlswede & Winter 2002] to obtain the following bound for the CGF

$$\mathbb{E}\left(\text{trace}(\exp(\sum_k \mathbf{X}_k))\right) \leq n \exp(\sum_k \lambda_{\max}(\log \mathbb{E}(e^{\mathbf{X}_k}))),$$

where $\mathbf{X}_k \in \mathbb{R}^{n \times n}$.

In [Tropp 2012a], the authors use Lieb's Theorem [Lieb 1973, Theorem 6] which states that the function $\text{trace}(e^{\mathbf{H}+\log(\mathbf{A})})$ is concave in $\mathbf{A}$ to show that

$$\mathbb{E}\left(\text{trace}(\exp(\sum_k \mathbf{X}_k))\right) \leq \text{trace}(\exp(\sum_k \log \mathbb{E} e^{\theta \mathbf{X}_k})).$$

We state here an example of Matrix Chernoff bound derived in [Tropp 2012a], using the above method.

**Theorem 2.3.** *[Tropp 2012a, Theorem 5.1.1] Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random Hermitian matrices that satisfy*

$$\mathbf{X}_k \succeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq R.$$

*Define the matrix $\mathbf{Y} = \sum_k \mathbf{X}_k$. Define $\mu_{\max} = \lambda_{\max}(\mathbb{E}(\mathbf{Y}))$. Then*

$$\mathbb{P}\left(\lambda_{\max}(\mathbf{Y}) \geq (1+\delta)\mu_{\max}\right) \leq n \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R} \quad \text{for} \quad \delta \geq 0.$$

For a more detailed treatment of Matrix Concentration Inequalities, refer to [Tropp 2012b].

### 2.1.3 Other results

The results we described in this chapter only represent a small fraction of the available results in Random Matrix Theory. A few other avenues and research directions in Random Matrix Theory are the following

- Local eigenvalue distributions [Erdős 2011, Tao & Vu 2011]
- Distribution of extremal eigenvalues and rate of convergence of e.s.d. [Alon *et al.* 2002, Bai 1999]
- Eigenvalues of heavy-tailed random matrices [Bordenave & Guionnet 2013]
- Non-asymptotic results for Random Matrix Theory [Vershynin 2011]
- Free probability [Nica & Speicher 2006]

### 2.1.4 Distribution of Eigenvectors

In this section, we summarize some results on the eigenvectors of the centered adjacency matrix of SBM. To the best of our knowledge, this area is, so far, little explored in comparison to results on eigenvalues.

Let us take the case of real gaussian Wigner Matrices, namely the Gaussian Orthogonal ensemble (Definition 2). Since orthogonal projections of gaussian random vectors are also gaussian, for a GOE matrix $\mathbf{X}$, $\mathbf{U}^T\mathbf{X}\mathbf{U}$ is also a GOE matrix [Anderson *et al.* 2009]. By (2.1), the eigenvalue distribution does not change when a unitary transformation is applied. It is argued in [Anderson *et al.* 2009] that the eigenvectors of $\mathbf{X}$ are uniformly distributed on the sphere $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$. This distribution is called the Haar distribution.

Similar to the work on eigenvalues, it can be investigated whether this property is universal, i.e., whether the fact that a random eigenvector is Haar distributed extends to general entry distributions other than gaussian.

Most works in this direction fall into one of the two following categories: works on delocalization properties of eigenvectors, and works on gaussianity properties of functionals of eigenvectors, based on properties of Haar vectors. For a Haar distributed unit vector $\mathbf{v}$, two properties hold:

- It is delocalized, i.e., the vector mass is not concentrated on any particular component. This property can be expressed in terms of bounds on the moments of vector such as the max moment (i.e., $\max_i |v_i|$) or the $p$-moment; i.e., $\|\mathbf{v}\|_p = (\sum_{i=1}^n |v_i|^p)^{\frac{1}{p}}$. We give an example of this in Theorem 2.4.

- Additionally, a Haar distributed vector can be modeled as $\frac{\mathbf{z}}{\|\mathbf{z}\|_2}$, where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. Thus for large $n$, individual components of $\mathbf{v}$ when normalized appropriately starts to resemble a gaussian random variable. This is exemplified by Theorem 2.7.

**Theorem 2.4.** *[O'Rourke et al. 2016, Theorem 2.1] Let $\mathbf{v}$ be a Haar distributed random vector on $S^{n-1}$. Then, for any $C > 1$ with probability at least $1 - 2n^{1-C} - \exp(-\frac{(C-1)^2}{4C^2}n)$,*

$$\max_i |v_i| \leq \sqrt{\frac{2C^3 \log(n)}{n}},$$

*and $\exists c_p$ s.t.*

$$\|\mathbf{v}\|_p = n^{1-p/2} c_p + o(n^{1-p/2}),$$

*almost surely.*

The above theorem makes sense intuitively because, if a unit vector is completely delocalized, then all its components are approximately $O(1/\sqrt{n})$ and no component can have a huge contribution to the total mass.

As for eigenvectors of Wigner matrices with subgaussian entries, a similar property was shown to be true, thus establishing that the eigenvectors satisfy some of the properties of Haar vectors.

**Theorem 2.5.** *[O'Rourke et al. 2016, Corollary 5.4] Let $\mathbf{X}$ be a Wigner matrix with subgaussian random entries with zero mean and let the non-diagonal entries have unit variance. Then for any $1 \leq p \leq 2$, there exist constants $C, c, C_0, c_0$ such that*

$$c_0 n^{1/p-1/2} \leq \min_{1 \leq j \leq n} \|\mathbf{v}_j\|_p \leq \max_{1 \leq j \leq n} \|\mathbf{v}_j\|_p \leq C_0 n^{1/p-1/2},$$

*where $\mathbf{v}_j$ is an eigenvector of $\mathbf{X}$.*

Similarly, there are bounds on the max-norm of an eigenvector to establish delocalization property of a typical eigenvector of a Wigner matrix.

**Theorem 2.6.** *[O'Rourke et al. 2016, Theorem 6.1] Let $\mathbf{X}$ be a Wigner matrix with subgaussian entries with zero mean and unit variance. Then for any $C_1 > 0$ and any $0 < \varepsilon < 1$, there exists a constant $C_2 > 0$ such that the following holds:*

- *For any $\varepsilon n \leq i \leq (1-\varepsilon)n$,*

$$\max_i |\mathbf{v}_i| \leq C_2 \sqrt{\frac{\log(n)}{n}}$$

  *with probability at least $1 - n^{-C_1}$.*

- *For $1 \leq i \leq \varepsilon n$ or $(1-\varepsilon)n \leq i \leq n$,*

$$\max_i |\mathbf{v}_i| \leq C_2 \frac{\log(n)}{\sqrt{n}}$$

  *with probability at least $1 - n^{-C_1}$.*

Next, we look at gaussianity properties of $\mathbf{v}$. The following result is on a random orthogonal matrix $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) \in \mathbb{R}^{n \times n}$ composed of independent Haar-distributed columns.

**Theorem 2.7.** *[Jiang et al. 2006] [Tao & Vu 2012] Let* $\mathbf{V}$ *be as defined above. Let* $\psi_{i,p} := \mathcal{N}(0,1)$ *for* $1 \leq i,p \leq k$ *be independent normal random variables. Then* $k = o(\sqrt{n})$*, then* $(\sqrt{n}V_{ip})_{1\leq i,p\leq k}$ *and* $(\psi_{i,p})_{1\leq i,p\leq k}$ *are close in total variation norm.*

*In other words, if* $F : \mathbb{R}^{k^2} \to \mathbb{R}$ *is a bounded measurable function, then*

$$|\mathbb{E}F\left((\sqrt{n}V_{i,p})_{1\leq i,p\leq k}\right) - \mathbb{E}F((\psi_{1\leq i,p\leq k}))| \leq o(1),$$

*because the above is bounded by the TV distance between the two distributions.*

In [Tao & Vu 2012], the above result is shown for $k = n^{\delta}$ for some $\delta > 0$.

**Theorem 2.8.** *[Tao & Vu 2012] Given* $C, \delta, C_0 > 0$ *and consider* $\mathbf{X}$ *with off-diagonal elements are independent and distributed as* $\xi$ *and diagonal elements are independently distributed as* $\zeta$*, which satisfy*

- $\xi, \zeta$ *are sub-exponential random variables*
- $\mathbb{E}(\xi) = \mathbb{E}(\zeta) = \mathbb{E}(\xi^3) = 0$*, and*
- $\mathbb{E}(\xi^2) = 1, \mathbb{E}(\xi^4) = 3, \mathbb{E}(\zeta^2) = 2.$

*For* $1 \leq i,j \leq n$*, let* $Z_{ij}$ *be independent random variables with* $Z_{ij} \sim \mathcal{N}(0,1)$ *for* $j > 1$ *and* $Z_{i,1} \sim |\mathcal{N}(0,1)|$*. Let* $1 \leq k \leq n^{\delta}$*, and let* $1 \leq i_1 < \ldots < i_k \leq n$ *and* $1 \leq j_1 < \ldots < j_k \leq n$ *be indices. Then*

$$|\mathbb{E}(F\left((\sqrt{n}v_{i_1}(j_b))_{1\leq a,b\leq k}\right) - \mathbb{E}(F\left((Z_{i_a,j_b})_{1\leq a,b\leq k}\right))| \leq C_0 n^{-\delta}$$

*whenever* $F : \mathbb{R}^{k^2} \to R$ *is a smooth function obeying the bounds*

$$|F(x)| \leq C$$

*and*

$$|\nabla^j F(x)| \leq Cn^{\delta}$$

*for all* $x \in \mathbb{R}^{k^2}$ *and* $0 \leq j \leq 5$.

In [Bai *et al.* 2007], the authors took another approach to characterizing the distribution of Wigner eigenvectors. They defined a function $Q_n(t)$ as follows [Bai *et al.* 2007]

$$Q_n(t) = \sqrt{\frac{n}{2}} \sum_{i=1}^{[nt]} \left(|u_i|^2 - \frac{1}{n}\right).$$

When $\mathbf{u}$ is Haar-distributed, then $Q_n(t)$ converges to a Brownian-bridge. For a general Wigner matrix $\mathbf{X}$, they take $\mathbf{u} = \mathbf{U}\mathbf{y}$, where $\mathbf{U}$ is a unitary matrix whose columns are the eigenvectors of $\mathbf{X}$ and $\mathbf{y}$ is a unit vector. They consider a rescaled form of the above function $Q_n(F^{\mathbf{X}_n}(x))$, where $F^{\mathbf{X}_n}(x)$ is the e.s.d. of $\mathbf{X}$, which is given as follows

$$Q_n(F^{\mathbf{X}_n}(x)) = \sqrt{\frac{n}{2}} \left(\sum_{i=1}^{n} |u_i|^2 \chi(\lambda_i \leq x) - F^{\mathbf{X}_n}(x)\right). \tag{2.8}$$

Let us denote

$$F_1^{\mathbf{X}_n}(x) = \sum_{i=1}^{n} |u_i|^2 \chi(\lambda_i \leq x).$$

We know $F^{\mathbf{X}_n}(x)$ converges by the semicircle law to a continuous function by Theorem 2.1, thus the convergence of $Q_n(F^{\mathbf{X}_n}(x))$ implies the convergence of $Q_n(t)$.

Consider $g(x)$ and

$$X_n(g) = \sqrt{n} \int g(x) d(F_1^{\mathbf{X}_n}(x) - F(x)).$$

We state the following result from [Bai *et al.* 2007].

**Theorem 2.9.** *[Bai et al. 2007, Theorem 1.1, Theorem 1.2] Assume that* $\{X_{ij}, i > j = 1, 2, \ldots, n\}$ *are i.i.d. real random variables with* $\mathbb{E}(X_{12}) = 0, \mathbb{E}(|X_{12}|^2) = 1$ *and* $\mathbb{E}(|X_{12}|^4) < \infty$, *that* $\{X_{ii}, i = 1, \ldots, n\}$ *are i.i.d. real random variables with* $\mathbb{E}(X_{11}) = 0$ *and* $\mathbb{E}(|X_{11}|^2) = 1$. *Let* $\mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}_n\| = 1$. *Then,*

$$Q_n(F^{\mathbf{X}_n})(x) \to 0, \quad a.s.$$

*where* $F(x)$ *is the distribution function of the semicircular law. In addition if* $\max_{1 \le k \le n} |x_k| \to 0$ *and* $\mathbb{E}(X_{12}^3) = 0$

$$X_n(g) \to \mathcal{N}(0, \sigma^2),$$

*where* $\sigma^2 = 2(\int g^2(x) dF(x) - (\int g(x) dF(x))^2)$.

It is indeed shown in [Bai *et al.* 2007, Theorem 1.2] that the vector $(X_n(g_1), X_n(g_2), \ldots, X_n(g_k))$ for $k$ functions $g_{i,i=1,\ldots,k}$ converges to a jointly gaussian random variable in distribution, hence showing that the process $Q_n(t)$ has the properties of a Brownian bridge.

## 2.2 Spectral Properties of Erdős-Rényi Graphs

As defined in Chapter 1, an ER graph is a random graph with $n$ nodes where all the pairs of nodes have equal probability $p_n$ of being connected by an edge, independently of all other pairs. Various interesting properties of ER graphs have been discovered since its introduction in 1959 [Erdős & Rényi 1959]. In this section, we review some spectral properties of the ER graph obtained using tools from Random Matrix Theory.

### Properties of Eigenvalues

Consider the ER graph adjacency matrix $\mathbf{A}^{\text{ER}}$. It is Hermitian with independent and identically distributed (i.i.d.) upper diagonal elements distributed as $\mathcal{B}(p_n)$, a Bernoulli distribution with parameter $p_n$. Consider the normalized form of the matrix $\widehat{\mathbf{A}}^{\text{ER}}$, defined as

$$\widehat{\mathbf{A}}^{\text{ER}} = \gamma(n) \mathbf{A}^{\text{ER}},$$

with

$$\gamma(n) = \frac{1}{\sqrt{n p_n (1 - p_n)}}.$$

The latter is not a Wigner matrix since the entries have non-zero mean. We therefore consider its centered version denoted as $\widetilde{\mathbf{A}}^{\text{ER}}$. We have,

$$\widehat{\mathbf{A}}^{\text{ER}} = \overline{\mathbf{A}}^{\text{ER}} + \widetilde{\mathbf{A}}^{\text{ER}},$$

where $\overline{\mathbf{A}}^{\text{ER}} := \mathbb{E}(\widehat{\mathbf{A}}^{\text{ER}}) = \gamma(n) p_n \mathbf{J}_n$, where $\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n^T - \mathbf{I}_n$.

The average degree of each graph node is given by

$$d_{\text{av}} = \mathbb{E}\left(\sum_j A_{i,j}\right) = np_n, \tag{2.9}$$

for any $i$. Based on the average node degree $d_{\text{av}}$, the ER graphs are classified as *dense* if $d_{\text{av}} = \Theta(n)$, *sparse* if $d_{\text{av}} = o(n)$ and $d_{\text{av}} \to \infty$, and *diluted* if $d_{\text{av}} = O(1)$ [Bordenave & Lelarge 2010].

### 2.2.1 Limiting Spectral Distribution

The result in Theorem 2.1 can be immediately specialized to normalized centered ER adjacency matrices $\widetilde{\mathbf{A}}^{\text{ER}}$. Since for the matrix $\widetilde{\mathbf{A}}^{\text{ER}}$ it holds $\sigma_{ij}^2 = n^{-1}$, for $i, j = 1, \dots n$, the conditions of Theorem 2.1 are satisfied if the limit (2.6) holds, i.e. for any $\tau > 0$

$$\lim_{n\to+\infty} (1-p)\chi\left(1-p \geq \tau\sqrt{np(1-p)}\right) + p\chi\left(p \geq \tau\sqrt{np(1-p)}\right) = 0. \tag{2.10}$$

It is straightforward to verify that this condition is equivalent to the condition $p \geq (\tau^2 n + 1)^{-1}$ for any $\tau > 0$, i.e. if $p = \omega(1/n)$. Then, we can state the following corollary.

**Corollary 2.1.** *Let us consider the normalized centered ER adjacency matrix $\widetilde{\mathbf{A}}^{ER}$ with $p_n \in \omega(n^{-1})$ as $n \to \infty$. Then, the sequence of the e.s.d. converges weakly to a the Wigner semicircle law in the almost sure sense, i.e. for any bounded continuous function $f$*

$$\int f(x) F^{\widetilde{\mathbf{A}}^{ER}}(x)dx \xrightarrow{a.s.} \int f(x)\mu_{sc}(x,1)dx.$$

The above result can also be found in [Ding *et al.* 2010].

According to this result, whether the e.s.d. of a centered ER adjacency matrix converges to a semicircle distribution depends on how fast $p_n$ decays to zero as $n \to +\infty$. Theorem 2.1 does not apply, for e.g., when $p_n = \frac{c}{n}$ because, for this probability, Assumption 2.1 does not hold. For diluted graphs, it is known that there exists a limiting spectral distribution, for which an explicit expression is not known [Bordenave & Lelarge 2010]. For this reason, in the following, we limit our attention to probabilities $p_n \geq \frac{\log(n)}{n}$.

### 2.2.2 Spectral Norm of the Centered Adjacency Matrix

If the multiplicity of an eigenvalue does not scale with $n$, the definition of the e.s.d. implies that, in the limit for $n \to +\infty$, the e.s.d. is not able to capture the existence of this eigenvalue in the spectrum matrix. Then, Corollary 2.1 can only provide a lower bound of the spectral norm of the normalized centered ER adjacency matrix $\widetilde{\mathbf{A}}^{\text{ER}}$. Hence, it is important to find an upper bound on the spectral norm of $\widetilde{\mathbf{A}}^{\text{ER}}$ to better understand its spectral properties.

By applying Theorem 2.2 to the normalized centered adjacency matrix $\widetilde{\mathbf{A}}^{\text{ER}}$ we obtain the following concentration result.

**Lemma 2.6.** *Let us consider the normalized centered adjacency matrix $\widetilde{\mathbf{A}}^{ER}$. If the probability $p_n$ satisfies the inequality $p_n \geq C' \log^4(n) n^{-1}$ for some constant $C' > 0$, then there exists a constant $C > 0$ such that almost surely*

$$\|\widetilde{\mathbf{A}}^{ER}\|_2 \leq 2 + C\sqrt[4]{\frac{1-p_n}{np_n}}\log n. \tag{2.11}$$

*Proof.* From the definition of $\widetilde{\mathbf{A}}^{\mathrm{ER}}$ it results $\sigma = n^{-1/2}$. Then, condition $\sigma \geq C^* n^{-1/2} K \log^2(n)$ implies $K \leq (C^* \log^2 n)^{-1}$. Additionally, the bound on the elements $\widetilde{A}_{ij}^{\mathrm{ER}}$ implies $\dfrac{1-p}{\sqrt{n(1-p)p}} \leq K$. Thus,

$$\sqrt{\frac{1-p}{np}} \leq K \leq (C^* \log^2 n)^{-1}. \tag{2.12}$$

Then, $K$ exists if $\sqrt{\dfrac{1-p}{np}} \leq (C^* \log^2 n)^{-1}$ or if $p$ satisfies the more stringent constraint

$$p \geq C' n^{-1} \log^4 n,$$

where $C'$ is a constant depending on $C^*$. The inequality in (2.11) is obtained from (2.7) by setting $K = \sqrt{\dfrac{1-p}{np}}$. □

### Spectrum of the Non-centered Adjacency Matrix

In Sections 2.2.1 and 2.2.2, we focused on the spectral properties of the normalized centered ER adjacency matrix $\widetilde{\mathbf{A}}^{\mathrm{ER}}$. In this section, we analyze the spectral properties of the normalized ER adjacency matrix $\widehat{\mathbf{A}}^{\mathrm{ER}}$ and the effect of the mean component $\overline{\mathbf{A}}^{\mathrm{ER}}$ on it. The following Lemma plays a key role to establish a fundamental relation between the eigenvalue e.d.f. $F^{\widehat{\mathbf{A}}^{\mathrm{ER}}}$ studied in the previous sections and $F^{\widehat{\mathbf{A}}^{\mathrm{ER}}}$.

We recall that $\overline{\mathbf{A}}^{\mathrm{ER}} = \widehat{\mathbf{A}}^{\mathrm{ER}} - \widetilde{\mathbf{A}}^{\mathrm{ER}}$ has unit rank for any $n$. Then, by Lemma 2.2, asymptotically for $n \to \infty$, the limiting eigenvalue distribution of the matrix $\widehat{\mathbf{A}}^{\mathrm{ER}}$ converges to the semicircular law, just like the limiting eigenvalue distribution of the matrix $\widetilde{\mathbf{A}}^{\mathrm{ER}}$.

Thus the asymptotic spectrum of the adjacency matrix is the same as that of the centered adjacency matrix. However, the spectral norm is different because the largest eigenvalue changes when a unit rank matrix is added to a Hermitian matrix. From Weyl's identities for Hermitian matrices [Saad 1992], we have:

$$|\lambda_i(\widehat{\mathbf{A}}^{\mathrm{ER}}) - \lambda_i(\overline{\mathbf{A}}^{\mathrm{ER}})| \leq \|\widetilde{\mathbf{A}}^{\mathrm{ER}}\|_2 \tag{2.13}$$

for $1 \leq i \leq n$.

However, we have, $\lambda_1(\overline{\mathbf{A}}^{\mathrm{ER}}) = n\gamma(n)p_n$ and $\lambda_i(\overline{\mathbf{A}}^{\mathrm{ER}}) = 0$ for $i \geq 2$. Also, from above, we have asymptotically $\|\widetilde{\mathbf{A}}^{\mathrm{ER}}\|_2 = 2$ a.s. Thus we get the following concentration result for the largest eigenvalue of the full adjacency matrix $\widehat{\mathbf{A}}^{\mathrm{ER}}$:

$$|\lambda_1(\widehat{\mathbf{A}}^{\mathrm{ER}}) - n\gamma(n)p_n| \leq 2 \tag{2.14}$$

We notice that $n\gamma(n)p_n = \sqrt{\frac{np_n}{1-p_n}} >> 2$. Hence the above result implies that $\lambda_1(\widehat{A}^{\mathrm{ER}}) \to n\gamma(n)p_n$, or in terms of the adjacency matrix $A$, for $np_n \to \infty$, the above implies that $\lambda_1(A^{\mathrm{ER}}) \to np_n$. Thus we have the following lemma [Ding *et al.* 2010]:

**Lemma 2.7.** *For the adjacency matrix of the ER graph as described above, the largest eigenvalue $\lambda_1(A)$ satisfies the following limit theorem:*

$$\lim_{n\to\infty} \lambda_1^n(\mathbf{A}^{ER}) = np_n \ \ a.s.$$

*I.e., the largest eigenvalue of $\mathbf{A}$ tends to the largest eigenvalue of the mean matrix $\overline{\mathbf{A}}^{ER}$ as $n \to \infty$.*

## Properties of Eigenvectors

In this section we discuss some relevant results on the eigenvectors of the adjacency matrix of the ER graph available in the literature. We discussed some properties of eigenvectors of classical Wigner matrices in Section 2.1.4. However, these results only hold for matrices with zero mean entries, and thus do not apply to the adjacency matrix of the ER graph [O'Rourke *et al.* 2016]. In addition, most of these results require matrix entries with unit variance that are either bounded or have well-behaved tails (i.e., they satisfy some form of Lindeberg condition), and thus cannot be directly applied to ER graphs, when the edge probability $p_n = o(1)$. In this section, we discuss some results in the literature that handle this scenario. We discuss two results that provide a bound on the variation of the norm of the eigenvectors, i.e., show the delocalization property of eigenvectors. Next, we also discuss a result on the gaussianity of the principal eigenvector of the ER graph, which is relevant to our work on anomaly detection (Chapter 4).

In [Mitra 2009], the author derived, by means of random matrix theoretic and graph theoretic arguments, entry-wise bounds on the principal eigenvector of the adjacency matrix of an ER graph. It shows that, as expected, the principal eigenvector is close to the all one vector. We state their result in the following theorem.

**Theorem 2.10.** *[Mitra 2009, Theorem 1] Consider an ER graph $G(n,p)$ with $p \geq \log^6(n)/n$ with adjacency matrix $\mathbf{A}$. Let $\mathbf{v}$ be its principal eigenvector corresponding to the largest eigenvalue. Then,*

$$\max_i |\mathbf{v}_i - \frac{1}{\sqrt{n}}| \leq c \frac{\log(n)}{\log(np)} \frac{1}{\sqrt{n}} \sqrt{\frac{\log(n)}{np}},$$

*with high probability.*

In [Erdős *et al.* 2013], the authors make use of results on the number of eigenvalues over small intervals, known as *local semicircle law* to prove closer bounds on the variation of the principal eigenvector around the all one vector, as well as bounds on maximum component of all other eigenvectors. We state the following theorem proven in [Erdős *et al.* 2013].

**Theorem 2.11.** *[Erdős et al. 2013, Theorem 2.16] [O'Rourke et al. 2016, Theorem 6.2] For $G(n,p)$, when $p \geq (\log(n))^{6\alpha}/n$ then there exist constants such that*

$$\max_{1 \leq i \leq n-1} \|\mathbf{v}_i\|_\infty \leq \frac{(\log(n))^{4\alpha}}{\sqrt{n}}$$

*and*

$$\|\mathbf{v}_n - \frac{1}{\sqrt{n}}\mathbf{1}_n\|_\infty \leq C \frac{\log(n)^\alpha}{\sqrt{n}} \frac{1}{\sqrt{n}}$$

*with probability at least $1 - C\exp(-c\log^\alpha(n))$.*

Finally, we would like to state an important result on the gaussianity of the fluctuations of the dominant eigenvector of the adjacency matrix of ER graph around the all one vector. In [Athreya *et al.* 2013], the authors derive a central limit theorem for the dominant eigenvectors of a random graph model called the Random Dot Product graphs. The random graph product graph is a generalization of the Stochastic Block Model, of which the ER graph is a special case. We state the result below for the ER graph.

Let $\lambda, \mathbf{v}$ be the largest eigenvalue and largest eigenvector respectively of the adjacency matrix $\mathbf{A}$. Define $\mathbf{x} := \sqrt{\lambda}\mathbf{v}$. Then the following central limit theorem holds for $\mathbf{x}$.

**Proposition 2.1.** *[Athreya et al. 2013, Corollary 3.4] For an Erdős-Rényi graph, the following central limit theorem holds*

$$\sqrt{n}(\mathbf{x}_i - \sqrt{p}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1-p),$$

*when $p$ is a constant.*

In Chapter 4 we show a similar result for sparse graphs, i.e., when $p = o(1)$.

## 2.3 Introduction to Message Passing and Belief Propagation on Graphs

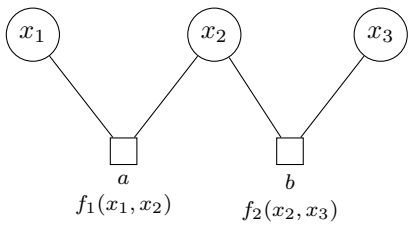
In this thesis, we develop a message passing algorithm to perform subgraph detection in the presence of side-information in Chapter 5. On the random graph model considered in that chapter, the subgraph detection problem can be solved via a bit-wise MAP detection problem on a pair-wise Ising model, which can be solved by an iterative and local algorithm called Belief Propagation (BP). In this section, we give a general introduction to BP and general message passing algorithms. The case where BP is applied to the specific problem of subgraph detection on graphs is discussed in Chapter 5.

### 2.3.1 Belief Propagation Fundamentals

Consider a graph $G = (V, E)$ with vertex set $V$ and edge set $E$. The graph is random and the edges are probabilistic functions of some variable associated with the vertices. Let $\mathbf{x} = (x_1, x_2, \dots x_n)$ be a vector such that $x_i \in M$, $M$ being a finite size alphabet, is a realization of the random variable associated with node $i$. The graph can be considered to encode the dependence structure of the random vector $\mathbf{x}$. A typical estimation problem on graphs is to find an estimate of $\mathbf{x}$ based on an observation of the graph. To solve this problem, one can estimate the marginal probabilites of each $x_i$ and perform a component-wise MAP decoding at each vertex [Mezard & Montanari 2009].

However, a naive algorithm to compute the marginal probabilities would involve summing over all $M^{|V|}$ configurations, and hence is not practical on large graphs. Nevertheless, in most practical graphs, the joint distribution has a definite structure, which can be exploited to simplify the computations. In particular, in most graphs the joint distribution of the variables can be factored into simpler terms containing a subset of the nodes. This means that instead of summing over all the variables at once, one can sum over small subsets of variables and then combine these terms in a suitable way. We demonstrate this methodology with a simple example.

Figure 2.1: Factor graph for the three variable problem



Consider a distribution function $f(x_1, x_2, x_3)$ on three variables $x_1, x_2$ and $x_3$. Say we want to compute the marginal distribution of the variable $x_2$, i.e., we want to integrate over the variables $x_1$ and $x_3$. Normally, this would take $|M|^2$ computations for each value of $x_2$. However, assume that $f(x_1, x_2, x_3)$ factors as $f(x_1, x_2, x_3) = f_1(x_1, x_2)f_2(x_2, x_3)$. Then by the distributive property of sum and product we have

$$\sum_{x_1, x_3} f(x_1, x_2, x_3) = \sum_{x_1} f_1(x_1, x_2) \sum_{x_3} f_1(x_2, x_3), \tag{2.15}$$

i.e., we can perform the same computation in $2|M|$ operations.

This dependence structure can be represented in terms of a factor graph as given in Figure 2.1. A factor graph consists of two kinds of nodes: the circles represent hidden variables and the squares represent function nodes. A function node determines the relationship between two variable nodes. In Figure 2.1, the two function nodes represent the two factors containing $x_1, x_2$ and $x_2, x_3$.

In our example, factor $a$ sends to node $x_2$ the value $\sum_{x_1} f_1(x_1, x_2)$ for each value of $x_2$, and similarly, factor $b$ sends the value $\sum_{x_3} f_1(x_2, x_3)$. At node $x_2$ these two quantities are multiplied. On the other hand, if we are interested in the marginal distribution at $x_1$, the node $x_2$ sends to node $a$, the message it received from node $x_3$, which is $\sum_{x_3} f_1(x_2, x_3)$. The factor graph is thus a graphical way of performing the distributed summation shown in (2.15).

This procedure can be extended to a general number of variables using factor graphs and the resulting algorithm is called Belief Propagation. It can be shown that BP is exact on tree graphs, i.e., graphs with no cycles [Mezard & Montanari 2009]. In this way, a computation that normally takes exponential time can be performed in linear time on trees [Mezard & Montanari 2009]. This basic idea, can in general be used, in addition to marginilization, to sample from a multi-variate distribution, to perform optimization, and to determine free-energy [Mezard & Montanari 2009]. Similarly, one can extend this procedure to other operations admitting the distributive property (2.15), such as max-product or min-sum.

We now present the general Belief Propagation iterations for computing the marginals of graphical models. Consider a general graphical model where the joint distribution $p(\mathbf{x})$ can be decomposed as follows

$$p(\mathbf{x}) \cong \prod_{a=1}^{N} \psi_a(\mathbf{x}_{\partial a}), \tag{2.16}$$

where $\cong$ represents equality up to a normalization factor, and $a$ denotes a factor with $\psi_a$ denoting the function associated with the factor and $\mathbf{x}_{\partial a}$ denotes the variables involved in the factor.

As in the previous example, we can represent the above equation in terms of a factor graph, and BP iterations can be written to compute the marginals on this graph [Mezard & Montanari 2009]. BP iterations consists of two types of messages: $\nu^t_{j\to a}$, the message sent by variable node $j$ to factor node $a$ at time $t$ and $\widehat{\nu}^t_{a\to j}$, the message sent to variable node $j$ by factor node $a$. They are defined by the following iterations [Mezard & Montanari 2009]:

$$\nu^{t+1}_{j\to a}(x_j) \cong \prod_{b\in \delta j \setminus a} \widehat{\nu}^t_{b\to j}(x_j) \tag{2.17}$$

$$\widehat{\nu}^t_{a\to j}(x_j) \cong \sum_{\mathbf{x}_{\partial a \setminus j}} \psi_a(\mathbf{x}_{\partial a}) \prod_{k\in \delta a \setminus j} \nu^{(t)}_{k\to a}(x_k), \tag{2.18}$$

where $\nu^0_{i\to a}(x_i)$ and $\widehat{\nu}^0_{a\to j}(x_j)$ are initialized to uniform distributions.

After $t$ steps, one can find an estimate of the marginal distribution at $x_i$ as

$$\widehat{p}^{\,t}(x_i) \cong \prod_{a \in \delta i} \widehat{\nu}^{\,t}_{a \to i}(x_i).$$

We provide an illustration of these messages in Figure 2.2. In this figure, we show a part of a factor graph with two factors $a$ and $b$ involving three variables each, namely $\psi_a(x_k, x_j, x_i)$ and $\psi_b(x_i, x_l, x_n)$ respectively. The messages transmitted to node $i$ by factor $a$ denoted as $\widehat{\nu}_{a \to i}(x_i)$ along with its computation from the messages transmitted to $a$ by the other two nodes $k$ and $j$ is displayed. Similarly, factor node $b$ sends its message to $i$ and finally the likelihood estimate at node $i$ is computed as $\widehat{p}(x_i)$ shown in the figure.

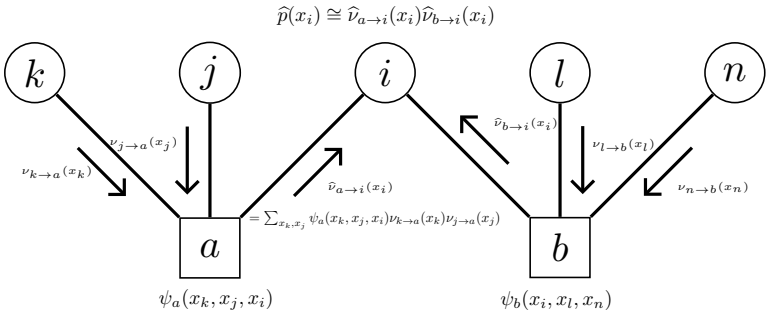



Figure 2.2: BP messages on a factor graph

A pair-wise Ising model is a specialization of (2.16) where each factor node $a$ consists of only two variables, i.e., it corresponds to a graph edge. Subgraph detection considered in this thesis has this form. In that case, BP updates can be expressed in terms of a single kind of message, for e.g. function to variable message $\widehat{\nu}^{\,t}_{a \to j}(x_j)$. We can write [Mezard & Montanari 2009]

$$\widehat{\nu}^{\,t}_{i \to j}(x_j) \cong \sum_{x_i} \psi_a(x_i, x_j) \prod_{k \in \delta i \setminus j} \widehat{\nu}^{\,(t)}_{k \to i}(x_i),$$

where we used $\widehat{\nu}^{\,t}_{i \to j}(x_j) := \widehat{\nu}^{\,t}_{(i,j) \to j}(x_j)$.

Two questions arise:

- Do the BP equations (2.17,2.18) converge on a general graph?
- If yes, are the final converged messages equal to the true marginals?

It is known that when the factor graph is a tree, the updates in (2.17, 2.18) converge to the true marginal in number of steps equal to the depth of the tree [Mezard & Montanari 2009]. However, on a general graph, these questions are still unanswered, but some heuristics are known. In general, for the messages to converge to the correct marginals, the following two graph properties must be satisfied. The details are in [Mezard & Montanari 2009].

1. *Locally tree-like property*: If a graph is locally tree-like, then the neighbourhood of a certain size $t$ of any vertex has no cycles. This implies that the messages transmitted by the neighbours of the node will be independent up to depth $t$.

2. *Small long-range correlations*: Since BP messages are functions on a local neighbourhood of a node up to depth $t$, for BP to be able to give approximately the correct marginals, it must be true that the local marginals are not affected by variables far away from the node, i.e., the correlations between variables should die down fast enough.

Chapter 3

# Spectral Functions of the Stochastic Block Model

## 3.1 Introduction

One of first random graph models to be deeply explored is the Erdős-Rényi graph [Erdős & Rényi 1959] where edges between nodes appear with equal probabilities. As described in the past chapters, this model has many appealing analytical properties, but it does not model important features of many real complex networks. In particular, the Erdős-Rényi graph fails to describe clustering and the presence of communities in complex networks. To mitigate this shortcoming, the more refined Stochastic Block Model (SBM) was introduced in [Holland *et al.* 1983]. In SBM, the nodes are classified into subsets which model communities. Two nodes that belong to the same community are connected to each other with a higher probability than two nodes belonging to different communities.

SBM has been used to show consistency of community detection algorithms such as *spectral clustering* [Rohe *et al.* 2011, Sussman *et al.* 2012]. To detect the communities in a graph with, say $M$ communities, spectral clustering works in two steps. First it computes $M$ eigenvectors corresponding to the $M$ largest eigenvalues of some appropriate graph matrix, usually the adjacency matrix or its normalized form. The nodes of the graph are then embedded on $\mathbb{R}^M$, by taking the coordinate of each vertex to be the corresponding elements of the $K$ eigenvectors computed in the previous step. On this space, the algorithm then performs k-means clustering or Expectation Maximization (EM) to determine the clusters [Filippone *et al.* 2008].

The success of spectral clustering is hinged on the presence of a few eigenvalues considerably larger than all others. Thus it is important to study the spectral properties of an SBM graph, more importantly, the dominant eigenvalues and corresponding eigenvectors of the adjacency matrix [Nadakuditi & Newman 2012]. Spectral properties of a graph has also been applied to the study of epidemic processes on a graph; for e.g., the cost of epidemic spread is characterized by the spectral properties of the adjacency matrix [Bose *et al.* 2013].

In this chapter, using well known tools in Random Matrix Theory, we analyze the limiting empirical distribution of the eigenvalues of the adjacency matrix of SBM. We derive a fixed point equation for the Stieltjes transform of the limiting eigenvalue empirical distribution function (e.d.f.) and provide an explicit expression of the asymptotic eigenvalue distribution in the case of symmetric communities. Further, we obtain concentration bounds on the extreme eigenvalues. Additionally, we derive parallel results for the normalized Laplacian Matrix and discuss potential applications of the general results in epidemics and random walks. Furthermore, we analyze a modified spectral function that takes into account the eigenvectors of the adjacency matrix of SBM.

SBM in a symmetric setting (i.e., probability of interconnections within communities are identical) has been studied in the context of community detection for two-community graphs in [Heimlicher *et al.* 2012, Decelle *et al.* 2011, Nadakuditi & Newman 2012]. In [Decelle *et al.* 2011], the authors investigate the detectability of communities in a two-community

SBM graph by analysing the phase transition in the spectrum of the adjacency matrix using methods from statistical physics. In [Heimlicher *et al.* 2012], the authors analyze a similar problem in the context of labelled Stochastic Matrices and provide theoretical evidence for detectability thresholds in [Decelle *et al.* 2011]. Here the nodes are randomly categorized into communities and the goal is to find the correct community to which a node belongs. In [Nadakuditi & Newman 2012] the authors derive the spectrum of the adjacency matrix of two community SBM when the probabilities of connection within a community is the same for all communities(symmetric scenario), and show the existence of a threshold beyond which community detection is impossible. Stieltjes transforms of spectral measures of various random graphs was also studied in [Bordenave & Lelarge 2010].

## 3.2 Stochastic Block Model and its Representations

We consider a SBM with $n$ nodes and $M$ communities $C_m$, for $m = 1, \dots, M$, of equal sizes $K$. Each node belongs to one of the $M$ communities such that $n = MK$. If two nodes belong to two different communities, then there is an edge between them with probability $p_0(n)$, which, in general, is a function of the size of the network $n$. Given two nodes belonging to the same community $C_m$, there exists an edge between them with probability $p_m(n)$, $1 \leq m \leq M$. Throughout this chapter, for the sake of conciseness, we adopt the short notation $p_m$ for the probabilities $p_m(n)$ where the dependence on $n$ is implicit. For a random graph as defined above, we can define a number of related random matrices whose spectral characteristics are relevant to capture related properties of the network. In this work we focus on the spectra of two types of random matrices: the adjacency matrix and the normalized Laplacian matrix.

*SBM adjacency matrix* $\mathbf{A}$

Without loss of generality, we assume that nodes belonging to the same community are clustered together and ordered from community 1 to community $M$, i.e. node $i$ belongs to community $C_m$ if $\left\lceil \frac{i}{K} \right\rceil = m$. Natually, $\mathbf{A}$ is symmetric matrix and its component $A_{ij}$ is a Bernoulli random variable (rv) with parameter $p_m$, $m = 1, \dots M$, if the corresponding nodes $i$ and $j$ belong to the community $C_m$, i.e. $\left\lceil \frac{i}{K} \right\rceil = \left\lceil \frac{j}{K} \right\rceil = m$, and with parameter $p_0$ otherwise. Let us denote by $\mathcal{B}(p_m)$ a Bernoulli probability distribution with parameter $p_m$. Then,

$$\begin{cases} A_{ij} = A_{ji} \sim \mathcal{B}(p_m), & \text{if } i, j \in C_m \\ A_{ij} = A_{ji} \sim \mathcal{B}(p_0), & \text{if } i \in C_\ell \text{ and } j \in C_m \text{ with } \ell \neq m. \end{cases} \tag{3.1}$$

In our definition of $\mathbf{A}$ we allow it to have non-zero diagonal elements, i.e., self-loops are permitted. It is worth noting that the results on asymptotic spectrum of adjacency matrices in this contribution hold independent of the assumption on the diagonal elements, since their contribution is $O(1)$.

Henceforth, it is convenient to normalize $\mathbf{A}$ by a scaling factor[1] $\gamma(n)$, depending on $n$, such that the support of the limiting eigenvalue distribution function stay finite and positive. Then, we consider the normalized SBM adjacency matrix $\widehat{A} = \gamma(n)A$ and we express it as the sum of a deterministic matrix $\overline{\mathbf{A}} = \mathbb{E}(\mathbf{A})$ a random matrix with zero mean random entries denoted as $\widetilde{\mathbf{A}}$, i.e.

$$\widehat{\mathbf{A}} = \overline{\mathbf{A}} + \widetilde{\mathbf{A}}. \tag{3.2}$$

[1] We use the short notation $\gamma$ when it is not necessary to emphasize the dependency on $n$.

In accordance with the definitions in (3.1) and (3.2), $\overline{\mathbf{A}}$ is a finite rank matrix of the following form:

$$\overline{\mathbf{A}} = \mathbf{B} \otimes \mathbf{J}_K,$$

where $\mathbf{B}$ is an $M \times M$ matrix given as

$$\mathbf{B} = \gamma(n) \begin{pmatrix} p_1 & p_0 & \dots & p_0 \\ p_0 & p_2 & \ddots & p_0 \\ \dots & & \ddots & \dots \\ p_0 & \dots & \dots & p_M \end{pmatrix}.$$

In general, since $p_m \neq p_0$ for $m = 1, \dots, M$, matrix $\mathbf{B}$ has rank $M$, and so does $\overline{\mathbf{A}}$.

The random centered SBM adjacency matrix is also a symmetric matrix whose elements follow the distributions

$$\mathcal{C}(p_m, \gamma) \quad = \quad \begin{cases} \gamma(1-p_m), & \text{w.p.} \quad p_m; \\ -\gamma p_m, & \text{w.p.} \quad 1-p_m; \end{cases} \qquad m \quad = \quad 0, 1, \dots, M, \tag{3.3}$$

with zero mean and variance $\sigma_m^2 = \gamma^2(1-p_m)p_m$. Consistently, with the definitions in (3.1) and (3.2),

$$\begin{cases} \widetilde{A}_{ij} = \widetilde{A}_{ji} \sim \mathcal{C}(p_m, \gamma) & \text{if } i, j \in C_m \\ \widetilde{A}_{ij} = \widetilde{A}_{ji} \sim \mathcal{C}(p_0, \gamma) & \text{if } i \in C_\ell \text{ and } m \in C_m \text{ with } \ell \neq m. \end{cases} \tag{3.4}$$

*Normalized Laplacian matrix* $\boldsymbol{\mathcal{L}}$

Let define the degree of node $i$ as

$$D_i = \sum_{j=1}^{n} A_{ij} \tag{3.5}$$

Then, the symmetric random SBM normalized Laplacian matrix $\boldsymbol{\mathcal{L}}$ is defined as

$$\mathcal{L}_{ij} = \mathcal{L}_{ji} = \begin{cases} 1 - \dfrac{A_{ii}}{D_i}, & \text{if } i = j; \\ -\dfrac{A_{ij}}{\sqrt{D_i D_j}}, & \text{otherwise.} \end{cases} \tag{3.6}$$

## 3.3 Empirical Spectral Distribution: Distribution of Eigenvalues

In this section we analyze the empirical spectral distribution (e.s.d.) of $\mathbf{A}$, as defined in (3.1). In the following subsection we deal with the adjacency matrix, and later with the normalized Laplacian matrix.

### 3.3.1 Results for Adjacency Matrix of M community Model

As done for ER graphs, we do this in two stages. First, we characterize the centralized adjacency matrix and then the full adjacency matrix.

#### 3.3.1.1 Finding the spectrum of centered Adjacency Matrix

We apply the following theorem to the centered Adjacency matrix.

**Theorem 3.1.** *[Girko 2001, Girko 1990, Chapter 1] For a symmetric random matrix* $\mathbf{W}$*, with* $\mathbb{E}W_{ij} = 0$ *and* $\mathbb{E}|W_{ij}|^2 = \sigma_{ij}$*, where* $W_{ij}$ *are independent random variables for* $1 \leq i \leq j \leq n$*, satisfying :* $\sup_n \max_{i=1,2,..n} \sum_j \sigma_{ij}^2 < \infty$ *and* $n\sigma_{ij} = c > 0$*, and Lindeberg condition: for any* $\tau > 0$

$$\lim_{n\to\infty} \max_{i=1,2,3..n} \sum_{j=1}^{n} \mathbb{E}|W_{ij}|^2 \chi_{|W_{ij}|\geq\tau} = 0 \tag{3.7}$$

*then, the e.s.d. is the inverse Stieltjes transform of* $s(z)$ *given by:*

$$s(z) = \frac{1}{n}\sum_{i=1}^{n} c_i(z)$$

*where* $c_i(z)$ *satisfies:*

$$c_i(z) = \left\{ \left[ -zI - \left( \delta_{pl} \sum_{s=1}^{n} c_s(z)\sigma_{sl} \right)_{p,l=1}^{n} \right]^{-1} \right\}_{ii},$$

$$i = 1, 2, ..n$$

*Note: The matrix* $W$ *in our example has mean 0, unlike the matrix in Girko's theorem.* Then we have the following corollary.

**Lemma 3.1.** *Let* $\widetilde{\mathbf{A}}$ *be the normalized centered SBM adjacency matrix with* $\gamma(n) = (np_1(1-p_1))^{-1}$*. If* $p_m(n) \in \omega(n^{-1})$ *and* $p_m(n) = O(p_0(n))$ *for all* $m = 1, \dots M$*, then, almost surely, the eigenvalue e.d.f. converge weakly to a distribution function whose Stieltjes transform is given by*

$$s(z) = c_1(z) + c_2(z) \tag{3.8}$$

$c_1(z), c_2(z)$ *being the unique solutions to the system of equations:*

$$c_i(z) = \frac{-1/2}{z + \varsigma_i c_i(z) + \varsigma_0 c_j(z)}, \tag{3.9}$$

*with* $i, j = 1, 2$ *and* $i, j = 2, 1$*, and where* $\varsigma_i = \lim_{n\to+\infty} \frac{p_i(1-p_i)}{p_1(1-p_1)}, i = 1, 2$ *that satisfies the condition that for each* $i = 1, 2$*,*

$$\Im(c_i(z))\Im(z) > 0 \text{ for } \Im z > 0. \tag{3.10}$$

*Proof.* For the matrix under consideration, if all variances of the SBM, $\sigma_i^2 = p_i(1-p_i)$ satisfy, $n\sigma_i^2 \to \infty$, and if we choose $\gamma(n) = 1/\sqrt{np^*(1-p^*)}$, where $p^*$ is such that $\sigma^{*2} = p^*(1-p^*) > \sigma_i^2$ for all $i = 0, 1, 2..M$, then the conditions in the above theorem are satisfied and we have $s(z)$, the Stieltjes transform of limiting e.s.d. of centralized adjacency matrix $\widetilde{A}$ (after undoing the scaling by $\sigma^*$ ) [2], given as

$$s(z) = \sum_{i=1}^{M} c_i(z) \tag{3.11}$$

[2] If we do not undo the scaling the variances below would be scaled by $\sigma^*$

where the following relation exists between $c_i(z)$'s.

$$c_i(z) = \frac{-1/M}{z + \sigma_i^2 c_i(z) + \sigma_0^2 \sum_{j \neq i} c_j(z)} \tag{3.12}$$

The valid solution must satisfy [Girko 2001]:

$$\Im c_i(z) z > 0 \text{ for } \Im z > 0 \tag{3.13}$$

Solving these equations would yield the Stieltjes transform of the limiting e.s.d. of the Adjacency matrix. *Note*: We need that each probability $p_i$ grows at the same rate: i.e., $\frac{p_i}{p_j} = O(1)$ for any $i, j$. □

#### 3.3.1.2 Spectrum of the full Adjacency Matrix

The result above gives the spectrum of matrix $\widetilde{A}$. Recall that

$$\mathbf{A} = \widetilde{\mathbf{A}} + \overline{\mathbf{A}}$$

where $\overline{\mathbf{A}}$ is the normalized mean matrix. Using Lemma 2.2 on the finite rank perturbation of a matrix, we deduce that the asymptotic spectrum of $\overline{\mathbf{A}}$ is the same as that of $\widetilde{\mathbf{A}}$. The difference however lies in the extreme eigenvalues.

#### 3.3.1.3 Extreme Eigenvalue of Adjacency Matrix

For matrices $\overline{\mathbf{A}}$, $\mathbf{A}$ and $\overline{\mathbf{A}}$, we have:

$$|\lambda_i(\mathbf{A}) - \lambda_i(\overline{\mathbf{A}})| \leq \|\widetilde{\mathbf{A}}\|_2,$$

by Weyl's identities. This is useful in getting an asymptotic characterization for the top $M$ eigenvalues of $\mathbf{A}$. Since $\overline{\mathbf{A}}$ has exactly $M$ non-zero eigenvalues, this result says that the $M$ largest eigenvalues of $\mathbf{A}$ are concentrated around these eigenvalues, within an error of the spectral norm of $\widetilde{\mathbf{A}}$. The rest of the eigenvalues are confined to the bulk of the spectrum of $\widetilde{\mathbf{A}}$. To use this result, we need a bound on the spectral norm of $\widetilde{\mathbf{A}}$, the zero mean part of $\mathbf{A}$. We use the methodology in [Vu 2007, Theorem 1.4] to derive a bound on the spectral norm of this matrix. The result is in the following lemma.

**Lemma 3.2.** *There are constants $C$ and $C'$ such that the following holds. Let $\widetilde{\mathbf{A}}$ be the centered and scaled adjacency matrix of a graph with $M$ communities such that $\sigma^2 = M^{-1}(\max_i\{\sigma_i^2\} + (M-1)\sigma_o^2)$. Then if $\sigma^2 \geq C' n^{-1} K \log^4(n)$*

$$\|\widetilde{\mathbf{A}}\|_2 \leq (2\sigma\sqrt{n} + C(K\sigma)^{1/2} n^{1/4} \log(n))\gamma(n)$$

Where $\gamma(n) = (\sqrt{np^*(1-p^*)})^{-1}$. This gives a direct relationship between the individual variances and the value of the edge in a way analogous to the result for the standard Wigner Matrix.

*Proof.* We follow the ideas in [Vu 2007, Theorem 1.3]. We make use of the moment method to bound the spectral norm of $\widetilde{\mathbf{A}}' = \widetilde{\mathbf{A}}/\gamma(n)$. We use the idea that spectral norm, which is the largest dominating eigenvalue in absolute value, can be bounded by the trace of the matrix raised to an even exponent, and that the larger the exponent, the sharper the bound:

$$\|\widetilde{\mathbf{A}}'\|_2^{2k} = \max_{1 \leq i \leq n} |\lambda_i(\widetilde{\mathbf{A}}')|^{2k} \leq (\sum_{i=1}^{n} |\lambda_i(\widetilde{\mathbf{A}}')|^{2k}) = (\text{tr}(\widetilde{\mathbf{A}}')^{2k})$$

Once we obtain a bound on the expected spectral norm, we can use Markov inequality, to bound the tail probabilities.

$$\mathbb{P}\left(\|\widetilde{\mathbf{A}}'\|_2 \geq \lambda\right) \leq \frac{\mathbb{E}\|\widetilde{\mathbf{A}}'\|_2^{2k}}{\lambda^{2k}} \leq \frac{\mathbb{E}\mathrm{tr}(\widetilde{\mathbf{A}}')^{2k}}{\lambda^{2k}} \tag{3.14}$$

If $\mathbf{A}$ is a standard Wigner matrix, for fixed $k$, the right hand side of the above equation is $n$ times the $2k^{\text{th}}$ moment of the empirical spectral distribution, which by the semicircle law tends to $C_k$. Therefore, for such matrices, if $k$ were chosen to be a fixed number independent of $n$, the right hand side tends to infinity for large $n$, making it useless. Therefore we choose $k$ to be a function of $n$ [Vu 2007]. The idea is that when $k$ is a slowly increasing function of $n$, the semicircle law still holds, and since $C_k \leq 4^k$, the upper bound tends to 0, for any $\lambda \geq 2$. Here, we extend this idea to a Wigner matrix displaying community structure.

We need to find a bound on the quantity $\mathbb{E}\mathrm{tr}(\widetilde{\mathbf{A}}')^{2k}$. To do this we expand the trace as a summation of expectation over cycles of length $2k$ of vertices in the set $\{1, 2, 3, ...n\}$

$$\mathbb{E}\mathrm{tr}(\widetilde{\mathbf{A}}')^{2k} = \mathbb{E} \sum_{i_1, i_2, i_3, \ldots i_{2k}} \widetilde{\mathbf{A}}'_{i_1,i_2} \widetilde{\mathbf{A}}'_{i_2,i_3} \ldots \widetilde{\mathbf{A}}'_{i_{2k},i_1}$$

where $\{i_1, i_2 ... i_{2k}, i_1\}$ form a cycle over edges such that $i_j \in \{1, 2, 3, ...n\}$, for each $1 \leq j \leq 2k$. Each edge $\{i_{j-1}, i_j\}$ corresponds to a random variable $\widetilde{\mathbf{A}}'_{i_{j-1},i_j}$.

We can partition the graphs based on the number of unique vertices that appear in the graph, called the weight of the graph, denoted by $t$, $1 \leq t \leq 2k$. We can then represent the original graph on $2k$ edges and $2k$ vertices equivalently by using a condensed undirected connected graph on $t$ vertices and the specific order of edges traversed can be represented as a walk on these $t$ vertices. An edge exists in this walk if and only if it exists in the original graph and if it exists more than once in the walk, then this edge has a weight equal to the number of times this edge is traversed by the walk. Since each such random variable $A_{ij}$ is zero mean and by independence, if an undirected edge $\{i_{j-1}, i_j\}$ has a weight equal to 1, i.e., it appears only once in the walk, then the corresponding term is trivially zero. So we need only consider the contribution of those walks that have every edge appearing at least twice.

By virtue of independence and zero mean property, if $t$ is greater than $k+1$, and because the number of edges required for connectivity is greater than or equal to $t+1$, there must at least be $k+1$ edges in the graph. Since the total number of edges is $2k$ in the walk, this means there exists an edge that appears only once, making the contribution of such a term zero. Hence we must have $1 \leq t \leq k+1$. Thus we can bound the quantity of interest as below:

$$\mathbb{E}\mathrm{tr}(\widetilde{\mathbf{A}}')^{2k} \leq \sum_{t=1}^{k+1} \sum_{G \in \mathcal{G}_{t,n,2k}} \mathbb{E}\widetilde{\mathbf{A}}'_G$$

where $\mathcal{G}_{t,n}$ represents a set of graphs on $t$ vertices drawn from $\{1, 2, ..n\}$ with $2k$ edges.

Similar to [Vu 2007], an edge $e = \{i_{j-1}, i_j\}$ such as described above, is called an innovation edge, if the vertex $i_j$ is such that $i_j \notin \{i_1, i_2, ...i_{j-1}\}$, i.e., it appears for the first time in $e$. The other edges in the graph either overlap the innovation edges or are interconnections between two vertices that already exist in the graph. Since in our case the random variables are bounded in absolute value by 1 (since they are Bernoulli), the contribution over all the edges other than the innovation edges can be bounded by 1. For any graph on $t$ vertices, there must be exactly $t-1$ innovation edges and each edge must have at least weight 2. The contribution to the expectation of each such edge would have a weight of atmost $\sigma_i^2$, $0 \leq i \leq M$ depending on whether the edge is between two communities or within

some community $i$. This bound is exact if each such edge has a weight two; if it has weight more than two, then this is an upper bound. Then, by independence, the contribution of the group of edges can be bounded by the product. Therefore, the following is true:

$$|\mathbb{E}\widetilde{A}'_{i1,i2}\widetilde{A}'_{i2,i3}\dots\widetilde{A}'_{i_{2k},i_1}|\leq (\max_{1\leq i\leq M}\sigma_i^2)^{(t-1-i)}(\sigma_0^2)^{(i)}$$

where $i$ are integers such that $0\leq i\leq t-1$.

For every term such as the above, there are $\binom{t-1}{i}$ ways in which the edges can be chosen to be mapped to $\sigma_0^2$ or $\max_i\sigma_i^2$. This corresponds to choosing out of $t$ edges, $i$ edges such that the vertices of those edges belong to two different communities. Once we choose such $i$ edges we need to choose the communities from which the corresponding vertices emerge. For convenience, we can assume the vertices have a preferred ordering. The first such vertex can then be chosen from any of the $M$ communities. Once such a community is chosen, if the next edge is upper bounded by $\max_i\sigma_i^2$, the next vertex of the edge can be chosen only in 1 way, because this corresponds to an edge belonging to the same community. If the edge is bounded by $\sigma_o^2$, then the community to which the next vertex belongs can be chosen in atmost $M-1$ ways, since this edge corresponds to an edge between communities. Corresponding to each selection of a community to which the edge can belong, the vertices can be chosen in $(n/M)^{(t)}$ ways.

Therefore, we can finally bound the full term as follows:

$$\mathbb{E}\text{tr}\mathbf{A}^{2k}\leq\sum_{t=1}^{k+1}M(n/M)^t\sum_{i=0}^{t-1}\binom{t-1}{i}(M-1)^{(t-1-i)}(\sigma_0^2)^{(t-1-i)}(\max_{1\leq j\leq m}\sigma_j^2)^i W(k,t).$$

The inner summation on the variable $i$ turns out to be the Binomial expansion. $W(k,t)$ is the number of equivalent graphs of $t$ fixed vertices with $2k$ edges and is related to the number of the Catalan number $C_t$. We use a bound on this quantity from [Vu 2007]:

$$W(k,t)\leq\binom{2k}{2p-2}p^N 2^{k+N+1}(N+2)^N$$

where $N=2k-2(t-1)$. Finally we get:

$$\mathbb{E}\text{tr}\mathbf{A}^{2k}\leq\sum_{t=0}^{k+1}n^t(\sigma^2 n)^{t-1}W(k,t)$$

where $\sigma^2=\frac{1}{M}(\max_i\sigma_i^2+(M-1)\sigma_o^2)$. As in [Vu 2007] it can be shown that when $2k=a\sigma^{1/2}n^{1/4}$, for some $a$, the term within the summation is bounded by a geometric series with growth factor $1/2$. Using this fact we finally get:

$$\mathbb{E}\text{tr}(\widetilde{\mathbf{A}})^{2k}\leq 2n(2\sigma\sqrt{n})^{2k}$$

Substituting in equation (3.14), and using
$\lambda=2\sigma\sqrt{n}+C(\sigma)^{1/2}n^{1/4}\log(n)$ we have:

$$
\begin{aligned}
&\mathbb{P}\left(\|\widetilde{\mathbf{A}}\|_2 \geq 2\sigma\sqrt{n} + C(\sigma)^{1/2}n^{1/4}\log(n)\right) \\
&\leq 2n\left(\frac{2\sigma\sqrt{n}}{2\sigma\sqrt{n} + C(K\sigma)^{1/2}n^{1/4}\log(n)}\right)^{2k} \\
&= 2n(1 - \frac{C\sigma^{1/2}n^{1/4}\log(n)}{2\sigma\sqrt{n} + C\sigma^{1/2}n^{1/4}\log(n)})^{2k} \\
&\leq 2n(1 - \frac{C\sigma^{1/2}n^{1/4}\log(n)}{3\sigma\sqrt{n}})^{2k} \\
&\leq 2n\exp(-\frac{c\log(n)k}{3\sqrt{\sigma}n^{1/4}}) \\
&= 2n\exp(-ca\log(n)/3),
\end{aligned}
\tag{3.15}
$$

where the second inequality above follows from the assumption that $\sigma \geq C^{'}n^{-1/2}\log^2(n)$, the third inequality because $e^{-x} \geq 1 - x$, and the last equality by the form of $k$. Now since the right hand side is summable in $n$ for appropriate constants $c$ and $a$, by Borel-Cantelli Lemma [Billingsley 2008], we have:

$$\|\widetilde{\mathbf{A}}\| \leq 2\sigma\sqrt{n} + C(\sigma)^{1/2}n^{1/4}\log(n) \text{ a.s.}$$

for $\sigma \geq C^{'}n^{-1/2}K\log^2(n)$.

This means that for the above $\sigma$:

$$\|\widetilde{\mathbf{A}}\| \leq 2\sigma\sqrt{n}(1+\delta) \tag{3.16}$$

where $\delta$ is a vanishing error, for large $n$, whenever $\sigma >> C^{'}n^{\frac{-1}{2}}\log^2(n)$, or $p >> C^{'}n^{-1}\log^4(n)$. Thus it follows that the spectral norm of the zero mean matrix $\widetilde{\mathbf{A}}^{'}$ is bounded by $2\sigma\sqrt{n}$ asymptotically analogously to the Wigner case. □

#### 3.3.1.4 Eigenvalues of the mean matrix

By the above result on the spectral norm of the zero mean matrix, we know that the largest eigenvalue of the matrix is somewhere close to the edge of the spectrum. But when the mean matrix is added to this matrix, the largest eigenvalue escapes the bounded spectrum. Namely, since the mean matrix has rank $M$, by interlacing inequalities on the sum of two Hermitian matrices, we can see that there are exactly $M$ eigenvalues outside the bounded spectrum.

$$\mathbf{A} = \widetilde{\mathbf{A}}^{'} + \overline{\mathbf{A}}/\gamma(n)$$

By Weyl's identities we have

$$|\lambda_i(\mathbf{A}) - \lambda_i(\overline{\mathbf{A}})/\gamma(n)| \leq \|\widetilde{\mathbf{A}}^{'}\| \tag{3.17}$$

We have that $||\widetilde{\mathbf{A}}^{'}|| \leq 2\sigma\sqrt{n}(1+\delta)$ asymptotically a.s. as shown above. For $i > M$, $\lambda_i(\overline{\mathbf{A}}) = 0$. Therefore, we see that $\lambda_i(\mathbf{A})$ for $i > M$ lies within the continuous band.

#### 3.3.1.5 Eigenvalues of $\overline{\mathbf{A}}$

The mean matrix can be written as a Kronecker product between a matrix that is a perturbed diagonal matrix, and an all one matrix of size $n/M$ ($\mathbf{J}_K$).

$$\overline{\mathbf{A}} = \gamma(n)\mathbf{B} \otimes \mathbf{J}_K,$$

where $\mathbf{D}_p$ is defined as : $\mathbf{B} = \mathbf{D} + p_0\mathbf{I}$, where $D_{ij} = (p_i - p_0)\delta_{ij}$.

Let the eigenvalues of $\mathbf{B}$ be $\mu_i, 1 \leq i \leq M$. They depend on the probabilities $p_i$, $0 \leq i \leq M$. Then the eigenvalues of $\overline{\mathbf{A}}$ are given as:

$$\lambda_i(\overline{\mathbf{A}}) = \frac{n\mu_i}{M}\gamma(n)$$

So we have the following relationship on the eigenvalues of $\mathbf{A}$:

**Lemma 3.3.** *The $M$ eigenvalues of $\mathbf{A}$, outside the continuous spectrum of $\mathbf{A}$ are given as:*

$$|\lambda_i(\mathbf{A}) - \frac{n}{M}\mu_i| \leq 2\sigma\sqrt{n}(1+\delta) \tag{3.18}$$

*for* $1 \leq i \leq M$

To complete this argument, we need approximate locations of $\mu_i$'s. By Gershgorin disc theorem [Saad 1992], the $\mu_i$'s should satisfy:

$$|\mu_i - p_i| \leq p_0(M-1) \tag{3.19}$$

In addition, if $p_0$ is small enough so that the individual discs do not overlap, then we must have by Gershgorin theorem, that there is a single eigenvalue in each of the discs [Saad 1992]. Additionally, the variation of the eigenvalues around the diagonal entries, can be further controlled by using Kato-Temple's theorem [Saad 1992]. Hence we have:

**Lemma 3.4.** *Approximate Eigenvalues of $\mathbf{B}$ When $p_0$ is such that $p_0 < \frac{\min_i|p_i - p_0|}{2(M-1)}$, then the following is true about the eigenvalues of $D_p$, $\mu_i$:*

$$|\mu_i - p_i| \leq p_0 \tag{3.20}$$

### 3.3.2 Spectral Distribution of Normalized Laplacian Matrix

The normalized Laplacian Matrix as defined in (1.3) is

$$\boldsymbol{\mathcal{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} \tag{3.21}$$

For the sake of simplicity, we consider the case of two blocks, i.e., $M = 2$, and probabilities $p_i, 0 \leq i \leq 2$ that are not dependent on the matrix size $n$.

Let $\mathbf{P}^{'} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$.

We show that asymptotically, the e.s.d. of matrix $\frac{1}{2}\sqrt{n}\mathbf{P}^{'}$ converges to that of matrix $\frac{1}{\sqrt{n}}\mathbf{A}^{''}$, which is derived from the adjacency matrix by scaling its elements by appropriate constants in each block.

$$A^{''}_{ij} = \begin{cases} A_{ij}/(p_1+p_0), & \text{if } i,j \in C_1 \\ A_{ij}/(p_2+p_0), & \text{if } i,j \in C_2 \\ A_{ij}/\sqrt{(p_1+p_0)(p_2+p_0)} & \text{otherwise} \end{cases}$$

Matrix $\mathbf{A}^{'}$ is similar to $\mathbf{A}$, i.e., it is a symmetric matrix with independent upper triangular entries, and is a $2\times2$ block matrix with each block containing elements of the same variance.

Consequently, the following holds:

**Lemma 3.5.** *The distribution of matrix $\frac{1}{2}\sqrt{n}\mathbf{P}^{'}$ is given by:*

$$c_i = \frac{-1/2}{z + \sigma^{'}_i c_i + \sigma^{'}_0 c_j} \tag{3.22}$$

*for* $i,j = 1,2$ *and* $i,j = 2,1$ *respectively. where* $\sigma_1^{'} = \frac{\sigma_1^2}{(p_1+p_0)^2}$, $\sigma_2^{'} = \frac{\sigma_2^2}{(p_2+p_0)^2}$, $\sigma_0^{'} = \frac{\sigma_0^2}{(p_0+p_1)(p_0+p_2)}$ *and the limiting distribution has a spectrum whose Stieltjes transform is given by* $c(z) = c_1(z) + c_2(z)$.

Since $\frac{\sqrt{n}}{2}\boldsymbol{\mathcal{L}} = \frac{\sqrt{n}}{2} - \frac{\sqrt{n}}{2}\mathbf{P}'$, its distribution has a bulk component that lies around $\sqrt{n}/2$, with an approximate width of $2\sqrt{\max(\sigma_1^{'}, \sigma_2^{'}) + \sigma_0^{'}}$. This matrix also has an eigenvalue at 0, by the property of Laplacian.

*Proof.* We follow the steps in the proof of Theorem 1.1 in [Bordenave *et al.* 2010]. The first step is a form of uniform strong law of large numbers called Kolomogorov-Marcinkiewicz-Zygmund strong law of large numbers. Since the elements of the matrix $\mathbf{A}$ are independent and have finite fourth moments from Lemma 2.3 in [Bordenave *et al.* 2010] we have the following as true:

$$\sum_{j=1}^{n} A_{ij} = \frac{n}{2}(p_1 + p_0 + \delta_i^{(1)}) \tag{3.23}$$

where $\max_i|\delta_i^1| = o(1)$ for $1 \le i \le n/2$ and

$$\sum_{j=1}^{n} A_{ij} = \frac{n}{2}(p_2 + p_0 + \delta_i^{(2)}) \tag{3.24}$$

where $\max_i|\delta_i^2| = o(1)$ for $1 + n/2 \le i \le n$.

From equations (3.23) and (3.24) we have uniform convergence for $1 \le i \le n$, with the error, $\max_i(\delta_i^{(1)}, \delta_i^{(2)}) = o(1)$:

$$d_i = \sum_{j=1}^{n} A_{ij} = \frac{n}{2}(p_k + p_0) + \varepsilon_i \tag{3.25}$$

where $p_k = p_1$ if $i \in C_1$ and $p_k = p_2$ if $i \in C_2$, and $\max_i|\varepsilon_i| = \varepsilon = o(1)$ uniformly.

Next step is to use Hoffman-Wielandt inequality [Anderson *et al.* 2009] to bound the error between the e.s.d. of $\frac{A^{''}}{\sqrt{n}}$ and $\frac{\sqrt{n}}{2}P^{'}$.

We have, using Hoffman-Wielandt inequality and the bound on Stieltjes transforms found in [Bai 1999],

$$|s_{F^{\frac{1}{\sqrt{n}}\mathbf{A}''}}(z) - s_{F^{\frac{\sqrt{n}}{2}P'}}(z)| \le \frac{c}{n\Im z}\sum_{ij}|\frac{\sqrt{n}}{2}P_{ij}^{'} - \frac{1}{\sqrt{n}}A_{ij}^{''}|^2$$

where $z \in \mathbb{C}^+$.
For any $i,j$, we have:

$$\begin{aligned}
\frac{\sqrt{n}}{2}P_{ij}^{'} &= \frac{\sqrt{n}}{2}\frac{A_{ij}}{\sqrt{d_i d_j}} \\
&= \frac{\sqrt{n}}{2}\frac{A_{ij}}{\sqrt{n/2(p_k + p_0 + \varepsilon_i)n/2(p_l + p_0 + \varepsilon_j)}} \\
&= \frac{A_{ij}}{\sqrt{n(p_k + p_0)(p_l + p_0)}}(1 + O(\varepsilon_i))(1 + O(\varepsilon_j)) \\
&= \frac{A_{ij}^{''}}{\sqrt{n}}(1 + O(\varepsilon)),
\end{aligned}$$

where $i \in C_m$ and $j \in C_n$ and $\varepsilon$ is infinitesimally small. The last equality follows because $\varepsilon_i$ and $\varepsilon_j$ tend to 0 uniformly for large $n$. Thus by Hoffman Wielandt inequality we have:

$$|s_{F^{\frac{1}{\sqrt{n}}\mathbf{A}''}}(z) - s_{F^{\frac{\sqrt{n}}{2}\mathbf{P}'}}(z)| \leq \frac{c}{n^2 \Im z} \sum_{ij} |A''_{ij}|^2 O(\varepsilon^2) \to 0 \text{ a.s.}$$

The last relation follows from the strong law of large numbers on variables $A''_{ij}$ and since $\varepsilon = o(1)$. Hence we have the result. □

## 3.4 Modified Empirical Spectral Distribution: Eigenvector Distribution

In this section, our contribution is to study the properties analyzed in [Bai & Pan 2012] for the eigenvectors of the centered SBM adjacency matrix. Although properties of extremal eigenvectors of the SBM adjacency matrix are well studied, not much attention has been given to the eigenvectors of the centered matrix, which represent the bulk eigenvectors. We consider, as in [Bai & Pan 2012], $Q(x, \mathbf{y})$, a modified empirical spectral density function of the eigenvalues, where the contribution of each eigenvalue is weighted by the magnitude of the projection of the corresponding eigenvector to an arbitrary, deterministic unit vector $\mathbf{y}$.

First we show that when the link probabilities within communities are different, i.e., the case of *asymmetric* SBM, the weighted spectral function $Q(x, \mathbf{y})$ has different limits depending on the unit vector $\mathbf{y}$, and we determine the asymptotic limits. From this we conclude that the eigenvectors of the asymmetric SBM are not Haar distributed. In contrast, when the link probabilities within all the communities are the same, i.e., in the case of *symmetric* SBM, the modified empirical spectral distribution $Q(x, \mathbf{y})$ has the same asymptotic limit as the empirical spectral distribution (e.s.d.) of eigenvalues, $\forall \mathbf{y}$ with $\|\mathbf{y}\|_2 = 1$. This is a necessary condition for Haar distribution of eigenvectors. In contrast, we show that, the variation around this mean cannot shown to be a Brownian Bridge. This is because the atom distribution is not symmetric, i.e., its third moment is not zero and thus the bulk eigenvectors of the symmetric SBM are not Haar distributed.

### 3.4.1 Asymptotic Results on Eigenvectors of SBM

In this section we analyze the asymptotic properties of the eigenvectors of $\widetilde{\mathbf{A}}$. To recall, matrix $\widetilde{\mathbf{A}}$ is obtained from the adjacency matrix $\mathbf{A}$ by subtracting the mean and dividing by $\sqrt{n}$. The variances of the components of $\widetilde{\mathbf{A}}$ are then

$$\mathbb{E}\widetilde{A}^2_{ij} = \begin{cases} \frac{\sigma_1^2}{n} & \text{if } 1 \leq i, j \leq n/2 \\ \frac{\sigma_2^2}{n} & \text{if } 1 + n/2 \leq i, j \leq n \\ \frac{\sigma_0^2}{n} & \text{otherwise,} \end{cases}$$

where $\sigma_1^2 = p_1(1-p_1)$, $\sigma_2^2 = p_2(1-p_2)$ and $\sigma_0^2 = p_0(1-p_0)$. We consider dense graphs, i.e., the probabilities $p_1, p_2$ and $p_0$ are constants independent of $n$.

Following the ideas in [Silverstein 1990] we consider the following spectral function

$$Q(x, \mathbf{y}) = \sum_{i=1}^{n} |\mathbf{u}_i^T \mathbf{y}|^2 \chi_{\{\lambda_i \leq x\}}, \ x \in (-\infty, \infty), \tag{3.26}$$

where $\mathbf{y} \in \mathbb{R}^n$ is an arbitrary deterministic unit vector. Notice that $Q(x, \mathbf{y}) \geq 0, \forall x$, $\lim_{x \to -\infty} Q(x, \mathbf{y}) = 0$, $\lim_{x \to \infty} Q(x, \mathbf{y}) = 1$ and $Q(x, \mathbf{y})$ is right continuous in $x$. Therefore $Q(x, \mathbf{y})$ satisfies all properties of a cumulative distribution function (cdf). In [Bai

*et al.* 2007], the authors study the above function and observe that if the eigenvectors are Haar-distributed $Q(x, \mathbf{y})$ satisfies the following two properties:

- **Property I**
$$\lim_{n\to\infty} |Q(x, \mathbf{y}) - F^{\widetilde{\mathbf{A}}}(x)| = 0,$$
where $F^{\widetilde{\mathbf{A}}}(x)$ is the e.s.d. of $\widetilde{\mathbf{A}}$. This property has to be satisfied if $\mathbf{u}_i$ are uniformly distributed on the unit sphere in $\mathbb{R}^n$.

- **Property II**
$\sqrt{\frac{n}{2}}(Q(x, \mathbf{y}) - F^{\widetilde{\mathbf{A}}}(x))$ converges to a Brownian Bridge.
Indeed, a vector uniformly distributed on the unit sphere in $\mathbb{R}^n$ is equivalent in distribution to a vector $\mathbf{z} \in \mathbb{R}^n$ with independent standard gaussian components normalized such that $\|\mathbf{z}\|_2 = 1$.

Instead of analyzing $Q(x, \mathbf{y})$ directly we can analyze its Stieltjes transform given as [Bai *et al.* 2007]

$$s_Q(z, \mathbf{y}) = \mathbf{y}^T(\widetilde{\mathbf{A}} - z\mathbf{I})^{-1}\mathbf{y}. \tag{3.27}$$

By the Stieltjes inversion formula, the convergence of the Stieltjes transform of a sequence of functions, implies the convergence of the original sequence [Anderson *et al.* 2009].

### 3.4.2 Asymptotic Limit of $Q(x, \mathbf{y})$ for general SBM

In this section we present a result about $Q(x, \mathbf{y})$ for the special case when $\mathbf{y} = \mathbf{e}_i$. The analysis adopts the same method as the one followed in [Girko *et al.* 1994]. In this case $s_Q(z, \mathbf{y}) = \left[(\widetilde{\mathbf{A}} - z\mathbf{I})^{-1}\right]_{ii}$, the diagonal component of the resolvent matrix of $\widetilde{\mathbf{A}}$.

Let us define $\mathbf{\Psi} := (\widetilde{\mathbf{A}} - z\mathbf{I})^{-1}$ and denote the diagonal component as $\Psi_{ii}$.
We have for $\widetilde{\mathbf{A}}$, if $\sigma_{ij}^2$ is the variance of the entry $\widetilde{A}_{ij}$,

$$\forall i \quad \exists c \text{ such that } \sum_j \sigma_{ij}^2 \leq c. \tag{3.28}$$

In addition, for $1 \leq i, j \leq n/2$,

$$\sum_k \sigma_{jk}^2 = \sum_k \sigma_{ik}^2, \tag{3.29}$$

and similarly for $n/2 + 1 \leq i, j \leq n$.
We have the following result.

**Proposition 3.1.** *For a centered adjacency matrix $\widetilde{\mathbf{A}}$ of SBM with constant probabilities $p_0, p_1, p_2$, independent of $n$, and $\mathbf{y} = \mathbf{e}_i, s_Q(z, \mathbf{e}_i)$, the Stieltjes transform of the spectral function $Q(x, \mathbf{y})$, converges in probability as follows.*

$$\lim_{n\to\infty} s_Q(z, \mathbf{e}_i) = \begin{cases} d_1, & \text{if } i \leq \frac{n}{2} \\ d_2, & \text{if } \frac{n}{2} + 1 \leq n, \end{cases} \tag{3.30}$$

*where $d_1, d_2$ are unique solutions to the following set of fixed point equations*

$$d_1 = \frac{1}{-z - \frac{d_1}{2}\sigma_1^2 - \frac{d_2}{2}\sigma_0^2},$$

$$d_2 = \frac{1}{-z - \frac{d_1}{2}\sigma_0^2 - \frac{d_2}{2}\sigma_2^2}.$$

*Proof.* From [Girko *et al.* 1994], under conditions (3.28) and (3.29), we have:

$$[\Psi_{ii} - \frac{1}{-z - \sum_{k=1}^{n} \Psi_{kk}\sigma_{ki}^2}] \to 0,$$

in probability, $\forall i$, i.e.,

$$\Psi_{ii} = \frac{-1}{z + \sum_{k=1}^{n} \Psi_{kk}\sigma_{ki}^2} + \mathrm{o}_p(1),$$

where $o_p(1)$ is a quantity that tends to zero in probability, as $n \to \infty$.
Consider $i \neq j$ such that $1 \leq i, j \leq n/2$. Then,

$$\begin{aligned}
\Psi_{ii} - \Psi_{jj} &= \frac{-1}{z + \sum_{k=1}^{n} \Psi_{kk}\sigma_{ki}^2} + \frac{1}{z + \sum_{k=1}^{n} \Psi_{kk}\sigma_{kj}^2} + \mathrm{o}_p(1) \\
&= \frac{\sum_{k=1}^{n} \Psi_{kk}(\sigma_{ki}^2 - \sigma_{kj}^2)}{(z + \sum_{k=1}^{n} \Psi_{kk}\sigma_{ki}^2)(z + \sum_{k=1}^{n} \Psi_{kk}\sigma_{kj}^2)} + \mathrm{o}_p(1) \\
&= o_p(1).
\end{aligned}$$

The last equality follows because $\sigma_{ki}^2 = \sigma_{kj}^2, \forall k$, if $1 \leq i, j \leq n/2$, and from the boundedness of the denominator.
Hence we have

$$\max_{1 \leq i,j \leq n/2} |\Psi_{ii} - \Psi_{jj}| \to 0 \quad \text{in probability,} \tag{3.31}$$

and similarly,

$$\max_{n/2+1 \leq i,j \leq n} |\Psi_{ii} - \Psi_{jj}| \to 0, \quad \text{in probability.} \tag{3.32}$$

Now, consider the following difference:

$$\begin{aligned}
&\left| \frac{-1}{z + \sum_{k=1}^{n} \Psi_{kk}\sigma_{ki}^2} + \frac{1}{z + \sum_{k=1}^{n/2} \Psi_{11}\sigma_{ki}^2 + \sum_{k=1+n/2}^{n} \Psi_{n/2+1,n/2+1}\sigma_{ki}^2} \right| \\
&= \left| \frac{\sum_{k=1}^{n/2}(\Psi_{kk} - \Psi_{11})\sigma_{ki}^2 + \sum_{k=1+n/2}^{n}(\Psi_{kk} - \Psi_{n/2+1,n/2+1})\sigma_{ki}^2}{(z + \sum_{k=1}^{n} \Psi_{kk}\sigma_{ki}^2)(z + \sum_{k=1}^{n/2} \Psi_{11}\sigma_{ki}^2 + \sum_{k=1+n/2}^{n} \Psi_{n/2+1,n/2+1}\sigma_{ki}^2)} \right| \\
&\leq K \max_{1 \leq k \leq n/2} |\Psi_{kk} - \Psi_{11}| \sum_{k=1}^{n/2} \sigma_{ki}^2 + \max_{1+n/2 \leq k \leq n} |\Psi_{kk} - \Psi_{n/2+1,n/2+1}| \sum_{k=1+n/2}^{n} \sigma_{ki}^2 \\
&= o_p(1),
\end{aligned}$$

by (3.31) and (3.31) since,

$$\begin{aligned}
\sum_{k=1}^{n/2} \sigma_{ki}^2 &= \frac{1}{2}\sigma_1^2 \quad , i \leq n/2 \qquad \sum_{k=1}^{n/2} \sigma_{kj}^2 = \frac{1}{2}\sigma_0^2 \quad , j \geq n/2 \\
\sum_{k=1+n/2}^{n} \sigma_{ki}^2 &= \frac{1}{2}\sigma_0^2 \quad , i \leq n/2 \qquad \sum_{k=1+n/2}^{n} \sigma_{kj}^2 = \frac{1}{2}\sigma_2^2 \quad , j \geq n
\end{aligned}$$

and each of the above is a constant.
So we finally have:

$$\Psi_{ii} \to d_1 = \frac{-1}{z + \frac{d_1}{2}\sigma_1^2 + \frac{d_2}{2}\sigma_0^2} \text{for } i \leq n/2$$

and

$$\Psi_{ii} \to d_2 = \frac{-1}{z + \frac{d_1}{2}\sigma_0^2 + \frac{d_2}{2}\sigma_2^2} \text{ for } i > n/2.$$

□

From the above result, we see that the eigenvectors of a general SBM are not Haar distributed, because the asymptotic limit is a function of the vector $\mathbf{y}$. In fact, $\mathbf{y} = \mathbf{e}_i$, the asymptotic limit of the spectral function is different when $i \leq n/2$ and $i > n/2$.

We make the following observation as a corollary to Proposition 3.1.

**Corollary 3.1.** *For an asymmetric SBM with $p_1 \neq p_2$,*

$$\lim_{n\to\infty} \frac{1}{2}\left(s_Q(z, \mathbf{e}_{k_1}) + s_Q(x, \mathbf{e}_{k_2})\right) = s_\sigma(z),$$

*in probability with $s_\sigma(z)$ is the Stieltjes transform of the semicircle distribution with variance parameter $\sigma^2$.*

In Proposition 3.2, we present a result for symmetric SBM, i.e., $p_1 = p_2$, that holds for any unit vector $\mathbf{y}$ (sometime denoted $\mathbf{y}_n$ to emphasize the dependance on $n$). More specifically, we show the convergence of the Stieltjes transform $s_Q(z, \mathbf{y})$ to $s_\sigma(z)$, the Stieltjes transform of the semicircle distribution with variance parameter $\sigma^2$.

Following [Bai & Pan 2012] we define $\widetilde{\mathbf{R}} = \widetilde{\mathbf{A}} - z\mathbf{I}$ and $\widetilde{\mathbf{R}}_k = \widetilde{\mathbf{A}_\mathbf{k}} - z\mathbf{I}$, where $\widetilde{\mathbf{A}_k}$ is obtained by removing the $k^{\text{th}}$ row and column of $\widetilde{\mathbf{A}}$ and the $k^{\text{th}}$ column is $\widetilde{\mathbf{a}}_k$. In addition, we use the following notation from [Bai & Pan 2012]

$$\begin{aligned}
\eta_k^{(')} &= \widetilde{\mathbf{a}}_k^T \widetilde{\mathbf{R}}_k^{-1} \mathbf{y_n} \mathbf{y_n}^T \widetilde{\mathbf{R}}_k^{-1} \widetilde{\mathbf{a}}_k, \\
\eta_k &= \eta_k^{(')} - \mathbb{E}_{\widetilde{\mathbf{a}}_k} \eta_k^{(')}, \\
\omega_k &= \frac{1}{1 + \frac{1}{nz}\left(\sigma_1^2 \sum_{i=1}^{n/2} \Psi_{ii}^k + \sigma_0^2 \sum_{i=1+n/2}^{n} \Psi_{ii}^k\right)}, \quad 1 \leq k \leq \frac{n}{2}, \\
\omega_k &= \frac{1}{1 + \frac{1}{nz}\left(\sigma_0^2 \sum_{i=1}^{n/2} \Psi_{ii}^k + \sigma_1^2 \sum_{i=1+n/2}^{n} \Psi_{ii}^k\right)}, \quad 1 + \frac{n}{2} \leq k \leq n,
\end{aligned} \tag{3.33}$$

where $\mathbb{E}_{\widetilde{\mathbf{a}_k}}$ denotes expectation w.r.t. $\widetilde{\mathbf{a}_k}$ and $\Psi_{ii}^k = \left[\widetilde{\mathbf{R}}_k^{-1}\right]_{ii}$. Additionally, we define

$$\begin{aligned}
\zeta_k &= \widetilde{\mathbf{a}}_k^T \widetilde{\mathbf{R}}_k^{-1} \widetilde{\mathbf{a}}_k - \frac{1}{n}\left(\sigma_1^2 \sum_{i=1}^{n/2} \Psi_{ii}^k + \sigma_0^2 \sum_{i=n/2+1}^{n} \Psi_{ii}^k\right), \quad 1 \leq k \leq n/2, \\
\zeta_k &= \widetilde{\mathbf{a}}_k^T \widetilde{\mathbf{R}}_k^{-1} \widetilde{\mathbf{a}}_k - \frac{1}{n}\left(\sigma_0^2 \sum_{i=1}^{n/2} \Psi_{ii}^k + \sigma_1^2 \sum_{i=n/2+1}^{n} \Psi_{ii}^k\right), \quad 1 + n/2 \leq k \leq n.
\end{aligned}$$

Ancillary to Proposition 3.2 is the following lemma.

**Lemma 3.6.** *For the centered SBM adjacency matrix $\widetilde{\mathbf{A}}$, with probabilities $p_0, p_1, p_2$,*

$$\max_{1\leq k\leq n/2} \mathbb{E}|\alpha_k + 2zc_1(z)|^4 \to 0,$$

$$\max_{n/2+1\leq k\leq n} \mathbb{E}|\alpha_k + 2zc_2(z)|^4 \to 0;$$

*where*

$$\alpha_k = \frac{1}{1 + z^{-1}\widetilde{\mathbf{a}}_k^T \widetilde{\mathbf{R}}_k^{-1} \widetilde{\mathbf{a}}_k}, \tag{3.34}$$

*and* $c_1(z), c_2(z)$ *are analytic functions that satisfy the fixed point equations in (3.9):*

$$c_1(z) = \frac{-0.5}{z + \sigma_1^2 c_1(z) + \sigma_0^2 c_2(z)}$$
$$c_2(z) = \frac{-0.5}{z + \sigma_2^2 c_2(z) + \sigma_0^2 c_1(z)}$$

*For the symmetric case, we recall that* $c_1(z) = c_2(z) = s_\sigma(z)/2$. *Also when* $p_1 = p_2$,

$$\max_{1 \le k \le n} |\omega_k + z s_\sigma(z)| \to 0$$

*in probability.*

*Proof.* This lemma follows along the same lines as the proof of Lemma 8.1 in [Bai & Pan 2012] without significant modifications because the random variables are bounded. □

By invoking the above lemma, we can prove the following proposition.

**Proposition 3.2.** *(Symmetric SBM) Let us consider the centered adjacency matrix* $\widetilde{\mathbf{A}}$ *of SBM, with* $p_1 = p_2$. *For any unit vector* $\mathbf{y}$, *the spectral function* $Q(x, \mathbf{y})$ *converges to the semicircle law.*

*Proof.* The proof consists of two steps:

- Showing $\mathbf{y}^T \widetilde{\mathbf{R}}^{-1} \mathbf{y} \to \mathbf{y}^T \mathbb{E}\widetilde{\mathbf{R}}^{-1} \mathbf{y}$. Since we consider the case where probabilities $p_1$ and $p_0$ are constants independent of $n$, the random variables are bounded, and therefore this results follows directly from the proof of the first part of Theorem 1.1 in [Bai & Pan 2012], without significant modifications.
- Showing $\mathbf{y}^T \mathbb{E}\widetilde{\mathbf{R}}^{-1} \mathbf{y} \to s_\sigma(z)$, where $s_\sigma(z)$ is the Stieltjes transform of the semicircle distribution. This is shown below.

We introduce the following notation and results.

Following [Bai & Pan 2012], we have the following bounds: $|\alpha_k| \le \frac{|z|}{v}, |\omega_k| \le \frac{|z|}{v}$. Using concentration bounds for quadratic forms [Bai & Pan 2012] we also have for $p \ge 2$

$$\mathbb{E}|\eta_k|^p = O(n^{-p/2-1}), \tag{3.35}$$
$$\mathbb{E}|\zeta_k|^p = O(n^{-p/2}), \tag{3.36}$$
$$\tag{3.37}$$

From the definition of $\alpha_k$, one can see $\alpha_k = \omega_k - z^{-1}\alpha_k \omega_k \varsigma_k$.

Similar to the procedure followed in [Bai & Pan 2012], we can decompose the above term as

$$z\mathbf{y}^T \mathbb{E}\widetilde{\mathbf{R}}^{-1} \mathbf{y} + 1 := L_1 + L_2,$$

where

$$L_1 = -\sum_{k=1}^{n} z^{-1} y_k \mathbb{E} y_n^T \widetilde{\mathbf{a}}_k \alpha_k, \quad L_2 = \sum_{k=1}^{n} z^{-1} \mathbb{E}\mathbf{y}_n^T \widetilde{\mathbf{a}}_k \widetilde{\mathbf{a}}_k^T \widetilde{\mathbf{R}}_k^{-1} \mathbf{y}_n \alpha_k$$

and and we analyze them one by one. First we look at $L_2$.

Using the result on $\alpha_k$ from Lemma 3.6 along with the fact that

$$\mathbb{E}|\mathbf{y}_n^T\widetilde{\mathbf{a}}_k| = O(\frac{1}{\sqrt{n}}) \quad \mathbb{E}|\widetilde{\mathbf{a}}_k^T\mathbf{R}_k^{-1}\mathbf{y}_n| = O(\frac{1}{\sqrt{n}})$$

and Holder's inequality [3] we can show,

$$L_2 = -s_\sigma(z)\sum_{k=1}^{n}\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k\widetilde{\mathbf{a}}_k^T\widetilde{\mathbf{R}}_k^{-1}\mathbf{y} + o(1),$$

where $o(1)$ is a term that goes to zero in probability. Notice that since $\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{R}}_k$ form an independent pair, we can take the expectation inside and use the fact that $\mathbb{E}\widetilde{\mathbf{a}}_k\widetilde{\mathbf{a}}_k^T = \mathbf{W}_1/n$, for $1 \le k \le n/2$, and $\mathbf{W}_2/n$, for $n/2+1 \le k \le n$, where $\mathbf{W}_1 \in \mathbb{R}^{n\times n}$ is a diagonal matrix such that

$$(\mathbf{W}_1)_{ii} = \begin{cases}\sigma_1^2 & \text{for } 1 \le i \le n/2,\\ \sigma_0^2, & \text{for } 1+n/2 \le i \le n\end{cases}$$

and similarly, $\mathbf{W}_2 \in \mathbb{R}^{n\times n}$ is diagonal with

$$(\mathbf{W}_2)_{ii} = \begin{cases}\sigma_0^2 & \text{for } 1 \le i \le n/2\\ \sigma_1^2, & \text{for } 1+n/2 \le i \le n.\end{cases}$$

Therefore $\frac{1}{2}(\mathbf{W}_1+\mathbf{W}_2) = \sigma^2\mathbf{I}$, where $\sigma^2 = (\sigma_0^2+\sigma_1^2)/2$.
Now we use the fact that $\widetilde{\mathbf{R}}_k^{-1} = \widetilde{\mathbf{R}}^{-1} + \widetilde{\mathbf{R}}_k^{-1}(\widetilde{\mathbf{a}}_k\mathbf{e_k}^T + \mathbf{e_k}\widetilde{\mathbf{a}}_k^T)\widetilde{\mathbf{R}}^{-1}$, to get

$$L_2 = -s_\sigma(z)\frac{1}{n}\left(\sum_{k=1}^{n/2}\mathbb{E}\mathbf{y}^T\mathbf{W}_1(\widetilde{\mathbf{R}}^{-1} + \widetilde{\mathbf{R}}_k^{-1}(\widetilde{\mathbf{a}}_k\mathbf{e_k}^T + \mathbf{e_k}\widetilde{\mathbf{a}}_k^T)\widetilde{\mathbf{R}}^{-1})\mathbf{y} + \sum_{k=1+n/2}^{n}\mathbb{E}\mathbf{y}^T\mathbf{W}_2(\widetilde{\mathbf{R}}^{-1} + \widetilde{\mathbf{R}}_k^{-1}(\widetilde{\mathbf{a}}_k\mathbf{e_k}^T + \mathbf{e_k}\widetilde{\mathbf{a}}_k^T)\widetilde{\mathbf{R}}^{-1})\mathbf{y}\right) + o(1).$$

Since $\|\widetilde{\mathbf{R}}_k^{-1}\mathbf{y}\|_2$ is bounded, we have $\frac{1}{n}\sum_{k=1}^{n}|\mathbf{e}_k^T\widetilde{\mathbf{R}}_k^{-1}\mathbf{y}| = O(\frac{1}{\sqrt{n}})$. We also use the fact that $\mathbb{E}|\mathbf{x}^T\widetilde{\mathbf{R}}_k^{-1}\mathbf{y}| < C$ for some $C$ for any unit $\mathbf{x},\mathbf{y}$. Also, we have $\mathbf{y}^T\mathbf{W}_1\widetilde{\mathbf{R}}_k^{-1}\mathbf{e_k} = -cy_k/z$ and $\frac{1}{n}\sum_{k=1}^{n}|y_k| = O(\frac{1}{\sqrt{n}})$, we get

$$L_2 = -s_\sigma(z)\frac{1}{n}\left(\sum_{k=1}^{n/2}\mathbb{E}\mathbf{y}^T\mathbf{W}_1\widetilde{\mathbf{R}}^{-1}\mathbf{y} + \sum_{k=1+n/2}^{n}\mathbb{E}\mathbf{y}^T\mathbf{W}_2\widetilde{\mathbf{R}}^{-1}\mathbf{y}\right) + o(1).$$

Then using the fact that $\frac{1}{2}(\mathbf{W}_1+\mathbf{W}_2) = \sigma^2\mathbf{I}$ we finally get

$$L_2 = -s_\sigma(z)\sigma^2\mathbf{y}^T\mathbb{E}\widetilde{\mathbf{R}}^{-1}\mathbf{y} + o(1).$$

Now we move on to show $L_1 \to 0$.

$$L_1 = -\sum_{k=1}^{n} z^{-1}y_k\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k\alpha_k$$

using $\alpha_k = \omega_k - z^{-1}\alpha_k\omega_k\varsigma_k$ [Bai & Pan 2012]

$$L_1 = -\sum_{k=1}^{n} z^{-1}y_k\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k\omega_k + \sum_{k=1}^{n} z^{-2}y_k\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k\alpha_k\omega_k\zeta_k.$$

[3] $\mathbb{E}(|XY|) \le (\mathbb{E}(|X|^p))^{\frac{1}{p}}(\mathbb{E}(|X|^q))^{\frac{1}{q}}$, if $\frac{1}{p}+\frac{1}{q} = 1$.

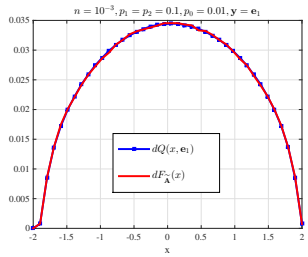


Figure 3.1: Asymptotic spectral function for symmetric SBM, $\mathbf{y} = \mathbf{e}_1$.

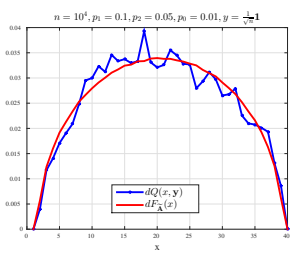


Figure 3.2: Asymptotic spectral function for symmetric SBM, $\mathbf{y} = 1/\sqrt{n}\mathbf{1}$.

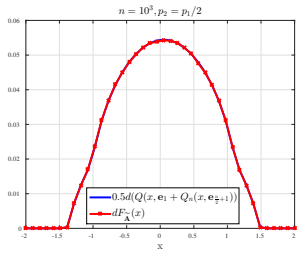


Figure 3.3: Asymptotic spectral function for asymmetric SBM, $\mathbf{y} = \mathbf{e}_1$.

The first term is zero, because $\widetilde{\mathbf{a}}_k, \omega_k$ are independent, and $\mathbb{E}\widetilde{\mathbf{a}}_k = \mathbf{0}$. Then we use the bounds on $\alpha_k, \omega_k$, and that $\mathbb{E}\zeta_k^2 = O(n^{-1})$ and $\mathbb{E}|\mathbf{y}^T\widetilde{\mathbf{a}}_k|^2 = O(n^{-1})$ to show $L_1 = \mathrm{O}(1/\sqrt{n})$.

Thus we have $z\mathbf{y}^T\mathbb{E}\widetilde{\mathbf{R}}^{-1}\mathbf{y} + 1 = -s_\sigma(z)\sigma^2\mathbf{y}^T\mathbb{E}\widetilde{\mathbf{R}}^{-1}\mathbf{y} + o(1)$,

Since $s_\sigma(z)$ satisfies

$$s_\sigma(z) = \frac{-1}{z + \sigma^2 s_\sigma(z)}$$

we have $\mathbf{y}^T\mathbb{E}\widetilde{\mathbf{R}}^{-1}\mathbf{y} = s_\sigma(z) + o(1)$. □

### 3.4.3 Gaussianity of the fluctuations

In this section we focus on the symmetric SBM and verify whether and if not, to what extent Property II is satisfied. The convergence of the process $\sqrt{n}\left(\mathbf{y}^T\widetilde{\mathbf{R}}^{-1}\mathbf{y} - s_\sigma(z)\right)$ to a Brownian Bridge in distribution is shown in two steps:

- The process $Y_n(z) = \sqrt{n}\left(\mathbf{y}^T\widetilde{\mathbf{R}}^{-1}\mathbf{y} - \mathbf{y}^T\mathbb{E}(\widetilde{\mathbf{R}})^{-1}\mathbf{y}\right)$ is shown to converge to a gaussian process of mean zero in distribution;
- $\sqrt{n}\left(\mathbf{y}^T\mathbb{E}(\widetilde{\mathbf{R}})^{-1}\mathbf{y} - s_\sigma(z)\right) \to 0$.

In the following proposition we show that for SBM matrices the second part does not converge to zero but instead, there is a bias term which is a function of $z$. This shows that the process does not converge to a Brownian Bridge, and consequently the eigenvectors of the adjacency matrix of the symmetric SBM are not Haar distributed, even though they display some of the required properties as shown above. This is due to the fact that the entries of the matrix $\widetilde{\mathbf{A}}$ does not have vanishing third moment, as required by [Bai & Pan 2012, Theorem 1.2]. In the following proposition we bound this bias term.

**Proposition 3.3.** *For the centered adjacency matrix $\widetilde{\mathbf{A}}$ of a symmetric SBM with probabilities $p_0, p_1$, the fluctuation of the mean, $\sqrt{n}\left(\mathbf{y}^T\mathbb{E}\widetilde{\mathbf{R}}^{-1}\mathbf{y} - s_\sigma(z)\right)$ when $y = \frac{1}{\sqrt{n}}\mathbf{1}$ does not converge to 0, but it is bounded as follows*

$$\left|\sqrt{n}\left(\mathbf{y}^T\mathbb{E}\widetilde{\mathbf{R}}^{-1}\mathbf{y} - s_\sigma(z)\right)\right| \leq C|s_\sigma(z)| + o(1),$$

*where $C$ is a constant that depends on the third moment of the elements of $\widetilde{\mathbf{A}}$.*

*Proof.* In [Bai & Pan 2012], it is shown that this term goes to zero whenever the matrix entries have zero third moment. However in our case, since the matrix entries are binomial random variables, this is not true unless all probabiities are equal to half. However, we can show the above upper bound.
As in [Bai & Pan 2012], we decompose the above term as

$$\mathbf{y}^T\mathbb{E}(\widetilde{\mathbf{A}} - z\mathbf{I})^{-1}\mathbf{y} - s_\sigma(z) = L_1 + L_2 \tag{3.38}$$

Ideally, we would like both $L_1$ and $L_2$ to be $o(1/\sqrt{n})$. However, $L_2$ can be shown to obey this, but not $L_1$. We skip the proof for $L_2$, because it is follows the same steps as the last part of section 5 in [Bai & Pan 2012]. We deal with $L_1$ below. In the setting treated in [Bai & Pan 2012] $L_1$ tends to zero because the matrix entries in his case are symmetric. However, in our case, since we are dealing with binary random variables, this is not true.
We have $L_1 = \sum_{k=1}^{n} z^{-1}y_k\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k\zeta_k\alpha_k\omega_k$. From this we get: $|L_1| = |\sum_k y_k\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k\zeta_k\gamma_k^2| + O(1/n)$. Observe that $\omega_k$ only depends on $\widetilde{\mathbf{R}}_k$, is independent of $\widetilde{\mathbf{a}}_k$, and is bounded. We have

$$\zeta_k = \widetilde{\mathbf{a}}_k^T\widetilde{\mathbf{R}}_k^{-1}\widetilde{\mathbf{a}}_k - \mathbb{E}_{\widetilde{\mathbf{a}}_k}(\widetilde{\mathbf{a}}_k^T\widetilde{\mathbf{R}}_k^{-1}\widetilde{\mathbf{a}}_k).$$

Then $\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k$ is zero, and $\widetilde{\mathbf{R}}_k^{-1}$ is independent of $\widetilde{\mathbf{a}}_k$, and so the contribution due to the second part in the definition of $\zeta_k$ is zero. Therefore we are left with $\mathbb{E}\mathbf{y}^T\widetilde{\mathbf{a}}_k\widetilde{\mathbf{a}}_k^T\widetilde{\mathbf{R}}_k^{-1}\widetilde{\mathbf{a}}_k = \mathbb{E}\frac{1}{n^{3/2}}\sum_{l,m,n\neq k} y_l A'_{lk}A'_{mk}(\widetilde{\mathbf{R}}_k)^{-1}A'_{nk}$, where $A'_{ij} = \widetilde{A}_{ij}\sqrt{n}$. Because of zero mean condition ($\mathbb{E}\widetilde{A}_{ij} = 0$,), the only terms that survive are such that $l = m = n (l \neq k, m \neq k, n \neq k)$. So we get $|L_1| \leq |\frac{1}{n^{3/2}}\sum_m y_m(\widetilde{\mathbf{R}}_k^{-1})_{nn}\mathbb{E}(A'_{nk})^3|$.

For binomial random variables with probability of $1 < 1/2$, the third moment is always positive. Therefore we have:

$$|L_1| \leq |C/n^{3/2}\sum_k y_k \sum_l y_l(\widetilde{\mathbf{R}}_k^{-1})_{ll}|$$

We consider the special case where $y_i = \frac{1}{\sqrt{n}}, \forall i$. We get

$$\frac{1}{n^{3/2}}\sum_k\sum_l y_k y_l(\widetilde{\mathbf{R}}_k^{-1})_{ll}\mathbb{E}(A'^{\,3}_{lk}) = \frac{1}{nn^{3/2}}\sum_l(\widetilde{\mathbf{R}}_k^{-1})_{ll}\sum_k \mathbb{E}A'^{\,3}_{lk}$$

Note that $\frac{1}{n}\sum_k \mathbb{E}A'^{\,3}_{lk} = K_p$, some constant that depends on $p_1, p_0$ and $\frac{1}{n}\sum_l(\widetilde{\mathbf{R}}_k^{-1})_{ll} \to \frac{1}{n}\text{trace}(\mathbf{A}^{-1}) \to s_\sigma(z)$ (Using Lemma 3.1). Therefore, $|L_1| \leq C\frac{1}{\sqrt{n}}s_\sigma(z)$, for some $C$. □

## 3.5 Example Application: Epidemic Spreading

In this section, we discuss an application of the result we derived above for adjacency matrices of SBM, in the topic of epidemic spreading. In [Bose *et al.* 2013], the authors study an epidemic process over a random network of nodes. The spread of the epidemic from one node to another is governed by the random graph, i.e., a node can only infect another if there exists an edge between the two nodes. We have the following result delineating the relationship between the expected cost of the epidemic per node denoted by $C_D(n)$ (disease cost) and the largest eigenvalue of the graph adjacency matrix $\mathbf{A}$ [Bose *et al.* 2013],

$$C_D(n) \leq \frac{\alpha c_d}{1 - \lambda_1(\mathbf{M})} \tag{3.39}$$

where $\mathbf{M} = (1-\delta)\mathbf{I} + \beta\mathbf{A}$ is the matrix which governs the dynamics of the system [Bose *et al.* 2013], with $\beta$ being the probability of infection, $\delta$ is the probability of recovery of any node, and $c_d$ is the cost parameter. We direct the reader to the original paper for more details.

We examine the epidemic spread on an SBM graph with $M$ communities. We know that in this case $\lambda_1(\mathbf{A}) \to \frac{n}{M}\mu_1$ as $n \to \infty$ a.s. under certain conditions. Also by (3.19) we have that $\mu_1 \leq p_1 + (M-1)p_0$, therefore we have:

$$\lambda_1(\mathbf{M}) = (1-\delta) + \beta\lambda_1(\mathbf{A}) \leq 1 - \delta + \beta(n/M\mu_1)$$

Thus we have:

$$C_D(n) \leq \frac{\alpha c_d}{\delta - \beta n/M(p_1 + (M-1)p_0)} \tag{3.40}$$

If $p_1 >> p_i$, for $i \geq 2$, then we can venture to say that this bound is tight, and that the community with the largest edge probability governs the disease cost.

## 3.6 Numerical Results

### 3.6.1 Asymptotic Eigenvalue Distribution

In this section we provide simulation results to demonstrate the results obtained earlier in this chapter. More specifically, we corroborate our results on the spectrum of adjacency matrix by comparing the spectrum obtained by simulating a 2-community SBM with the distribution obtained by inverting the Stieltjes transform, which is an explicit solution of the simultaneous equations (3.12). In the simulations, we use a matrix of size $n = 10^4$. For a 2-community system, the solution amounts to solving explicity the resulting quartic equation and choosing the solution branch that satisfies the conditions (3.13). The inverse relationship between the limiting e.s.d. and the Stieltjes transform thus obtained, is given by the well known Stieltjes inversion formula:

$$f(x) = \lim_{y \to 0} \Im s_F(x + \sqrt{-1}y)/\pi \tag{3.41}$$

where $f(x)$ is the p.d.f. corresponding to the c.d.f. $F(x)$, whenever the limit exists.
Figure 3.4 shows the histogram of normalized adjacency matrix $\frac{1}{\sqrt{n}}\mathbf{A}$ and compares it to the theoretical spectrum obtained as above for $n = 10^4$, and several values of edge probabilities.

In the second part of this section we turn our attention to the extremal eigenvalues of the adjacency matrix for a 3-community SBM of size $n = 999$. Over several independent runs, we get values of the top 4 eigenvalues of the matrix $\mathbf{A}$, for $0.3 \leq p_1 \leq 0.48$, $0.15 \leq p_2 \leq 0.33$, $0.08 \leq p_3 \leq 0.26$ and $0.03 \leq p_0 \leq 0.031$. We note that in Figure 3.5, as expected , there are three eigenvalues outside the bulk, which agree very well with the expected values, i.e., the non-zero eigenvalues of the mean matrix $\overline{A}$. In addition, it can also be seen that the upperbound in (3.16) is remarkably tight for the simulated probabilities.

Next, we consider the spectrum of the normalized Laplacian matrix. In fact, we consider the spectrum of the shifted normalized Laplacian matrix, which we denote $\widetilde{\boldsymbol{\mathcal{L}}}$, defined as $\widetilde{\boldsymbol{\mathcal{L}}} := \sqrt{n}/2 - \sqrt{n}/2\boldsymbol{\mathcal{L}}$. By Lemma 3.5, its spectrum is given by the solution of the equation (3.22). We explicitly solve this equation for SBM with two-communities, and compare it the result obtained by simulations for a graphs with $n = 999$ for various values of the probabilities $p_1$, $p_2$ and $p_0$. The comparison is shown in Fig.3.6.

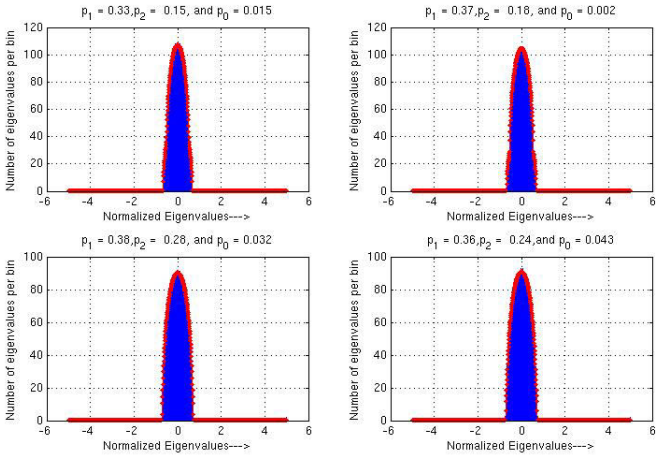



Figure 3.4: Comparison plot between empirically obtained spectrum(bar graph), and explicit solution(line plot) of 2-community SBM adjacency matrix

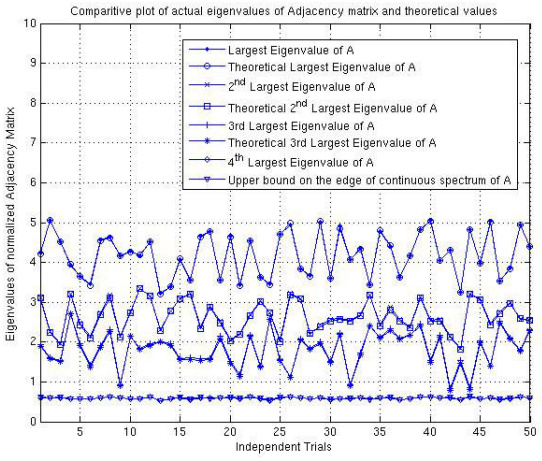



Figure 3.5: Extremal eigenvalues of 3-community SBM normalized matrix compared to expected values.

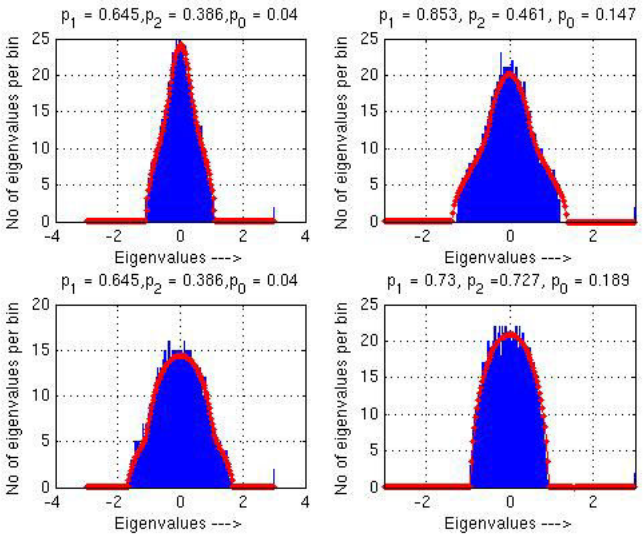



Figure 3.6: Histogram of 2-community $\tilde{\boldsymbol{\mathcal{L}}}$ for various edge probabilities compared to theoretical spectrum

### 3.6.2 Asymptotic Eigenvector Distribution

In this section, we perform simulations to corroborate our theoretical results. First, we consider a realization of a random symmetric SBM of size $n = 10^3$, with $p_0 = 10^{-2}$ and $p_1 = p_2 = 10^{-1}$. For the results in Figure 3.1, we set $\mathbf{y} = \mathbf{e}_1$ and we plot two histograms of the eigenvalues: the first one denoted by $dF_{\tilde{\mathbf{A}}}(x)$ is obtained giving a unit weight to each eigenvalue falling in a histogram bin $[x, x + \Delta)$; the second one, denoted by $dQ(x, \mathbf{y})$ is obtained by giving the weight $|\mathbf{u}_i^T \mathbf{y}|^2$ to an eigenvalue $\lambda_i \in [x, x + \Delta)$. In other words, let $\lambda_i, \lambda_{i+1}, \ldots \lambda_j$ be the $j - i + 1$ eigenvalues in the bin $[x, x + \Delta)$. Then $dF_{\tilde{\mathbf{A}}}(x) = \frac{j-i+1}{n}$, and $dQ(x, \mathbf{y}) = \sum_{k=i}^{j} |\mathbf{u}_k^T \mathbf{y}|^2$. From Figure 3.1, it can be seen that both $dF_{\tilde{\mathbf{A}}}(x)$ and $dQ(x, \mathbf{y})$ approximate the semicircle law very well, consistent with Lemma 3.1 and Proposition 3.2, respectively. In Figure 3.2 we repeat the same experiment as in Figure 3.1, but for a slightly different setting. In this case $n = 10^4$, and $\mathbf{y} = \frac{1}{\sqrt{n}}\mathbf{1}$. Although the size of the matrix is an order of magnitude higher, the histogram $dQ(x, \mathbf{y})$ approximate the semicircle law quite roughly, suggesting a much slower convergence of $dQ(x, \mathbf{y})$ compared to the case where $\mathbf{y} = \mathbf{e}_i$.

Next, we consider an asymmetric SBM with $n = 10^3, p_0 = 10^{-2}, p_1 = 0.1$, and $p_2 = 0.05$. In Figure 3.3, we plot $dF_{\tilde{\mathbf{A}}}(x)$ and $\frac{1}{2}(dQ(x, \mathbf{e}_1) + dQ(x, \mathbf{e}_{\frac{n}{2}+1}))$. They match very well, consistently with Corollary 3.1.

Finally, we aim at validating our theoretical results on the gaussianity of the eigenvector fluctuations. To this end, we generate 4000 independent realizations of a symmetric SBM centered adjacency matrix $\tilde{\mathbf{A}}$ with $p_0 = 0.01$, and $p_1 = p_2 = 0.1$. Using these realizations, we calculate the empirical cdf of $\sqrt{n} dQ(x, \frac{1}{\sqrt{n}}\mathbf{1})$, for $x_1 = -1.0538$ and $x_2 = 1.0489$. In both Figure 3.7 and Figure 3.8, the solid red line show the cdf of the centered variables, $\sqrt{n}(dQ(x_i, \frac{1}{\sqrt{n}}\mathbf{1}) - \mathbb{E}dQ(x, \frac{1}{\sqrt{n}}\mathbf{1})), i = 1, 2$. We compare them to the cdfs of a zero mean gaussian variable with variance properly adapted to the empirical variance of our processes. These are plotted with solid line with crosses as marker. The perfect match between the solid line and the solid lines with markers confirms the gaussianity of the perturbations of $Q(x, \frac{1}{\sqrt{n}}\mathbf{1})$. The dashed lines in Figures 3.7 and 3.8, correspond to the

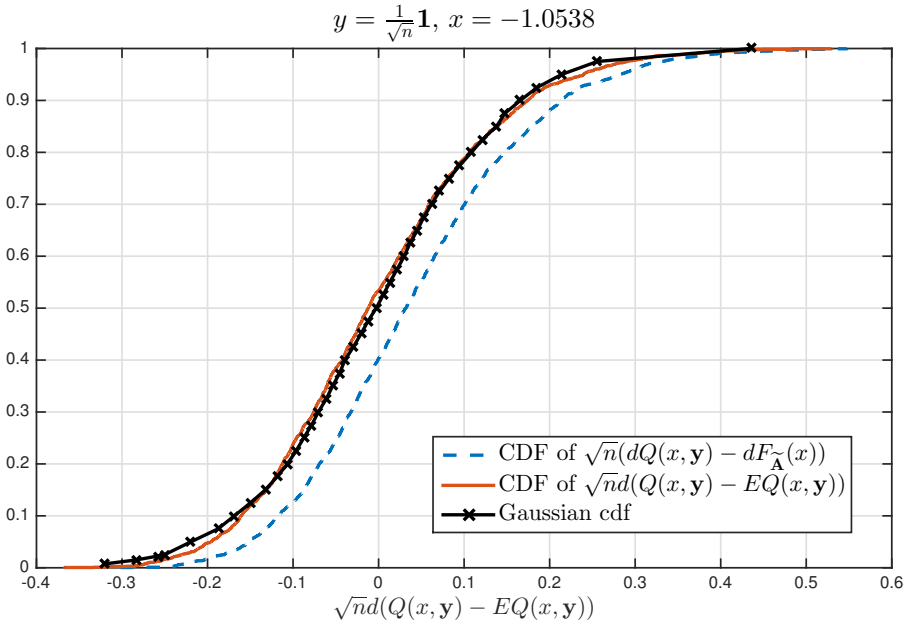



Figure 3.7: gaussianity of the fluctuations of $Q(x, \mathbf{y})$

function $\sqrt{n}(dQ(x, \frac{1}{\sqrt{n}}\mathbf{1}) - dF_{\widetilde{\mathbf{A}}}(x))$. The shift of these lines w.r.t the solid lines confirms the presence of the bias pointed out in Proposition 3.3.

## 3.7 Conclusions and Perspectives

In this chapter, we studied in detail the spectrum of adjacency matrix of a SBM with $M$ communities and derived the limiting form of the bulk spectrum of the normalized Laplacian matrix. We observed that these results can be potentially of application in varying fields such as community detection, in addition to presenting one application in the field of epidemic spreading. We obtained simulation results to substantiate the theoretical results obtained. As future work, we need to consider SBM models where sizes of communities are not equal. More general models of edge probabilities also can be studied.

We also analyzed the bulk eigenvectors of the centered adjacency matrix of SBM graphs. Following a classical approach, we studied the spectral function $Q(x, \mathbf{y})$, which depends on the eigenvectors and its fluctuations around the e.s.d. $F_{\widetilde{\mathbf{A}}}(x)$. We show that for the centered adjacency matrix $\widetilde{\mathbf{A}}$ of symmetric SBM, $Q(x, \mathbf{y})$ converges almost surely to $F_{\widetilde{\mathbf{A}}}(x)$ for any $\mathbf{y}$. This suggests that $|\mathbf{u}_i^T \mathbf{y}| \approx \frac{1}{\sqrt{n}}$, for any $\mathbf{y}$. Additionally we show that the fluctuations $\sqrt{n}\left(Q(x, \mathbf{y}) - F_{\widetilde{\mathbf{A}}}(x)\right)$ converge in distribution to a gaussian process. However, this process has non-zero mean, and hence is not a Brownian bridge. Therefore, the eigenvectors of the centered SBM adjacency matrix violates a property required for them to be Haar distributed.

We also consider the eigenvectors of the centered adjacency matrix of the asymmetric SBM, and show that the asymptotic limit of the spectral function $Q(x, \mathbf{y})$ depends on $\mathbf{y}$, as opposed to a matrix with Haar distributed eigenvectors.

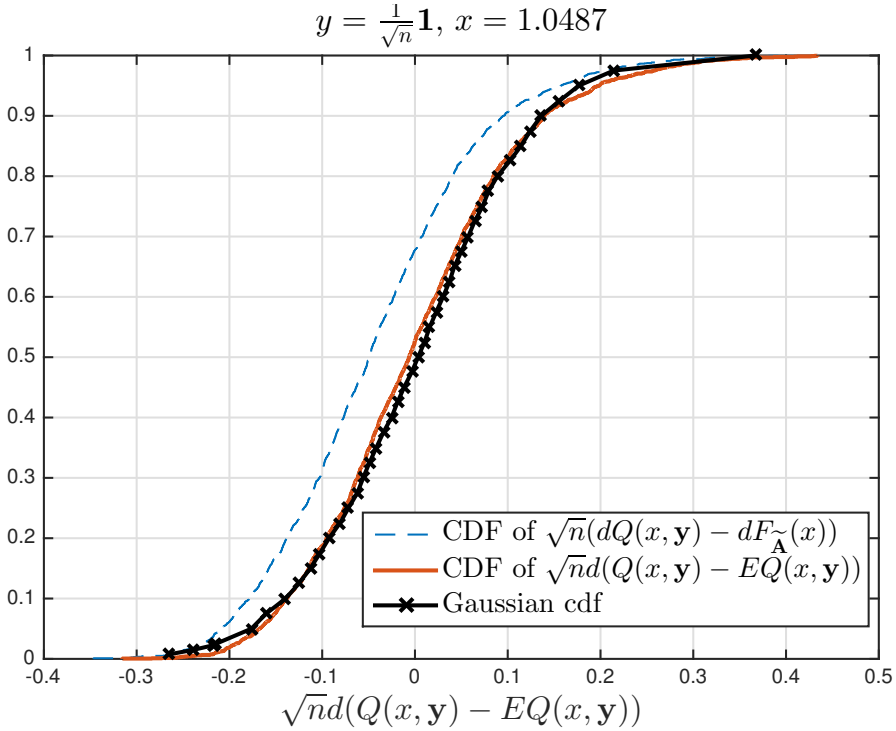



Figure 3.8: gaussianity of the fluctuations of $Q(x, \mathbf{y})$

Chapter 4

# Anomaly Detection in Erdős-Rényi Graphs

## 4.1 Introduction

Anomaly detection represents a class of important problems in Machine Learning and Data Mining. In data-driven problems, an anomaly represents a *rare* artifact in the data under consideration. The specific traits of what constitutes rare depends on the data being analyzed. Unsurprisingly, this problem area has significance in important fields such as security, forensics, network maintenance and others [Heard *et al.* 2010, Chen *et al.* 2012].

A subfield of anomaly detection is graph-based anomaly detection, which is concerned with detecting rare occurences in data instances modelled as graphs. Graphs can efficiently capture long-range correlations among data-objects in many fields such as physics, social sciences, biology, and information systems. Example problems in graph anomaly detections are detecting edge deletions or additions and/or node deletions or additions from or to an expected baseline configuration.

A survey of graph-based anomaly detection methods can be found in [Akoglu *et al.* 2015] for a wide range of real-world applications in telecommunications, auction, account, opinion, social and computer networks.

In this chapter we take a random graph based approach to anomaly detection. Consider a network modeled as a random graph, where nodes represent computers or people. A link is present between two nodes when there is an exchange between them, which happens at some expected rate. However, if there is anomaly in the network, this edge probability changes. In general, the anomaly could be a breakdown in communication, in which case the edge probability in the affected node subset is smaller than the background. On the other hand, when the anomaly corresponds to spurious elements in the network, such as terrorist transactions, the edge probability will be higher than the background. Our goal is to detect the presence of such a sub-network of spurious users. This problem was first studied in [Mifflin *et al.* 2004], where an ER graph with a planted subgraph was proposed to model the anomalous network.

Consider an Erdos-Renyi graph with $n$ nodes with edge probability $q$. When an anomaly is present, the edge probability between a random subset of nodes of size $K$ is changed to $p$. In this work, we assume that $p > q$, but the treatment of the other case is similar. The problem we address is to decide whether there exists a subset of graph vertices with an edge probability greater than $q$, given one realization of the graph.

In [Mifflin *et al.* 2004] the authors used likelihood ratio based techniques to detect the presence of the anomalous subgraph, assuming knowledge of $p, q$ and $K$. In [Miller *et al.* 2010], the authors proposed the use of $L^1$-norm of eigenvectors to detect the presence of anomalies in a graph. They validated this method on several real-world networks. However, the question of theoretical guarantees was left open. In this work, we address this question.

In our analysis, we make use of a shifted adjacency matrix of the graph defined as follows. It is the difference between the adjacency matrix of the graph and the edge probability when

there is no subgraph. A crucial observation is that in the absence of an anomaly this matrix is a symmetric matrix with independent upper triangular entries with zero mean. The eigenvectors of such a matrix have been shown to be approximately Haar distributed [Tao & Vu 2012, Bai & Pan 2012], under certain conditions on the moments of the entries. This means that a typical eigenvector of the shifted adjacency matrix is delocalized, meaning its $L^1$-norm is large.

Note that the $L^1$-norm of a unit vector $\mathbf{v}$ satisfies $1 \leq \|\mathbf{v}\|_1 \leq \sqrt{n}$, where the upper bound corresponds to the case of complete delocalization, i.e., all the entries of the vector are of the same order of magnitude, and the lower bound corresponds to the completely localized case, i.e., only one entry is non-zero. On the other hand, when there is a subgraph embedded onto the random graph, we hypothesize that there will exist an eigenvector that is "localized", i.e., a fraction of components possess most of the mass of the eigenvector. This principle is similar to that of community detection algorithms that use the dominant eigenvectors of the graph matrices to perform clustering [Newman & Girvan 2004], [Von Luxburg 2007]. Delocalization properties of eigenvectors of random matrices under a variety of distributions have been studied recently in a series of works [Bordenave & Guionnet 2013, Erdős *et al.* 2009, Rudelson & Vershynin 2015], by studying the $L^p$-norms of graph eigenvectors for $p > 2$.

Anomaly detection based on norms has been studied empirically in [Miller *et al.* 2010, Miller *et al.* 2015a]. In [Miller *et al.* 2015a], the authors look for the presence of an eigenvector whose $L^1$-norm deviates from the mean of $L^1$-norms of all the eigenvectors of the modulariy matrix, by more than a factor of the standard deviation. The subgraph is declared to be present if there exists such an eigenvector. In our work, we provide theoretical validation on a random graph model for anomaly detection based on the $L^1$-norm of only the dominant eigenvector, and show that it is possible to detect the anomaly in this way, under certain conditions on the subgraph size. Through our analysis, we find the approximate distributions of the test statistic with and without the embedded subgraph.

Our contribution in this chapter is as follows. We derive the distribution of the dominant eigenvector components of the shifted adjacency matrix when there is an embedded subgraph. We use this result to derive the asymptotic distribution of the $L^1$-norm of this eigenvector. We also look at the case where there is no subgraph embedded and use the properties of the eigenvectors of Wigner matrices as explored in [Tao & Vu 2012, Benaych-Georges 2011], to derive the $L^1$-norm of the eigenvectors when there is no subgraph embedded. Using these distributions we then devise a statistical test to detect the presence of the extraneous subgraph.

In Section 4.2 we formulate the detection problem, first in general terms, and then in the more specific case studied in this chapter. In Section 4.3, we present our anomaly detection algorithm, which is formulated as the solution to a hypothesis test problem with fixed false alarm probability. In Section 4.3.1, we describe the spectral properties of the shifted adjacency matrix $\boldsymbol{\mathcal{A}}$ under $\mathcal{H}_0$, and characterize the distribution of the $L^1$-norm of its eigenvectors. Proposition 4.1 gives the main result on the asymptotic distribution of $\chi$ under $\mathcal{H}_0$. In Section 4.3.2 we analyze the spectral properties under $\mathcal{H}_1$, and in Theorem 4.1, derive a Central Limit Theorem (CLT) for the individual components of the dominant eigenvector of $\boldsymbol{\mathcal{A}}$. Using this distribution, we compute the approximate asymptotic distribution of the $L^1$-norm statistic under $\mathcal{H}_1$ in Section 4.3.2.2. Finally in Section 4.5 we describe our conclusions and directions for future research.

## 4.2 Anomaly detection problem and statement

In this section, we formulate the general problem of anomaly detection and later, we describe the specific problem we want to analyze. Let $G = (V, E)$ denote the observed graph, where $V$ is the set of vertices, with cardinality $|V| = n$, and $E \subset V \times V$ is the set of edges. When there is no anomalous subgraph, $G = G_b$, where $G_b = (V, E_b)$ is the background graph with $E_b$ used to denote the edge set of the background graph. Let us denote the subgraph by $G_s = (V_s, E_s)$ with $V_s \subset V$, and $|V_s| = K$. When there is an embedded subgraph we have $E = E_b \cup E_s$. We desire to perform the following detection problem based on an observation of the graph $G$,

$$\mathcal{H}_0 : E = E_b \tag{4.1}$$
$$\mathcal{H}_1 : E = E_b \cup E_s. \tag{4.2}$$

In other words, Null Hypothesis $\mathcal{H}_0$ corresponds to the case when there is no embedded subgraph and all the edges of the observed graph belong to the background graph, and Hypothesis $\mathcal{H}_1$ corresponds to the case where the edges of the observed graph belong to either the background graph or the subgraph.

In this work, we focus on a specific case of the above problem where both the background graph and the embedded subgraph are independently drawn from an ER graph ensemble. For simplicity of the analysis, we allow self-loops, but this does not impact the asymptotic results. We assume $G_b = \mathcal{G}(n, q)$, and $G_s = \mathcal{G}(K, p_s)$, where $\mathcal{G}(l, q)$ denotes the class of ER random graphs of size $l$ and edge probability $q$. Under $\mathcal{H}_1$, the probability of two nodes within $V_s$ being connected is therefore $p = 1 - (1 - q)(1 - p_s) = q + p_s - qp_s$ and elsewhere the edge probability is $q$. Under $\mathcal{H}_0$, the edge probability is uniformly $q$. Without loss of generality, we assume that $V_s = \{1, 2, \dots K\}$.

It can be observed that under $\mathcal{H}_1$ the graph is an instance of the Stochastic Block Model (SBM) with two communities of size $K$ and $n - K$, within community link probabilities $p_1 = q + p_s - qp_s$ and $p_2 = q$; and outlink probability $p_0 = q$. Properties of SBM have been studied extensively in several works in the literature under the assumption of linearly increasing block sizes; see e.g. [Decelle *et al.* 2011, Avrachenkov *et al.* 2015].

The problem of subgraph detection that we consider has also been studied in [Hajek *et al.* 2015b]. In this work the authors study in detail the fundamental information theoretic limits in subgraph detection and subgraph recovery and identify easy, hard and impossible regimes with respect to the subgraph size and probability parameters under the hypothesis that a clique of size $k = o(\sqrt{n})$ cannot be detected by any polynomial time solvers. We note that the model of sublinear subgraph size was also studied in [Chen & Xu 2016, Arias-Castro *et al.* 2014].

The adjacency matrix $\mathbf{A}$ of $G$ is given as below

$$A_{ij} = A_{ji} \sim \begin{cases} \mathcal{B}(p_a) & \text{if } i, j \leq K \\ \mathcal{B}(q) & \text{otherwise} \end{cases} \tag{4.3}$$

where if a random variable (rv) $X \sim \mathcal{B}(p)$, then $X$ is a Bernouli random variable such that

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p. \end{cases}$$

We have, $p_a = p$ under $\mathcal{H}_1$ and $p_a = q$ under $\mathcal{H}_0$. Notice that $q$, $p_s$ and $K$ scale with the graph size $n$; the constraints on the actual scaling with respect to $n$ will be made explicit when the results are given. We also define $\boldsymbol{\mathcal{A}} = \mathbf{A} - q\mathbf{J}_n$. Since we are considering undirected

graphs, $\mathbf{A}$ is symmetric with independent upper diagonal entries and the same holds for $\boldsymbol{\mathcal{A}}$. Being a symmetric matrix it admits a spectral decomposition such that $\boldsymbol{\mathcal{A}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathbf{T}}$, where $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n\end{bmatrix}$, is an orthonormal matrix whose columns are made of the normalized eigenvectors with respective eigenvalues $\Lambda_{ii} = \lambda_i$, in decreasing order without loss of generality, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

## 4.3 Algorithm Description and Mathematical Analysis

In what follows, we present our anomaly detection algorithm. It is similar to the algorithm introduced in [Miller *et al.* 2010] based on finding the eigenvector of $\boldsymbol{\mathcal{A}}$ with the least $L^1$ norm.

**Algorithm 1** Anomaly Detection in a Random Graph

1: Inputs: Adjacency matrix $\mathbf{A}$, background probability $q$. Fix probability of false alarm $p_{FA}$.
2: Construct the matrix $\boldsymbol{\mathcal{A}} = \mathbf{A} - q\mathbf{J}$.
3: Compute the eigenvector $\mathbf{u}_1$ corresponding to eigenvalue $\lambda_1$, and find $\chi = \|\mathbf{u}_1\|_1$.
4: Find $\tau$, such that (s.t.) $\mathbb{P}_{\mathcal{H}_0}(\chi < \tau) = p_{FA}$, i.e., $\tau = \mu_{(0)} + \sigma_{(0)}\Phi^{-1}(p_{FA})$
5: If $\chi < \tau$, declare $\mathcal{H}_1$, otherwise $\mathcal{H}_0$, where $\Phi$ is the Cumulative Density Function (CDF) of $\mathcal{N}(0,1)$.

### 4.3.1 Statistic Distribution under $\mathcal{H}_0$

Under $\mathcal{H}_0$, $\boldsymbol{\mathcal{A}}$ is a symmetric matrix with independent zero mean upper triangular entries as given below

$$\mathcal{A}_{ij} = \mathcal{A}_{ji} = \begin{cases} 1-q & \text{w.p. } q \\ -q & \text{w.p. } 1-q \end{cases}$$

i.e., the components of $\boldsymbol{\mathcal{A}}$ are independent on and above the diagonal, with zero mean, and variance $q(1-q)$. Thus the matrix $\boldsymbol{\mathcal{A}}$ under $\mathcal{H}_0$ is a standard Wigner matrix. Its spectral properties such as the empirical spectral distribution and the spectral radius are well-studied in the literature under different scaling laws on $q$, see e.g., [Ding *et al.* 2010], and also Chapter 3. The eigenvectors of Wigner matrices are approximately Haar-distributed on the space of unitary matrices in $\mathbb{R}^{n\times n}$ as suggested by partial results on universality of eigenvector statistics in [Tao & Vu 2012, Bai & Pan 2012]. In other words, a typical eigenvector $\mathbf{u}_i$ is approximately uniformly distributed on the hypersphere, $\mathbf{S}^{n-1} = \{\mathbf{s} : \|\mathbf{s}\|_2 = 1\}$. A random unit vector on the hypersphere can be modelled as a gaussian eigenvector normalized to have unit $L^2$ norm, i.e., $\frac{\mathbf{x}}{\|\mathbf{x}\|}$, with $\mathbf{x}$ being a $\mathbb{R}^n$ gaussian vector with covariance matrix $\mathbf{I}$, i.e., $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$. We assume the following fact, which is a widely held conjecture about the asymptotic distribution of the eigenvectors of a Wigner matrix. This holds exacly for Wigner matrices with gaussian entries such as the Gaussian Unitary ensemble and the Gaussian Orthogonal Ensemble [Anderson *et al.* 2009].

**Approximation 4.1.** *(Haar distribution of Eigenvectors of a Wigner matrix) A typical eigenvector* $\mathbf{u}_i$ *of* $\boldsymbol{\mathcal{A}}$ *under hypothesis* $\mathcal{H}_0$ *is distributed uniformly on the hypersphere on* $\mathbf{S}^{(n-1)}$. *The distribution of a typical eigenvector* $\mathbf{u}_i$ *is identical to the distribution of* $\mathbf{x}/\|\mathbf{x}\|$, *where* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

Let us define $g(\mathbf{x}) = \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|}$. Below we derive a central limit theorem for $g(\mathbf{x})$, when $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix in $\mathbb{R}^{n\times n}$ such that $\Sigma_{ii} = \mathbb{E}x_i^2 = \sigma_i^2$.

**Lemma 4.1.** *(Central Limit Theorem for $\|\mathbf{x}\|_1/\|\mathbf{x}\|$) Let* $\mathbf{x}$ *be a gaussian random vector with i.i.d. components, then* $g(\mathbf{x})$ *satisfies a central limit theorem with the limit distribution being gaussian with mean* $\mu_0 = \sqrt{\frac{n}{\alpha_2}}\alpha_1$ *and variance* $\sigma_0^2 = \frac{1}{\alpha_2}\left(C_{11} + (\frac{\alpha_1}{2\alpha_2})^2 C_{22} - \frac{\alpha_1}{\alpha_2}C_{12}\right)$, *where* $\alpha_1 = \mathbb{E}(|x_1|)$, $\alpha_2 = \mathbb{E}(|x_1|^2)$, $C_{11} = \mathrm{Var}(|x_1|), C_{22} = \mathrm{Var}(|x_1|^2), C_{12} = \mathbb{E}((|x_1| - \mathbb{E}(|x_1|))(|x_1|^2 - \mathbb{E}(|x_1|^2)))$, *i.e.,* $g(\mathbf{x}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mu_0, \sigma_0^2)$).

*Proof.* Consider the two dimensional vector $\mathbf{z}_i = \begin{pmatrix} |x_i| \\ |x_i|^2 \end{pmatrix}$, and $\mathbf{z}^{(n)} = \sum_{i=1}^n \mathbf{z}_i$. Note that $\mathbf{z}_i$ are i.i.d. random vectors in $\mathbb{R}^2$, with mean $\mathbf{m} = \mathbf{E}\mathbf{z}_i = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$, and covariance matrix

$$\mathbf{C} = \begin{bmatrix} \mathbb{E}|x_i|^2 - (\mathbb{E}|x_i|)^2 & \mathbb{E}|x_i|^3 - \mathbb{E}|x_i|^2\mathbb{E}|x_i| \\ \mathbb{E}|x_i|^3 - \mathbb{E}|x_i|^2\mathbb{E}|x_i| & \mathbb{E}|x_i|^4 - (\mathbb{E}|x_i|^2)^2 \end{bmatrix}.$$

Hence, by applying the multidimensional CLT , see [Billingsley 2008], we conclude that the distribution of $\mathbf{r}^{(n)} = \frac{1}{\sqrt{n}}\left(\mathbf{z}^{(n)} - n\mathbf{m}\right)$ converges to $\mathcal{N}(0, \mathbf{C})$. Now the function $g(\mathbf{x})$ can be represented as a function of the vector $\mathbf{z}^{(n)}$, which we denote as $g$ for brevity. By the Skorohod representation theorem see [Billingsley 2008] there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ where we can construct a sequence of random vectors $\mathbf{r}^{(n)}$ that converges in the almost sure sense to the random vector $\mathbf{r}$ with distribution $\mathcal{N}(\mathbf{0}, \mathbf{C})$. Therefore,

$$\begin{aligned}
g &= \frac{z_1^{(n)}}{\sqrt{z_2^{(n)}}} \\
&= (\sqrt{n}r_1 + n\alpha_1)(\sqrt{n}r_2 + n\alpha_2)^{-1/2} \\
&= \frac{1}{\alpha_2^{1/2}}(r_1 + \sqrt{n}\alpha_1)(1 - \frac{1}{2}\frac{r_2}{\alpha_2\sqrt{n}} + \mathrm{o}_p(n^{-1/2})) \\
&= \frac{1}{\alpha_2^{1/2}}(\sqrt{n}\alpha_1 - \frac{r_2}{2\alpha_2}\alpha_1 + r_1 - \mathrm{O}_p(n^{-1/2}) + \mathrm{o}_p(n^{-1/2})) \\
&= \sqrt{n}\frac{\alpha_1}{\sqrt{\alpha_2}} + \frac{1}{\sqrt{\alpha_2}}(r_1 - \frac{r_2}{2}\frac{\alpha_1}{\alpha_2}) + \mathrm{o}_p(1),
\end{aligned}$$

Therefore we obtain

$$g - \sqrt{n}\frac{\alpha_1}{\sqrt{\alpha_2}} = \frac{1}{\sqrt{\alpha_2}}(r_1 - \frac{\alpha_1}{2\alpha_2}r_2) + \mathrm{o}_p(1). \tag{4.4}$$

Since the vector $\mathbf{r}^{(n)}$ almost surely converges to the vector $\mathbf{r}$, any continuous function $f(\mathbf{r}^{(n)})$ converges to $f(\mathbf{r})$) almost surely by the Continuous Mapping Theorem, where in our case $f(\mathbf{r}) = \frac{1}{\sqrt{\alpha_2}}(r_1 - \frac{\alpha_1}{2\alpha_2}r_2)$. But this is a linear combination of two jointly gaussian random variables, and hence is also a gaussian rv with mean 0, and variance $\beta_1 + \beta_2\frac{\alpha_1^2}{4} - \alpha_1\beta_{12}$. Also, by the fact that if $x_n, y_n$ are two random variable sequences such that $x_n \to x$ a.s. and $y_n \to y$ in probability, then $x_n + y_n \to x + y$ in probability, the right hand side of (4.4) is a random variable that converges in probability to a gaussian random variable with mean 0, and variance $\sigma_{(0)}^2 = \frac{1}{\alpha_2}\left(C_{11} + (\frac{\alpha_1}{2\alpha_2})^2 C_{22} - \frac{\alpha_1}{\alpha_2}C_{12}\right) = 1 - 3/\pi$, and hence $g$ converges to a gaussian random variable with mean $\mu_{(0)} = \sqrt{n}\frac{\alpha_1}{\sqrt{\alpha_2}} = \sqrt{\frac{2n}{\pi}}$ and variance $\sigma_{(0)}^2$. Now $g(\mathbf{x})$ has the same distribution as $g$. Therefore $g(\mathbf{x})$ converges in distribution to $\mathcal{N}(\mu_0, \sigma_0^2)$. □

**Proposition 4.1.** *Under* $\mathcal{H}_0$, $\chi \sim \mathcal{N}(\mu_0, \sigma_0^2)$, *asymptotically in distribution, where* $\mu_0 = \sqrt{\frac{2n}{\pi}}$, *and* $\sigma_0^2 = 1 - \frac{3}{\pi}$.

*Proof.* The proof uses Approximation 4.1 and follows from Lemma 4.1, where $\alpha_1 = \mathbb{E}(|x_1|) = \sqrt{\frac{2}{\pi}}$, $\alpha_2 = \mathbb{E}(|x_1|^2) = 1$, $C_{11} = \mathrm{Var}(|x_1|) = 1 - 2/\pi, C_{22} = \mathrm{Var}(|x_1|^2) = 2, C_{12} = \mathbb{E}((|x_1| - \mathbb{E}(|x_1|))(|x_1|^2 - \mathbb{E}(|x_1|^2))) = \sqrt{\frac{2}{\pi}}$. □

### 4.3.2 Statistic Distribution under $\mathcal{H}_1$

Under hypothesis $\mathcal{H}_1$, Matrix $\boldsymbol{\mathcal{A}}$ is given as below

$$\mathcal{A}_{ij} = \begin{cases} \begin{cases} 1-q & \text{w.p. } p \\ -q & \text{w.p. } 1-p \end{cases}, & \text{if } 1 \le i,j \le K, \\ \begin{cases} 1-q & \text{w.p. } q \\ -q & \text{w.p. } 1-q \end{cases} & \text{if } i > K \text{ or } j > K, \end{cases}$$

Thus under $\mathcal{H}_1$, Matrix $\boldsymbol{\mathcal{A}}$ has a non-zero mean $\overline{\boldsymbol{\mathcal{A}}} = \mathbb{E}_{\mathcal{H}_1}(\boldsymbol{\mathcal{A}})$ given by

$$\overline{\boldsymbol{\mathcal{A}}} = \begin{bmatrix} (p-q)\mathbf{J}_K & \mathbf{0}_{K\times n-K} \\ \mathbf{0}_{n-K\times K} & \mathbf{0}_{n-K\times n-K} \end{bmatrix}. \tag{4.5}$$

This matrix has rank 1, and with a single non-zero eigenvalue $K\delta_p$, with eigenvector $\frac{1}{\sqrt{K}}\begin{bmatrix} \mathbf{1}_K \\ \mathbf{0}_{n-K\times 1} \end{bmatrix}$.

Also, note that the components $\mathcal{A}_{ij}$ such that $1 \le i,j \le K$ have a variance of $p(1-p)$, while the other components have a variance of $q(1-q)$. Let $\delta_p \coloneqq p - q$.

Intuitively, $\overline{\boldsymbol{\mathcal{A}}}$ is the subgraph component, and when the subgraph component is large enough, we can conceivably detect the subgraph from the observed graph. Specifically, if the eigenvalue of $\overline{\boldsymbol{\mathcal{A}}}$ is large to be separate enough from the spectrum of $\boldsymbol{\mathcal{A}} - \overline{\boldsymbol{\mathcal{A}}}$, we expect to be able to detect the embedded subgraph. We have the following proposition on the asymptotic bound on $\|\boldsymbol{\mathcal{A}} - \overline{\boldsymbol{\mathcal{A}}}\|$.

**Proposition 4.2.** *Under the condition that* $\max(q(1-q), p(1-p)) > C\frac{\log^4(n)}{n}$ *for some* $C, \exists c$ *such that*

$$\|\boldsymbol{\mathcal{A}} - \overline{\boldsymbol{\mathcal{A}}}\| \le c\sqrt{\max(q(1-q), p(1-p))n} \text{ almost surely (a.s.).}$$

If $q$ does not scale with $n$, the condition in the proposition is immediately satisfied. Let us consider the case where the embedded subgraph is a clique, i.e., $p_s = p = 1$. Then $\sigma^2 = \sigma_0^2 n = q(1-q)n$, and the condition is satisfied when $nq \gg \log(n)$; similarly when both $p, q$ are decreasing functions of $n$, the condition is easily verified to be satisfied when $nq \gg \log(n)$. From now on, we assume that $q \le p \le 0.5$ and hence the bound in the above proposition becomes $c\sqrt{np}$.

*Proof.* The result is an application of Theorem 2.2. In this theorem, by taking $\sigma^2 = \max(p(1-p), q(1-q))$, we have the required result. □

**Definition 4.** *(Spectral gap G) We define the spectral gap* $\Delta$ *as the difference between the maximum eigenvalue of the mean matrix and edge of the spectrum*

$$\begin{aligned} G &= K\delta_p - \|\boldsymbol{\mathcal{A}} - \overline{\boldsymbol{\mathcal{A}}}\| \\ &\ge K\delta_p - c\sqrt{np} \text{ a.s.} \\ &= G_0. \end{aligned}$$

Let

$$\Delta := c\frac{\sqrt{np}}{K\delta_p}. \tag{4.6}$$

In Lemma 4.2, we will show that a.s.,

$$K\delta_p(1-\Delta) \leq \lambda \leq K\delta_p(1+\Delta), \tag{4.7}$$

and in Proposition 4.2, we will prove that a.s.

$$|\lambda_i| \leq c\sqrt{np} = K\delta_p\Delta, \text{ a.s.} \tag{4.8}$$

for $i \geq 2$.

#### 4.3.2.1 Eigenvector distribution under $\mathcal{H}_1$

We are interested in the dominant eigenvector of $\boldsymbol{\mathcal{A}}$, the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\mathcal{A}}$. We develop a CLT for the components of this eigenvector. We use the ideas in [Athreya *et al.* 2013], where the authors derived a CLT for the components of the eigenvector of a single dimensional Random Dot Product Graph(RDPG). However, while the result in [Athreya *et al.* 2013] only holds for dense graphs (i.e., $nq = \Theta(n)$), our result holds more generally for sparse graphs, i.e., $nq = n^{o(1)}$. Similarly, our result also contains an extension to the case $K = o(n)$, which does not follow directly from the results in [Athreya *et al.* 2013], which are limited to the case $K = cn$, for some constant $c > 0$. Throughout this section the distributions of the random variables correspond to those under $\mathcal{H}_1$, and this fact is not explicitly noted from here onwards.

Let $\mathbf{u} := \mathbf{u}_1(\boldsymbol{\mathcal{A}})$, be the normalized dominant eigenvector corresponding to the eigenvalue $\lambda := \lambda_1(\boldsymbol{\mathcal{A}})$. Observe that the mean matrix $\overline{\boldsymbol{\mathcal{A}}}$ can be written as $\overline{\mathbf{x}}\overline{\mathbf{x}}^T$, where $\overline{\mathbf{x}} = \sqrt{\delta_p}\begin{bmatrix}\mathbf{1}_K^T & \mathbf{0}_{n-K}^T\end{bmatrix}^T$, with a single non-zero eigenvalue $\overline{\lambda} = K\delta_p$ and its eigenvector as $\overline{\mathbf{u}} = \frac{\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|}$.

Let us define $\mathbf{x}$ as $\mathbf{x} = \lambda^{1/2}\mathbf{u}$, and so $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|_2$. As we will soon show, when there is a non-diminishing spectral gap $G$, $\lambda \approx \overline{\lambda} = K\delta_p$, for large $n$, and in addition, a random realization of $\mathbf{x}$ would be close to $\overline{\mathbf{x}}$. Therefore the $i^{\text{th}}$ component of $\mathbf{x}$ would have a limiting distribution with mean $\overline{x}_i$. We can then derive the limiting distribution of the $L^1$-norm statistic from the distribution of $\mathbf{x}$. We state below the conditions under which our results hold.

**Assumption 4.1.**

$$q > C\frac{\log^4(n)}{n}$$

**Assumption 4.2.**

$$\frac{p}{q} = O(1)$$

**Assumption 4.3.**

$$K\delta_p = \omega((nq)^{2/3})$$

*Discussion of the Conditions:*
Assumption 4.1 is needed to ensure that the graph is dense enough to apply the concentration results we use in the proofs. Assumption 4.2 implies that $p$ and $q$ are of the same order, and thus we are in the hard regime of detection. The next assumption, Assumption 4.3 is required so that the spectral gap $G$ is large enough to prove the results on the CLT of the eigenvector components presented in this chapter. Notice that Assumption 4.3 implies

$$Kq = \omega(1). \tag{4.9}$$

We need the following concentration lemma for the eigenvalue $\lambda$, based on the Bauer-Fike lemma ( [Saad 1992]).

**Lemma 4.2.** *Under Condition 4.3, $\lambda \to K\delta_p$ a.s. as $n \to \infty$.*

*Proof.* By Weyl's identities ( [Saad 1992]) and Proposition 4.2,

$$|\lambda - K\delta_p| \leq c\sqrt{np},$$

a.s. Therefore,

$$\left|\frac{\lambda}{K\delta_p} - 1\right| \leq c\frac{\sqrt{np}}{K\delta_p} \tag{4.10}$$

which implies $\lambda \to K\delta_p$, a.s., by Condition 4.3, where in (4.10) we used the fact that $K\delta_p < nq$, which follows from Condition 4.2. □

We present below our main theorem on the CLT of the components of the dominant eigenvectors.

**Theorem 4.1.** *Under Assumptions 4.2 and 4.3, the following CLT holds true for the entries of the unnormalized eigenvector $\mathbf{x} = \lambda^{1/2}\mathbf{u}$.*

$$\sqrt{\frac{K\delta_p}{p(1-p)}}\left(x_i - \sqrt{\delta_p}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{4.11}$$

*for $1 \leq i \leq K$, and*

$$\sqrt{\frac{K\delta_p}{q(1-q)}}x_i \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{4.12}$$

*for $1 + K \leq i \leq n$.*

**Note on Subgraph recovery**

Using Theorem 4.1 we can design a detector that approximately detects the dense subgraph vertices by thresholding the components of $x_i$. Define a threshold $\tau := \frac{\sqrt{\delta_p}}{2}$. Let the detector be $T_i, i \in V$ such that $T_i = 1$ implies $i \in \overline{S}$ for a subgraph estimate $\overline{S}$ and zero otherwise, where

$$T_i = \chi_{\{x_i > \tau\}}.$$

We can show that this detector approximately recovers the subgraph nodes in the following sense:

$$\lim_{n\to\infty} \mathbb{P}(T_i = 1 | i \notin S) + \mathbb{P}(T_i = 0 | i \in S) = 0,$$

i.e., the sum of false detection and missed detection probabilities tends to zero as $n \to \infty$. We first consider the first term in the above sum, i.e., the probability of false alarm. We get

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(T_i = 1 | i \notin S) &= \lim_{n\to\infty} \mathbb{P}(x_i > \tau | i \notin S) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\sqrt{\frac{K\delta_p}{q(1-q)}}x_i > \sqrt{\frac{K\delta_p}{q(1-q)}}\tau\right) \\
&= Q\left(\lim_{n\to\infty} \frac{1}{2}\sqrt{K\delta_p}\right) = 0,
\end{aligned}$$

since by assumption $K\delta_p = \omega(1)$, where $Q(x)$ is the c.c.d.f. of a standard gaussian rv. Similarly

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(T_i = 1 | i \notin S) &= \lim_{n\to\infty} \mathbb{P}(x_i < \tau | i \in S) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\sqrt{\frac{K\delta_p}{p(1-p)}}\left(x_i - \sqrt{\delta_p}\right) < -\frac{\sqrt{\delta_p}}{2}\sqrt{\frac{K\delta_p}{p(1-p)}}\right) \\ &= \Phi(\lim_{n\to\infty} -\frac{1}{2}\sqrt{K\delta_p}) = 0,\end{aligned}$$

where $\Phi(\cdot)$ is the c.d.f. of a standard gaussian r.v. Therefore under Assumptions 4.1 to 4.3, a thresholding of the scaled eigenvector can approximately recover the subgraph nodes.

We compare the recovery threshold of this algorithm with the limits given in Figure 2 of [Hajek *et al.* 2015b]. Under the scaling used in [Hajek *et al.* 2015b] we take $q = n^{-\alpha}$ and $K = n^{\beta}$. Then under Assumptions 4.1 to 4.3 we must have $\alpha \in [0, 1)$ and $\beta > \frac{2}{3} + \frac{\alpha}{3}$, which is in the *easy* regime for subgraph recovery given in [Hajek *et al.* 2015b]. We recognize that this is suboptimal in view of the regime of recovery achievable by known polynomial-time algorithms [Ames 2013, Chen & Xu 2016]; we leave it to future work to investigate if it is possible to improve the performance of this simple algorithm.

*Proof.* Define $\gamma_i = \sqrt{\frac{K\delta_p}{p(1-p)}}$ for $1 \le i \le K$ and $\gamma_i = \sqrt{\frac{K\delta_p}{q(1-q)}}$ for $K+1 \le i \le n$. Notice that $x_i = \frac{1}{\lambda^{1/2}}[\boldsymbol{\mathcal{A}}\mathbf{u}]_i$ and $\overline{x}_i = \frac{1}{\overline{\lambda}^{1/2}}\left[\overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}\right]_i = \sqrt{\delta_p}$ for $1 \le i \le K$ and $\overline{x}_i = 0$ for $K+1 \le i \le n$. In the following, $[\mathbf{z}]_i$ denotes the $i^{\text{th}}$ component of vector $\mathbf{z}$. We can write

$$\begin{aligned}\gamma_i(x_i - \overline{x}_i) = \gamma_i\left(\frac{1}{\lambda^{1/2}}[\boldsymbol{\mathcal{A}}(\mathbf{u} - \overline{\mathbf{u}})]_i\right) + \gamma_i\left(\frac{1}{\lambda^{1/2}}[\boldsymbol{\mathcal{A}}\overline{\mathbf{u}} - \overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}]_i\right) + \\ \gamma_i\left(\left(\frac{1}{\lambda^{1/2}} - \frac{1}{\overline{\lambda}^{1/2}}\right)[\overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}]_i\right) \\ := T_1 + T_2 + T_3.\end{aligned}$$

We treat each of the above three terms separately as below.

- We show that $T_1 = \gamma_i\left(\frac{1}{\lambda^{1/2}}[\boldsymbol{\mathcal{A}}(\mathbf{u} - \overline{\mathbf{u}})]_i\right) \to 0$ in probability, in Lemma 4.5.
- We show $T_2 = \gamma_i\left(\frac{1}{\lambda^{1/2}}[\boldsymbol{\mathcal{A}}\overline{\mathbf{u}} - \overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}]_i\right)$ satisfies a CLT and is asymptotically distributed as $\mathcal{N}(0, 1)$, in Lemma 4.3.
- Finally we show that $T_3 = \gamma_i\left((\frac{1}{\lambda^{1/2}} - \frac{1}{\overline{\lambda}^{1/2}})[\overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}]_i\right) \to 0$, for $1 \le i \le K$ in probability in Lemma 4.2, by showing a concentration result for the dominant eigenvalue $\lambda$. Notice that $T_3 = 0$ for $i > K$.

The result then follows by an application of Slutsky's thereom [Billingsley 2008]. □

Let us define $\mathbf{y} \in \mathbb{R}^n$ as follows

$$\mathbf{y} = \frac{1}{\lambda^{1/2}}\boldsymbol{\mathcal{A}}\overline{\mathbf{u}}.$$

In the following Lemma, we prove a CLT for the components of $\mathbf{y}$.

**Lemma 4.3.** *Under Assumptions 4.2 and 4.3 we have*

$$\sqrt{\frac{K\delta_p}{p(1-p)}}\left(y_i - \frac{\|\overline{\mathbf{x}}\|\overline{x}_i}{\lambda^{1/2}}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{4.13}$$

*for* $1 \leq i \leq K$, *and*

$$\sqrt{\frac{K\delta_p}{q(1-q)}}y_i \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{4.14}$$

*for* $1+K \leq i \leq n$.

*Proof:*
We prove (4.13) and the proof for (4.14) follows along the same lines. Observe that

$$\begin{aligned} y_i - \frac{\|\overline{\mathbf{x}}\|\overline{x}_i}{\lambda^{1/2}} &= \frac{1}{\lambda^{1/2}}\sum_{j=1}^{n}\mathcal{A}_{ij}\overline{u}_j - \frac{\|\overline{\mathbf{x}}\|\overline{x}_i}{\lambda^{1/2}} \\ &= \frac{1}{\lambda^{1/2}}\left(\sum_{j=1}^{n}\mathcal{A}_{ij}\overline{x}_j/\|\overline{\mathbf{x}}\| - \overline{x}_i\|\overline{\mathbf{x}}\|\right) \\ &= \frac{1}{\lambda^{1/2}\|\overline{\mathbf{x}}\|}\left(\sum_{j=1}^{K}\mathcal{A}_{ij}\overline{x}_j - \overline{x}_i\|\overline{\mathbf{x}}\|^2\right) \qquad (4.15) \\ &= \frac{1}{\lambda^{1/2}\|\overline{\mathbf{x}}\|}\left(\sum_{j=1}^{K}(\mathcal{A}_{ij} - \overline{x}_i\overline{x}_j)\overline{x}_j\right), \end{aligned}$$

where in (4.15) we used the fact that $\overline{x}_i = 0$, for $i > K$.

Notice $\|\overline{\mathbf{x}}\| = \sqrt{\overline{x}_1^2 + \overline{x}_2^2 + \overline{x}_3^2 + \ldots \overline{x}_n^2} = \sqrt{K\delta_p}$ deterministically. Thus we obtain

$$\begin{aligned} \sqrt{\frac{K\delta_p}{p(1-p)}}(y_i - \frac{\|\overline{\mathbf{x}}\|\overline{x}_i}{\lambda^{1/2}}) &= \frac{\sqrt{K\delta_p}}{\lambda^{1/2}\sqrt{K\delta_p(p(1-p))}}\left(\sum_{j=1}^{K}(\mathcal{A}_{ij} - \overline{x}_i\overline{x}_j)\overline{x}_j\right) \\ &= \frac{\sqrt{K\delta_p}}{\sqrt{Kp(1-p)}\lambda^{1/2}}\left(\sum_{j=1}^{K}(\mathcal{A}_{ij} - \delta_p)\right), \end{aligned}$$

since $\overline{x}_i = \sqrt{\delta_p}$ for $1 \leq i \leq K$. We invoke the Lindeberg Central Limit Theorem [Billingsley 2008] to determine the asymptotic distribution of the above.

**Theorem 4.2.** *(Lindeberg Central Limit Theorem) Suppose that for each* $n$,

$$X_{n1}, X_{n2}, \ldots X_{nr_n}$$

*are independent, with* $\mathbb{E}X_{nk} = 0, \sigma_{nk}^2 = \mathbb{E}X_{nk}^2$, *and define* $s_n^2 = \sum_{k=1}^{r_n}\sigma_{nk}^2$. *Define* $S_n = \sum_{k=1}^{r_n}X_{nk}$. *Then* $S_n/s_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$, *if*

$$\lim_{n\to\infty}\sum_{k=1}^{r_n}\frac{1}{s_n^2}\mathbb{E}X_{nk}^2\mathbb{I}\{|X_{nk}| \geq \varepsilon s_n\} = 0, \tag{4.16}$$

$\forall \varepsilon > 0$.

Now take $S_n = \sum_{j=1}^{K}(\mathcal{A}_{ij} - \delta_p)\sqrt{\delta_p}$, then $X_{nk} \coloneqq (\mathcal{A}_{ij} - \delta_p)\sqrt{\delta_p}$, and $\mathbb{E}X_{nk} = 0$, and $\sigma_{nk}^2 = \mathbb{E}X_{nk}^2 = \delta_p p(1-p)$, giving $s_n = K\delta_p p(1-p)$. Then the left hand side of condition (4.16) becomes

$$\lim_{n\to\infty} \frac{K}{K\delta_p p(1-p)} \mathbb{E}X_{nk}^2 \mathbb{I}\{|X_{nk}|/\sqrt{K\delta_p p(1-p)} \geq \varepsilon\},$$

because $X_{nk}$ are i.i.d. random variables. The above is equivalent to

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}\left(\frac{X_{nk}}{\sqrt{\delta_p p(1-p)}}\right)^2 \mathbb{I}\left\{\frac{|X_{nk}|}{\sqrt{\delta_p p(1-p)}} \geq \varepsilon\sqrt{K}\right\} \\ &\coloneqq \lim_{n\to\infty} \mathbb{E}\tilde{X}_{nk}^2 \mathbb{I}\left\{|\tilde{X}_{nk}| \geq \varepsilon\sqrt{K}\right\}, \end{aligned} \tag{4.17}$$

where $\tilde{X}_{nk} = X_{nk}/\sqrt{\delta_p p(1-p)}$ is given as

$$\tilde{X}_{nk} = \begin{cases} \frac{1-p}{\sqrt{p(1-p)}} & \text{w.p. } p \\ \frac{-p}{\sqrt{p(1-p)}} & \text{w.p. } 1-p. \end{cases}$$

Therefore we can write (4.17) as

$$\frac{1-p}{p}\chi\left(\sqrt{\frac{1-p}{pK}} \geq \varepsilon\right) + \frac{p}{1-p}\chi\left(\sqrt{\frac{p}{Kp}} \geq \varepsilon\right).$$

Clearly, if $Kp = \omega(1), \exists N$, s.t. the above is zero $\forall n > N$, and $\varepsilon > 0$. Hence Lindeberg condition is satisfied, and we obtain that

$$\lim_{n\to\infty} \frac{1}{\sqrt{K\delta_p p(1-p)}} \sum_{j=1}^{K}(\mathcal{A}_{ij} - \delta_p)\sqrt{\delta_p} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

or equivalently,

$$\lim_{n\to\infty} \frac{1}{\sqrt{Kp(1-p)}} \sum_{j=1}^{K}(\mathcal{A}_{ij} - \delta_p) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{4.18}$$

Thus by applying Slutsky's theorem with Lemma 4.2 and (4.18) we obtain the result for $1 \leq i \leq K$.

Similarly, for $K+1 \leq i \leq n$,

$$\begin{aligned} \sqrt{\frac{K\delta_p}{q(1-q)}} y_i &= \sqrt{\frac{K\delta_p}{q(1-q)}} \lambda^{-1/2} \sum_{j=1}^{K} \mathcal{A}_{ij}\overline{u}_j, \\ &= \sqrt{\frac{K\delta_p}{\lambda}} \frac{1}{\sqrt{Kq(1-q)}} \sum_{j=1}^{K} \mathcal{A}_{ij} \\ &\xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \end{aligned}$$

where the proof follows from another application of Theorem 4.2, Lemma 4.2 and Slutsky's Theorem, provided that $Kq \to \infty$, which follows from (4.9). To complete the proof of Theorem 4.1, we need to first derive an entry-wise error bound between the eigenvector $\overline{\mathbf{u}}$ of $\overline{\boldsymbol{\mathcal{A}}}$ and the dominant eigenvector $\mathbf{u}$, of $\boldsymbol{\mathcal{A}}$ which we present in the following lemma.

Armed with the results we have thus far, we are now prepared to prove the main central limit theorem, a CLT for each individual component of the non-normalized dominant eigenvector $\mathbf{x}$ of $\boldsymbol{\mathcal{A}}$.

In order to prove Theorem 4.1, we need an error bound between $\mathbf{u}$ and $\bar{\mathbf{u}}$. To derive this we use the traditional Davis-Kahan theorem from [Bhatia 2013], which we quote below.

**Theorem 4.3.** *(Davis-Kahan Theorem [Bhatia 2013]) Let* $\mathbf{C}$ *and* $\mathbf{D}$ *be two Hermitian operators, and let* $S_1, S_2$ *be any two subsets of* $\mathbb{R}$ *such that the distance between the two subsets,* $d(S_1, S_2) = \delta > 0$. *Let* $\mathbf{E} = P_{\mathbf{C}}(S_1)$, *the projection matrix on to the space spanned by the eigenvectors of* $\mathbf{C}$ *whose eigenvalues fall in* $S_1$, *and similarly,* $\mathbf{F} = P_{\mathbf{D}}(S_2)$. *Then, for every unitarily invariant matrix norm* [1] $|||\cdot|||$,

$$|||\mathbf{EF}||| \leq \frac{c}{\delta}|||\mathbf{C} - \mathbf{D}|||$$

*where* $c$ *is a fixed constant. In fact,* $c = \pi/2$.

Using the above, we derive the following result.

**Lemma 4.4.** *Let* $\Delta$ *be as defined in (4.6). Then a.s.,*

$$\|\mathbf{u} - \bar{\mathbf{u}}\|_2 \leq \frac{c\Delta}{1 - 2\Delta},$$

*where* $c$ *is a constant independent of* $n$.

*Proof:*
In the notation of Theorem 4.3, choose $\mathbf{C} := \overline{\boldsymbol{\mathcal{A}}}$, and $\mathbf{D} := \boldsymbol{\mathcal{A}}$. Let us take $S_1 = [-a_n, a_n]$, where $a_n = K\delta_p\Delta$. Then $S_1$ does not contain the non-zero eigenvalue $\overline{\lambda}$ of $\mathbf{C}$, and hence $\mathbf{E} = \mathbf{P}_{\mathbf{C}}(S_1)$ is the projection matrix on to the orthogonal space of $\bar{\mathbf{u}}$, and therefore, $\mathbf{E} = \mathbf{I} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T$. Let $S_2 = [K\delta_p(1-\Delta), -\infty)$, such that, for a sufficiently large $n$, it only contains the dominant eigenvalue of $\boldsymbol{\mathcal{A}}$. Therefore, $\mathbf{F} = \mathbf{P}_{\mathbf{D}}(S_2) = \mathbf{u}\mathbf{u}^T$. Demonstrably, $\delta$ in Theorem 4.3 satisfies $\delta > K\delta_p(1-\Delta) - K\delta_p\Delta = K\delta_p(1-2\Delta)$. Also, we choose $|||\cdot||| := \|\cdot\|_2$, the induced $L^2$-norm on matrices, which is unitarily invariant. From Proposition 4.2 it holds that $\|\overline{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_2 \leq K\delta_p\Delta$. Also,

$$\begin{aligned}
\|\mathbf{EF}\|_2 &= \|(\mathbf{I} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T)\mathbf{u}\mathbf{u}^T\|_2 \\
&= \|\mathbf{u}\mathbf{u}^T - \bar{\mathbf{u}}(\bar{\mathbf{u}}^T\mathbf{u})\mathbf{u}^T\|_2 \\
&= \|(\mathbf{u} - \alpha\bar{\mathbf{u}})\,\mathbf{u}^T\|_2 \qquad (4.19)\\
&= \|\mathbf{u} - \alpha\bar{\mathbf{u}}\|_2 \qquad (4.20)\\
&= (1-\alpha^2)^{1/2},
\end{aligned}$$

where in (4.19) we used the notation $\alpha := \bar{\mathbf{u}}^T\mathbf{u}$. In obtaining (4.20) we used the fact that $\|\mathbf{x}\mathbf{y}^T\|_2 = \|\mathbf{x}\|_2\|\mathbf{y}\|_2$, for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and in the last line we used the fact that $\|\mathbf{u}\|_2 = \|\bar{\mathbf{u}}\|_2 = 1$. Therefore by Theorem 4.3

$$\begin{aligned}
(1-\alpha^2)^{1/2} &\leq \frac{\sqrt{2}c\Delta K\delta_p}{K\delta_p(1-2\Delta)} \\
&= c\frac{\Delta}{1-2\Delta} \qquad (4.21)
\end{aligned}$$

[1] A unitarily invariant matrix norm is such that $|||\mathbf{UAV}||| = |||\mathbf{A}|||$, for any matrix $\mathbf{A}$, where $\mathbf{U}, \mathbf{V}$ are two unitary matrices

Thus we obtain

$$\begin{aligned}\|\mathbf{u}-\bar{\mathbf{u}}\|_2 &= \sqrt{2}(1-\alpha)^{1/2} \\ &< \sqrt{2}(1-\alpha^2)^{1/2} \qquad (4.22)\\ &\leq c\frac{\Delta}{1-2\Delta}, \qquad (4.23)\end{aligned}$$

where in (4.22) we used the fact that $\mathbf{u}$ is only fixed up to a scale factor of $\pm 1$, and so $\alpha$ can be chosen to be non-negative, and in (4.23) we used (4.21). □

We finally need the following lemma and the subsequent observations.

**Lemma 4.5.** *$\exists$ a constant $C$ s.t. $\|\mathbf{y}-\frac{1}{\lambda^{1/2}}\boldsymbol{\mathcal{A}}\mathbf{u}\|\leq C\sqrt{K\delta_p}\Delta^2 = C\frac{np}{(K\delta_p)^{3/2}}$, a.s.*

*Proof*: Observe that can write $\boldsymbol{\mathcal{A}}=\sum_{i\geq 2}\lambda_i\mathbf{u}_i\mathbf{u}_i^T+\lambda\mathbf{u}\mathbf{u}^T=\widetilde{\boldsymbol{\mathcal{A}}}+\lambda\mathbf{u}\mathbf{u}^T$, where $\|\widetilde{\boldsymbol{\mathcal{A}}}\|_2=\max_{i>2}|\lambda_i|\leq K\delta_p\Delta$, a.s. Hence we have

$$\begin{aligned}\|\mathbf{y}-\frac{1}{\lambda^{1/2}}\boldsymbol{\mathcal{A}}\mathbf{u}\| &= \frac{1}{\lambda^{1/2}}\|\boldsymbol{\mathcal{A}}(\mathbf{u}-\bar{\mathbf{u}})\|_2 \\ &= \frac{1}{\lambda^{1/2}}\|\left(\widetilde{\boldsymbol{\mathcal{A}}}+\lambda\mathbf{u}\mathbf{u}^T\right)(\mathbf{u}-\bar{\mathbf{u}})\|_2 \\ &\leq \frac{\|\widetilde{\boldsymbol{\mathcal{A}}}(\mathbf{u}-\bar{\mathbf{u}})\|_2}{\lambda^{1/2}}+\lambda^{1/2}\|\mathbf{u}-\bar{\mathbf{u}}\|^2 \\ &\leq c\frac{\sqrt{K\delta_p}\Delta^2}{(1-\Delta)^{1/2}(1-2\Delta)}+\frac{C\sqrt{K\delta_p}\Delta^2(1+\Delta)^{1/2}}{(1-2\Delta)^2} \qquad (4.24)\\ &\leq C\sqrt{K\delta_p}\Delta^2,\end{aligned}$$

a.s., where in (4.24), we used the bound in Lemma 4.4.

Notice that the eigenvector components $\mathbf{u}_i$, are exchangeable for $1\leq i\leq K$, and similarly for $\mathbf{u}_i, 1+K\leq i\leq n$. (This is clear since we have $\boldsymbol{\mathcal{A}}\mathbf{u}=\lambda\mathbf{u}$, and the distribution of $\mathcal{A}_{ij}$ being the same for $1\leq i\leq K$, and for $i>K$.)

**Lemma 4.6.** *For $1\leq i\leq K$, we have $\sqrt{\frac{K\delta_p}{p(1-p)}}\left|y_i-\frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}\right|\to 0$, and for $K+1\leq i\leq n$, we have $\sqrt{\frac{K\delta_p}{q(1-q)}}\left|y_i-\frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}\right|\to 0$, in probability.*

*Proof:*
For $1\leq i\leq K$, using Markov inequality,

$$\begin{aligned}\mathbb{P}\{\sqrt{\frac{K\delta_p}{p(1-p)}}\left|y_i-\frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}\right|>\varepsilon\} &\leq C\frac{\mathbb{E}K\frac{\delta_p}{p}|y_i-\frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}|^2}{\varepsilon^2} \\ &= C\frac{\sum_{i=1}^K\mathbb{E}|y_i-\frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}|^2}{\varepsilon^2} \qquad (4.25)\\ &\leq C\frac{\mathbb{E}\|\mathbf{y}-\frac{1}{\lambda^{1/2}}\boldsymbol{\mathcal{A}}\mathbf{u}\|^2}{\varepsilon^2} \\ &\leq C\left(\frac{np}{(K\delta_p)^{3/2}}\right)^2\to 0, \qquad (4.26)\end{aligned}$$

where (4.25) follows from $\frac{\delta_p}{p}=\frac{p-q}{p}\leq C$, for some $C, N$, $n>N$, and exchangeability, and

the last step follows from Lemma 4.5. Similarly for $1+K \le i \le n$,

$$\begin{aligned}
\mathbb{P}\{\sqrt{\frac{K\delta_p}{q(1-q)}}|y_i - \frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}| > \varepsilon\} &\le \frac{\mathbb{E}K|y_i - \frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}|^2}{\varepsilon^2}\frac{\delta_p}{q(1-q)} \\
&\le C\frac{\mathbb{E}(n-K)|y_i - \frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}|^2}{\varepsilon^2} \qquad (4.27) \\
&= C\frac{\sum_{i=1+K}^{n}\mathbb{E}|y_i - \frac{(\boldsymbol{\mathcal{A}}\mathbf{u})_i}{\lambda^{1/2}}|^2}{\varepsilon^2} \qquad (4.28) \\
&\le C\frac{\mathbb{E}\|\mathbf{y} - \frac{1}{\lambda^{1/2}}\boldsymbol{\mathcal{A}}\mathbf{u}\|^2}{\varepsilon^2} \\
&\le C\left(\frac{np}{(K\delta_p)^{3/2}}\right)^2 \to 0,
\end{aligned}$$

where in (4.27), we use Condition 4.2, and in 4.28, we used exchangeability of $u_i, K+1 \le i \le n$.

Finally, in the following Lemma we show that $T_3 \to 0$ in probability.

**Lemma 4.7.** *Under Condition 4.3,* $\sqrt{\frac{K\delta_p}{p(1-p)}}\left(\frac{1}{\overline{\lambda}^{1/2}} - \frac{1}{\lambda^{1/2}}\right)[\overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}]_i \to 0.$

*Proof.* Since $[\overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}]_i = 0$ for $i > K$, we only need to consider $1 \le i \le K$. We have $[\overline{\boldsymbol{\mathcal{A}}}\overline{\mathbf{u}}]_i = \sqrt{\delta_p}$. Thus the result amounts to

$$\sqrt{K\delta_p}\left(\frac{1}{\lambda^{1/2}} - \frac{1}{\overline{\lambda}^{1/2}}\right) \to 0$$

in probability. We have

$$\left|\frac{1}{\lambda^{1/2}} - \frac{1}{\overline{\lambda}^{1/2}}\right| = \frac{|\overline{\lambda}^{1/2} - \lambda^{1/2}|}{\lambda^{1/2}\overline{\lambda}^{1/2}}. \qquad (4.29)$$

Since $\overline{\lambda} = K\delta_p$, to prove the result we need to show that

$$\frac{|\overline{\lambda}^{1/2} - \lambda^{1/2}|}{\lambda^{1/2}} \to 0.$$

We have

$$|\overline{\lambda}^{1/2} - \lambda^{1/2}| = \frac{|\lambda - \overline{\lambda}|}{\overline{\lambda}^{1/2} + \lambda^{1/2}}.$$

By Lemma 4.2, we have $|\lambda - \overline{\lambda}| \le c\sqrt{np}$ asymptotically almost surely (a.a.s) and we have $\overline{\lambda} > cm\delta_p$ with high probability (whp). Therefore

$$|\overline{\lambda}^{1/2} - \lambda^{1/2}| \le c\frac{\sqrt{np}}{\sqrt{K\delta_p}},$$

whp. Therefore

$$\frac{|\overline{\lambda}^{1/2} - \lambda^{1/2}|}{\lambda^{1/2}} \le c\frac{\sqrt{np}}{K\delta_p} \to 0,$$

by Assumption 4.3. □

#### 4.3.2.2 Distribution of $\chi$ under $\mathcal{H}_1$

We use the CLT derived in Thereom 4.1 to derive an approximate CLT for our test statistic $\chi = \|\mathbf{u}\|_1$ under $\mathcal{H}_1$. The distribution is approximate since we make the assumption that the components of $\mathbf{x}$ are independently distributed and have the gaussian distribution derived in thereom 4.1 for finite $n$ as opposed to the asymptotic regime in which Theorem 4.1 holds.

**Proposition 4.3.** *Under the assumption that the components of* $\mathbf{x}$ *are independent and gaussian with the distribution derived in theorem 4.1,* $\frac{\chi-\mu_{(1)}}{\sigma_{(1)}}$ *is asymptotically distributed as* $\mathcal{N}(0,1)$.

*To simplify the presentation of the formulae we introduce the following notation. Let* $r = \frac{K\delta_p^2}{2p(1-p)}, s = \frac{K\delta_p^2}{2q(1-q)}$. *Also,* $\beta_1 = \sqrt{\frac{\delta_p}{\pi r}}e^{-r} + \sqrt{\delta_p}\left(1 - 2Q(\sqrt{2r})\right)$, *and* $\beta_2 = \sqrt{\frac{\delta_p}{\pi s}}$. *In addition we also define*

$$E_1 = \frac{1}{\sqrt{\pi}}\left(\frac{\delta_p}{r}\right)^{3/2} M(-\frac{3}{2}, \frac{1}{2}, -r)$$

$$E_2 = \frac{3}{4}\left(\frac{\delta_p}{r}\right)^{2} M(-2, 1/2, -r)$$

*where* $M(a,b,z)$ *is the confluent hypergeometric gamma function [Abramowitz & Stegun 1964]. Then*

$$\mu_{(1)} = \frac{N_{\alpha_1}}{\sqrt{N_{\alpha_2}}}$$

*and*

$$\sigma^2_{(1)} = \frac{1}{N_{\alpha_2}}\left(C_{11} + \left(\frac{N_{\alpha_1}}{2N_{\alpha_2}}\right)^2 C_{22} - \frac{N_{\alpha_1}}{N_{\alpha_2}}C_{12}\right),$$

*where* $N_{\alpha_1} = K\beta_1 + (n-K)\beta_2$, *and* $N_{\alpha_2} = K\left(\delta_p(1+\frac{1}{2r})\right) + (1-\frac{2}{\pi})\frac{\delta_p(n-K)}{2s}$. *Finally,*

$$C_{11} = K\left(\delta_p(1+\frac{1}{2r}) - \beta_1^2\right) + (1-\frac{2}{\pi})\frac{\delta_p(n-K)}{2s}$$

$$C_{12} = K\left(E_1 - \beta_1\delta_p(1+\frac{1}{2r})\right) + \frac{n-K}{\sqrt{4\pi}}\left(\frac{\delta_p}{s}\right)^{3/2}$$

$$C_{22} = K\left(E_2 - \delta_p^2(1+\frac{1}{2r})^2\right) + \frac{3(n-K)}{4}\left(\frac{\delta_p}{s}\right)^2$$

The CLT result stated in Proposition 4.3 is approximate, since in deriving the result we assumed that the components of the scaled dominant eigenvector are gaussian for finite $n$, whereas in truth the distribution is only gaussian in the asymptotic limit. On the other hand, from simulations we see that the distribution indeed matches our prediction. We provide approximate expressions of $\mu_{(1)}$ and $\sigma^2_{(1)}$ derived above, using the fact that $r = \omega(1)$, and $s = \omega(1)$. For the parameter values we choose under the Conditions 4.1 and 4.3, and using asymptotic approximations for the Q-function and $M(a,b,x)$, [Abramowitz & Stegun 1964] we can show that for large $n$,

$$\mu_{(1)} \approx \sqrt{K}\left(1 - \frac{1}{4r} - \frac{\rho}{4s}\right)\left(1 + \frac{\rho}{\sqrt{\pi s}}\right),$$

where $\rho := \frac{n-K}{K}$. For large $n$, the fractions in the braces are o(1) implying that the expected value of $\chi$ is close to $\sqrt{K} \ll \mu_{(0)}$. This agrees with our intuition that asymptotically the

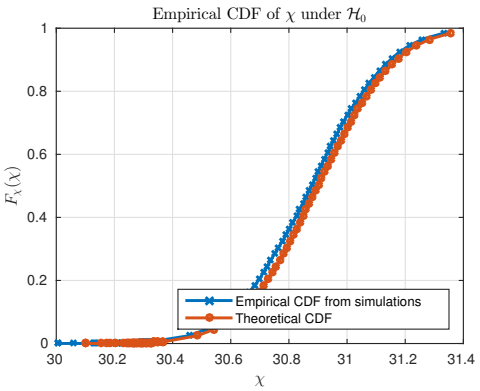



Figure 4.1: CDF of $\chi$ under $\mathcal{H}_0$

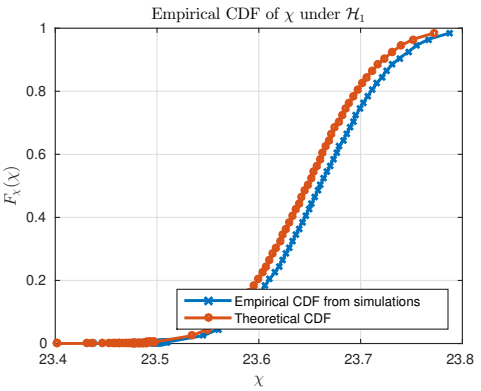



Figure 4.2: CDF of $\chi$ under $\mathcal{H}_1$.

eigenvector $\mathbf{u}$ is localized to the nodes belonging to the subgraph. Similarly using the asymptotic approximation for $M(a, b, x)$ for large $x$ [Abramowitz & Stegun 1964], one can show that for large $n$, and $K, \delta_p$ satisfying Condition 4.3,

$$\sigma^2_{(1)} \approx \frac{1}{2}(1 - \frac{2}{\pi})\frac{\rho}{s}(1 - \frac{1}{2r} - \frac{\rho}{2s})$$

Thus we see that $\sigma^2_{(1)} \sim \frac{\rho}{s} = \frac{2(n-K)q(1-q)}{(K\delta_p)^2} \sim \frac{(n-K)q}{(K\delta_p)^2}$. This is interesting because it says that the variance of $\chi$ under $\mathcal{H}_1$ is inversely proportional to the strength of the signal $K\delta_p$ and in addition it is inversely proportional to $\Delta$, the spectral gap ratio, indicating that smaller the spectral gap, the harder it is to detect the presence of the subgraph. In addition $\sigma^2_{(1)}$ is several orders of magnitude less than $\mu_{(1)}$ and so the concentration is quite sharp.

## 4.4 Numerical Results

We present simulations to validate the distributions of the statistic under $\mathcal{H}_0$ and $\mathcal{H}_1$. We choose values of $K, n, \delta_p$ and $q$ so that the Conditions 4.1, 4.2, and 4.3 are satisfied. First we generate an ER graph of size $n = 1500$ and edge probability $q = 0.15$, and calculate the dominant eigenvector of the shifted adjacency matrix. We compute its $L^1$-norm and repeat the experiment $10^4$ times and compute the empirical CDF $F_\chi(\chi)$, which is the solid blue line with "x" marker in figure 4.1. In the same figure we plot the CDF of a gaussian rv with mean $\mu_{(0)}$ and variance $\sigma^2_{(0)}$ (red solid line with "o" marker). This verifies that $\chi$ indeed has a distribution close to a gaussian with the predicted mean and variance. Next we embed a subgraph in this ER graph with $K = 450$ and $\delta_p = 0.25$, and compute the $L^1$-norm of the dominant eigenvector and repeat the experiment $10^4$ times to obtain the empirical CDF. The results are plotted in figure 4.2. We indeed can observe that the empirical CDF (blue solid line with "x" marker), matches quite well with the gaussian CDF (red solid line with "o" marker whose mean and variance are $\mu_{(1)}$ and $\sigma^2_{(1)}$ respectively, thus corroborating our theoretical findings. Notice that because the distributions are far apart in the parameter regime under consideration, we obtain practically error free detection.

In addition, we also compare the probability of subgraph detection of our algorithm with the edge thresholding algorithm in [Hajek *et al.* 2015b]. We consider $n = 10^3, 5 \times 10^3$ and $K = \lceil c\sqrt{n} \rceil$ with $p = 0.2, q = 0.1$. We observe from Tables 4.1 and 4.2 that the two algorithms have similar error performance for the parameter values considered. However our algorithm requires fewer parameters than the algorithm in [Hajek *et al.* 2015b].

| $n = 10^3, K = \lceil c\sqrt{n} \rceil$ | $L^1$-norm algorithm | Edge thresholding algo |
|---|---|---|
| c = 5 | 1 | 0.9960 |
| c= 4 | 1 | 0.9820 |
| c= 3.5 | 0.9940 | 0.98 |

Table 4.1: Probability of Subgraph Detection for $n = 10^3$

| $n = 5 \times 10^3, K = \lceil c\sqrt{n} \rceil$ | $L^1$-norm algorithm | Edge thresholding algo |
|---|---|---|
| c = 5 | 1 | 1 |
| c= 4.5 | 1 | 1 |
| c= 4 | 1 | 1 |

Table 4.2: Probability of Subgraph Detection for $n = 5 \times 10^3$

## 4.5 Conclusions and Future Work

In this work we studied an algorithm for detecting the presence of a denser subgraph in an ER background graph based on thresholding $L^1$-norm of the leading eigenvector of a shifted adjacency matrix. This algorithm was also considered in a general form in [Miller *et al.* 2010], however in our work we define the threshold in terms of the graph parameters. Our detection algorithm only requires the knowledge of $n$, the graph size and $q$, the edge probability of the background graph, and does not require the knowledge of $K$ and $p$ unlike the algorithms analyzed in [Hajek *et al.* 2015b]. We compare our algorithm with the latter numerically and conclude that they are similar in performance. In addition to the above detection algorithm we also develop a subgraph recovery algorithm for a graph containing a hidden subgraph and we show that it approximately recovers the subgraph under certain assumptions on the subgraph parameters. The regime of recovery for this algorithm is however suboptimal with respect to available works such as in [Chen & Xu 2016] and we would like to investigate this algoithm further to make it competitive with the literature or to modify its analysis to improve this detectability threshold.

Chapter 5

# Hidden Community Recovery with Side-information

## 5.1 Introduction

### 5.1.1 Problem Motivation

We consider the problem of hidden community recovery in graphs in the presence of side-information. In various disciplines graphs have been used to model, in a parsimonious fashion, relationships between heterogenous data. The presence of a dense hidden community in such graphs is usually indicative of interesting phenomena in the associated real-world network.

An example application of dense subgraph recovery in Signal Processing is the problem of Correlation Mining [Firouzi *et al.* 2013]. Given a network of correlated signals, a graph is formed with nodes representing signals, and weighted links representing pairwise correlations. The problem of detecting a group of closely correlated signals is then a dense subgraph recovery problem on the constructed graph [Firouzi *et al.* 2013]. Dense subgraph recovery also finds application in real-world computer and social networks; for e.g., in detecting fraudulent activity [Chau *et al.* 2006, Beutel *et al.* 2013, Smith *et al.* 2014]. It can, in addition, be viewed as a signal recovery problem on graphs [Chen *et al.* 2015, Wang *et al.* 2015].

A majority of subgraph recovery algorithms try to find a subset of nodes that maximizes some objective such as the average link density within the subset [Lee *et al.* 2010]. A good way to benchmark the performance of various community recovery algorithms is to validate them on generative graph models with inherent community structure. In this work, we model the hidden community as a small but well-connected Erdős-Rényi graph embedded within a larger but sparser Erdős-Renyi graph. This model was used in [Mifflin *et al.* 2004] to capture terrorist transactions in a computer network. It is a special case of the Stochastic Block Model (SBM), which has been widely used to assess the performance of different community recovery algorithms [Rohe *et al.* 2011].

The study of subgraph recovery on generative models is interesting in itself from an algorithmic perspective. Recent works on hidden community recovery and related problems demonstrate the presence of sharp phase transitions in the range of parameter values between three regimes: easy (recovery achievable with relatively small computational costs), hard (computationally taxing, but detectable), and impossible to detect [Hajek *et al.* 2015a, Montanari 2015, Caltagirone *et al.* 2016]. We provide more details on these phenomena while reviewing prior works in the next subsection. The novel aspect of our work is a theoretical study of the impact of side-information on this computational barrier. The form of side-information we consider is the identity of special nodes called cues that are known to belong to the subgraph, either deterministically or with some level of certainty. One often has access to such prior knowledge in real-world applications [Avrachenkov *et al.* 2012, Zhou *et al.* 2004, Zhu *et al.* 2003].

By developing and analyzing the asymptotic performance of a local algorithm based on Belief Propagation (BP), we show that even a small amount of side-information can lead to the disappearance of the computational barrier. BP is an efficient way to perform approximate ML recovery on certain types of graphs using distributed and local message passing [Mezard & Montanari 2009]. It belongs to the class of guilt-by-association schemes [Koutra *et al.* 2011] and has been successfully applied to many practical problems in graphs such as fraud recovery [Chau *et al.* 2006] and data mining [Kang *et al.* 2011].

### 5.1.2 Review of Existing Works

Consider a graph with $n$ nodes that contains a hidden community of size $K$. The edge probability between any two nodes within the community is $p$ and it is $q$ otherwise, such that $p > q$. The parameters $p, q$ and $K$ can in general be functions of $n$. This model, denoted by $G(K, n, p, q)$, was already considered in [Mifflin *et al.* 2004, Miller *et al.* 2010, Kadavankandy *et al.* 2016, Hajek *et al.* 2015b] and references therein in the context of anomaly detection.

A special case of the above model is the hidden clique model with $p = 1$ and $q = 1/2$. The study of clique detection algorithms demonstrate the presence of phase transitions in the subgraph size $K$ between impossible, hard and easy regimes. If $K \leq 2(1-\varepsilon)\log_2(n)$, the clique is impossible to detect; however, an exhaustive search detects the clique nodes when $K \geq 2(1+\varepsilon)\log_2(n)$. In contrast, the smallest clique size that can be detected in polynomial time is believed to be $c\sqrt{n}$ [Alon *et al.* 1998] for some $c > 0$, and the minimum clique-size that can be detected in nearly-linear time is believed to be $\sqrt{n/e}$ [Deshpande & Montanari 2015].

The computational barriers for subgraph recovery in a sparse graph without cues were studied in [Montanari 2015, Hajek *et al.* 2015a, Hajek *et al.* 2016b]. In [Montanari 2015] the author investigated the performance of Maximum Likelihood (ML) detection and BP, and analyzed the phase transition with respect to an effective signal-to-noise ratio (SNR) parameter $\lambda$ defined as

$$\lambda = \frac{K^2(p-q)^2}{(n-K)q}. \tag{5.1}$$

The larger the $\lambda$, the easier it is to detect the subgraph. Subgraph recovery was considered under a parameter setting where $K = \kappa n, p = a/n$ and $q = b/n$, where $\kappa, a$ and $b$ are constants independent of $n$. It was shown under this setting that, for any $\lambda > 0$, an exhaustive search can detect the subgraph with success probability approaching one as $\kappa \to 0$. However BP, which has quasi-linear time complexity, achieves non-trivial success probability only when $\lambda > 1/e$ in the same regime. Further, for $\lambda < 1/e$, the success probability of the algorithm is bounded away from one. This demonstrates the existence of a computational barrier for local algorithms.

In [Hajek *et al.* 2016b] the authors show that when $K = o(n)$, i.e., when $\kappa \to 0$, and $p, q$ are such that $a = np = n^{o(1)}$ and $p/q = O(1)$, ML detection succeeds when $\lambda = \Omega(\frac{K}{n}\log(\frac{n}{K}))$, i.e., detection is possible even when the SNR parameter goes to zero so long as it does not go to zero too fast. Under the same parameter setting, it was shown that BP succeeds in detecting the subgraph with the fraction of misdetected nodes going to zero, only when $\lambda > 1/e$ [Hajek *et al.* 2015a]. Therefore, $\lambda = 1/e$ represents a computational barrier for BP in the subgraph detection problem without side-information.

In the present work, we examine the impact of side-information on the above computational barrier. To the best of our knowlege, ours is the first theoretical study of the performance of local algorithms for subgraph detection in the presence of side-information in $G(K, n, p, q)$. In [Miller *et al.* 2015b], the authors compared, but only empirically, several guilt-by-association schemes for subgraph detection with cues.

There exist many works on the effect of side-information in the context of identifying multiple communities [Allahverdyan *et al.* 2010, Caltagirone *et al.* 2016, Cai *et al.* 2016, Mossel & Xu 2016]. These works considered a different variant of the SBM where nodes are partitioned into two or more communities, with dense links inside communities and sparse links across communities. The authors of [Cai *et al.* 2016] and [Mossel & Xu 2016] consider a BP algorithm to detect two equal-sized communities. In [Mossel & Xu 2016], the side-information is such that all nodes indicate their community information after passing it through a binary symmetric channel with error rate $\alpha$. They show that when $\alpha < 1/2$, i.e., when there is non-trivial side-information, there is no computational barrier and BP works all the way down to the detectability threshold called the Kesten-Stigum threshold [Abbe & Sandon 2015b]. In [Cai *et al.* 2016], a vanishing fraction $n^{-o(1)}$ of nodes reveal their true communities. Again, there is no computational barrier and BP works all the way down to the detectability threshold. A fuller picture is available in [Caltagirone *et al.* 2016], which considers asymmetric communities and asymmetric connection probabilities within communities. In this setting, the authors of [Caltagirone *et al.* 2016] demonstrate the presence of all three regimes (easy to detect, hard to detect but possible via exhaustive search, and impossible to detect) as a function of the size of the smallest community. In contrast, [Mossel & Xu 2016] and [Cai *et al.* 2016] consider equal-sized communities with the same edge probability within each community. In [Caltagirone *et al.* 2016, Cai *et al.* 2016, Mossel & Xu 2016], the parameters are chosen such that node degrees alone are not informative. Our work is different from the above settings, in that we deal with a single community, and the degrees can be informative in revealing node identities, i.e., the average degree of a node within the subgraph $Kp + (n-K)q$ is greater than $nq$, the average degree of a node outside the subgraph. In this setting we show that the computational barrier disappears when side-information is available. We emphasize that our results cannot be obtained as a special case of the results in [Allahverdyan *et al.* 2010, Caltagirone *et al.* 2016, Cai *et al.* 2016, Mossel & Xu 2016].

### 5.1.3 Summary of Results

We consider subgraph detection in $G(K, n, p, q)$ with two types of side-information:

1. A fraction $\alpha$ of subgraph nodes are revealed to the detector, which we call reliable cues. This represents the case of perfect side-information.

2. A similar number of nodes are marked as cues, but they are unreliable, i.e., imperfect side-information.

These two types of side-information are typical in semi-supervised clustering applications [Avrachenkov *et al.* 2012, Zhou *et al.* 2004, Zhu *et al.* 2003].

We use BP for subgraph detection to handle these two kinds of side-information. Our computations are local and distributed and require only neighbourhood information for each node in addition to the graph parameters $p, q$ and $K$.

We analyze the detection performance of our algorithm when $p = a/n, q = b/n$ with $a, b$ fixed and $K = \kappa n$ with $\kappa$ fixed, as in the regime of [Montanari 2015]. Under this setting, we derive recursive equations for the distributions of BP messages in the limit as the graph size $n$ tends to infinity. These recursions allow for numerical computation of the error rates for finite values of $a, b$ and $\kappa$.

Based on these recursions, we obtain closed form expressions for the distributions when $a, b \to \infty$. We then show that when there is non-trivial side-information, the expected fraction of misclassified nodes goes to zero as $\kappa \to 0$, for any positive value of the respective

SNR parameter $\lambda_\alpha$ or $\lambda$, for perfect or imperfect side-information, made explicit later. Thus the computational barrier of $\lambda = 1/e$ for BP without side-information disappears when there is side-information.

We validate our theoretical findings by simulations. To demonstrate the practical usefulness of our algorithm we also apply it to subgraph detection on real-world datasets.

## 5.2 Model and Problem Definition

Let $G(K, n, p, q)$ be a random undirected graph with $n$ nodes and a hidden community $S$ such that $|S| = K$. Let $\mathcal{G} = (V, E)$ be a realization of $G(K, n, p, q)$. An edge between two nodes appears independently of other edges such that $\mathbb{P}((i,j) \in E | i, j \in S) = p$ and $\mathbb{P}((i,j) \in E | i \in S, j \notin S) = \mathbb{P}((i,j) \in E | i, j \notin S) = q$. We assume that $S$ is chosen uniformly from $V$ among all sets of size $K$. Additionally let $p = a/n$ and $q = b/n$, where $a$ and $b$ are constants independent of $n$. Such graphs, with average degree $O(1)$, are called diluted graphs. We use a function $\sigma : V \to \{0,1\}^n$ to denote community membership such that $\sigma_i = 1$ if $i \in S$ and 0 otherwise. Next we describe the model for selecting $C$, the set of cues. To indicate which nodes are cues, we introduce a function $c : V \to \{0,1\}^n$ s.t. $c_i = 1$ if $i$ is a cued vertex and $c_i = 0$ otherwise. The model for cues depends on the type of side-information: perfect or imperfect.

The side-information models are as follows:

1. **Perfect side-information**: In this case the cues are reliable, i.e., they all belong to the subgraph. To construct $C$ we sample nodes as follows

$$\mathbb{P}(c_i = 1 | \sigma_i = x) = \begin{cases} \alpha & \text{if } x = 1 \\ 0 & \text{if } x = 0, \end{cases}$$

for some $\alpha \in (0,1)$. Under this model we have

$$\begin{aligned} n\mathbb{P}(c_i = 1) &= \sum_{i \in V} \mathbb{P}(c_i = 1 | \sigma_i = 1)\mathbb{P}(\sigma_i = 1) \\ &= \alpha K. \end{aligned} \tag{5.2}$$

2. **Imperfect side-information**: Under imperfect side-information, the cues are unreliable. We generate $C$ by sampling nodes from $V$ as follows using a fixed $\beta \in (0,1]$. For any $i \in V$:

$$\mathbb{P}\left(c_i = 1 | \sigma_i = x\right) = \begin{cases} \alpha\beta & \text{if } x = 1, \\ \frac{\alpha K(1-\beta)}{(n-K)} & \text{if } x = 0. \end{cases} \tag{5.3}$$

Under this model we have for any $i \in V$,

$$\begin{aligned} \mathbb{P}(c_i = 1) &= \mathbb{P}(\sigma_i = 1)\mathbb{P}(c_i = 1 | \sigma_i = 1) \\ &\quad + \mathbb{P}(\sigma_i = 0)\mathbb{P}(c_i = 1 | \sigma_i = 0) \\ &= \frac{K}{n}\alpha\beta + \frac{(n-K)}{n}\frac{\alpha K(1-\beta)}{(n-K)} \\ &= \alpha K/n; \end{aligned}$$

hence it matches with (5.2) of the perfect side-information case. It is easy to verify that under the above sampling

$$\mathbb{P}\left(\sigma_i = 1 | c_i = 1\right) = \beta, \tag{5.4}$$

which provides us with the interpretation of $|\log(\beta/(1-\beta))|$ as a reliability parameter for cue information.

Given $G, C$ our objective is to infer the labels $\{\sigma_i, i \in V \backslash C\}$. The optimal detector that minimizes the expected number of misclassified nodes is the per-node MAP detector given as [Hajek *et al.* 2016b]:

$$\hat{\sigma}_i = \chi\left(R_i > \log \frac{\mathbb{P}(\sigma_i = 0)}{\mathbb{P}(\sigma_i = 1)}\right),$$

where

$$R_i = \log\left(\frac{\mathbb{P}(G, C|\sigma_i = 1)}{\mathbb{P}(G, C|\sigma_i = 0)}\right)$$

is a log-likelihood ratio of the detection problem. Observe that this detector requires the observation of the whole graph. Our objective then is to compute $R_i$ for each $i$ using a local Belief Propagation (BP) algorithm and identify some parameter ranges for which it is useful. Specifically, we want to show that a certain barrier that exists for BP when $\alpha = 0$ disappears when $\alpha\beta > 0$.

## 5.3 Subgraph Detection with Perfect Side-information

In this section we present the BP algorithm, Algorithm 2, which performs detection in the presence of perfect side-information. We provide here a brief overview of the algorithm. At step $t$ of Algorithm 2, each node $u \in V \backslash C$ updates its own log-likelihood ratio based on its $t$-hop neighbourhood:

$$R_u^t := \log\left(\frac{\mathbb{P}(G_u^t, C_u^t|\sigma_u = 1)}{\mathbb{P}(G_u^t, C_u^t|\sigma_u = 0)}\right), \tag{5.5}$$

where $G_u^t$ is the set of $t$-hop neighbours of $u$ and $C_u^t$ is the set of cues in $G_u^t$, i.e., $C_u^t = G_u^t \cap C$. The beliefs are updated according to (5.8). The messages transmitted to $u$ by the nodes $i \in \delta u$, the immediate neighbourhood of $u$, are given by

$$R_{i\to u}^t := \log\left(\frac{\mathbb{P}(G_i^t \backslash u, C_i^t \backslash u|\sigma_i = 1)}{\mathbb{P}(G_i^t \backslash u, C_i^t \backslash u|\sigma_i = 0)}\right), \tag{5.6}$$

where $G_i^t \backslash u$ and $C_i^t \backslash u$ are defined as above, but excluding the contribution from node $u$. Node $i$ updates $R_{i\to u}^t$ by acquiring messages from its neighbours, except $u$, and aggregating them according to (5.7). If node $u$ is isolated, i.e., $\delta u = \emptyset$, there are no updates for this node. It can be checked that the total computation time for $t_f$ steps of BP is $O(t_f|E|)$.

The detailed derivation of the algorithm can be found in Appendix A.1. The derivation consists of two steps. First we establish a coupling between $G_u^t$, the $t$-hop neighbourhood of a node $u$ of the graph and a specially constructed Galton-Watson (G-W) tree[1] $T_u^t$ of depth $t$ rooted on $u$. This coupling ensures that for a carefully chosen $t = t_f$ the neighbourhood $G_u^{t_f}$ of the node is a tree with probability tending to one as $n \to \infty$ (i.e., with high probability (w.h.p)). The second step of the derivation involves deriving the recursions (5.7) and (5.8) to compute (5.6) and (5.5) respectively, using the tree coupling.

The output of the algorithm is $C$ along with the set of $K - |C|$ nodes with the largest value of log-likelihoods $R_i^{t_f}$. In the following section we derive the asymptotic distributions of the BP messages as the graph size tends to infinity, so as to quantify the error performance of the algorithm.

[1] Detailed in Appendix A.1

**Algorithm 2** BP with perfect side-information

1: Initialize: Set $R^0_{i\to j}$ to 0, for all $(i,j)\in E$ with $i,j\notin C$. Let $t_f < \frac{\log(n)}{\log(np)}+1$. Set $t=0$.
2: For all directed pairs $(i,u)\in E$, such that $i,u\notin C$:

$$R^{t+1}_{i\to u} = -K(p-q) + \sum_{l\in C^1_i, l\neq u} \log\left(\frac{p}{q}\right) + \sum_{l\in \delta i\setminus C^1_i, l\neq u} \log\left(\frac{\exp(R^t_{l\to i}-\upsilon)(p/q)+1}{\exp(R^t_{l\to i}-\upsilon)+1}\right), \tag{5.7}$$

where $\upsilon = \log(\frac{n-K}{K(1-\alpha)})$.
3: Increment $t$, if $t < t_f - 1$ go back to 3, else go to 3
4: Compute $R^{t_f}_u$ for every $u\in V\setminus C$ as follows:

$$R^{t+1}_u = -K(p-q) + \sum_{l\in C^1_u} \log\left(\frac{p}{q}\right) + \sum_{l\in \delta u\setminus C^1_u} \log\left(\frac{\exp(R^t_{l\to u}-\upsilon)(p/q)+1}{\exp(R^t_{l\to u}-\upsilon)+1}\right) \tag{5.8}$$

5: The output set is the union of $C$ and the $K-|C|$ set of nodes in $V\setminus C$ with the largest values of $R^{t_f}_u$.

## 5.4 Asymptotic Error Analysis

In this section we analyze the distributions of BP messages $R^t_{i\to u}$ given $\{\sigma_i=1\}$ and given $\{\sigma_i=0\}$ for $i\in V\setminus C$. First, we derive a pair of recursive equations for the asymptotic distributions of the messages $R^t_{i\to u}$ given $\{\sigma_i=0, c_i=0\}$ and given $\{\sigma_i=1, c_i=0\}$ in the limit as $n\to\infty$ in Lemma 5.1. In Proposition 5.1 we present the asymptotic distributions of the messages in the large degree regime where $a,b\to\infty$. This result will enable us to derive the error rates for detecting the subgraph in the large degree regime (Theorem 5.1). Finally, we contrast this result with Proposition 5.2 from [Montanari 2015], which details the limitation of local algorithms.

Instead of studying $R^t_{i\to u}$ directly, we look at the log-likelihood ratios of the posterior probabilities of $\sigma_i$ given as

$$\widetilde{R}^t_i = \log\left(\frac{\mathbb{P}(\sigma_i=1|G^t_i, C^t_i, c_i=0)}{\mathbb{P}(\sigma_i=0|G^t_i, C^t_i, c_i=0)}\right)$$

and the associated messages $\widetilde{R}^t_{i\to u}$. By Bayes rule, $\widetilde{R}^t_{i\to u} = R^t_{i\to u} - \upsilon$, where

$$\upsilon = \log\left(\frac{\mathbb{P}(\sigma_i=0|c_i=0)}{\mathbb{P}(\sigma_i=1|c_i=0)}\right) = \log\left(\frac{n-K}{K(1-\alpha)}\right).$$

Let $\xi^t_0, \xi^t_1$ be rvs with the same distribution as the messages $\widetilde{R}^t_{i\to u}$ given $\{\sigma_i=0, c_i=0\}$ and given $\{\sigma_i=1, c_i=0\}$, respectively in the limit as $n\to\infty$. Based on the tree coupling in Lemma A.1 of Appendix A.1, it can be shown that these rvs satisfy the recursive distributional evolutionary equations given in the following lemma.

**Lemma 5.1.** *The random variables $\xi^t_0$ and $\xi^t_1$ satisfy the following recursive distributional equations with initial conditions $\xi^0_0 = \xi^0_1 = \log\left(\kappa(1-\alpha)/(1-\kappa)\right)$.*

$$\xi^{(t+1)}_0 \overset{D}{=} h + \sum_{i=1}^{L_{0c}} \log(\rho) + \sum_{i=1}^{L_{00}} f(\xi^{(t)}_{0,i}) + \sum_{i=1}^{L_{01}} f(\xi^{(t)}_{1,i}) \tag{5.9}$$

$$\xi^{(t+1)}_1 \overset{D}{=} h + \sum_{i=1}^{L_{1c}} \log(\rho) + \sum_{i=1}^{L_{10}} f(\xi^{(t)}_{0,i}) + \sum_{i=1}^{L_{11}} f(\xi^{(t)}_{1,i}), \tag{5.10}$$

*where* $\stackrel{D}{=}$ *denotes equality in distribution,* $h = -\kappa(a-b) - \upsilon$, $\rho := p/q = a/b$, *and the function* $f$ *is defined as*

$$f(x) := \log\left(\frac{\exp(x)\rho + 1}{\exp(x) + 1}\right). \tag{5.11}$$

*The rvs* $\xi_{0,i}^t, i = 1, 2, \ldots$ *are independent and identically distributed (i.i.d.) with the same distribution as* $\xi_0^t$. *Similarly* $\xi_{1,i}^t, i = 1, 2, \ldots$ *are i.i.d. with the same distribution as* $\xi_1^t$. *Furthermore,* $L_{00} \sim Poi((1-\kappa)b), L_{01} \sim Poi(\kappa b(1-\alpha)), L_{10} \sim Poi((1-\kappa)b), L_{11} \sim Poi(\kappa a(1-\alpha)), L_{0c} \sim Poi(\kappa b\alpha)$ *and* $L_{1c} \sim Poi(\kappa p\alpha)$.

*Proof.* This follows from (5.7) and the tree coupling in Lemma A.1 of Appendix A.1. □

We define the effective SNR for the detection problem in the presence of perfect side-information as:

$$\lambda_\alpha = \frac{K^2(p-q)^2(1-\alpha)^2}{(n-K)q} = \frac{\kappa^2(a-b)^2(1-\alpha)^2}{(1-\kappa)b}, \tag{5.12}$$

where the factor $(1-\alpha)^2$ arises from the fact that we are now trying to detect a smaller subgraph of size $K(1-\alpha)$.

We now present one of our main results, on the distribution of BP messages in the limit of large degrees as $a, b \to \infty$ such that $\lambda_\alpha$ is kept fixed.

**Proposition 5.1.** *In the regime where* $\lambda_\alpha$ *and* $\kappa$ *are held fixed and* $a, b \to \infty$, *we have*

$$\xi_0^{t+1} \xrightarrow{D} \mathcal{N}\left(-\log\frac{1-\kappa}{\kappa(1-\alpha)} - \frac{1}{2}\mu^{(t+1)}, \mu^{(t+1)}\right)$$

$$\xi_1^{t+1} \xrightarrow{D} \mathcal{N}\left(-\log\frac{1-\kappa}{\kappa(1-\alpha)} + \frac{1}{2}\mu^{(t+1)}, \mu^{(t+1)}\right).$$

*The variance* $\mu^{(t)}$ *satisfies the following recursion with initial condition* $\mu^{(0)} = 0$ :

$$\mu^{(t+1)} = \lambda_\alpha\alpha\frac{1-\kappa}{(1-\alpha)^2\kappa} + \lambda_\alpha\mathbb{E}\left(\frac{(1-\kappa)}{\kappa(1-\alpha) + (1-\kappa)\exp(-\mu^{(t)}/2 - \sqrt{\mu^{(t)}}Z)}\right), \tag{5.13}$$

*where the expectation is taken w.r.t.* $Z \sim \mathcal{N}(0,1)$.

Before providing a short sketch of the proof of the above proposition, we state a Lemma from [Hajek *et al.* 2015a], which we need for our derivations.

**Lemma 5.2.** *[Hajek et al. 2015a, Lemma 11] Let* $S_\gamma = X_1 + X_2 + \ldots + X_{N_\gamma}$, *where* $X_i$, *for* $i = 1, 2, \ldots N_\gamma$, *are independent, identically distributed rv with mean* $\mu$, *variance* $\sigma^2$ *and* $\mathbb{E}(|X_i^3|) \le g^3$, *and for some* $\gamma > 0$, $N_\gamma$ *is a* $Poi(\gamma)$ *rv independent of* $X_i : i = 1, 2, \ldots, N_\gamma$. *Then*

$$\sup_x \left|\mathbb{P}\left(\frac{S_\gamma - \gamma\mu}{\sqrt{\gamma(\mu^2+\sigma^2)}}\right) - \Phi(x)\right| \le \frac{C_{BE}g^3}{\sqrt{\gamma(\mu^2+\sigma^2)^3}},$$

*where* $C_{BE} = 0.3041$.

We now provide a sketch of the proof of Proposition 5.1; the details can be found in Appendix A.2.

*Sketch of Proof of Proposition 5.1.* The proof proceeds primarily by applying the expectation and variance operators to both sides of (5.9) and (5.10) and applying various reductions. First notice that when $a, b \to \infty$ and $\lambda$ and $\kappa$ are held constant, we have $\rho \to 1$ as follows:

$$\rho = a/b = 1 + \sqrt{\frac{\lambda_\alpha(1-\kappa)}{(1-\alpha)^2\kappa^2 b}}. \tag{5.14}$$

Then using Taylor's expansion of $\log(1+x)$ we can expand the function $f(x)$ in (5.11) up to second order as follows:

$$f(x) = (\rho-1)\frac{e^x}{1+e^x} - \frac{1}{2}(\rho-1)^2(\frac{e^x}{1+e^x})^2 + O(b^{-3/2}). \tag{5.15}$$

We use these expansions to simplify the expressions for the means and variances of (5.9) and (5.10). Then, by a change of measure, we express them in terms of functionals of a single rv, $\xi_1^t$. We then use induction to show that the variance $\mu^{(t+1)}$ satisfies the recursion (5.13) and use Lemma 5.2 to prove gaussianity. □

In the following subsection, we use Proposition 5.1 to derive the asymptotic error rates of the detector in Algorithm 1.

### 5.4.1 Detection Performance

Let us use the symbol $\overline{S}$ to denote the subgraph nodes with the cued nodes removed, i.e., $\overline{S} = S \backslash C$. This is the set that we aim to detect. The output of Algorithm 2, $\widehat{S}$ is the set of nodes with the top $K - |C|$ beliefs. We are interested in bounding the expected number of misclassified nodes $\mathbb{E}(|\overline{S}\Delta\widehat{S}|)$. Let $\widehat{S}$ be the output set of the algorithm excluding cues since the cues are always correctly detected. Note that $|\overline{S}| = |\widehat{S}| = K - |C|$. To characterize the performance of the detector, we need to choose a performance measure. In [Montanari 2015], a rescaled probability of success was used to study the performance of a subgraph detector without cues, defined as

$$P_{\text{succ}}(\widehat{\sigma}) = \mathbb{P}(i \in \widehat{S} | i \in S) + \mathbb{P}(i \notin \widehat{S} | i \notin S) - 1, \tag{5.16}$$

where $\widehat{\sigma}_i = \chi(i \in \widehat{S})$, and the dependence of $P_{\text{succ}}(\widehat{\sigma})$ on $n$ is implicit. In our work, we study the following error measure, which is the average fraction of misclassified nodes, also considered in [Hajek *et al.* 2015a], which for the uncued case is defined as

$$\mathcal{E} := \frac{\mathbb{E}(|S\Delta\widehat{S}|)}{K}.$$

Observe that $0 \le \mathcal{E} \le 2$. In particular $\mathcal{E} = 2$ if the algorithm misclassifies all the subgraph nodes. We now show that these two measures are roughly equivalent. For simplicity we consider the case where there are no cues, but the extension to the cued case is straightforward. Since our algorithm always outputs $K$ nodes as the subgraph, i.e., $|\widehat{S}| = K$, the following is true for any estimate $\widehat{\sigma}$ of $\sigma$ :

$$r_n := \sum_{i=1}^{n} \chi(\widehat{\sigma}_i = 0, i \in S) = \sum_{i=1}^{n} \chi(\widehat{\sigma}_i = 1, i \notin S), \tag{5.17}$$

i.e., the number of misclassified subgraph nodes is equal to the number of misclassified nodes outside the subgraph. We can rewrite the error measure $\mathcal{E}$ in terms of $r_n$, since

$$\frac{|S\Delta\widehat{S}|}{K} = \frac{2r_n}{K}. \tag{5.18}$$

Next notice that we can rewrite $P_{\text{succ}}(\widehat{\sigma})$ as follows.

$$\begin{aligned} P_{\text{succ}}(\widehat{\sigma}) &= 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{P}(\widehat{\sigma}_i = 0 | i \in S) + \mathbb{P}(\widehat{\sigma}_i = 1 | i \notin S)\right) \\ &\overset{(a)}{=} 1 - \sum_{i=1}^{n}\left(\frac{\mathbb{P}(\widehat{\sigma}_i = 0, i \in S)}{K} + \frac{\mathbb{P}(\widehat{\sigma}_i = 1, i \notin S)}{n-K}\right) \\ &\overset{(b)}{=} 1 - \left(\frac{\mathbb{E}(r_n)}{K} + \frac{\mathbb{E}(r_n)}{n-K}\right) = 1 - \frac{n\mathbb{E}(r_n)}{K(n-K)}, \end{aligned} \tag{5.19}$$

where in step (a) we used Bayes rule with $\mathbb{P}(i \in S) = \frac{K}{n}$. Since $1 \leq \frac{n}{n-K} \leq 2$, we get

$$1 - 2\mathbb{E}(r_n)/K \leq P_{\text{succ}}(\widehat{\sigma}) \leq 1 - \mathbb{E}(r_n)/(K). \tag{5.20}$$

Hence from (5.18) and (5.20), $P_{\text{succ}}(\widehat{\sigma}) \to 1$ if and only if $\frac{\mathbb{E}(|S\Delta\widehat{S}|)}{K} \to 0$.

In the following proposition, we state and prove the main result concerning the asymptotic error performance of Algorithm 2.

**Theorem 5.1.** *For any $\lambda_\alpha > 0, \alpha > 0$,*

$$\lim_{b\to\infty}\lim_{n\to\infty}\frac{\mathbb{E}(|\overline{S}\Delta\widehat{S}|)}{K(1-\alpha)} \leq 2\sqrt{\frac{1-\kappa}{\kappa(1-\alpha)}}e^{-\frac{1}{8}\frac{\alpha\lambda_\alpha(1-\kappa)}{\kappa(1-\alpha)^2}}. \tag{5.21}$$

*Consequently,*

$$\lim_{\kappa\to 0}\lim_{b\to\infty}\lim_{n\to\infty}\frac{\mathbb{E}(|\overline{S}\Delta\widehat{S}|)}{K(1-\alpha)} = 0.$$

*Proof.* Let $\widehat{S}_0$ be the MAP estimator given by

$$\widehat{S}_0 = \left\{i : R_i^t > \log\frac{1-\kappa}{\kappa(1-\alpha)}\right\}.$$

Since $\widehat{S}$ is the set of nodes with the top $K - |C|$ beliefs, we have either $\widehat{S} \subset \widehat{S}_0$ or $\widehat{S}_0 \subset \widehat{S}$. Therefore,

$$\begin{aligned} |\overline{S}\Delta\widehat{S}| &\leq |\overline{S}\Delta\widehat{S}_0| + |\widehat{S}\Delta\widehat{S}_0| \\ &= |\overline{S}\Delta\widehat{S}_0| + |K - |C| - |\widehat{S}_0|| \\ &= |\overline{S}\Delta\widehat{S}_0| + ||\overline{S}| - |\widehat{S}_0|| \\ &\leq 2|\overline{S}\Delta\widehat{S}_0|, \end{aligned} \tag{5.22}$$

where the last step follows because the set difference between two sets is lower bounded by the difference of their sizes. If we can bound $\frac{\mathbb{E}(|\overline{S}\Delta\widehat{S}_0|)}{K(1-\alpha)}$ by one-half the expression in (5.21) the result of the Proposition follows. The proof of this upper bound uses Proposition 5.1 and is given in Appendix A.3. □

Theorem 5.1 states that the detectability threshold does not exist for Belief Propagation with cues.

This is in stark contrast to the performance of BP when there is no side-information. In that case, as stated in the following theorem from [Montanari 2015], the performance of any local algorithm suffers when the SNR parameter $\lambda < 1/e$. In the following LOC denotes the class of all local algorithms, i.e., algorithms that take as input the local neighbourhood of a node.

**Proposition 5.2.** *[Montanari 2015, Theorem 1] If $\lambda < 1/e$, then all local algorithms have success probability uniformly bounded away from one; in particular,*

$$\sup_{T\in \mathrm{LOC}} \lim_{n\to\infty} P_{\mathrm{succ}}(T) \le \frac{e-1}{4},$$

*and therefore*

$$\sup_{T\in \mathrm{LOC}} \lim_{n\to\infty} \mathcal{E}(T) \ge \frac{5-e}{4} > 1/2.$$

## 5.5 Subgraph Detection with Imperfect Side Information

In this section, we develop a BP algorithm under the more realistic assumption of imperfect side information, where the available cue information is not completely reliable. This is true of humanly classfied data available for many semi-supervised learning problems.

Our BP algorithm can easily take into account imperfection in side information. Suppose we know the parameters $\alpha$ and $\beta$ defined in (5.2) and (5.4) respectively, or their estimates thereof. We remark that unlike Algorithm 2, which only has to detect the uncued subgraph nodes, our algorithm needs to explore the whole graph, since we do not know a priori which cues are correct. As before, for a node $u$, we wish to compute the following log-likelihood ratio in a distributed manner:

$$R_u^t = \log\left(\frac{\mathbb{P}(G_u^t, c_u, C_u^t|\sigma_u=1)}{\mathbb{P}(G_u^t, c_u, C_u^t|\sigma_u=0)}\right),$$

where $c_u$ is the indicator variable of whether $u$ is a cued node, and $C_u^t$ is the cued information of the $t$-hop neighbourhood of $u$, excluding $u$. Note that we can expand $R_u^t$ as follows

$$\begin{aligned} R_u^t &= \log\left(\frac{\mathbb{P}(G_u^t, C_u^t|\sigma_u=1, c_u)}{\mathbb{P}(G_u^t, C_u^t|\sigma_u=0, c_u)}\right) + \log\left(\frac{\mathbb{P}(c_u|\sigma_u=1)}{\mathbb{P}(c_u|\sigma_u=0)}\right) \\ &= \log\left(\frac{\mathbb{P}(G_u^t, C_u^t|\sigma_u=1)}{\mathbb{P}(G_u^t, C_u^t|\sigma_u=0)}\right) + \log\left(\frac{\mathbb{P}(c_u|\sigma_u=1)}{\mathbb{P}(c_u|\sigma_u=0)}\right), \end{aligned} \tag{5.23}$$

where in the second step we dropped the conditioning w.r.t. $c_u$ because $(G_u^t, C_u^t)$ is independent of the cue information of node $u$ given $\sigma_u$. Let $h_u = \log\left(\frac{\mathbb{P}(c_u|\sigma_u=1)}{\mathbb{P}(c_u|\sigma_u=0)}\right)$. Then it is easy to see from (5.3) that

$$h_u = \begin{cases} \log\left(\frac{\beta(1-\kappa)}{(1-\beta)\kappa}\right), & \text{if } u \in C, \\ \log\left(\frac{(1-\alpha\beta)(1-\kappa)}{(1-\kappa-\alpha\kappa+\alpha\kappa\beta)}\right), & \text{otherwise.} \end{cases} \tag{5.24}$$

The recursion for the first term in (B.11) can be derived along the same lines as the derivation of Algorithm 2 and is skipped. The final BP recursions are given in Algorithm 3.

In order to analyze the error performance of this algorithm we derive the asymptotic distributions of the messages $R_{u\to i}^t$, for $\{\sigma_u = 0\}$ and $\{\sigma_u = 1\}$. Note that, since we now assume that we do not know the exact classification of any of the subgraph nodes, we need to detect $K$ nodes, and hence the effective SNR parameter is defined as

$$\lambda = \frac{K^2(p-q)^2}{(n-K)q}. \tag{5.27}$$

The following proposition presents the asymptotic distribution of the messages $R_{u\to i}^t$ in the limit of $n \to \infty$ and in the large degree regime where $a, b \to \infty$.

**Algorithm 3** BP with imperfect cues

1: Initialize: Set $R^0_{i\to j}$ to 0, for all $(i,j)\in E$. Let $t_f < \frac{\log(n)}{\log(np)}+1$. Set $t=0$.
2: For all directed pairs $(i,u)\in E$:

$$R^{t+1}_{i\to u} = -K(p-q) + h_i + \sum_{l\in\delta i, l\neq u} \log\left(\frac{\exp(R^t_{l\to i}-\nu)(p/q)+1}{\exp(R^t_{l\to i}-\nu)+1}\right), \quad (5.25)$$

where $\nu = \log(\frac{n-K}{K})$.
3: Increment t; if $t < t_f - 1$ go back to 3, else go to 3
4: Compute $R^{t_f}_u$ for every $u\in V$ as follows:

$$R^{t+1}_u = -K(p-q) + h_u + \sum_{l\in\delta u} \log\left(\frac{\exp(R^t_{l\to u}-\nu)(p/q)+1}{\exp(R^t_{l\to u}-\nu)+1}\right) \quad (5.26)$$

5: Output $\widehat{S}$ as $K$ set of nodes in $V$ with the largest values of $R^{t_f}_u$.

**Proposition 5.3.** *Let $n\to\infty$. In the regime where $\lambda$ and $\kappa$ are held fixed and $a,b\to\infty$, the message $R^t_{u\to i}$ given $\{\sigma_u = j\}$, where $j=\{0,1\}$ converges in distribution to $\Gamma^t_j + h_u$ where $h_u$ is defined in (5.24). The rvs $\Gamma^t_j$ have the following distribution:*

$$\Gamma^t_0 \sim \mathcal{N}(-\mu^{(t)}/2, \mu^{(t)}), \text{and}$$
$$\Gamma^t_1 \sim \mathcal{N}(\mu^{(t)}/2, \mu^{(t)}),$$

*where $\mu^{(t)}$ satisfies the following recursion with $\mu^{(0)}=0$,*

$$\mu^{(t+1)} = \alpha\beta^2\lambda\mathbb{E}\left(\frac{(1-\kappa)/\kappa}{\beta+(1-\beta)e^{(-\mu^{(t)}/2-\sqrt{\mu^{(t)}}Z)}}\right) + (1-\alpha\beta)^2\lambda \mathbb{E}\left(\frac{(1-\kappa)}{\kappa(1-\alpha\beta)+(1-\kappa-\alpha\kappa+\alpha\kappa\beta)e^{(-\mu^{(t)}/2-\sqrt{\mu^{(t)}}Z)}}\right), \quad (5.28)$$

*and the expectation is with respect to (w.r.t.) $Z\sim\mathcal{N}(0,1)$.*

*Proof.* The proof proceeds by deriving the recursive distributional equations that the message distributions satisfy in the limit $n\to\infty$, and then applying the large degree limit of $a,b\to\infty$ to these recursions.The details are in the supplementary material. □

The above proposition immediately leads to the following result on the asymptotic error rate of Algorithm 3.

**Theorem 5.2.** *For any $\lambda>0,\alpha>0,\beta>0$,*

$$\lim_{b\to\infty}\lim_{n\to\infty}\frac{\mathbb{E}(|\hat{S}\Delta S|)}{K} \leq 2\left(\alpha\sqrt{\beta(1-\beta)}+\sqrt{(1-\alpha\beta)(\frac{1-\kappa}{\kappa}-\alpha(1-\beta))}\right)e^{-\frac{\lambda\alpha\beta^2(1-\kappa)}{8\kappa}}.$$

*Consequently,*

$$\lim_{\kappa\to 0}\lim_{b\to\infty}\lim_{n\to\infty}\frac{\mathbb{E}(|S\Delta\widehat{S}|)}{K} = 0.$$

*Proof.* The proof essentially analyzes the properties of the recursion (5.28) and is similar to the proof of Theorem 5.1. See supplementary material for details. □

## 5.6 Numerical Experiments

In this section we provide numerical results to validate our theoretical findings on the synthetic model as well as on two real-world datasets. We compare the performance of BP to another seed-based community detection algorithm, the personalized PageRank, which is widely used for local community detection [Andersen & Chung 2007].

### 5.6.1 Synthetic dataset

First we show that the limitation of local algorithms described in Proposition 5.2 is overcome by BP when there is non-trivial side-information. Proposition 5.2 says that when $\lambda < 1/e$, $\mathcal{E}(T) > 1/2$ for any local algorithm $T$. We run our Algorithm 2, on a graph generated with $\alpha = 0.1, \kappa = 5 \times 10^{-4}, b = 100$ and $n = 10^6$. For $\lambda = 1/4 < 1/e$, we get an average value of $\mathcal{E} = 0.228 < 1/2$. Thus it is clear that our algorithm overcomes the computational threshold of $\lambda = 1/e$.

Next, we study the performance of Algorithm 3 when there is noisy side-information with $\beta = 0.8$. For $\lambda = 1/3 < 1/e$, we get an average error rate of $0.3916 < 1/2$ clearly beating the threshold of $\lambda = 1/e$. Thus we have demonstrated that both with perfect and imperfect side-information, our algorithm overcomes the $\lambda = 1/e$ barrier of local algorithms.

Next, we verify that increasing $\alpha$ improves the performance of our algorithm as expected. In Figure 5.1, we plot the variation of $\mathcal{E}$ of Algorithm 2 as a function of $\alpha$. Our parameter setting is $\kappa = 0.01, b = 100$, and $\lambda = 1/2$ with $n = 10^4$. In the figure, we also plot the error rate $\mathcal{E}$ obtained by personalized PageRank under the same setting, with damping factor $\alpha_{pr} = 0.9$ [Andersen & Chung 2007]. The figure demonstrates that BP benefits more as the amount of side-information is increased than PageRank does.

Next, we compare the performance of BP algorithm without side-information given in [Montanari 2015] to our algorithm with varying amounts of side-information. We choose the setting where $n = 10^4, b = 140$ and $\kappa = 0.033$ for different values of $\lambda$ by varying $p$. In Figure 5.2 we plot the metric $\mathcal{E}$ against $\lambda$ for different values of $\beta$, with $\alpha = 0.1$. For $\beta = 1$ we use Algorithm 2. We can see that even BP with noisy side-information performs better than standard BP with no side-information. In addition, as expected increasing $\beta$ improves the error performance.

### 5.6.2 Real-world datasets

We consider two real-world networks: The USPS dataset and the Reuters-911 dataset. For these two datasets we compare the performance of BP with personalized PageRank in terms of recall rate $\mathcal{R}$ defined as

$$\mathcal{R} = \frac{|S \cap \hat{S}|}{|S|},$$

where $S$ is the true community and $\hat{S}$ is its estimate. This is a commonly used metric for community detection applications [Yang & Leskovec 2015]. We use $\alpha_{\text{pr}} = 0.9$ as the damping factor of PageRank. We describe the datasets and the results obtained by our algorithms below.

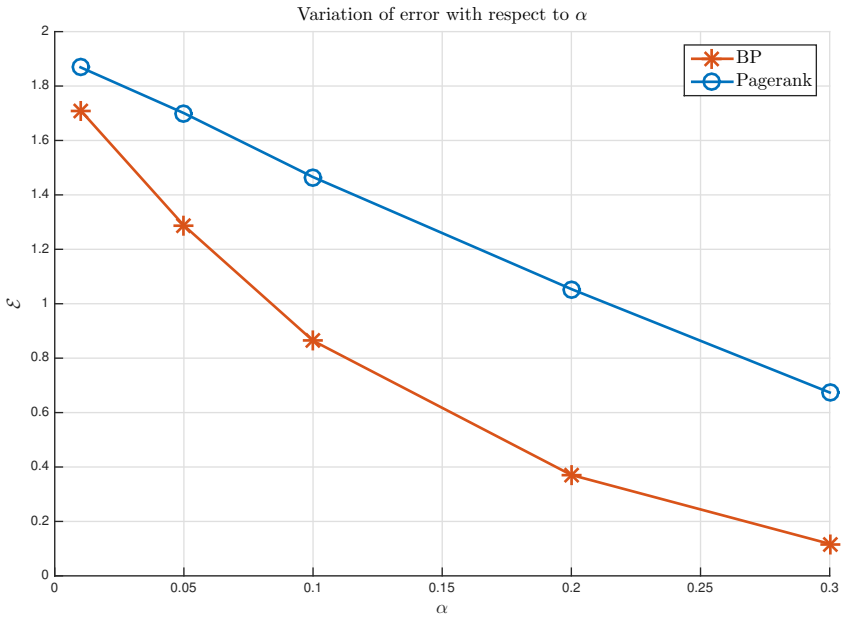



Figure 5.1: Performance of BP Algo 2 as a function of $\alpha$

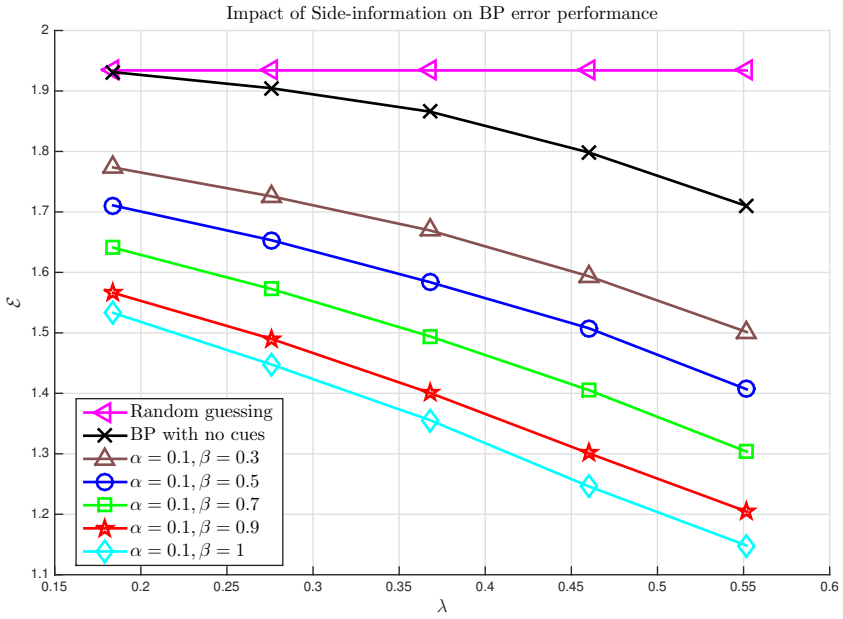



Figure 5.2: Comparison of BP for subgraph detection for different amounts of side-information

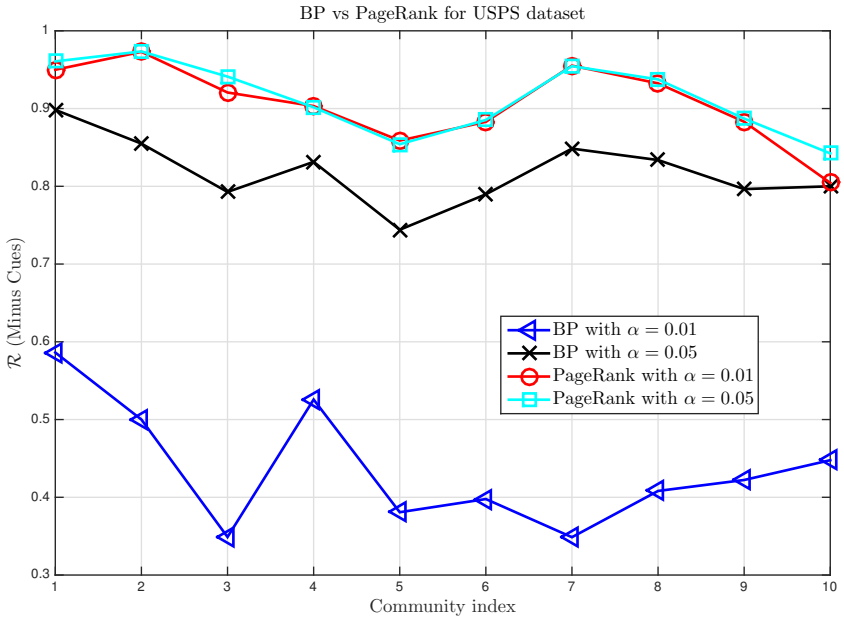



Figure 5.3: Comparison of BP for subgraph detection for different amounts of side-information

#### 5.6.2.1 USPS dataset

The USPS dataset contains 9296 scanned images of size 16 × 16, which can represented by a feature vector of size 256 × 1 with values from -1 to +1 [Zhou *et al.* 2004]. First, we construct a graph from this dataset, where nodes represent scanned images, by adding a link between a node and its three nearest neighbours, where the distance is defined as the euclidean distance between the images represented as feature vectors. The resulting graph is undirected with a minimum degree of at least 3. This is an instance of the $k$ nearest neighbour graph, with $k = 3$. On this graph we run BP and PageRank separately for each of the 10 communities for $\alpha = 0.01$ and $\alpha = 0.05$ (Figure 5.3). It can be seen from Figure 5.3, that the performance of BP is strictly worse than that of PageRank. This result points to the importance of having the correct initialization for the BP parameters. Indeed, in our underlying model for BP, we assumed that there is only one dense community in a sparse network, in which case, as demonstrated in Figure5.1, BP outperforms PageRank by a big margin. However in the USPS graph, there are ten dense communities, and therefore it deviates significantly from our underlying model.

#### 5.6.2.2 Reuters911 Dataset

In this subsection we consider a graph that is closer to our assumed model. We consider the Reuters911 dataset also used in [Chen & Saad 2012]. It is made up of words from all news released by Reuters for 66 days since September 11, 2001. Table 5 in [Chen & Saad 2012] shows a group of 99 collocated words in this dataset. This subset represents the largest dense community to be detected in this dataset. A graph of size $n = 13332$ is generated from this dataset by adding a link between two words if they appear together in a sentence. The resulting graph is undirected and unweighted. We compare BP and Pagerank on this

| Class 0 | #of cues = 1 | #of cues = 2 |
|---|---|---|
| BP | **0.7143** | **0.7216** |
| PageRank | 0.6327 | 0.6392 |

Table 5.1: Reuters911 recall results

dataset for one and two cues. The cues we use are the words *pentagon* and *11*. In Table 5.1 we show the recall values $\mathcal{R}$ of PageRank and BP, excluding cues. Clearly, BP performs better.

### 5.6.3 Comparison with simpler algorithms

We note that under the parameter setting we discussed there are simpler algorithms that can recover the community nodes when there is side-information. We disuss one such algorithm in what follows. As above let $C$ be the set of cues, and let us define $d_i(C)$ as the number of neighbours of any node $i$ in the set $C$, i.e.,

$$d_i(C) = |\{j \in C : j \sim i\}|.$$

Consider an estimator $T_s$ that declares nodes with the $K$ largest values of $d_i(C)$ as the subgraph. We can show that as $n \to \infty$ $d_i(C)$ has the following distribution:

$$d_i(C) \sim \begin{cases} \text{Poi}(\kappa\alpha b(1 + (\rho - 1)\beta) & \text{if } i \in S \\ \text{Poi}(\kappa\alpha b) & \text{otherwise.} \end{cases}$$

Using the above distribution we can show that this estimator achieves zero asymptotic error for any $\lambda, \alpha, \beta > 0$ as $n \to \infty$ and $b \to \infty$. However, in terms of the performance on a finite sized graph, it performs worse than Belief Propagation as shown in Figure 5.4. Here we simulated $G(K, n, p, q)$ with $n = 10^4, K = 200, p = 0.05$, and $q = 0.0046$. We fix $\alpha = 0.1$ and compute the error metric $\sum_{i \in S} \chi_{\{\hat{\sigma}_i = 0\}}/K$, i.e., the fraction of wrongly classified subgraph nodes. In Figure 5.4 we plot this metric against $\beta$. We can observe that Algorithm 2 outperforms the other two algorithms and in addition Belief Propagation far outperforms the simple algorithm described above. In closing we would like to note that the first step of BP in both Algorithm 2 and Algorithm 3 are similar to the simple algorithm discussed above and our results on the error performance of BP, Theorems 5.1, 5.2 apply to the first step of BP as well.

## 5.7 Conclusions and Future Extensions

In this work we developed a local distributed BP algorithm that takes advantage of side-information to detect a dense subgraph embedded in a sparse graph. We obtained theoretical results based on density evolution on trees to show that it achieves zero asymptotic error regardless of the SNR parameter $\lambda$, unlike BP without cues, where there is a non-zero detectability threshold. We then validated our theoretical results by simulating our algorithm on a synthetic dataset and showing that, in the presence of both noise-less and noisy side-information, our BP algorithm overcomes the error bound of local algorithms when $\lambda < 1/e$. We then applied our algorithm to two real-world datasets: USPS and Reuters911 and compared its performance with personalized PageRank. Our results indicate that the relative improvement in BP depends on the closeness of the dataset to the underlying graph model used to derive BP. In the future, we would like to do non-asymptotic analysis when

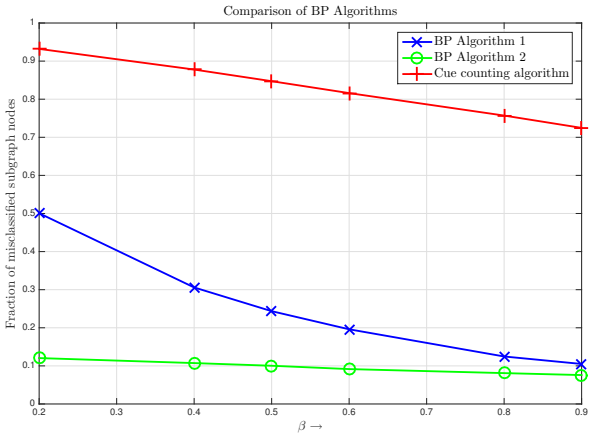



Figure 5.4: Comparison of Algorithm 3 and BP for exact cues Algorithm 2

$a, b$ and $\kappa$ are functions of $n$. Extension to dense graphs would also be interesting, where traditional BP and tree coupling-based analysis will not work owing to the presence of loops.

Chapter 6

# PageRank Analysis on Undirected Random Graphs

## 6.1 Introduction

PageRank has numerous applications in information retrieval [Haveliwala 2002, Page *et al.* 1999, Yeh *et al.* 2009], reputation systems [Gkorou *et al.* 2013, Kamvar *et al.* 2003], machine learning [Avrachenkov *et al.* 2008, Avrachenkov *et al.* 2012], and graph partitioning [Andersen *et al.* 2006, Chung 2009]. It is surprising that not many analytic studies are available for PageRank in random graph models. We mention the work [Avrachenkov & Lebedev 2006] where PageRank was analysed in preferential attachment models and the more recent works [Chen *et al.* 2014, Chen *et al.* 2016], where PageRank was analysed in directed configuration models. According to several studies [Ding *et al.* 2003, Fortunato *et al.* 2006, Litvak *et al.* 2007, Volkovich & Litvak 2010], PageRank and in-degree are strongly correlated in directed networks such as the Web graph.

Apart from some empirical studies [Boudin 2013, Page *et al.* 1999], to the best of our knowledge, there is no rigorous analysis of PageRank on basic undirected random graph models such as the Erdős-Rényi graph [Erdős & Rényi 1959] or the Chung-Lu graph [Chung & Lu 2002a]. In this chapter, we attempt to fill this gap and show that under certain conditions on the preference vector and the spectrum of the graphs, PageRank in these models can be approximated by a mixture of the preference vector and the vertex degree distribution when the size of the graph goes to infinity. First, we show the convergence in total variation norm for a general family of random graphs with expansion property. Then, we specialize the results for the Chung-Lu random graph model proving the element-wise convergence. We also analyse the asymptotics of PageRank on Stochastic Block Model (SBM) graphs, which are random graph models used to benchmark community detection algorithms. In these graphs the asymptotic expression for PageRank contains an additional correction term that depends on the community partitioning. This demonstrates that PageRank captures properties of the graph not visible in the stationary distribution of a simple random walk. We conclude the chapter with numerical experiments and several future research directions.

## 6.2 Definitions

Let $G^{(n)} = (V^{(n)}, E^{(n)})$ denote a family of random graphs, where $V^{(n)}$ is a vertex set, $|V^{(n)}| = n$, and $E^{(n)}$ is an edge set, $|E^{(n)}| = m$. Matrices and vectors related to the graph are denoted by bold letters, while their components are denoted by non-bold letters. We denote by $\mathbf{A}^{(n)}$ the associated adjacency matrix. In the interest of compactness of notation, the superscript $n$ is dropped when it is not likely to cause confusion. In this work, since we analyze PageRank on undirected graphs, we have $\mathbf{A}^T = \mathbf{A}$. The personalized PageRank is denoted by $\boldsymbol{\pi}$. We consider unweighted graphs; however our analysis easily extends to some families of weighted undirected graphs. Let $\mathbf{1}$ be a column vector of $n$ ones and let $\mathbf{d} = \mathbf{A}\mathbf{1}$

be the vector of degrees. It is helpful to define $\mathbf{D} = \text{diag}(\mathbf{d})$, a diagonal matrix with the degree sequence on its diagonal.

Let $\mathbf{P} = \mathbf{A}\mathbf{D}^{-1}$ be column-stochastic Markov transition matrix corresponding to the standard random walk on the graph and let $\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ be the symmetrized transition matrix, whose eigenvalues are the same as those of $\mathbf{P}$. Note that the symmetrized transition matrix is closely related to the normalized Laplacian $\boldsymbol{\mathcal{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{Q}$ [Chung 1997], where $\mathbf{I}$ is the identity matrix. Further we will also use the resolvent matrix $\mathbf{R} = [\mathbf{I} - \alpha\mathbf{P}]^{-1}$ and the symmetrized resolvent matrix $\mathbf{S} = [\mathbf{I} - \alpha\mathbf{Q}]^{-1}$.

Note that since $\mathbf{Q}$ is a symmetric matrix, its eigenvalues $\lambda_i$, $i = 1, ..., n$ are real and can be arranged in decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots$ . In particular, we have $\lambda_1 = 1$. The value $\delta = 1 - \max\{|\lambda_2|, |\lambda_n|\}$ is called the spectral gap.

In what follows, let $K, C$ be arbitrary constants independent of graph size $n$, which may change from one line to the next (of course, not causing any inconsistencies).

For two functions $f(n), g(n)$, $g(n) = O(f(n))$ if $\exists C, N$ such that $\left|\frac{g(n)}{f(n)}\right| \leq C$, $\forall n > N$ and $g(n) = o(f(n))$ if $\limsup_{n\to\infty} \left|\frac{g(n)}{f(n)}\right| = 0$. Also $f(n) = \omega(g(n))$ or $f(n) \gg g(n)$ if $g(n) = o(f(n))$.

We use $\mathbb{P}, \mathbb{E}$ to denote probability and expectation respectively. An event $E$ is said to hold with high probability (w.h.p.) if $\exists N$ such that (s.t.) $\mathbb{P}(E) \geq 1 - O(n^{-c})$ for some $c > 0$, $\forall n > N$. Recall that if a finite number of events hold true w.h.p., then so does their intersection. Furthermore, we say that a sequence of random variables $X_n = o(1)$ w.h.p. if there exists a function $\psi(n) = o(1)$ such that the event $\{X_n \leq \psi(n)\}$ holds w.h.p.

In the first part of this chapter, we study the asymptotics of PageRank for a family of random graphs with the following two properties:

**Property 1.** *For some $K$ w.h.p., $d^{(n)}_{max}/d^{(n)}_{min} \leq K$, where $d^{(n)}_{max}$ and $d^{(n)}_{min}$ are the maximum and minimum degrees, respectively.*

**Property 2.** *W.h.p.,* $\max\{|\lambda^{(n)}_2|, |\lambda^{(n)}_n|\} = o(1)$.

The above two properties can be regarded as a variation of the expansion property. In the standard case of an expander family, one requires the graphs to be regular and the spectral gap $\delta = 1 - \max\{|\lambda_2|, |\lambda_n|\}$ to be bounded away from zero (see, e.g., [Vadhan *et al.* 2012]). Property 1 is a relaxation of the regularity condition, whereas Property 2 is stronger than the requirement for the spectral gap to be bounded away from zero. These two properties allow us to consider several standard families of random graphs such as ER graphs, regular random graphs with increasing average degrees, and Chung-Lu graphs. For Chung-Lu graphs Property 1 imposes some restriction on the degree spread of the graph.

*Remark:* Property 2 implies that the graph is connected w.h.p., since the spectral gap is strictly greater than zero.

Later, we study the asymptotics of PageRank for specific classes of random graphs namely the Chung-Lu graphs, and the Stochastic Block Model. Recall that the Personalized PageRank vector with preference vector $\mathbf{v}$ is defined as the stationary distribution of a modified Markov chain with transition matrix

$$\widetilde{\mathbf{P}} = \alpha\mathbf{P} + (1 - \alpha)\mathbf{v}\mathbf{1}^T, \tag{6.1}$$

where $\alpha$ is the so-called damping factor [Haveliwala 2002]. In other words, $\boldsymbol{\pi}$ satisfies

$$\boldsymbol{\pi} = \widetilde{\mathbf{P}}\boldsymbol{\pi}, \tag{6.2}$$

or,

$$\boldsymbol{\pi} = (1-\alpha)[\mathbf{I} - \alpha\mathbf{P}]^{-1}\mathbf{v} = (1-\alpha)\mathbf{R}\mathbf{v}, \tag{6.3}$$

where (6.3) holds when $\alpha < 1$.

## 6.3 Convergence in total variation on Fast Expander Graphs

We recall that for two discrete probability distributions $u$ and $v$, the total variation distance $d_{\mathrm{TV}}(u,v)$ is defined as $d_{\mathrm{TV}}(u,v) = \frac{1}{2}\sum_i |u_i - v_i|$. This can also be thought of as the $L^1$-norm distance measure in the space of probability vectors, wherein for $\mathbf{x} \in \mathbb{R}^n$, the $L^1$-norm is defined as $\|\mathbf{x}\|_1 = \sum_i |x_i|$. Since for any probability vector $\boldsymbol{\pi}$, $\|\boldsymbol{\pi}\|_1 = 1\ \forall n$, it makes sense to talk about convergence in 1-norm or TV-distance. Also recall that for a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_2 = \sqrt{\sum_i |x_i|^2}$ is the $L^2$-norm. Now we are in a position to formulate our first result.

**Theorem 6.1.** *Let a family of graphs $G^{(n)}$ satisfy Properties 1 and 2. If, in addition, $\|\mathbf{v}\|_2 = O(1/\sqrt{n})$, PageRank can be asymptotically approximated in total variation norm by a mixture of the restart distribution $\mathbf{v}$ and the vertex degree distribution. Namely, w.h.p.,*

$$d_{TV}(\boldsymbol{\pi}^{(n)}, \overline{\boldsymbol{\pi}}^{(n)}) = o(1) \quad \text{as } n \to \infty,$$

*where*

$$\overline{\boldsymbol{\pi}}^{(n)} = \frac{\alpha \mathbf{d}^{(n)}}{\mathrm{vol}(G^{(n)})} + (1-\alpha)\mathbf{v}, \tag{6.4}$$

*with* $\mathrm{vol}(G^{(n)}) = \sum_i d_i^{(n)}$.

*Observations:*

1. This result says that PageRank vector asymptotically behaves like a convex combination of the preference vector and the stationary vector of a standard random walk with transition matrix $\mathbf{P}$; with the weight being $\alpha$, and that it starts to resemble the random walk stationary vector as $\alpha$ gets close to 1.

2. One of the possible intuitive explanations of the result of Theorem 6.1 is based on the observation that when Properties 1 & 2 hold, as $n \to \infty$, the random walk mixes approximately in one step and so for any probability vector $\mathbf{x}$ $\mathbf{P}\mathbf{x}$ is roughly equal to $\mathbf{d}/\mathrm{vol}(G)$, the stationary distribution of the simple random walk. The proposed asymptotic approximation for PageRank can then be seen to follow from the series representation of PageRank if we replace $\mathbf{P}\mathbf{v}$ by $\mathbf{d}/\mathrm{vol}(G)$. Note that since $\mathbf{d}/\mathrm{vol}(G)$ is the stationary vector of the simple random walk, if $\mathbf{P}\mathbf{v} = \mathbf{d}/\mathrm{vol}(G)$, it also holds that $\mathbf{P}^k\mathbf{v} = \mathbf{d}/\mathrm{vol}(G), \forall k \geq 2$. Making these substitutions in the series representation of PageRank, namely

$$\boldsymbol{\pi} = (1-\alpha)\left(\mathbf{I} + \alpha\mathbf{P} + \alpha^2\mathbf{P}^2 + \ldots\right)\mathbf{v}, \tag{6.5}$$

we obtain

$$\begin{aligned}\boldsymbol{\pi} &= (1-\alpha)\mathbf{v} + (1-\alpha)\alpha(1+\alpha+\alpha^2+\ldots)\frac{\mathbf{d}}{\mathrm{vol}(G)} \\ &= (1-\alpha)\mathbf{v} + \alpha\frac{\mathbf{d}}{\mathrm{vol}(G)}.\end{aligned}$$

3. The condition on the 2-norm of the preference vector $\mathbf{v}$ can be viewed as a constraint on its allowed localization.

*Proof of Theorem 6.4.* First observe from (6.1) that when $\alpha = 0$, we have $\widetilde{\mathbf{P}} = \mathbf{v}\mathbf{1}^T$, hence from (6.2) we obtain $\boldsymbol{\pi} = \mathbf{v}$, since $\mathbf{1}^T\boldsymbol{\pi} = 1$. Similarly for the case $\alpha = 1$, $\widetilde{\mathbf{P}} = \mathbf{P}$ and so $\boldsymbol{\pi}$ in this case is just the stationary distribution of the original random walk, which is well-defined and equals $\frac{\mathbf{d}}{\mathrm{vol}(G)}$ since by Property 2 the graph is connected. Examining (6.4) for these two cases we can see that the statement of the theorem holds trivially for both $\alpha = 0$ and $\alpha = 1$. In what follows, we consider the case $0 < \alpha < 1$. We first note that the matrix $\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ can be written as follows by Spectral Decomposition Theorem [Bhatia 2013]:

$$\mathbf{Q} = \mathbf{u}_1\mathbf{u}_1^T + \sum_{i=2}^{n} \lambda_i \mathbf{u}_i\mathbf{u}_i^T, \tag{6.6}$$

where $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_n\}$ with $\mathbf{u}_i \in \mathbf{R}^n$ and $\|\mathbf{u}_i\|_2 = 1$ are the corresponding orthogonal eigenvectors of $\mathbf{Q}$. Recall that $\mathbf{u}_1 = \mathbf{D}^{1/2}\mathbf{1}/\sqrt{\mathbf{1}^T\mathbf{D}\mathbf{1}}$ is the Perron–Frobenius eigenvector. Next, we rewrite (6.3) in terms of the matrix $\mathbf{Q}$ as follows

$$\boldsymbol{\pi} = (1-\alpha)\mathbf{D}^{1/2}[\mathbf{I} - \alpha\mathbf{Q}]^{-1}\mathbf{D}^{-1/2}\mathbf{v}. \tag{6.7}$$

Substituting (6.6) into (6.7), we obtain

$$\begin{aligned}
\boldsymbol{\pi} &= (1-\alpha)\mathbf{D}^{1/2}\left(\frac{1}{1-\alpha}\mathbf{u}_1\mathbf{u}_1^T + \sum_{i=2}^{n}\frac{1}{1-\alpha\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\right)\mathbf{D}^{-1/2}\mathbf{v} \\
&= \mathbf{D}^{1/2}\mathbf{u}_1\mathbf{u}_1^T\mathbf{D}^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}^{1/2}\left(\sum_{i\neq 1}\frac{1}{1-\alpha\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\right)\mathbf{D}^{-1/2}\mathbf{v}.
\end{aligned}$$

Let us denote the error vector by $\boldsymbol{\varepsilon} = \boldsymbol{\pi} - \overline{\boldsymbol{\pi}}$. Note that since $\mathbf{u}_1 = \frac{\mathbf{D}^{1/2}\mathbf{1}}{\sqrt{\mathrm{vol}(G)}}$, we can write $\overline{\boldsymbol{\pi}}$ as

$$\begin{aligned}
\overline{\boldsymbol{\pi}} &= \alpha\frac{\mathbf{d}}{\mathrm{vol}(G)} + (1-\alpha)\mathbf{v} \\
&\overset{\text{(a)}}{=} \alpha\frac{\mathbf{D}\mathbf{1}\mathbf{1}^T\mathbf{v}}{\mathrm{vol}(G)} + (1-\alpha)\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\mathbf{v} \\
&= \alpha\mathbf{D}^{1/2}\frac{\mathbf{D}^{1/2}\mathbf{1}}{\sqrt{\mathrm{vol}(G)}}\frac{\mathbf{1}^T\mathbf{D}^{1/2}}{\sqrt{\mathrm{vol}(G)}}\mathbf{D}^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\mathbf{v} \\
&= \alpha\mathbf{D}^{1/2}\mathbf{u}_1\mathbf{u}_1^T\mathbf{D}^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\mathbf{v},
\end{aligned}$$

where in (a) above we used the fact that $\mathbf{1}^T\mathbf{v} = 1$, since $\mathbf{v}$ is a probability vector. Then, we can write $\boldsymbol{\varepsilon}$ as

$$\begin{aligned}
\boldsymbol{\varepsilon} &= \boldsymbol{\pi} - \alpha\mathbf{D}^{1/2}\mathbf{u}_1\mathbf{u}_1^T\mathbf{D}^{-1/2}\mathbf{v} - (1-\alpha)\mathbf{D}^{1/2}\mathbf{I}\mathbf{D}^{-1/2}\mathbf{v} \\
&= (1-\alpha)\mathbf{D}^{1/2}\left(\sum_{i\neq 1}\frac{\mathbf{u}_i\mathbf{u}_i^T}{1-\alpha\lambda_i} - (\mathbf{I} - \mathbf{u}_1\mathbf{u}_1^T)\right)\mathbf{D}^{-1/2}\mathbf{v} \\
&= (1-\alpha)\mathbf{D}^{1/2}\left(\sum_{i\neq 1}\mathbf{u}_i\mathbf{u}_i^T\frac{\alpha\lambda_i}{1-\alpha\lambda_i}\right)\mathbf{D}^{-1/2}\mathbf{v}.
\end{aligned} \tag{6.8}$$

Now let us bound the $L^1$-norm $\|\boldsymbol{\varepsilon}\|_1$ of the error:

$$\begin{aligned}
\|\boldsymbol{\varepsilon}\|_1/(1-\alpha) &\overset{(a)}{\leq} \sqrt{n}\|\boldsymbol{\varepsilon}\|_2/(1-\alpha) \\
&\overset{(b)}{\leq} \sqrt{n}\|\mathbf{D}^{1/2}\|_2 \left\| \sum_{i\neq 1} \mathbf{u}_i \mathbf{u}_i^T \frac{\alpha\lambda_i}{1-\alpha\lambda_i} \right\|_2 \|\mathbf{D}^{-1/2}\|_2 \|\mathbf{v}\|_2 \\
&\overset{(c)}{\leq} \sqrt{d_{max}/d_{min}}\sqrt{n} \max_{i>1} \left| \frac{\alpha\lambda_i}{1-\alpha\lambda_i} \right| \|\mathbf{v}\|_2 \\
&\leq C\sqrt{d_{max}/d_{min}} \max(|\lambda_2|, |\lambda_n|) \qquad (6.9)
\end{aligned}$$

where in (a) we used the fact that for any vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$ by Cauchy-Schwartz inequality. In (b) we used the submultiplicative property of matrix norms, i.e., $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$. We obtain (c) by noting that the norm of a diagonal matrix is the leading diagonal value and the fact that for a symmetric matrix the 2-norm is the largest eigenvalue in magnitude. The last inequality is obtained by noting that the assumption $\lambda_i = o(1)$ w.h.p. $\forall i > 1$ implies that $\exists N$ s.t. $\forall n > N$, $|1 - \alpha\lambda_i| > C$ for some constant C and the fact that $\|\mathbf{v}\|_2 = O(1/\sqrt{n})$.
Observing that $d_{max}/d_{min}$ is bounded w.h.p. by Property 1 and $\max(|\lambda_2|, |\lambda_n|) = o(1)$ w.h.p. by Property 2 we obtain our result. □

Note that in the case of standard PageRank, $v_i = 1/n, 1 \leq i \leq n$, and hence $\|\mathbf{v}\|_2 = O(1/\sqrt{n})$, but Theorem 6.1 also admits more general preference vectors than the uniform one.

**Corollary 6.1.** *The statement of Theorem 6.1 also holds with respect to the weak convergence, i.e., for any function $f$ on $V$ such that* $\max_{x\in V}|f(x)| \leq 1$,

$$\sup \left\{ \sum_v f(v)\pi_v - \sum_v f(v)\overline{\pi}_v \right\} = o(1) \quad w.h.p.$$

*Proof.* This follows from Theorem 6.1 and the fact that the left-hand side of the above equation is upper bounded by $2\, d_{\mathrm{TV}}(\boldsymbol{\pi}_n, \overline{\boldsymbol{\pi}}_n)$ [Levin *et al.* 2009]. □

## 6.4 Chung-Lu random graphs

In this section, we study the PageRank for the Chung-Lu model [Chung & Lu 2002a] of random graphs. These results naturally hold for w.h.p. graphs also. The spectral properties of Chung-Lu graphs have been studied extensively in a series of papers by Fan Chung et al [Chung *et al.* 2003, Chung & Radcliffe 2011].

### 6.4.1 Chung-Lu Random Graph Model

Let us first provide a definition of the Chung-Lu random graph model.

**Definition 5. Chung-Lu Random Graph Model** *A Chung-Lu graph $\mathcal{G}(w)$ with an expected degree vector $\mathbf{w} = (w_1, w_2, \ldots w_n)$, where $w_i$ are positive real numbers, is generated by drawing an edge between any two vertices $v_i$ and $v_j$ independently of all other pairs, with probability $p_{ij} = \frac{w_i w_j}{\sum_k w_k}$. To ensure that the probabilities $p_{ij}$ are well-defined, we need* $\max_i w_i^2 \leq \sum_k w_k$.

In the following, let $w_{\max} = \max_i w_i$ and $w_{\min} = \min_i w_i$. Below we specify a corollary of Theorem 6.1 as applied to these graphs. But before that we need the following lemmas about Chung-Lu graphs mainly taken from [Chung *et al.* 2003, Chung & Radcliffe 2011].

**Lemma 6.1.** *If the expected degrees $w_1, w_2, \ldots w_n$ satisfy $w_{\min} \gg \log(n)$, then in $\mathcal{G}(w)$ we have, w.h.p., $\max_i |\frac{d_i}{w_i} - 1| = o(1)$.*

In the proof we use Bernstein Concentration Lemma [Billingsley 2008]:

**Lemma 6.2.** *(Bernstein Concentration Lemma [Billingsley 2008]) If $Y_n = X_1 + X_2 + \ldots X_n$, where $X_i$ are independent random variables such that $|X_i| \leq b$ and if $B_n^2 = \mathbb{E}(Y_n - \mathbb{E}(Y_n))^2$ then*

$$\mathbb{P}\{|Y_n - \mathbb{E}(Y_n)| \geq \varepsilon\} \leq 2 \exp \frac{-\varepsilon^2}{2(B_n^2 + b\varepsilon/3)},$$

*for any $\varepsilon > 0$.*

**Proof of Lemma 6.1:** This result is shown in the sense of convergence in probability in the proof of [Chung & Radcliffe 2011, Theorem 2]; using Lemma 6.2 we show the result holds w.h.p. By a straight forward application of Lemma 6.2 to the degrees $d_i$ of the Chung-Lu graph we obtain

$$\mathbb{P}\left(\max_{1 \leq i \leq n} \left|\frac{d_i}{w_i} - 1\right| \geq \beta\right) \leq \frac{2}{n^{c/4-1}}, \quad \text{if} \quad \beta \geq \sqrt{\frac{c \log(n)}{w_{\min}}} = o(1)$$

if $w_{\min} \gg \log(n)$. □We present below a perturbation result for the eigenvalues of Hermitian matrices, called Weyl's inequalities, which we will need for our proofs.

**Lemma 6.3.** *[Horn & Johnson 2012, Theorem 4.3.1] Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ be Hermitian and let the eigenvalues $\lambda_i(\mathbf{A})$, $\lambda_i(\mathbf{B})$ and $\lambda_i(\mathbf{A} + \mathbf{B})$ be arranged in decreasing order. For each $k = 1, 2, \ldots n$ we have*

$$|\lambda_k(\mathbf{A} + \mathbf{B}) - \lambda_k(\mathbf{A})| \leq \|\mathbf{B}\|_2,$$

*where $\|\mathbf{B}\|_2$ is the induced 2-norm or the spectral norm of $\mathbf{B}$.*

The following lemma is an application of Theorem 5 in [Chung *et al.* 2003].

**Lemma 6.4.** *If $w_{\max} \leq K w_{\min}$, for some $K > 0$ and $\overline{w} = \sum_k w_k/n \gg \log^6(n)$, then for $\mathcal{G}(w)$ we have almost surely (a.s.)*

$$\|\mathbf{C}\|_2 = \frac{2}{\sqrt{\overline{w}}}(1 + o(1)),$$

*where $\mathbf{C} = \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2} - \boldsymbol{\chi}^T\boldsymbol{\chi}$, $\mathbf{W} = diag(\mathbf{w})$, and $\boldsymbol{\emptyset}_i = \sqrt{w_i/\sum_k w_k}$ is a row vector.*

**Proof:** It can be verified that when $w_{\max} \leq K w_{\min}$ and $\overline{w} \gg \log^6(n)$, the condition in [Chung *et al.* 2003, Theorem 5], namely, $w_{\min} \gg \sqrt{\overline{w}} \log^3(n)$, is satisfied and hence the result follows. □

**Lemma 6.5.** *For $\mathcal{G}(w)$ with $w_{\max} \leq K w_{\min}$, and $\overline{w} \gg \log^6(n)$,*

$$\max(\lambda_2(\mathbf{P}), -\lambda_n(\mathbf{P})) = o(1) \quad w.h.p.,$$

*where $\mathbf{P}$ is Markov matrix.*

**Proof:** Recall that $\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ is the normalized adjacency matrix. We want to be able to bound the eigenvalues $\lambda_i, i \geq 2$ of $\mathbf{Q}$. We do this in two steps. Using Lemmas 6.1 and 6.3 we first show that if we replace the degree matrix $\mathbf{D}$ in the expression for $\mathbf{Q}$ by the expected degree matrix $\mathbf{W} = \mathbb{E}(\mathbf{D})$, the eigenvalues of the resulting matrix are close to those of $\mathbf{Q}$. Then, using Lemma 6.4 we show that the eigenvalues of $\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}$ roughly coincide with those of $\boldsymbol{\chi}^T\boldsymbol{\chi}$, which is a unit rank matrix and hence only has a single non-zero eigenvalue. Thus we arrive at the result of Lemma 6.5. Now we give the detailed proof.

The first step, $\|\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}\|_2 = o(1)$ w.h.p. follows from Lemma 6.1 and the same argument as in the last part of the proof of Theorem 2 in [Chung & Radcliffe 2011]. We present the steps in the derivation here for the sake of completeness.

Since the 2-norm of a diagonal matrix is the maximum diagonal in absolute value, we have

$$\|\mathbf{W}^{-1/2}\mathbf{D}^{1/2} - \mathbf{I}\|_2 = \max_{\{i=1,2,\ldots\}} \left|\sqrt{\frac{d_i}{w_i}} - 1\right| \leq \max_{\{i=1,2,\ldots\}} \left|\frac{d_i}{w_i} - 1\right| = o(1), \tag{6.10}$$

by Lemma 6.1. Also observe that

$$\|\mathbf{Q}\|_2 = \max_{\{i=1,2,\ldots n\}} |\lambda_i(\mathbf{Q})| = \max_{\{i=1,2,\ldots n\}} |\lambda_i(\mathbf{P})| = 1. \tag{6.11}$$

We now proceed to bound the norm of the difference $\|\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}\|$ as follows

$$\begin{aligned}
&\|\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}\|_2 \\
&= \|\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\|_2 \\
&= \|\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\|_2 \\
&= \|\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q} + \mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\|_2 \\
&\overset{(a)}{=} \|(\mathbf{I} - \mathbf{W}^{-1/2}\mathbf{D}^{1/2})\mathbf{Q}\|_2 + \|\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}(\mathbf{I} - \mathbf{D}^{1/2}\mathbf{W}^{-1/2})\|_2 \\
&\overset{(b)}{\leq} \|(\mathbf{I} - \mathbf{W}^{-1/2}\mathbf{D}^{1/2})\|_2\|\mathbf{Q}\|_2 + \|\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\|_2\|\mathbf{Q}\|_2\|\mathbf{I} - \mathbf{D}^{1/2}\mathbf{W}^{-1/2}\|_2 \\
&\overset{(c)}{=} o(1) + (1 + o(1))o(1) = o(1) \quad w.h.p.,
\end{aligned} \tag{6.12}$$

where (a) follows from triangular inequality of norms, in (b) we used submultiplicativity of matrix norms, and (c) follows from (6.10), (6.11) and the fact that $\|\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\|_2 \leq \|\mathbf{I}\|_2 + \|\mathbf{W}^{-1/2}\mathbf{D}^{1/2} - \mathbf{I}\|_2 = (1 + o(1))$.

By Lemma 6.3 we have for any $i$,

$$|\lambda_i(\mathbf{Q}) - \lambda_i(\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2})| \leq \|\mathbf{Q} - \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}\|_2 = o(1), \tag{6.13}$$

by (6.12). Furthermore, using Lemma 6.3 and the fact that $\lambda_i(\boldsymbol{\chi}^T\boldsymbol{\chi}) = 0$ for $i > 1$, we have for $i \geq 2$,

$$\begin{aligned}
&|\lambda_i(\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2})| \\
&= |\lambda_i(\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}) - \lambda_i(\boldsymbol{\chi}^T\boldsymbol{\chi})| \leq \|\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2} - \boldsymbol{\chi}^T\boldsymbol{\chi}\|_2 \\
&= o(1),
\end{aligned} \tag{6.14}$$

where the last inequality follows from Lemma 6.4.
Now recall that $\max(\lambda_2(\mathbf{P}), -\lambda_n(\mathbf{P})) = \max_{\{i\geq 2\}}|\lambda_i(\mathbf{Q})|$. We have for any $i$,

$$|\lambda_i(\mathbf{Q})| \leq |\lambda_i(\mathbf{Q}) - \lambda_i(\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2})| + ||\lambda_i(\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2})|, \tag{6.15}$$

which implies from (6.13) and (6.14):

$$\max_{\{i\geq 2\}}|\lambda_i(\mathbf{Q})|= o(1).$$

□Armed with these lemmas we now present the following corollary of Theorem 6.1 in the case of Chung-Lu graphs.

**Corollary 6.2.** *Let* $\|\mathbf{v}\|_2 = O(1/\sqrt{n})$, *and* $\alpha \in (0,1)$. *Then PageRank* $\boldsymbol{\pi}$ *of the Chung-Lu graph* $\mathcal{G}(w)$ *can asymptotically be approximated in TV distance by* $\overline{\boldsymbol{\pi}}$, *defined in Theorem 6.1, if* $\overline{w} \gg \log^6(n)$ *and* $w_{\max} \leq K w_{\min}$ *for some* $K$ *that does not depend on* $n$.

**Proof:** Using Lemma 6.1 and the condition that $w_{\max} \leq K w_{\min}$, one can show that $\exists K^{'}$ s.t. $\frac{d_{max}}{d_{min}} \leq K^{'}$ w.h.p. Then the result is a direct consequence of Lemma 6.5 and the inequality from (6.9). □

We further note that this result also holds for ER graphs $\mathcal{G}(n,p_n)$ with $n$ nodes and edge probability $p_n$ such that $np_n \gg \log^6(n)$, where we have $(w_1, w_2, \dots w_n) = (np_n, np_n, \dots np_n)$.

### 6.4.2 Element-wise Convergence of PageRank

In Corollary 6.2 we proved the convergence of PageRank in TV distance for Chung-Lu random graphs. Note that since each component of PageRank could decay to zero as the graph size grows to infinity, this does not necessarily guarantee convergence in an element-wise sense. In this section, we provide a proof for our convergence conjecture to include the element-wise convergence of the PageRank vector. Here we deviate slightly from the spectral decomposition technique and eigenvalue bounds used hitherto, and instead rely on well-known concentration bounds to bound the error in convergence.

Let $\overline{\boldsymbol{\Pi}} = \text{diag}\{\overline{\pi}_1, \overline{\pi}_2, \dots \overline{\pi}_n\}$ be a diagonal matrix whose diagonal elements are made of the components of the approximated PageRank vector and $\widetilde{\boldsymbol{\delta}} = \overline{\boldsymbol{\Pi}}^{-1}(\boldsymbol{\pi} - \overline{\boldsymbol{\pi}})$, i.e., $\widetilde{\delta}_i = (\pi_i - \overline{\pi}_i)/\overline{\pi}_i = \varepsilon_i/\overline{\pi}_i$, where $\boldsymbol{\varepsilon}$ is the unnormalized error defined in Section 6.3. Then using (6.8) we obtain

$$\widetilde{\delta}_i = \left((1-\alpha)v_i + \alpha\frac{d_i}{\text{vol(G)}}\right)^{-1}\left[\mathbf{D}^{1/2}\sum_{j\neq 1}\frac{\alpha\lambda_j}{1-\alpha\lambda_j}\mathbf{u}_j\mathbf{u}_j^T\mathbf{D}^{-1/2}\mathbf{v}\right]_i.$$

Therefore, using $\mathbf{v}'$ to denote $n\mathbf{D}^{-1/2}\mathbf{v}$ we can bound $\left\|\widetilde{\boldsymbol{\delta}}\right\|_\infty = \max_i|\widetilde{\delta}_i|$ as follows

$$\left\|\widetilde{\boldsymbol{\delta}}\right\|_\infty \leq \frac{1}{\min_i\left((1-\alpha)v_i + \alpha\frac{d_i}{\text{vol}(G)}\right)}\left\|\mathbf{D}^{1/2}\sum_{j\neq 1}\frac{\alpha\lambda_j}{1-\alpha\lambda_j}\mathbf{u}_j\mathbf{u}_j^T\mathbf{D}^{-1/2}\mathbf{v}\right\|_\infty \tag{6.16}$$

$$\leq \frac{\sum_i d_i/n}{\alpha d_{\min}}\sqrt{d_{max}}\left\|\sum_{j\neq 1}\frac{\alpha\lambda_j}{1-\alpha\lambda_j}\mathbf{u}_j\mathbf{u}_j^T\mathbf{v}'\right\|_\infty. \tag{6.17}$$

Here $d_{\min}$ denotes $\min_i d_i$. To obtain (6.17) we used the submultiplicativity property of matrix norms, the fact that $\|\mathbf{D}^{1/2}\|_\infty = \sqrt{\max_i d_i} = \sqrt{d_{\max}}$ and the fact that $v_i \geq 0, \forall i \in V$.

Define $\widetilde{\mathbf{Q}} = \mathbf{Q} - \mathbf{u}_1\mathbf{u}_1^T$, the restriction of the matrix $\mathbf{Q}$ to the orthogonal subspace of $\mathbf{u}_1$.

**Lemma 6.6.** *For a Chung-Lu random graph* $\mathcal{G}(w)$ *with expected degrees* $w_1, \dots w_n$, *where* $w_{\max} \leq K w_{\min}$ *and* $w_{\min} \gg \log(n)$, *we have w.h.p.,*

$$\left\|\widetilde{\mathbf{Q}}\mathbf{v}'\right\|_\infty = o(1/\sqrt{w_{\min}}),$$

*when* $v_i = O(1/n)\ \forall i$.

This lemma can be proven by a few applications of Lemma 6.1 and Bernstein's concentration inequality. To keep the train of thought intact, please refer to Appendix B.1 for a detailed proof of this lemma.

In the next lemma we prove an upper bound on the infinity norm of the matrix $\mathbf{S} = (\mathbf{I} - \alpha\mathbf{Q})^{-1}$.

**Lemma 6.7.** *Under the conditions of Lemma 6.6, $\|\mathbf{S}\|_\infty \leq C$ w.h.p., where $\mathbf{C}$ is a number independent of $n$ that depends only on $\alpha$ and $K$.*

**Proof:** Note that $\mathbf{S} = (\mathbf{I} - \alpha\mathbf{Q})^{-1} = \mathbf{D}^{-1/2}(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{D}^{1/2}$. Therefore, $\|\mathbf{S}\|_\infty \leq \sqrt{\frac{d_{max}}{d_{min}}}\left\|(\mathbf{I} - \alpha\mathbf{P})^{-1}\right\|_\infty$ and the result follows since $\left\|(\mathbf{I} - \alpha\mathbf{P})^{-1}\right\|_\infty \leq \frac{1}{1-\alpha}$ [Langville & Meyer 2004] and using Lemma 6.1. □Now we are in a position to present our main result in this section.

**Theorem 6.2.** *Let $v_i = O(1/n)\ \forall i$, and $\alpha < 1$. PageRank $\boldsymbol{\pi}$ converges element-wise to $\overline{\boldsymbol{\pi}} = (1-\alpha)\mathbf{v} + \alpha\mathbf{d}/\mathrm{vol}(G)$, in the sense that $\max_i\ (\pi_i - \overline{\pi}_i)/\overline{\pi}_i = o(1)$ w.h.p., on the Chung-Lu graph $\mathcal{G}(w)$ with expected degrees $\{w_1, w_2, \ldots w_n\}$ such that $w_{\min} > \log^c(n)$ for some $c > 1$ and $w_{\max} \leq K w_{\min}$, for some $K$, a constant independent of $n$.*

**Proof:** Define $\mathbf{Z} = \sum_{i\neq 1} \frac{\alpha\lambda_i}{1-\alpha\lambda_i}\mathbf{u}_i\mathbf{u}_i^T$. We then have:

$$\begin{aligned}
\mathbf{Z} &= \sum_{i=1}^{n} \frac{\alpha\lambda_i}{1-\alpha\lambda_i}\mathbf{u}_i\mathbf{u}_i^T - \frac{\alpha}{1-\alpha}\mathbf{u}_1\mathbf{u}_1^T \\
&= (\mathbf{I} - \alpha\mathbf{Q})^{-1}\alpha\mathbf{Q} - \frac{\alpha}{1-\alpha}\mathbf{u}_1\mathbf{u}_1^T \\
&= \mathbf{S}\left[\alpha\mathbf{Q} - \frac{\alpha}{1-\alpha}(\mathbf{I} - \alpha\mathbf{Q})\mathbf{u}_1\mathbf{u}_1^T\right] \\
&= \alpha\mathbf{S}\widetilde{\mathbf{Q}} \qquad (6.18)
\end{aligned}$$

Now from (6.17) we have

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\delta}}\right\|_\infty &\leq C\frac{\sum_i d_i/n}{d_{\min}}\sqrt{d_{max}}\|\mathbf{S}\widetilde{\mathbf{Q}}\mathbf{v}'\|_\infty \\
&\overset{(a)}{\leq} C\frac{\sum_i d_i/n}{d_{\min}}\sqrt{d_{max}}o(1/\sqrt{w_{\min}}) \\
&\leq C\frac{d_{\max}}{d_{\min}}\sqrt{w_{\max}(1+o(1))}\frac{1}{\sqrt{w_{\min}}}o(1) \\
&= C\frac{w_{\max}}{w_{\min}}\sqrt{\frac{w_{\max}}{w_{\min}}}(1+o(1))o(1) \\
&= C\left(\frac{w_{\max}}{w_{\min}}\right)^{\frac{3}{2}}o(1) \\
&\leq Co(1) \quad \text{w.h.p.},
\end{aligned}$$

where in (a) we used (6.18) and Lemmas 6.6 and 6.7. The rest of the inequalities are obtained by repeatedly using the fact that $d_{\max} = w_{\max}(1+o(1))$ and $d_{min} = w_{\min}(1+o(1))$, from Lemma 6.1. The last step follows from the assumption that $w_{\max} \leq K w_{\min}$ for some constant $K$. □

**Corollary 6.1** (ER Graphs)**.** *For an ER graph $\mathcal{G}(n, p_n)$ such that $np_n \gg \log(n)$, we have that asymptotically the personalized PageRank $\boldsymbol{\pi}$ converges pointwise to $\overline{\boldsymbol{\pi}}$ for $\mathbf{v}$ such that $v_i = O(1/n)$.*

## 6.5 Asymptotic PageRank for the Stochastic Block Model

In this section, we extend the analysis of PageRank to Stochastic Block Models (SBM) with constraints on average degrees. The SBM is a random graph model that reflects the community structure prevalent in many online social networks. It was first introduced in [Holland *et al.* 1983] and has been analyzed subsequently in several works, specifically in the community detection literature, including [Condon & Karp 1999], [Karrer & Newman 2011], [Rohe *et al.* 2011] and several extensions thereof as in [Heimlicher *et al.* 2012] and [Zhao *et al.* 2012], and the references therein.

For the sake of simplicity we focus on an SBM graph with two communities, but the idea of the proof extends easily to generalizations of this simple model.

**Definition 6.** *[Stochastic Block Model (SBM) with two communities]: An SBM graph $\mathcal{G}(m, n-m, p, q)$ with two communities is an undirected graph on a set of disjoint vertices $C_1, C_2$ such that $C_1 \cup C_2 = V$, and let $|C_1| = m$ and $|C_2| = n-m$. Furthermore, if two vertices $i, j \in C_k, k = 1, 2$, then $\mathbb{P}((i,j) \in E) = p$, if $i \in C_1$ and $j \in C_2$, then $\mathbb{P}((i,j) \in E) = q$. The probabilities $p, q$ may scale with $n$ and we assume that $m > n/2$ and $p > q$; this last assumption is necessary for modeling the community structure of a network.*

*Remark:* For the sake of simplicity, we assume that the edge probabilities within both communities are equal to $p$, but this is a minor assumption and can be generalised so that community 1 has a different edge probability to community 2.

For an SBM graph we use $w_{\max}$ and $w_{\min}$ to denote the maximum and the minimum expected degrees of the nodes respectively. From Definition 6, by our assumption on $m, p$ and $q$, we have $w_{\max} = mp + (n-m)q$ and $w_{\min} = (n-m)p + mq$. Note that our results only depend on these two parameters. We present our main result on SBM graphs in the following theorem.

**Theorem 6.3.** *For a Stochastic Block Model with $w_{\min} = \omega(\log^3(n))$ and $\frac{w_{\max}}{w_{\min}} \leq C$, PageRank with preference vector $\mathbf{v}$ such that $\|\mathbf{v}\|_2 = O(\frac{1}{\sqrt{n}})$ satisfies*

$$\|\boldsymbol{\pi} - \overline{\boldsymbol{\pi}}_{\mathrm{SBM}}\|_{TV} = o(1)$$

*w.h.p., where*

$$\overline{\boldsymbol{\pi}}_{\mathrm{SBM}} = (1-\alpha)\left(\mathbf{I} - \alpha\overline{\mathbf{P}}\right)^{-1}\mathbf{v}. \tag{6.19}$$

*Here $\overline{\mathbf{P}}$ represents the "average" Markov matrix given as $\overline{\mathbf{P}} = \overline{\mathbf{A}}\mathbf{W}^{-1}$ where $\mathbf{W} = \mathbb{E}(\mathbf{D})$ and $\overline{\mathbf{A}} = \mathbb{E}(\mathbf{A})$.*

*Discussion:* Let us look at the permissible values of $m, p, q$ under the assumptions in the above theorem. Recall that we have $w_{\min} = (n-m)p + mq > nq$. Therefore the condition on the growth of minimum expected degree is met, for example, if $q = \omega(\log^3(n)/n)$. On the other hand we have

$$\frac{w_{\max}}{w_{\min}} = \frac{mp + (n-m)q}{(n-m)p + mq} = \frac{\frac{m}{n-m}\frac{p}{q} + 1}{\frac{m}{n-m} + \frac{p}{q}} \quad ,$$

which remains bounded if either $m/(n-m)$ or $p/q$ tends to infinity, but not both.

The following corollary of Theorem 6.3 gives an interesting expression for PageRank for an SBM graph with two equal-sized communities.

**Corollary 6.2.** *For an SBM graph as in Definition 6, with $m = n/2$, ($n$ assumed to be even) such that $p+q \gg \log^3(n)/n$ the PageRank vector $\boldsymbol{\pi}$ with preference vector $\mathbf{v}$ such that $\|\mathbf{v}\|_2 = O(\frac{1}{\sqrt{n}})$ satisfies*

$$\|\boldsymbol{\pi} - \overline{\boldsymbol{\pi}}_{\text{SBM}}\|_{TV} \to 0$$

*w.h.p as $n \to \infty$ where*

$$\overline{\boldsymbol{\pi}}_{\text{SBM}} = \alpha\frac{1}{n}\mathbf{1} + (1-\alpha)\left(\mathbf{v} + \frac{\alpha\beta}{1-\alpha\beta}(\mathbf{v}^T\mathbf{u})\mathbf{u}\right), \tag{6.20}$$

*where $\beta := \frac{p-q}{p+q}$, and $\mathbf{u} \in \mathbb{R}^n$ is a unit vector such that $u_i = \frac{1}{\sqrt{n}}$, for $i \in C_1$ and $u_i = -\frac{1}{\sqrt{n}}$ for $i \in C_2$.*

**Proof:** With equal-sized communities, i.e., $m = n/2$, we have $w_{\max} = w_{\min} = \frac{n}{2}(p+q)$. Therefore the conditions of Theorem 6.3 are satisfied if $p+q \gg \log^3(n)/n$. Observe that the expected adjacency matrix can be written as $\overline{\mathbf{A}} = \frac{p+q}{2}\mathbf{1}\mathbf{1}^T + \frac{n}{2}(p-q)\mathbf{u}\mathbf{u}^T$. Furthermore, $\mathbf{W} = \frac{n}{2}(p+q)\mathbf{I}$. Therefore $\overline{\mathbf{P}} = \overline{\mathbf{A}}\mathbf{W}^{-1} = \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{p-q}{p+q}\mathbf{u}\mathbf{u}^T$. From (6.19), the asymptotic PageRank $\overline{\boldsymbol{\pi}}_{\text{sbm}}$ is therefore given as

$$\overline{\boldsymbol{\pi}}_{\text{sbm}} = \alpha\overline{\mathbf{P}}\overline{\boldsymbol{\pi}}_{\text{sbm}} + (1-\alpha)\mathbf{v}.$$

Consequently, $\overline{\boldsymbol{\pi}}_{\text{sbm}} = \frac{\alpha}{n}\mathbf{1} + \alpha\beta\mathbf{u}\mathbf{u}^T\overline{\boldsymbol{\pi}}_{\text{sbm}} + (1-\alpha)\mathbf{v}$, or $\left[\mathbf{I} - \alpha\beta\mathbf{u}\mathbf{u}^T\right]\overline{\boldsymbol{\pi}}_{\text{sbm}} = \frac{\alpha}{n}\mathbf{1} + (1-\alpha)\mathbf{v}$. By Woodbury Matrix Inversion Lemma in [Horn & Johnson 2012], $\left[\mathbf{I} - \alpha\beta\mathbf{u}\mathbf{u}^T\right]^{-1} = \mathbf{I} + \frac{\alpha\beta}{1-\alpha\beta}\mathbf{u}\mathbf{u}^T$. Hence we obtain $\overline{\boldsymbol{\pi}}_{\text{sbm}} = \frac{\alpha}{n}\mathbf{1} + (1-\alpha)\left(\mathbf{v} + \frac{\alpha\beta}{1-\alpha\beta}(\mathbf{u}^T\mathbf{v})\mathbf{u}\right)$, using the fact that $\mathbf{u}$ and $\mathbf{1}$ are orthogonal vectors. □The above corollary asserts that on an SBM matrix the PageRank is well approximated in the asymptotic regime of large graph size by the convex combination of the uniform probability vector $\frac{1}{n}\mathbf{1}$, which is the asymptotic stationary distribution of a simple random walk on the SBM graph, and a linear combination of the preference vector $\mathbf{v}$ and the projection of the preference vector onto the community partitioning vector $\mathbf{u}$. Thus in this simple scenario of SBM graphs with equally sized communities, we observe that PageRank incorporates information about the community structure, in the form of a term correlated with the partition vector $\mathbf{u}$, as opposed to the usual random walk, which misses this information. It can also be inferred from (6.20) that if the correlation between the preference vector $\mathbf{v}$ and $\mathbf{u}$ is large, e.g., when the seed set of PageRank is chosen to be in one of the communities, the resulting PageRank will display a clear delineation of the communities. This provides a mathematical rationale for why PageRank works for semi-supervised graph partitioning [Avrachenkov *et al.* 2012], at least in the asymptotic regime.

To prove Theorem 6.3 we need the following Lemmas, whose proofs are given in Appendix B.2.

**Lemma 6.8.** *For an SBM graph $\mathcal{G}(m, n-m, p, q)$, when $w_{\min} = \omega(\log^3(n))$ it can be shown that for some $C$,*

$$\max_{1\le i\le n}\left|\frac{D_i}{\mathbb{E}(D_i)} - 1\right| \le C\sqrt{\frac{\log(n)}{w_{\min}}} \text{ w.h.p.}$$

The proof of this lemma follows from applying Bernstein's concentration lemma to the degrees of SBM graph. The proof is given in Appendix B.2.1.

For ease of notation, let $\overline{\mathbf{Q}} = \mathbf{W}^{-1/2}\mathbb{E}(\mathbf{A})\mathbf{W}^{-1/2}$, where $\mathbf{W} = \mathbb{E}(\mathbf{D})$. As before $\mathbf{Q} = \mathbf{D}^{1/2}\mathbf{A}\mathbf{D}^{1/2}$. We need the following concentration result on $\mathbf{Q}$.

**Lemma 6.9.** *For an SBM graph for which* $w_{\min} = \omega(\log^3(n))$, *and* $\frac{w_{\max}}{w_{\min}} \leq C$ *for some* $C$, *it can be shown that*

$$\|\mathbf{Q} - \overline{\mathbf{Q}}\|_2 = C\frac{\sqrt{\log(n)w_{\max}}}{w_{\min}} = o(1)$$

*w.h.p.*

We prove this lemma in Appendix B.2.2.

**Proof of Theorem 6.3:** We write the error between $\boldsymbol{\pi}$ and $\overline{\boldsymbol{\pi}}$ as follows

$$\begin{aligned}
\boldsymbol{\delta} &= \boldsymbol{\pi} - \overline{\boldsymbol{\pi}} \\
&= (1-\alpha)\left[\mathbf{D}^{1/2}(\mathbf{I}-\alpha\mathbf{Q})^{-1}\mathbf{D}^{-1/2} - \mathbf{W}^{1/2}(\mathbf{I}-\alpha\overline{\mathbf{Q}})^{-1}\mathbf{W}^{-1/2}\right]\mathbf{v} \\
&= (1-\alpha)\left[\mathbf{W}^{1/2}\left((\mathbf{I}-\alpha\mathbf{Q})^{-1} - (\mathbf{I}-\alpha\overline{\mathbf{Q}})^{-1}\right)\mathbf{W}^{-1/2}\right]\mathbf{v} + \\
&(1-\alpha)\left[\mathbf{D}^{1/2}(\mathbf{I}-\alpha\mathbf{Q})^{-1}\mathbf{D}^{-1/2} - \mathbf{W}^{1/2}(\mathbf{I}-\alpha\mathbf{Q})^{-1}\mathbf{W}^{-1/2}\right]\mathbf{v}, \qquad (6.21)
\end{aligned}$$

where in the last equality we added and subtracted $\mathbf{W}^{1/2}(\mathbf{I}-\alpha\mathbf{Q})^{-1}\mathbf{W}^{-1/2}$ and reordered terms. Now we analyse the two terms in square brackets in the last equality in (6.21), which we denote $T_1$ and $T_2$, respectively. Notice that we have $\|\boldsymbol{\delta}\|_1 \leq \|T_1\|_1 + \|T_2\|_1$. Next we show that as $n \to \infty$, $\|T_1\|_1$ and $\|T_2\|_1$ are $o(1)$ separately and consequently we obtain the result of the theorem.

Let us first consider $T_1$. We have

$$\begin{aligned}
T_1 &= (1-\alpha)\left[\mathbf{W}^{1/2}\left((\mathbf{I}-\alpha\mathbf{Q})^{-1} - (\mathbf{I}-\alpha\overline{\mathbf{Q}})^{-1}\right)\mathbf{W}^{-1/2}\right]\mathbf{v} \\
&= (1-\alpha)\mathbf{W}^{1/2}(\mathbf{I}-\alpha\mathbf{Q})^{-1}\left(\overline{\mathbf{Q}} - \mathbf{Q}\right)(\mathbf{I}-\alpha\overline{\mathbf{Q}})^{-1}\mathbf{W}^{-1/2}\mathbf{v},
\end{aligned}$$

which we obtained by factoring out $(\mathbf{I}-\alpha\mathbf{Q})^{-1}$ and $(\mathbf{I}-\alpha\overline{\mathbf{Q}})^{-1}$ on the left and right sides of the square brackets. Next we focus on the 2-norm of $T_1$.

$$\begin{aligned}
\|T_1\|_2 &\overset{(a)}{\leq} (1-\alpha)\sqrt{w_{\max}}\|(\mathbf{I}-\alpha\mathbf{Q})^{-1}\|_2\|\overline{\mathbf{Q}} - \mathbf{Q}\|_2\|(\mathbf{I}-\alpha\overline{\mathbf{Q}})^{-1}\|_2\frac{1}{\sqrt{w_{\min}}}\|\mathbf{v}\|_2 \\
&\overset{(b)}{\leq} \frac{1}{1-\alpha}\sqrt{\frac{w_{\max}}{w_{\min}}}\|\mathbf{Q} - \overline{\mathbf{Q}}\|_2\|\mathbf{v}\|_2 \\
&\overset{(c)}{\leq} C\frac{\sqrt{\log(n)w_{\max}}}{w_{\min}\sqrt{n}} \\
&= C\sqrt{\frac{\log(n)}{nw_{\max}}}\frac{w_{\max}}{w_{\min}}.
\end{aligned}$$

This proves $\|T_1\|_1 \leq \sqrt{n}\|T_1\|_2$[1] $\leq C\sqrt{\frac{\log(n)}{w_{\max}}}\frac{w_{\max}}{w_{\min}} = o(1)$, from the assumptions of the theorem. Here in (a) we used the submultiplicative property of matrix norms and the fact that 2-norm of diagonal matrices is the maximum diagonal element in magnitude. The inequality (b) follows because $\|(\mathbf{I}-\alpha\mathbf{Q})^{-1}\|_2 \leq \frac{1}{1-\alpha}$ and $\|(\mathbf{I}-\alpha\overline{\mathbf{Q}})^{-1}\|_2 \leq \frac{1}{1-\alpha}$ and step (c) follows from Lemma 6.9 and the assumption that $\|\mathbf{v}\|_2 = O(1/\sqrt{n})$.

[1] By Cauchy Schwartz inequality on norms.

Next we analyse the second term $T_2$. For ease of notation we denote $\widetilde{\mathbf{R}} = \mathbf{W}^{1/2}\left(\mathbf{I}-\alpha\mathbf{Q}\right)^{-1}\mathbf{W}^{-1/2}$. Then by simple algebraic manipulations

$$\begin{aligned}
T_2 &= (1-\alpha)\left[\mathbf{D}^{1/2}\left(\mathbf{I}-\alpha\mathbf{Q}\right)^{-1}\mathbf{D}^{-1/2} - \mathbf{W}^{1/2}\left(\mathbf{I}-\alpha\mathbf{Q}\right)^{-1}\mathbf{W}^{-1/2}\right]\mathbf{v} \\
&= (1-\alpha)\left(\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\widetilde{\mathbf{R}}\mathbf{W}^{1/2}\mathbf{D}^{-1/2} - \widetilde{\mathbf{R}}\right)\mathbf{v} \\
&= (1-\alpha)\left(\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\widetilde{\mathbf{R}}\left(\mathbf{W}^{1/2}\mathbf{D}^{-1/2}-\mathbf{I}\right) + \left(\mathbf{D}^{1/2}\mathbf{W}^{-1/2}-\mathbf{I}\right)\widetilde{\mathbf{R}}\right)\mathbf{v},
\end{aligned}$$

where the last step is obtained by adding and subtracting $\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\widetilde{\mathbf{R}}$. Now we have $\|\mathbf{D}^{1/2}\mathbf{W}^{-1/2}-\mathbf{I}\|_2 = \max_i\left|\sqrt{\frac{d_i}{w_i}}-1\right| \leq \max_i\left|\frac{d_i}{w_i}-1\right| \leq C\sqrt{\frac{\log(n)}{w_{\min}}}$ w.h.p. by Lemma 6.8 and similarly $\|\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\|_2 \leq \|\mathbf{D}^{1/2}\mathbf{W}^{-1/2}-\mathbf{I}\|_2+\|\mathbf{I}\|_2 \leq C\sqrt{\frac{\log(n)}{w_{\min}}}+1$. In addition $\|\mathbf{W}^{1/2}\mathbf{D}^{-1/2}-\mathbf{I}\|_2 = \max_i\left|\sqrt{\frac{w_i}{d_i}}-1\right| \leq \max_i\left|\frac{w_i}{d_i}-1\right|$. It can be shown that since $\max_i\left|\frac{d_i}{w_i}-1\right| \leq C\sqrt{\frac{\log(n)}{w_{\min}}}$ w.h.p. (by Lemma 6.8), then $\max_i\left|\frac{w_i}{d_i}-1\right| \leq C\sqrt{\frac{\log(n)}{w_{\min}}}$ w.h.p.[2] Therefore $\|\mathbf{W}^{1/2}\mathbf{D}^{-1/2}\|_2 \leq \|\mathbf{W}^{1/2}\mathbf{D}^{-1/2}-\mathbf{I}\|_2+\|\mathbf{I}\|_2 \leq C\sqrt{\frac{\log(n)}{w_{\min}}}+1$ w.h.p. Using the above facts and denoting $\delta = C\sqrt{\frac{\log(n)}{w_{\min}}}$ we obtain

$$\begin{aligned}
\|T_2\|_2 &\leq \left(\|\mathbf{D}^{\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}\|_2\|\widetilde{\mathbf{R}}\|_2\|\mathbf{W}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}-\mathbf{I}\|_2+\|\mathbf{D}^{\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}-\mathbf{I}\|_2\|\widetilde{\mathbf{R}}\|_2\right)\|\mathbf{v}\|_2 \\
&\leq C(\delta(\delta+1)\frac{1}{1-\alpha}+\delta)\frac{1}{1-\alpha}\sqrt{\frac{w_{\max}}{nw_{\min}}} \qquad (6.22) \\
&\leq C\delta\sqrt{\frac{w_{\max}}{nw_{\min}}}\text{w.h.p.} \qquad (6.23)
\end{aligned}$$

Hence we have $\|T_2\|_1 \leq \sqrt{n}\|T_2\|_2 \leq C\delta\sqrt{\frac{w_{\max}}{w_{\min}}}$ w.h.p., which from our assumptions is $o(1)$. Here in (6.22) we used the fact that

$$\|\widetilde{\mathbf{R}}\|_2 = \|\mathbf{W}^{1/2}\left(\mathbf{I}-\alpha\mathbf{Q}\right)^{-1}\mathbf{W}^{-1/2}\|_2 \leq \sqrt{\frac{w_{\max}}{w_{\min}}}\|\mathbf{I}-\alpha\mathbf{Q}\|_2 \leq \frac{1}{1-\alpha}\sqrt{\frac{w_{\max}}{w_{\min}}} \leq C,$$

and that $\|\mathbf{v}\|_2 \leq C/\sqrt{n}$, for some $C$. □ *Remark:* This method of proof can be extended to similar models like the Stochastic Block Model with multiple communities and their generalizations, e.g., Random Dot Product Graphs [Athreya *et al.* 2013].

## 6.6 Experimental Results

In this section, we provide experimental evidence to further illustrate the analytic results obtained in the previous sections. In particular, we simulated ER graphs with $p_n = C\frac{\log^7(n)}{n}$ and Chung-Lu graphs with the degree vector $w$ sampled from a geometric distribution so that the average degree $\overline{w} = cn^{1/3}$, clipped such that $w_{\max} = 7w_{\min}$, for various values of graph size, and plotted the maximum of normalized error $\widetilde{\delta}$ and TV distance error $\|\delta\|_1$, respectively, in Figures 6.1 and 6.2. As expected, both these errors decay as functions of $n$, which illustrates that the PageRank vector does converge to the asymptotic value.

[2] This follows since we can write $\frac{d_i}{w_i} = 1+\eta_i$, with $\max_i|\eta_i| = O\left(\sqrt{\frac{\log(n)}{w_{\min}}}\right) = o(1)$ w.h.p., then $\frac{w_i}{d_i} = \frac{1}{1+\eta_i} = 1-\eta_i+O(\eta_i^2)$, hence $\max_i|\frac{w_i}{d_i}-1| = O(\max_i|\eta_i|) = O\left(\sqrt{\frac{\log(n)}{w_{\min}}}\right) = o(1)$ w.h.p.

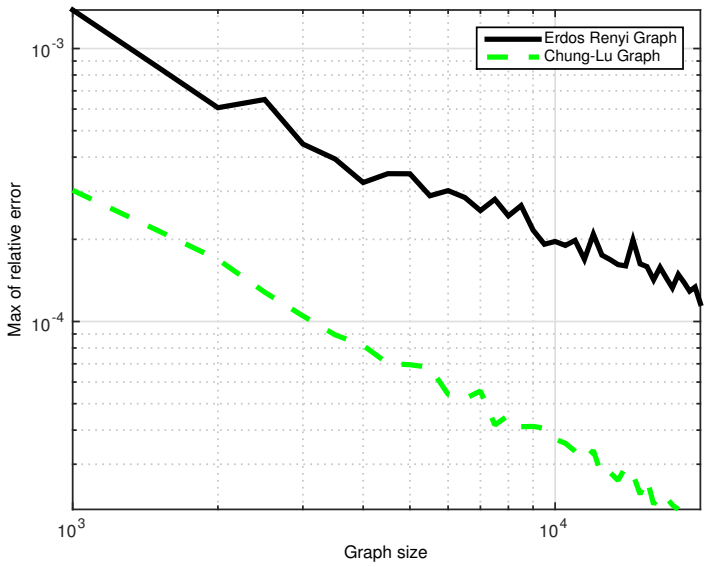



Figure 6.1: Log-log plot of maximum normalized error for ER and Chung-Lu graphs

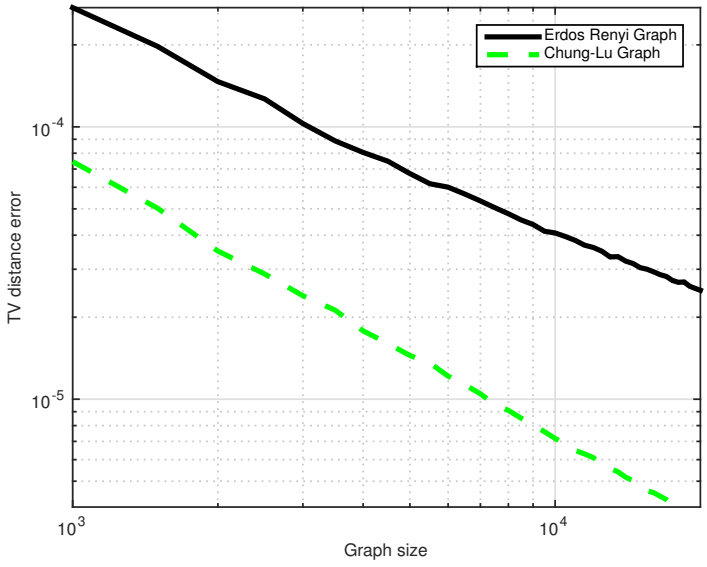



Figure 6.2: Log-log plot of TV distance error for ER and Chung-Lu graphs

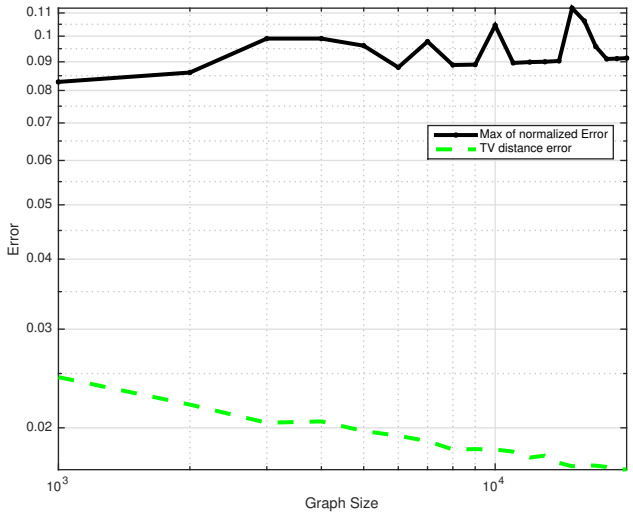



Figure 6.3: Log-log plot of TV distance and maximum error for power-law graphs

In the spirit of further exploration, we have also conducted simulations on power-law graphs with exponent $\beta = 4$ using the Chung-Lu graph model with $w_i = ci^{-1/(\beta-1)}$, for $i_0 \leq i \leq n + i_0$ with

$$c = \frac{\beta - 2}{\beta - 1} dn^{1/(\beta-1)},$$

$$i_0 = n \left[ \frac{d(\beta - 1}{m(\beta - 2)} \right]$$

Please refer to [Chung *et al.* 2003] for details. We set max degree $m = n^{1/3}$ and average degree $d = n^{1/6}$. In Figure 6.3 we observe that for this graph the max-norm of the relative error does not converge to zero. On the other hand the TV-norm seems to converge to zero with graph size, albeit very slowly. Note that these graphs satisfy Property 2 [Chung *et al.* 2003], but they do not satisfy Property 1. Therefore at this point, it is not possible to conclude whether the assumption of bounded variation of degrees is necessary for the convergence to hold. It might be interesting to investigate in detail the asymptotic behavior of PageRank in undirected power-law graphs.

Furthermore, we also see that in the case $\mathbf{v} = \mathbf{e}_i$, the standard unit vector, for some $i$ we do not have the conjectured convergence, as can be seen on Figure 6.4 in the case of ER graphs. It can also be seen from our analysis that if $v_k = 1$ for some $k$, the quantity $\left\| \widetilde{Q} D^{-1/2} v \right\|_\infty$, becomes:

$$\max_i \left| \sum_j \left( \frac{A_{ij}}{\sqrt{d_i d_j}} - \frac{\sqrt{d_i d_j}}{\sum_l d_l} \right) v_j / \sqrt{d_j} \right| = \max_i \frac{1}{\sqrt{d_i} d_k} \left| A_{ik} - \frac{d_i d_k}{\sum_l d_l} \right|,$$

which is $O\left(\frac{1}{\sqrt{w_{\min}} w_k}\right)$ and does not fall sufficiently fast. We simulated an SBM matrix with two communities of equal size, with $p = 0.1$ and $q = 0.01$. In Figure 6.5 we plot the maximum normalized error and the TV-distance error against graph size on a log-log plot. As expected both errors go to zero for large graph sizes.

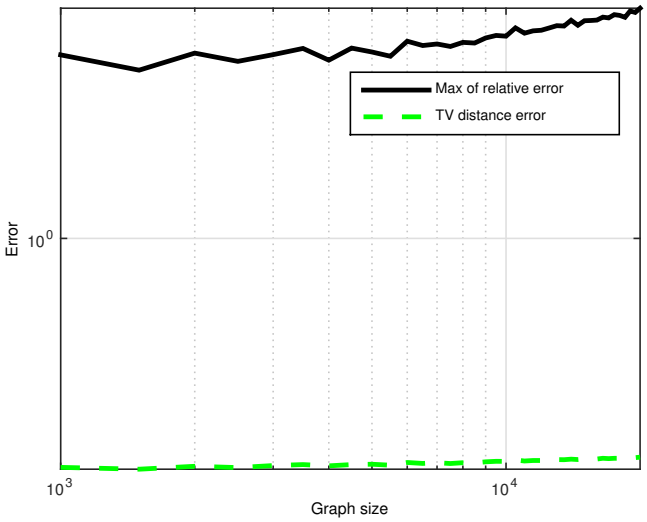



Figure 6.4: Log-log plot of TV distance and maximum relative error for ER-graph when $v = e_1$

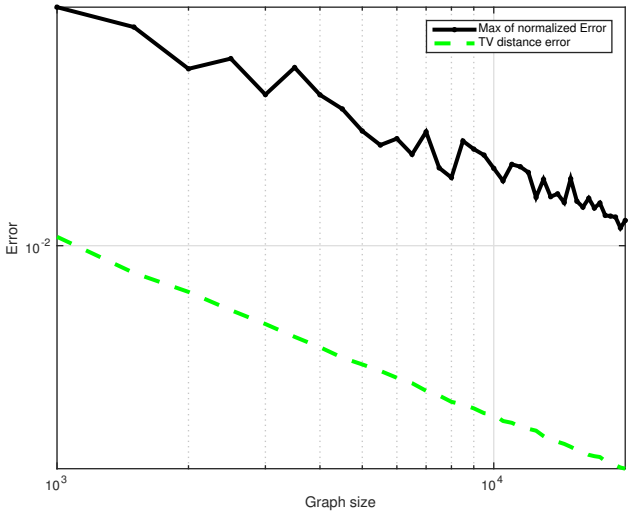



Figure 6.5: Log-log plot of maximum normalized error and TV-distance error for an SBM graph

## 6.7 Conclusions

In this work, we have shown that when the size of a graph tends to infinity, the PageRank vector lends itself to be approximated by a mixture of the preference vector and the degree distribution, for a class of undirected random graphs including the Chung-Lu graph. We expect that these findings will shed more light on the behaviour of PageRank on undirected graphs, and possibly help to optimize the PageRank operation, or suggest further modifications to better capture local graph properties. We also obtain an asymptotic expression for the PageRank on SBM graphs. It is seen that this asymptotic expression contains information about community partitioning in the simple case of SBM with equal-sized communities. It would be interesting to study the implications of our results for community detection algorithms.

Chapter 7

# Random-walk based methods for network average function estimation

## 7.1 Introduction

The prohibitive sizes of most practical networks make graph-processing that requires complete knowledge of the graph impractical. For instance, social networks like Facebook™ and Twitter™ have billions of edges and nodes. We address the problem of estimating global properties of such a large network. Some examples of potentially interesting properties include the size of the support base of a certain political party, the average age of users in an Online Social Network (OSN), the proportion of male-female connections with respect to the number of female-female connections in an OSN, and many others. Naturally, since graphs can be used to represent data in myriad disciplines and scenarios, finding a good estimate of graph function averages is of utmost importance.

Graph sampling can be used to solve the above problem. To collect information from an OSN, the sampler issues an Application Programming Interface (API) query for a particular user, which returns its one-hop neighborhood and the content published by that user. Though some OSNs, for instance Twitter, allow access to the complete database at an additional expense, we focus here on the typical case where a sampler can get information only about the neighbors of a particular user by means of API queries. There are several ways to collect representative samples in a network. One straightforward approach is to collect independent samples via uniform node or edge sampling. However, uniform sampling is not efficient because we do not know the user ID space beforehand. Consequently, the sampler may waste samples issuing invalid IDs, resulting in an inefficient and costly data collection method. Moreover, OSNs typically impose rate limitations on API queries, for e.g., Twitter with 313 million active users enforces a limit of a maximum of 15 requests in a 15-minute time window, for most of APIs.[1] We therefore resort to other, mostly random walk-based, techniques.

### Important Notation and Problem Formulation

Let $G = (V, E)$ be an undirected labeled network, with node set $V$ and edge set $E \subseteq V \times V$. Although the graph is undirected, in later use it would be more convenient to represent edges by ordered pairs $(u, v)$. Of course, if $(u, v) \in E$, it holds that $(v, u) \in E$, since $G$ is undirected. With a slight abuse of notation, $|E|$ denotes the total number of *undirected* edges.

Both edges and nodes can have function values defined on them. For instance, in an OSN, nodes are people, and the node function can be the age or number of friends and the

[1] https://dev.twitter.com/rest/public/rate-limits

edge function can be an indicator function when the end points of the edge are of same gender, if we are interesting in studying these properties.

Let us denote by $g : V \to \mathbb{R}$, where $\mathbb{R}$ is the real number space, a function on the vertices of the graph. We aim to estimate the following network function average:

$$\nu(G) = \frac{1}{|V|} \sum_{u \in V} g(u). \tag{7.1}$$

The constraint on the estimator is that it does not know the whole graph, and can only traverse the graph locally in steps. It does this by issuing API requests, where each API request furnishes the function value $g(\cdot)$ at the queried node and the list of its neighbors. Let $\hat{\nu}^{(n)}_{\mathrm{XY}}(G)$ be our estimate of $\nu(G)$ formed from $n$ samples using the scheme XY. We will occasionally drop the sub and superscripts whenever these are clear from the context.

A *simple RW* on a graph offers a viable solution to this problem that respects the above constraints. From an initial node, a simple RW proceeds by choosing one of the neighbors uniformly randomly. In general, a RW need not sample the neighbors uniformly and can take any transition probability compliant with the underlying graph, an example being the Metropolis-Hastings (MH) schemes [Robert & Casella 2013]. Random walk techniques are well-known (see, for instance, [Cooper *et al.* 2013, Massoulié *et al.* 2006, Avrachenkov *et al.* 2016, Nazi *et al.* 2015, Goel & Salganik 2009, Salganik & Heckathorn 2004, Volz & Heckathorn 2008, Gjoka *et al.* 2010, Dasgupta *et al.* 2014, Ribeiro & Towsley 2010] and references therein).

A drawback of random walk techniques is that they all suffer from the problem of initial burn-in, i.e., a number of initial samples roughly equivalent to the mixing time or *burn-in time* of the RW need to be discarded to get samples from the desired probability distribution. This poses serious limitations, especially in view of the stringent constraints on the number of samples imposed by API query rates. In addition, near-by samples of a RW are obviously not independent. To get independent samples, it is customary to take only samples apart by the mixing time and drop others [Levin *et al.* 2009]. *In this work, we focus on RW-based algorithms that bypass this burn-in time barrier.* We focus on two methods: *Reinforcement learning* and *tour-based Ratio-estimator.*

## Related Work and Contributions

The literature on RW-based sampling techniques is rich and diverse. The estimation techniques in [Cooper *et al.* 2013, Massoulié *et al.* 2006, Avrachenkov *et al.* 2016, Nazi *et al.* 2015] also propose methods to avoid the burn-in time drawback of random walks. The works [Cooper *et al.* 2013, Massoulié *et al.* 2006, Avrachenkov *et al.* 2016] are based on the idea of a random walk tour, which is a sample path of a random walk starting and ending at a fixed node. In [Massoulié *et al.* 2006], the authors estimate the size of a network based on the return times of RW tours. In [Cooper *et al.* 2013], the authors estimate the number of triangles, network size, and subgraph counts from weighted random walk tours using the results of Aldous and Fill [Aldous & Fill 2002, Chapters 2-3]. The work [Avrachenkov *et al.* 2016] extends these results to edge functions, provides real-time Bayesian guarantees for the performance of the estimator, and introduced some hypothesis tests using the estimator.

Instead of the estimators for the sum function of the form $\sum_{u \in V} g(u)$ proposed in these previous works, here we study the average function (7.1). Walk-Estimate proposed in [Nazi *et al.* 2015] aimed to reduce the overhead of burn-in period by considering short random walks and then using acceptance-rejection sampling to adjust the sampling probability of a node with respect to its stationary distribution. This work requires an estimate of

probability of hitting a node at time $t$, which introduces a computational overhead. It also needs an estimate of the graph diameter to work correctly. Our algorithms are completely local and do not require these global inputs.

There are also specific random walk methods tailored for certain forms of function $g(v)$ or criterion, for e.g., in [Dasgupta *et al.* 2014] the authors developed an efficient estimation technique for estimating the average degree, and Frontier sampling in [Ribeiro & Towsley 2010] introduced dependent multiple random walks in order to reduce estimation error.

Two well-known techniques for estimating network averages $\nu(G)$ are the Metropolis-Hastings MCMC (MH-MCMC) scheme [Brémaud 2013, Gjoka *et al.* 2010, Nummelin 2002, Robert & Casella 2013] and Respondent-Driven Sampling (RDS) [Goel & Salganik 2009, Salganik & Heckathorn 2004, Volz & Heckathorn 2008]. In our work, we present a theoretical comparison of the mean-squared error of these two estimators. It has been observed that RDS outperforms MH-MCMC in terms of asymptotic error in many practical cases. We confirm this observation by deriving theoretical expressions for the asymptotic mean-squared errors of the two estimators. We introduce a novel estimator for the network average based on reinforcement learning (RL). Using simulations on real networks, we demonstrate that, with a good choice of cooling schedule, RL can achieve similar asymptotic error performance to RDS but its trajectories have smaller fluctuations. Finally, we extend RDS to accommodate the idea of regeneration during revisits to a node or to a 'super-node', formed by aggregating several nodes, and propose the Ratio with Tours Estimator (RT) estimator, which does not suffer from burn-in period constraints and significantly outperforms the RDS estimator.

### Notational Conventions

Expectation w.r.t. to the MC given initial distribution $\eta$ is denoted by $\mathbb{E}_\eta$, and if this distribution degenerates at a particular node $j$, the expectation is $\mathbb{E}_j$. By $\mathcal{L}(X)$ we mean the law or the probability distribution of a random variable. The stationary distribution is denoted by the row vector $\pi$. In this chapter, $\mathbf{P}$ represents the row-symmetric Markov matrix. Let us define the *fundamental matrix* of a Markov chain as $\mathbf{Z} := (\mathbf{I} - \mathbf{P} + \mathbf{1}\vec{\pi}^\intercal)^{-1}$ [Brémaud 2013]. For two functions $f, g : \mathcal{V} \to \mathbb{R}$, we define $\sigma^2_{ff} := 2\langle \mathbf{f}, \mathbf{Zf}\rangle_\pi - \langle \mathbf{f}, \mathbf{f}\rangle_\pi - \langle \mathbf{f}, \mathbf{1}\vec{\pi}^\intercal\mathbf{f}\rangle_\pi$, and $\sigma^2_{fg} := \langle \mathbf{f}, \mathbf{Zg}\rangle_\pi + \langle \mathbf{g}, \mathbf{Zf}\rangle_\pi - \langle \mathbf{f}, \mathbf{g}\rangle_\pi - \langle \mathbf{f}, \mathbf{1}\vec{\pi}^\intercal\mathbf{g}\rangle_\pi$, where $\langle \mathbf{x}, \mathbf{y}\rangle_\pi := \sum_i x_i y_i \pi_i$, for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{|\mathcal{V}|\times 1}$, where we used $\mathbf{f}$ to denote the function $f$ as a vector indexed by the graph vertices.

## 7.2 MH-MCMC and RDS estimators

The utility of RW based methods comes from the fact that for any initial distribution $\nu$, as time progresses, the sample distribution of the RW at time $t$ starts to resemble a fixed distribution, which we call the stationary distribution of the RW, denoted by $\pi$.

We will study mean squared error and asymptotic variance of random walks based estimators in this chapter. For this purpose, following extension of the central limit theorem for Markov chains plays a significant role:

**Theorem 7.1** ( [Roberts & Rosenthal 2004]). *Let $f$ be a real-valued function $f : V \mapsto \mathbb{R}$ with $\mathbb{E}_\pi[f^2(X_0)] < \infty$. For a finite irreducible Markov chain $\{X_n, n \geq 0\}$ with stationary distribution $\pi$,*

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \mathbb{E}_\pi[f(X_0)]\right) \xrightarrow{D} \mathcal{N}(0, \sigma_f^2),$$

*irrespective of the initial distribution, where*

$$\sigma_f^2 = \lim_{n\to\infty} n \times \mathbb{E}\left[\left\{\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \mathbb{E}_\pi[f(X_0)]\right\}^2\right]$$
$$:= \lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\left[\sum_{k=0}^{n-1} f(X_k)\right]. \tag{7.2}$$

Note that both the above theorems hold for finite periodic chains also (with the existence of unique solution to $\vec{\pi}^\intercal \mathbf{P} = \vec{\pi}^\intercal$).

By [Brémaud 2013, Theorem 6.5] $\sigma_f^2$ in Theorem 7.1 is the same as $\sigma_{ff}^2$. We will also need the following theorem.

**Theorem 7.2** ( [Nummelin 2002, Theorem 3]). *If $f, g$ are two functions defined on the states of a random walk, define the vector sequence $\vec{Z}_k = \begin{bmatrix} f(X_k) \\ g(X_k) \end{bmatrix}$ the following central limit theorem holds*

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} \vec{Z}_k - \mathbb{E}_\pi(\vec{Z}_k)\right) \xrightarrow{D} \mathrm{Normal}(0, \vec{\Sigma}),$$

*where $\vec{\Sigma}$ is $2\times 2$ matrix such that $\vec{\Sigma}_{11} = \sigma_{ff}^2, \vec{\Sigma}_{22} = \sigma_{gg}^2$ and $\vec{\Sigma}_{12} = \vec{\Sigma}_{21} = \sigma_{fg}^2$.*

## Function average from RWs

The simple RW is biased towards higher degree nodes and by Theorem 7.1, the sample averages converge to the stationary average. Hence if the aim is to estimate an average function (7.1), the RW needs to have uniform stationary distribution. Alternatively the RW should be able to unbias it locally. In order to obtain the average, we modify the function $g$ by normalizing by the vertex degrees to get $g'(u) = g(u)/\vec{\pi}_u$, where $\pi_u = d_u/(2|E|)$. Since $\vec{\pi}(u)$ contains $|E|$ and the knowledge of $|E|$ is not available to us initially, it also needs to be estimated. To overcome this problem, we consider the following variation of simple RW.

### 7.2.1 Metropolis-Hastings random walk

We review here the Metropolis Hastings MCMC (MH-MCMC) algorithm. When the chain is in state $i$, it chooses the next state $j$ according to transition probability $p_{ij}$. It then jumps to this state with probability $q_{ij}$ or remains in the current state $i$ with probability $1 - q_{ij}$, where $q_{ij}$ is given as below

$$q_{ij} = \begin{cases} \min\left(\frac{p_{ji}}{p_{ij}}, 1\right) & \text{if } p_{ij} > 0, \\ 1 & \text{if } p_{ij} = 0. \end{cases} \tag{7.3}$$

Therefore the effective jump probability from state $i$ to state $j$ is $q_{ij}p_{ij}$, when $i \neq j$. It follows then that such a process represents a Markov chain with the following transition matrix $\mathbf{P}^{\mathrm{MH}}$

$$P_{ij}^{\mathrm{MH}} = \begin{cases} \frac{1}{\max(d_i, d_j)} & \text{if } (i,j) \in E \\ 1 - \sum_{k\neq i} \frac{1}{\max(d_i, d_k)} & \text{if } i = j \\ 0 & \text{if } (i,j) \notin E, i \neq j. \end{cases}$$

This chain is reversible with stationary distribution $\pi(i) = 1/n \ \forall i \in V$. Therefore the following estimate for $\nu(G)$ using MH-MCMC, where $\{X_n\}$ are MH-MCMC samples, is asymptotically consistent.

$$\widehat{\nu}^{(n)}_{\mathrm{MH}}(G) = \frac{1}{n}\sum_{k=1}^{n} g(X_k).$$

By using Theorem 7.1, we can show the following central limit theorem for MH-MCMC.

**Proposition 7.1** (Central Limit Theorem for MH-MCMC)**.** *For MCMC with uniform target stationary distribution it holds that*

$$\sqrt{n}\left(\widehat{\nu}^{(n)}_{\mathrm{MH}}(G) - \nu(G)\right) \xrightarrow{D} \mathrm{Normal}(0, \sigma^2_{\mathrm{MH}}),$$

*as* $n \to \infty$, *where* $\sigma^2_{\mathrm{MH}} = \sigma^2_{gg} = \frac{2}{|V|}\vec{g}^{\intercal}\mathbf{Z}^{\mathrm{MH}}\vec{g} - \frac{1}{|V|}\vec{g}^{\intercal}\vec{g} - \left(\frac{1}{|V|}\vec{g}^{\intercal}\mathbf{1}\right)^2$, *where* $\mathbf{Z}^{\mathrm{MH}} = (\mathbf{I} - \mathbf{P}^{\mathrm{MH}} + \frac{1}{|V|}\mathbf{1}\mathbf{1}^{\intercal})^{-1}$.

### 7.2.2 Respondent driven sampling technique (RDS-technique)

The estimator with respondent driven sampling uses the simple RW on graphs but applies a correction to the estimator to compensate for the non-uniform stationary distribution.

$$\widehat{\nu}^{(n)}_{\mathrm{RDS}}(G) = \frac{\sum_{k=1}^{n} g(X_k)/d(X_k)}{\sum_{k=1}^{n} 1/d(X_k)} \tag{7.4}$$

We define $h_{\mathrm{nm}}(X_k) := g(X_k)/d(X_k), h_{\mathrm{dm}}(X_k) := 1/d(X_k)$.

The asymptotic unbiasedness derives from the Ergodic Theorem and also as a consequence of the CLT given below.

Now we have the following CLT for the RDS Estimate.

**Proposition 7.2.** *The RDS estimate* $\widehat{\nu}^{(n)}_{\mathrm{RDS}}(G)$ *satisfies a central limit theorem given below*

$$\sqrt{n}\left(\widehat{\nu}^{(n)}_{\mathrm{RDS}}(G) - \nu(G)\right) \xrightarrow{D} \mathrm{Normal}(0, \sigma^2_{\mathrm{RDS}}),$$

*where* $\sigma^2_{\mathrm{RDS}}$ *is given by*

$$\sigma^2_{\mathrm{RDS}} = d^2_{\mathrm{av}}\left(\sigma_1^2 + \sigma_2^2\nu^2(G) - 2\nu(G)\sigma_{12}^2\right),$$

*where* $\sigma_1^2 = \frac{1}{|E|}\sum_{i,j\in V} g_i Z_{ij} g_j/d_j - \frac{1}{2|E|}\sum_{i\in V}\frac{g_i}{d_i} - (\frac{1}{2|E|}\sum_{i\in V} g_i)^2, \sigma_2^2 = \frac{1}{|E|}\sum_{i,j\in V} Z_{ij}/d_j - \frac{1}{2|E|}\sum_i \frac{1}{d_i} - (\frac{1}{d_{\mathrm{av}}})^2, \sigma_{12}^2 = \frac{1}{2|E|}\sum_{i,j\in V} g_i Z_{ij}/d_j + \frac{1}{2|E|}\sum_{i,j\in V} Z_{ij}/d_i - \frac{1}{2|E|d_{av}}\sum_i g_i$

*Proof.* Define the vector $\mathbf{z}_t = \begin{bmatrix} h_{\mathrm{nm}}(x_t) \\ h_{\mathrm{dm}}(x_t) \end{bmatrix}$, and let $\widetilde{\mathbf{z}}_n = \sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{z}_t - \mathbb{E}_\pi(\mathbf{z}_t)\right)$. Then by Theorem 7.2, $\widetilde{\mathbf{z}}_n \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the correlation matrix, whose formula given in Theorem 7.2. Let $\widetilde{\mathbf{z}}_n = (\widetilde{\mathbf{z}}^1_n, \widetilde{\mathbf{z}}^2_n)$. Then we have

$$\begin{aligned}\frac{\sum_{t=1}^{n} h_{\mathrm{nm}}(x_t)}{\sum_{t=1}^{n} h_{\mathrm{dm}}(x_t)} &= \frac{\frac{1}{\sqrt{n}}\widetilde{\mathbf{z}}^1_n + \mu_{h_{nm}}}{\frac{1}{\sqrt{n}}\widetilde{\mathbf{z}}^2_n + \mu_{h_{dm}}} \\ &= \frac{\widetilde{\mathbf{z}}^1_n + \sqrt{n}\mu_{h_{nm}}}{\widetilde{\mathbf{z}}^2_n + \sqrt{n}\mu_{h_{dm}}} = \frac{\widetilde{\mathbf{z}}^1_n + \sqrt{n}\mu_{h_{nm}}}{\sqrt{n}\mu_{h_{dm}}(1 + \frac{\widetilde{\mathbf{z}}^2_n}{\sqrt{n}\mu_{h_{dm}}})}\end{aligned}$$

$$= \frac{1}{\sqrt{n}\mu_{h_{dm}}}(\widetilde{\mathbf{z}}_n^1 - \frac{\widetilde{\mathbf{z}}_n^1\widetilde{\mathbf{z}}_n^2}{\sqrt{n}\mu_{h_{dm}}} + \sqrt{n}\mu_{h_{nm}} - \frac{\widetilde{\mathbf{z}}_n^2\mu_{h_{nm}}}{\mu_{h_{dm}}} + O_p(\frac{1}{\sqrt{n}})),$$

where $O_p(\frac{1}{\sqrt{n}})$ is a term that goes to zero in probability at least as fast as $\frac{1}{\sqrt{n}}$, and $\mu_{h_{nm}}, \mu_{h_{dm}}$ are respectively $\mathbb{E}_\pi(h_{nm})$ and $\mathbb{E}_\pi(h_{dm})$. Then

$$\lim_{n\to\infty} \mathcal{L}\left(\sqrt{n}\left(\frac{\sum_{t=1}^n f'(X_t)}{\sum_{t=1}^n g(X_t)} - \frac{\mu_{f'}}{\mu_g}\right)\right) = \lim_{n\to\infty} \mathcal{L}\left(\frac{1}{\mu_g}\left(\widetilde{\mathbf{z}}_n^1 - \widetilde{\mathbf{z}}_n^2\frac{\mu_{f'}}{\mu_g}\right)\right), \quad (7.5)$$

by Slutsky's lemma [Billingsley 2008]. The result then follows since $(\widetilde{\mathbf{z}}_n^1, \widetilde{\mathbf{z}}_n^2)$ converges to jointly gaussian rv, and by continuous mapping theorem. □

### 7.2.3 Comparing Random Walk Techniques

Two random walks can be compared in many ways. Two prominent techniques are in terms of their mixing times $t_{\text{mix}}$ and the asymptotic variance $\sigma_f^2$ (7.2) of the average estimator. Mixing time is relevant in the situations where the speed at which the RW approaches the stationary distribution matter. But many MCMC algorithms discard some initial samples (called burn-in period) to mitigate the dependence on the initial distribution and this amounts to the mixing time. After the burn-in period, the number of samples needed for achieving a certain estimation accuracy can be determined from the gaussian approximation given by the central limit theorem (see Theorem 7.1). Hence another measure for comparison of the random walks is the asymptotic variance in the gaussian approximation. The lower the asymptotic variance, the smaller the number of samples needed for a certain estimation accuracy. Many authors consider asymptotic unbiasedness as the principal parameter to compare RW based estimators. For instance, the authors in [Lee *et al.* 2012] prove that non-backtracking random walks perform better than the simple RW and MH-MCMC methods in terms of the asymptotic variance of the estimators. The asymptotic variance can be related to the eigenvalues of $\mathbf{P}$ as follows,

$$\sigma_f^2 = \sum_{i=2}^{|V|} \frac{1+\lambda_i^{\mathbf{P}}}{1-\lambda_i^{\mathbf{P}}} |\langle f, \mathbf{u}_i\rangle_\pi|^2,$$

where $\langle \mathbf{x}, \mathbf{y}\rangle_\pi = \sum_{i\in V} \vec{x}_i\vec{y}_i\vec{\pi}_i$ [Brémaud 2013, Chapter 6]. When the interest is in the speed of convergence to equilibrium, then only the second-largest eigenvalue modulus matters. However, if the aim is to compute $\mathbb{E}_\pi[f(X_0)]$ as the ergodic mean $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^n f(X_k)$, then all the eigenvalues become significant and this is captured when the quality of the ergodic estimator is measured by the asymptotic variance.

## 7.3 Network Sampling with Reinforcement Learning (RL-technique)

We will now introduce a reinforcement learning approach based on stochastic approximation to estimate $\nu(G)$. The underlying idea relies on the idea of tours and regeneration introduced in [Avrachenkov *et al.* 2016, Cooper *et al.* 2013, Massoulié *et al.* 2006]. We will compare the mean squared error of the new estimator with that of MH-MCMC and RDS, and see how the stability of the sample paths can be controlled.

### 7.3.1 Estimator

Let $V_0 \subset V$ with $|V_0| << |V|$. We assume that the nodes inside $V_0$ is known beforehand. Consider a simple random walk $\{X_n\}$ on $G$ with transition probabilities $p(j|i) = 1/d(i)$ if $(i,j) \in E$ and zero otherwise. A random walk tour is defined as the sequence of nodes visited by the random walk during successive return to the set $V_0$. Let $\tau_n :=$ successive times to visit $V_0$ and let $\xi_k := \tau_k - \tau_{k-1}$. We denote the nodes visited in the $k$th tour as $X_1^{(k)}, X_2^{(k)}, \ldots, X_{\xi_k}^{(k)}$. Note that considering $V_0$ helps to tackle a disconnected graph[2] with RW theory and makes tours shorter. Moreover the tours are independent to each other and the implementation can be made massively parallel. The estimators derived below and later in Section 7.4 exploit the independence of the tours and the result that expected sum of functions of nodes visited in a tour is proportional to $\sum_{u \in V} g(u)$ [Avrachenkov *et al.* 2016, Lemma 3].

Define $Y_n := X_{\tau_n}$. Then $\{(Y_n, \tau_n)\}$ is a semi-Markov process on $V_0$ [Ross 2013, Chapter 5]. In particular, $\{Y_n\}$ is a Markov chain on $V_0$ with transition probability matrix, say $[p_Y(j|i)]$. We have $\xi_1 := \min\{n > 0 : X_n \in V_0\}$. For a prescribed $g : V \mapsto \mathbb{R}$, define

$$\begin{aligned} T_i &:= \mathbb{E}_i[\xi_1], \\ h(i) &:= \mathbb{E}_i\left[\sum_{m=1}^{\xi_1} g(X_m)\right], \ i \in V_0. \end{aligned}$$

Consider an average cost Markov decision problem (MDP), then the Poisson equation for the semi-Markov process $\{(Y_n, \tau_n)\}$ is [Ross 2013, Chapter 7]

$$\mathcal{V}(i) = h(i) - \beta T_i + \sum_{j \in V_0} p_Y(j|i)\mathcal{V}(j), \ i \in V_0, \tag{7.6}$$

which is to be solved for the pair $(\mathcal{V}, \beta)$, where $\mathcal{V} : V_0 \mapsto \mathbb{R}$ and $\beta \in \mathbb{R}$. Under mild conditions, (7.6) has the solution $(V^*, \beta^*)$. The optimal $\beta^*$ is the average expected cost stationary average of $g$, $\mathbb{E}_\pi[g(X_1)]$ [Ross 2013, Theorem 7.6].

In the following, we provide numerical ways to solve (7.6). This could be achieved using the classical MDP methods like relative value iteration; instead we look for solutions from reinforcement learning in which the knowledge of transition probability $[p_Y(j|i)]$ is not needed. Stochastic approximation provides a simple and easily tunable solution as follows. The relative value iteration algorithm to solve (7.6) is

$$\mathcal{V}_{n+1}(i) = h(i) - \mathcal{V}_n(i_0)T_i + \sum_j p_Y(j|i)V_n(j). \tag{7.7}$$

We can implement this using stochastic approximation as follows. Let $\{Z_n, n \geq 1\}$ be i.i.d. uniform on $V_0$. Construct a tour for $n \geq 1$ by starting a simple RW $X_i^{(n)}, i \geq 0$, with $X_0^{(n)} = Z_n$ and observing its sample path until it returns to $V_0$.

A learning algorithm for (7.6) along the lines of [Abounadi *et al.* 2001] then is, for $i \in V_0$,

$$\begin{aligned} \mathcal{V}_{n+1}(i) = \mathcal{V}_n(i) \ &+ a(n)\chi(z = i)\times \\ &\left[\left(\sum_{m=1}^{\xi_n} g(X_m^{(n)})\right) - \mathcal{V}_n(i_0)\xi_n + \mathcal{V}_n(X_{\xi_n}^{(n)}) - \mathcal{V}_n(i)\right], \end{aligned} \tag{7.8}$$

[2]The underlying Markov chain of the RW requires to be irreducible in order to apply many results of the RWs and this is satisfied when the graph is connected. In case of a disconnected graph, taking at least one seed node from each of the components to form $V_0$ helps to achieve this.

where $a(n) > 0$ are stepsizes satisfying $\sum_n a(n) = \infty$, $\sum_n a(n)^2 < \infty$. (One good choice is $a(n) = 1/\lceil \frac{n}{N} \rceil$ for $N = 50$ or 100.) Also, $i_0$ is a prescribed element of $V_0$. One can use other normalizations in place of $\mathcal{V}_n(i_0)$, such as $\frac{1}{|V_0|}\sum_j \mathcal{V}_n(j)$ or $\min_i \mathcal{V}_n(i)$, see e.g., [Borkar *et al.* 2014]. Then this normalizing term ($\mathcal{V}_n(i_0)$ in (7.8)) converges to $\beta^*$, $\mathbb{E}_\pi[g(X_1)]$, as $n$ increases to $\infty$.

Taking expectations on both sides of (7.8), we obtain a deterministic iteration that can be viewed as an incremental version of the relative value iteration (7.7) with suitably scaled stepsize $\tilde{a}(n) := \frac{a(n)}{|V|}$. This can be analyzed the same way as the stochastic approximation scheme with the same o.d.e. limit and therefore the same (deterministic) asymptotic limit. This establishes the asymptotic unbiasedness of the RL estimator.

The normalizing term used in (7.8) ($\mathcal{V}_n(i_0)$, $\frac{1}{|V_0|}\sum_j \mathcal{V}_n(j)$ or $\min_i \mathcal{V}_n(i)$), along with the underlying random walk as the Metropolis-Hastings, forms our estimator $\widehat{\nu}_{\mathrm{RL}}(G)$ in RL based approach. The iteration in (7.8) is the stochastic approximation analog of it which replaces conditional expectation w.r.t. transition probabilities with an actual sample and then makes an incremental correction based on it, with a slowly decreasing stepwise that ensures averaging. The latter is a standard aspect of stochastic approximation theory. The smaller the stepwise the less the fluctuations but slower the speed, thus there is a trade-off between the two.

RL methods can be thought of as a cross between a pure deterministic iteration such as the relative value iteration above and pure MCMC, trading off variance against per iterate computation. The gain is significant if the number of neighbors of a node is much smaller than the number of nodes, because we are essentially replacing averaging over the latter by averaging over neighbors. The $\mathcal{V}$-dependent terms can be thought of as control variates to reduce variance.

### 7.3.2 Extension of RL-technique to uniform stationary average case

The stochastic approximation iteration in (7.8) converges to $\beta$, which is $\mathbb{E}_\pi[g(X_1)]$, where $\pi$ is the stationary distribution of the underlying walk. To make it converge to $\nu(G)$, we can use the Metropolis-Hastings random walk with uniform target distribution. However, we can avoid the use of Metropolis-Hastings algorithm by the following modification, motivated from importance sampling that achieves the convergence to $\nu(G)$ with the simple random walk. We have

$$\mathcal{V}_{n+1}(i) = \mathcal{V}_n(i) \; + a(n)\chi(z=i) \times \Gamma_{\xi_n}^{(n)} \times \left[ \left( \sum_{m=1}^{\xi_n} g(X_m^{(n)}) \right) - \mathcal{V}_n(i_0)\xi_n + \mathcal{V}_n(X_{\xi_n}^{(n)}) - \mathcal{V}_n(i) \right],$$

where

$$\Gamma_m^{(n)} = \prod_{k=1}^{m} \left( \frac{p(X_k^{(n)}|X_{k-1}^{(n)})}{q(X_k^{(n)}|X_{k-1}^{(n)})} \right).$$

Here $q(\cdot|\cdot)$ is the transition probability of the random walk with which we simulate the algorithm and $p(\cdot|\cdot)$ corresponds to the transition probability of the random walk with respect to which we need the the stationary average. The transition probability $p$ can belong to any random walk having uniform stationary distribution such that $q(\cdot|\cdot) > 0$ whenever $p(\cdot|\cdot) > 0$. One example is to use $p$ as the transition probability of Metroplis-Hastings algorithm with target stationary distribution as uniform and $q$ as the transition

probability of a lazy version of simple random walk, i.e., with transition probability matrix $(\mathbf{I}+\mathbf{P}_{\text{simple RW}})/2$. In comparison with basic Metropolis-Hastings sampling, such importance sampling avoids the API requests for probing the degree of all the neighboring nodes, instead requires only one such, viz., that of the sampled node. Note that the self-loops wherein the chain re-visits a node immediately are not wasted transitions, because it amounts to re-application of a map to the earlier iterate which is distinct from its single application.

The reinforcement learning scheme introduced above is the semi-Markov version of the scheme proposed in [Borkar 2009] and [Borkar *et al.* 2014].

### 7.3.3 Advantages

The RL-technique extends the use of regeneration, tours and super-node introduced in [Avrachenkov *et al.* 2016] to the average function $\nu(G)$. Even though the RL-technique is not non-asymptotically unbiased unlike the algorithm in [Avrachenkov *et al.* 2016], it has the following advantages:

1. It does not need to wait until burn-in time to collect samples;

2. Comparison with [Avrachenkov *et al.* 2016]: The super-node in [Avrachenkov *et al.* 2016] is a single node, an amalgamation of the node set $V_0$. But such a direction assumes that the contribution of all the edges inside the induced subgraph of $V_0$ to $\nu(G)$ completely known. It could have been avoided if we could make use of the techniques for partitioning state space of a Markov chain (called *lumpability* in [Kemeny & Snell 1983]). The conditions stated in [Kemeny & Snell 1983, Theorem 6.3.2] are not satisfied here and hence we can not invoke such techniques. But the RL-technique, without using the lumpability arguments, need not know the edge functions of the subgraph induced by $V_0$;

3. RL-technique along with the extension in Section 7.3.2 can further be extended to the directed graph case provided the graph is strongly connected. On the other hand, for estimators from other RW based sampling schemes, the estimator requires knowledge of the stationary distribution to unbias and thus to form the estimator. But in many cases including simple RW on directed graphs, the stationary distribution does not have a closed form expression unlike in undirected case, and this poses a big challenge for design of simple random walk based estimators;

4. As explained before, the main advantage of RL-estimator is its ability to control the stability of sample paths and its position as a cross between deterministic and MCMC iteration. We will see more about this in the numerical section.

## 7.4 Ratio with Tours Estimator (RT estimator)

In this section we use of the idea of regeneration and tours introduced in [Avrachenkov *et al.* 2016] to estimate the average function $\nu(G)$. However, since the tour estimator only gives an unbiased estimator for network sums namely $\sum_{i\in V} g(i)$, to find an estimate for $\nu(G)$ we use the same samples to get an estimate for $|V|$. Let $I_n$ be the set of initial nodes recruited for forming the super-node [Avrachenkov *et al.* 2016] and let $S_n$ be the single combined node corresponding to $I_n$. We emphasize that while in RL-technique, the set of selected nodes $I_n$ stays intact, in the RT estimator case, we shrink all these nodes in one super-node $S_n$. The estimator will compensate for network modification. With a sampling

budget $B$, the RT estimator is given by

$$\widehat{\nu}(\mathcal{D}_{m(B)}(S_n)) = \frac{\sum_{k=1}^{m(B)} \sum_{t=1}^{\xi_k - 1} \frac{f(X_t^{(k)})}{d_{X_t^{(k)}}} + \frac{\sum_{i \in I_n} g(i)}{d_{S_n}}}{\sum_{k=1}^{m(B)} \sum_{t=1}^{\xi_k - 1} \frac{1}{d_{X_t^{(k)}}} + \frac{n}{d_{S_n}}}, \tag{7.9}$$

where $m(B)$ is the number of tours until the budget $B$,

$$m(B) := \max\{k : \sum_{j=1}^{k} \xi_j \leq B\}.$$

The function $f(u) := g(u)$ if $u \notin I_n$, otherwise $f(u) = 0$.

This estimator is very close to RDS sampling, explained in Section 7.2.2, except that we miss $B - \sum_{k=1}^{m(B)} \xi_k$ samples for the estimation purpose. An advantage of RT estimator is that we could leverage all the advantages of super-node mentioned in [Avrachenkov *et al.* 2016, Section 2] and we claim that this would highly improve the performance. We show this via numerical simulations in the next section, and theoretical properties will be studied in future.

Note that the formulation of super-node is different from $V_0$ considered in the RL-technique, where the RW tours can start from any uniformly selected node inside $V_0$ and the tours end when it hit the set $V_0$. On the other hand, the super-node which is formed from $n$ nodes in $V$ is considered as a single node (removing all the edges in between the nodes in $S_n$) and this contract the original graph $G$. Both the formulations have advantages of their own: Super-node and its estimator is easy to form and compute, but one needs to know all the edges between the nodes in $S_n$, i.e., the induced subgraph from $S_n$ should be known a priori. The set $V_0$ in RL-technique does not demand this.

## 7.5 Numerical results

The algorithms RL-technique, RT-estimator, RDS and MH-MCMC (see Sections 7.2.1 and 7.2.2) are compared in this section using simulations on two real-world networks. For the figures given below, the x-axis represents the budget $B$ which is the number of allowed samples, and is the same for all the techniques. We use the normalized root mean squared error (NRMSE) for comparison for a given $B$ and is defined as

$$\text{NRMSE} := \sqrt{\text{MSE}}/\nu(G),$$

where

$$\text{MSE} = \mathbb{E}\left[\left(\widehat{\nu}^{(n)}(G) - \nu(G)\right)^2\right].$$

Recall that $\text{MSE} = \text{Var}[\widehat{\nu}^{(n)}(G)] + \left(\mathbb{E}[\widehat{\nu}^{(n)}(G)] - \nu(G)\right)^2$. We also study the asymptotic variance $\sigma_g^2$ (see (7.2)) of the random walk based estimators including RL-technique in terms of $n \times \text{MSE}$, since the bias $|\mathbb{E}[\widehat{\nu}^{(n)}(G)] - \nu(G)| \to 0$ as $n \to \infty$.

*Note that in the numerical results of RL-technique and MH-MCMC, we have not included burn-in time for calculating their budget $B$, and if it is added, their performance will be much worse than what we have shown below.*

### 7.5.1 Numerical Results for RL-technique

For the RL-technique we choose the initial or super-node $V_0$ by uniformly sampling nodes assuming the size of $V_0$ is given a priori.

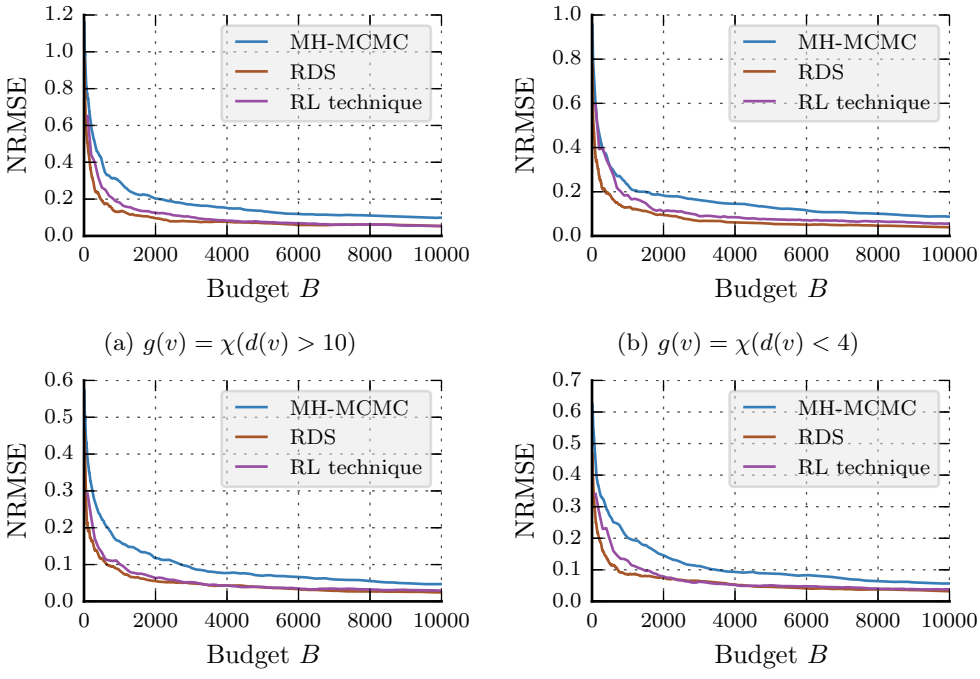
#### 7.5.1.1 Les Misérables network

In Les Misérables network, nodes are the characters of the novel and edges are formed if two characters appear in the same chapter in the novel. The number of nodes is 77 and number of edges is 254. We have chosen this rather small network in order to compare all the three methods in terms of theoretical limiting variance. Here we consider four demonstrative functions: a) $g(v) = \chi(d(v) > 10)$ b) $g(v) = \chi(d(v) < 4)$ c) $g(v) = d(v)$ and d) for calculating $\nu(G)$ as the average clustering coefficient

$$C := \frac{1}{|V|}\sum_{v\in V} c(v), \quad \text{where } c(v) = \begin{cases} t(v)/\binom{d_v}{2} & \text{if } d(v) \geq 2 \\ 0 & \text{otherwise,} \end{cases} \tag{7.10}$$

with $t(v)$ as the number of triangles that contain node $v$. Then $g(v)$ is taken as $c(v)$ itself.

The average in MSE is calculated from multiple runs of the simulations. The simulations on Les Misérables network is shown in Figure 7.1 with $a(n) = 1/\lceil \frac{n}{10} \rceil$ and the super-node size as 25.




(a) $g(v) = \chi(d(v) > 10)$ (b) $g(v) = \chi(d(v) < 4)$

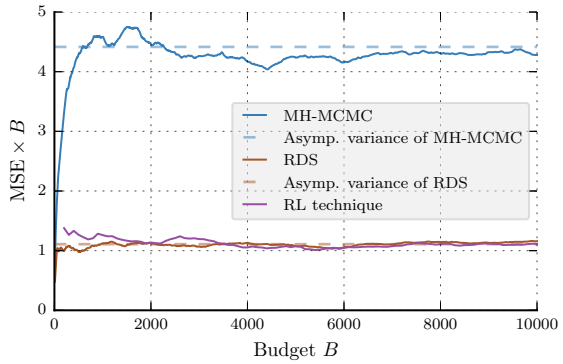
(c) $g(v) = d(v)$ (d) $g(v) = c(v)$, $c(v)$ defined in (7.10)

Figure 7.1: Les Misérables network: NRMSE comparisons

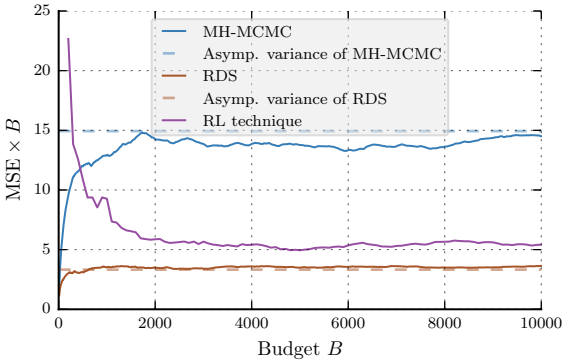
**Study of asymptotic MSE:**

In order to show the asymptotic MSE expressions derived in Propositions 7.1 and 7.2, we plot the sample MSE as MSE $\times$ $B$ in Figures 7.2a, 7.3b and 7.2c. These figures correspond

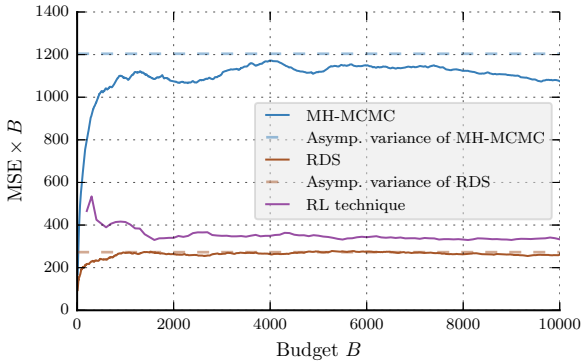
to the three different functions we have considered. It can be seen that asymptotic MSE expressions match well with the estimated ones.




(a) $g(v) = \chi(d(v) > 10)$




(b) $g(v) = \chi(d(v) < 4)$




(c) $g(v) = d(v)$

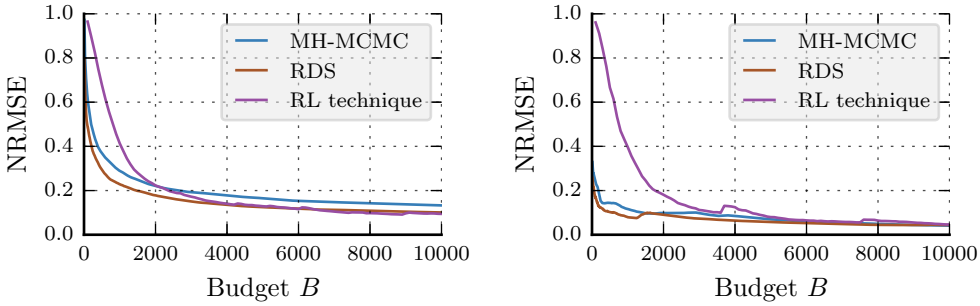
Figure 7.2: Les Misérables network: asymptotic MSE comparisons (contd.)

#### 7.5.1.2 Friendster network

We consider a larger graph here, a connected subgraph of an online social network called Friendster with 64,600 nodes and 1,246,479 edges. The nodes in Friendster are individuals and edges indicate friendship. We consider the functions a). $g(v) = \chi(d(v) > 50)$ and b). $g(v) = c(v)$ (see (7.10)) used to estimate the average clustering coefficient. The plot

in Figure 7.3a shows the results for Friendster graph with super-node size 1000. Here the sequence $a(n)$ is taken as $1/\lceil \frac{n}{25} \rceil$.

Now we concentrate on *single* sample path properties of the algorithms. Hence the numerator of NRMSE becomes absolute error. Figure 7.3c shows the effect of increasing super-node size while fixing step size $a(n)$ and Figure 7.3d shows the effect of changing $a(n)$ when super-node is fixed. In both the cases, the green curve of RL-technique shows much stability compared to the other techniques.




(a) $g(v) = \chi(d(v) > 50)$

(b) $g(v) = c(v)$, $c(v)$ defined in(7.10)

Figure 7.3: Friendster network: (a) & (b) NRMSE comparison

.

#### 7.5.1.3 Observations

Some observations from the numerical experiments are as follows:

1. With respect to the limiting variance, RDS always outperforms the other two methods tested. However, with a good choice of parameters the performance of RL is not far from that of RDS;

2. In the RL-technique, we find that the normalizing term $1/|V_0|\sum_j \mathcal{V}_n(j)$ converges much faster than the other two options, $\mathcal{V}_t(i_0)$ and $\min_i \mathcal{V}_t(i)$;

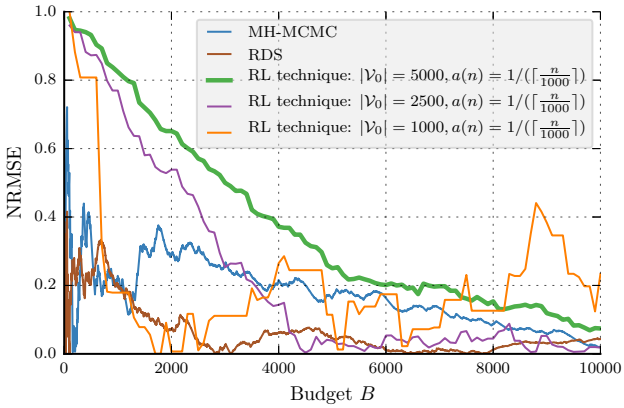
3. When the size of the super-node decreases, the RL-technique requires smaller step size $a(n)$. For instance in case of Les Misérables network, if the super-node size is less than 10, RL-technique does not converge with $a(n) = 1/(\lceil \frac{n}{50} \rceil + 1)$ and requires $a(n) = 1/(\lceil \frac{n}{5} \rceil)$;

4. If step size $a(n)$ decreases or the super node size increases, RL fluctuates less but with slower convergence. In general, RL has less fluctuations than MH-MCMC or RDS.

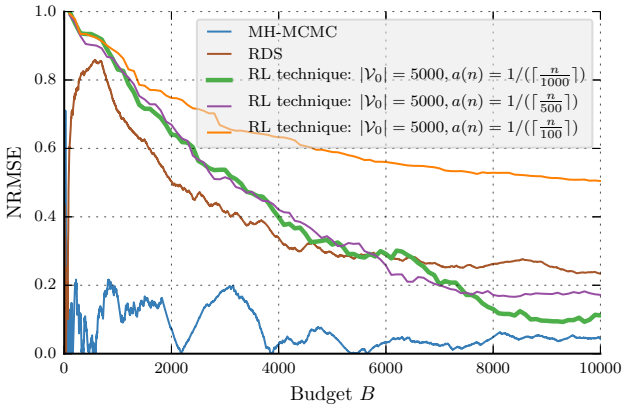
### 7.5.2 Numerical results for RT-estimator

Here we compare RDS and RT estimator. The choice of RDS for comparison is motivated by the results shown in the previous section that it outperforms other sampling schemes considered here, in terms of asymptotic variance and mean squared error. Moreover, RT-estimator can be regarded as a natural modification of RDS making use of the ideas of tours and super-node.

Figure 7.4 shows the results in Friendster network ($|V| = 64,600, |E| = 1,246,479$). One can see that the performance is improved even for small super-node size.




(c) Single sample path: Varying super-node size




(d) Single sample path: Varying step size

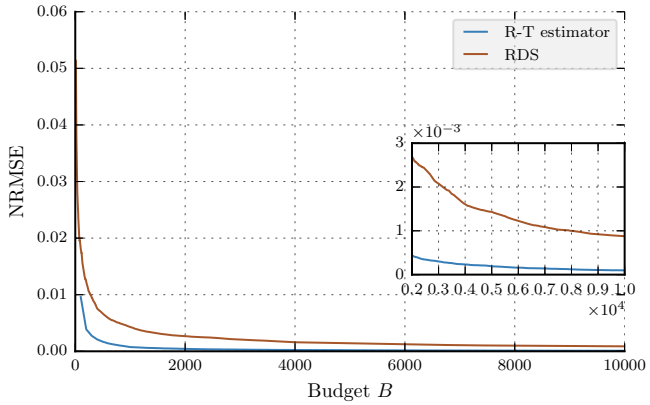
Figure 7.3: Friendster network (contd.): (c) & (d) Single sample path comparison with $g(v) = \chi(d(v) > 50)$

.

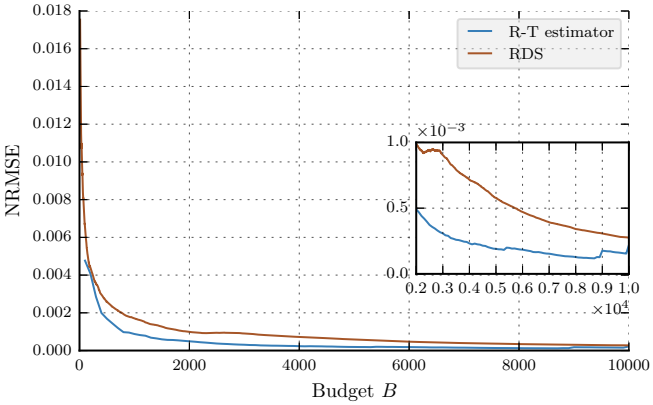
## 7.6 Conclusions

We addressed a critical issue in the study of random walks on graphs: the burn-in period. Our ideas are based on exploiting the tours (regenerations) and on the best use of the given seed nodes by making only short tours. These short tours or crawls, which start and return to the seed node set, are independent and can be implemented in a massively parallel way. The idea of regeneration allows us to construct estimators that are not marred by the burn-in requirement. We proposed two estimators based on this idea of regeneration. The first, the RL estimator, uses reinforcement learning and stochastic approximation to build a stable estimator by observing random walks returning to the seed set. We then proposed the RT estimator, which is a modification of the classical respondent driven sampling, making use of the idea of short crawls and super-node. These two schemes have advantages of their own: the reinforcement learning scheme offers more control on the stability of the sample path with varying error performance, and the modified RDS scheme based on short crawls is simple and has superior performance compared to the classical RDS.

In the future, our aim is to study deeply the theoretical performance of our algorithms. We have also left open the selection process for the initial seed set or the super-node, and




(a) $g(v) = \chi(d(v) > 50)$: super-node size = 5000




(b) $g(v) = c(v)$: super-node size = 1000

Figure 7.4: Friendster network: Comparison between RDS and RT estimators

.

this also suggests an interesting research topic to explore in the future.

Chapter 8

# Conclusions and Future Research

## 8.1 Summary and Conclusions

Random Matrix Theory is a rich field with a wealth of landmark and deep results on the eigenvalue and eigenvector properties of ensembles of large matrices with random entries. These results have been successfully applied to the developement of algorithms and theory in varied fields such as Telecommunications, Finance, Statistics and Compressive Sensing, just to name a few. The explosion in processing power and a simultaneous spurt in interest in algorithms on data modeled as graphs have, in recent times, made the theory of large random matrices attractive in graph theory and random graph analysis as well. The theory of random matrices applied to the study of graph algorithms and processes on graphs was the principal focus of this thesis.

Arguably, one of the most widely studied random graph model in Random Matrix Theory is the Stochastic Block Model due to its significance in the analysis and comparison of community partitioning algorithms. In the related literature, there is an almost exclusive focus on the properties of extremal eigenvalues and eigenvectors since the latter are key to the performance of spectral clustering and related algorithms. The literature on the properties of non-dominant eigenvectors and eigenvalues is hence lacking. In Chapter 3 we tried to mitigate this by studying the shape of the asymptotic e.s.d. of the adjacency and Laplacian matrices of an M-community SBM using classical Random Matrix Theoretic tools. We discovered that while in *symmetric SBM*, where the expected node degrees are the same, the asymptotic e.s.d. is the semicircle law as for the ER graph, when the expected node degrees are different, the spectral shape is different from the semicircle and can be determined as a solution to a set of fixed point equations. In addition, we used a similar approach to studying the asymptotic distribution of the *bulk* eigenvectors of the adjacency matrix via a modified e.s.d. We established that these eigenvectors satisfy some of the properties of the standard Wigner matrix eigenvectors, but not all. The properties of bulk eigenvectors could be leveraged in feature detection problems in graphs, where the bulk eigenvectors act as noise, and this remains to be explored.

In continuing the application of Random Matrix Theoretic tools to the study of processes on random graphs, we analysed the asymptotic behaviour of PageRank on undirected random graphs by leveraging asymptotic bounds on the eigenvalues of Markov matrix in Chapter 6. We found out that PageRank has simple asymptotic expressions on a class of random graphs with a large spectral gap called expanders. We also analyzed PageRank on SBM and found out that its limiting expression has a term that incorporates community partitioning. This fact could be leveraged to analyze the performance of community partioning algorithms based on PageRank.

Related to community partioning on graphs is the problem of anomaly detection on graphs, where the anomaly takes the form of a dense subgraph. In Chapter 4 we used spectral tools to perform a hypothesis test in a random graph model where there is a hidden community of nodes with a larger edge probability than the rest of the nodes. The graph we considered is relatively dense with the average degree growing as $\Omega(\text{polylog}(n))$.

We characterized the asymptotic distribution of the dominant eigenvector components of the modularity matrix of this problem and proved that for a certain minimum subgraph size, they are asymptotically gaussian. Using the latter, we derived an approximate CLT for the distribution of the test statistic based on the eigenvector $L^1$-norm, which was then used to devise a detection test.

While traditional Random Matrix Theoretic tools are quite efficient on relatively dense graphs, when the average degree scales to infinity, dilute graphs with average degree $O(1)$ require more complex matrix models and different tools. In Chapter 5 we apply Belief Propagation, a local algorithm that takes advantage of the locally-tree like property of dilute graphs, to the problem of hidden community detection. As opposed to the previous works, we tackled the problem of hidden community detection in a *semi-supervised* setting and discovered that an important *threshold phenomenon* present in local algorithms on graphs can be overcome in the presence of side-information under a certain scaling regime of the subgraph parameters. In other words, while it has been shown that the local BP algorithm for subgraph detection has a non-diminishing error rate whenever an *effective SNR* falls below a certain threshold, we show that in the presence of side-information, this computational threshold disappears. In the future, we would like to investigate this phenomenon under more general scaling laws.

In the last chapter of this thesis we looked at RW-based sampling algorithms for estimating averages of an arbitrary function defined on the nodes of the graph. We compare two algorithms, MH-MCMC and RDS estimators with respect to their asymptotic variance. We finally propose two algorithms based on re-inforcement learning and random walk tours that overcome the burn-in time barrier that afflicts many random-walk based estimation algorithms. In the future, it would be interesting to use Random Matrix Theoretic techniques to compare the mean-squared error performance of these algorithms on some representative random graph models.

In the course of this thesis we explored some topics that hitherto had not received sufficient attention in applied Random Matrix Theory such as the asymptotic gaussianity property of eigenvectors. Our work on PageRank is one of the first that analyzes PageRank properties using tools from Random Matrix Theory, and it would be interesting to deepen this research direction. The work on hidden community detection, although it uses known tools in Statistical Physics and Information Theory, provides insight into the dramatic role played by side-information in graph-based detection problems. In the next section, we discuss possible ways in which the different research directions considered in this thesis can be further explored and extended.

## 8.2 Future works and Perspectives

We discuss in this section some of the possible avenues for future work.

### More General Graph Models

In the community detection and anomaly detection algorithmic analysis in this thesis, we mainly focused on the basic Stochastic Block Model, which differs greatly from real-world graphs. The major drawback of the plain SBM is that the expected degree of all nodes are the same in a community. As discussed in Chapter 1, the degree corrected SBM mitigated this drawback. Another extension is the case of overlapping Stochastic Block Models [Latouche *et al.* 2009], where nodes do not belong to only one community, but can be members of multiple communities. In such cases one can investigate if the results of Chapters 4, 5 continue to hold, and if not, how they change.

### Different Kinds of Anomalies

In Chapter 4, we considered one kind of anomaly, i.e., the presence of a dense subgraph in a sparse graph. One can consider other kinds of anomalies; for e.g., deletion or addition of edges with some probability over all nodes, and not just a subset of nodes. One can investigate the effect this would have on the eigenvalues and thanks to this, if we can detect the anomaly. Anomaly detection, as we formulated in Chapter 4 is a hypothesis test problem between two different graphical models and hence can be readily extended. It can be investigated whether a graph spectrum-based framework can be developed to differentiate between more general random graph models.

### General Analysis of Belief Propagation

In Chapter 5, we analyzed the error performance of Belief Propagation for subgraph detection and showed that any $\lambda > 0$ leads to zero asymptotic error when $\kappa = K/n \to 0$. However, under the parameter setting we considered, $\lambda$ is a constant with respect to $n$. It would be interesting to characterize the minimum possible $\lambda$ as a function of $n$ that still ensures error-free detection when $K = o(n)$. For example, in [Hajek *et al.* 2015b], the authors showed that for ML detection, the detectability threshold for the subgraph is $\lambda = \Omega(\frac{K}{n}\log(n/K))$. It can be asked if BP with side-information comes close to this threshold. In addition, we considered the performance of BP when $a = np$ scales to infinity. What happens for a finite $a$? This scenario was for e.g. considered for community partitioning in SBM in [Cai *et al.* 2016]. In this case, the graph becomes disconnected with a large connected component, when $a > 1$.

### Different kinds of Side-Information

In Chapter 5, we considered two kinds of side-information and found that the asymptotic error can be bounded uniformly for any small amount of side-information and that the minimum computational threshold disappears. It can be investigated if this generalizes to other kinds of side-information. For e.g., one can consider clustering problems where cued nodes are not revealed, instead for a fraction of node pairs it is revealed whether they belong to the same community or different communities, known as must-link and cannot link constraints in semi-supervised clustering literature. This case represents a more realistic scenario because it reflects the type of side-information available in real-life cases, since in user-assisted classification where the exact nature of clusters is not known beforehand, only information such as whether two items are similar or dissimilar is available [Basu *et al.* 2006]. The BP algorithm to handle this side-information would be different and also its analysis.

### BP without parameter knowledge

One drawback of the BP algorithm considered in this thesis is that it requires the knowledge of the parameters $p, q, K$. One solution to this issue is to estimate these parameters beforehand. Algorithms are available for this in [Kloumann *et al.* 2016, Mossel *et al.* 2012]. It may also be possible to develop message passing that do not require the knowledge of these parameters. On the other hand, one can also investigate an algorithm that learns these parameters online as in expectation maximisation algorithms.

### BP with heterogenous graphs

As noted during the simulations on real graphs in Section 5.6.2, heterogenous graphs with more than a single community hidden in a sparse graph are challenging to our BP algorithm. Is it possible to develop an algorithm that handles heterogeneity such as overlapping communities, multiple communities in the same graph etc.?

### Closing the gap between BP and linear methods

We developed BP for sparse graphs, thanks to locally tree like property. On the other hand, spectral methods are close to optimal on dense graphs. Is it possible to construct an algorithm that takes advantage of these properties and achieves the best of both worlds. We can investigate if a clever linearization of BP can make it work for dense graphs also, without sacrificing the detection performance of BP.

### Pagerank analysis and Heat Kernel diffusion algorithms

We analyzed PageRank on SBM graphs in this thesis. PageRank is one of many existing diffusion algorithms on graphs. For example, there are Heat Kernel and Katz diffusion algorithms. There does not exist a comparitive analysis of these different algorithms on random graphs. It can be asked if random matrices provides a viable way of comparing these different techniques for community detection and other tasks on graphs.

### Higher order graph structures

In this thesis, we dealt with node communities, i.e., a group of nodes with a larger edge density than the rest of the graph. This is only one of the many possible conceptions of a graph community. In the paradigm considered in this thesis, a community can be defined as a set of nodes that has the smallest conductance defined in terms of the number of edges cut. This concept can be extended to more complicated motifs on graphs such as triangles and cycles. For example, one can ask how can a given graph be partitioned so that a minimum number of triangles are cut. A PageRank-based approach to solving this problem was undertaken in [Benson *et al.* 2015]. One can investigate the detection limits in this setting and spectral approaches to solving this problem.

### Hypergraphs

The graph concepts discussed here such as the graph Laplacian, modularity and others could be extended to more complicated graph structures such as hypergraphs [Lu & Peng 2013]. Hypergraphs model multiway relationship between nodes that normal graphs cannot. Although there have been many efforts to extend the ideas of spectral clustering to spectral graphs [Zhou *et al.* 2007,Louis 2015,Ghoshal *et al.* 2009], a principled approach to extending modularity and other concepts from graphs to hypergraphs is still missing. Applications are in community detection and anomaly detection on hypergraphs [Silva & Willett 2009].

Appendix A

# Appendix: Chapter 5

## A.1 Description of G-W tree and derivation of Algorithm 2

We derive Algorithm 2 by establishing a coupling formulation between a $t$-hop neighbourhood $G_u^t$ of node $u$ and a Galton-Watson (G-W) tree rooted at $u$ constructed as follows. Let $T_u^t$ be a labelled Galton-Watson (G-W) tree of depth $t$ rooted at node $u$ constructed as follows (as in [Hajek *et al.* 2015a]): The label $\tau_u$ at node $u$ is chosen at random in the following way:

$$\mathbb{P}(\tau_u = 1) = \frac{K}{n}, \qquad \mathbb{P}(\tau_u = 0) = \frac{n-K}{n}.$$

The number of children $N_u$ of the root $u$ is Poisson-distributed with mean $d_1 = Kp+(n-K)q$ if $\tau_u = 1$ and mean $d_0 = nq$ if $\tau_u = 0$. Each child is also assigned a label. The number of children $i$ with label $\tau_i = 1$ is Poisson distributed with mean $Kp$ if $\tau_u = 1$ and mean $Kq$ if $\tau_i = 0$. The number of children with label $\tau_i = 0$ is Poisson distributed with mean $(n-K)q$ for both $\tau_u = 0$ and $\tau_u = 1$. By the independent splitting property of Poisson random variables, this is equivalent to assigning the label $\tau_i = 1$ to each child $i$ by sampling a Bernoulli random variable with probability (w.p.) $Kp/d_1$ if $\tau_u = 1$ and $Kq/d_0$ if $\tau_u = 0$. Similarly $\tau_i = 0$ w.p. $(n-K)q/d_1$ and $(n-K)q/d_0$ for $\tau_u = 0$ and 1 respectively. Namely, if $i$ is a child of $u$,

$$\mathbb{P}(\tau_i = 1|\tau_u = 1) = \frac{Kp}{d_1}, \qquad \mathbb{P}(\tau_i = 1|\tau_u = 0) = \frac{Kq}{d_0}. \tag{A.1}$$

We then assign the cue indicator function $\widetilde{c}$ such that $\widetilde{c}_i = 1$ w.p. $\alpha$ if $\tau_i = 1$ and $\widetilde{c}_i = 0$ if $\tau_i = 0$. The process is repeated up to depth $t$ giving us $\widetilde{C}_u^t$, the set of cued neighbours. Now we have the following coupling result between $(G_u^t, \sigma^t, C_u^t)$, the neighbourhood of $u$ and the node labels of that neighbourhood and $(T_u^t, \tau^t, \widetilde{C}_u^t)$, the depth-$t$ tree $T_u^t$ and its labels due to [Hajek *et al.* 2015a].

**Lemma A.1.** *[Hajek et al. 2015a, Lemma 15] For $t$ such that $(np)^t = n^{o(1)}$, there exists a coupling such that $(G_u^t, \sigma^t, C_u^t) = (T_u^t, \tau^t, \widetilde{C}_u^t)$ with probability $1 - n^{-1+o(1)}$.*

We now derive the recursions for the likelihood ratios on the tree $T_u^t$. For large $n$ with high probability, by the coupling formulation, $R_u^t$ also satisfy the same recursions. For notational simplicity, from here onwards we represent the cue labels on the tree by $c$ and the set of cued neighbours by $C_u^t$, just as for the original graph. We use $\Lambda_u^t$ to denote the likelihood ratio of node $u$ computed on a tree defined as below:

$$\Lambda_u^{t+1} = \log\left(\frac{\mathbb{P}(T_u^{t+1}, C_u^{t+1}|\tau_u = 1)}{\mathbb{P}(T_u^{t+1}, C_u^{t+1}|\tau_u = 0)}\right).$$

By virtue of tree construction, if the node $u$ has $N_u$ children, the $N_u$ subtrees rooted on these children are jointly independent given $\tau_u$. We use this fact to split $\Lambda_u^{t+1}$ in two parts.

$$\begin{aligned}\Lambda_u^{t+1} &= \log\left(\frac{\mathbb{P}(T_u^{t+1}, C_u^{t+1}|\tau_u=1)}{\mathbb{P}(T_u^{t+1}, C_u^{t+1}|\tau_u=0)}\right) \\ &= \log\left(\frac{\mathbb{P}(N_u|\tau_u=1)}{\mathbb{P}(N_u|\tau_u=0)}\right) + \sum_{i\in\delta u}\log\left(\frac{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=1)}{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=0)}\right),\end{aligned} \tag{A.2}$$

by the independence property of subtress $T_i^t$ rooted on $i \in \delta u$. Since by Lemma A.1, the degrees are Poisson,

$$\mathbb{P}(N_u|\tau_u=1) = d_1^{N_u} e^{-d_1}/N_u!\,,$$

and similarly for $\mathbb{P}(N_u|\tau_u=0)$. Therefore we have

$$\begin{aligned}\log\left(\frac{\mathbb{P}(N_u|\tau_u=1)}{\mathbb{P}(N_u|\tau_u=0)}\right) &= N_u\log\left(\frac{d_1}{d_0}\right) - (d_1-d_0) \\ &= N_u\log\left(\frac{d_1}{d_0}\right) - K(p-q).\end{aligned} \tag{A.3}$$

Next we look at the second term in (A.2). We analyze separately the case of $c_i = 1$ and $c_i = 0$ for $i \in \delta_u$, i.e, the cued and uncued children are handled separately.

*Case 1* ( $c_i = 1$): We have

$$\begin{aligned}\log\left(\frac{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=1)}{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=0)}\right) &\overset{(a)}{=} \log\left(\frac{\mathbb{P}(T_i^t, c_i, C_i^t, \tau_i=1|\tau_u=1)}{\mathbb{P}(T_i^t, c_i, C_i^t, \tau_i=1|\tau_u=0)}\right) \\ &= \log\left(\frac{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_i=1)\mathbb{P}(\tau_i=1|\tau_u=1)}{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_i=1)\mathbb{P}(\tau_i=1|\tau_u=0)}\right) \\ &\overset{(b)}{=} \log\left(\frac{Kp/d_1}{Kq/d_0}\right),\end{aligned} \tag{A.4}$$

where in step (a) we applied the fact that $c_i = 1$ implies $\tau_i = 1$, and in (b) we used (A.1).

*Case 2* ($c_i = 0$): Observe that $\mathbb{P}(c_i = 0|\tau_i = 1) = 1-\alpha$ and $\mathbb{P}(c_i = 0|\tau_i = 0) = 1$. Note that

$$\begin{aligned}\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=1) &= \mathbb{P}(T_i^t, C_i^t|\tau_i=1)\mathbb{P}(c_i|\tau_i=1)\mathbb{P}(\tau_i=1|\tau_u=1) \\ &\quad + \mathbb{P}(T_i^t, C_i^t|\tau_i=0)\mathbb{P}(c_i|\tau_i=0)\mathbb{P}(\tau_i=0|\tau_u=1) \\ &= \mathbb{P}(T_i^t, C_i^t|\tau_i=1)(1-\alpha)\frac{Kp}{d_1} + \mathbb{P}(T_i^t, C_i^t|\tau_i=0)\frac{(n-K)q}{d_1}.\end{aligned} \tag{A.5}$$

Similarly, we can show

$$\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=0) = \mathbb{P}(T_i^t, C_i^t|\tau_i=1)\frac{Kq}{d_0}(1-\alpha) + \mathbb{P}(T_i^t, C_i^t|\tau_i=0)\frac{(n-K)q}{d_0}. \tag{A.6}$$

Let us define

$$\Lambda_{i\to u}^t := \log\left(\frac{\mathbb{P}(T_i^t, C_i^t|\tau_i=1)}{\mathbb{P}(T_i^t, C_i^t|\tau_i=0)}\right),$$

the message that $i$ sends to $u$ at step $t$.

Using the above definition, (A.5), and (A.6) we get

$$\begin{aligned}
&\log\left(\frac{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=1)}{\mathbb{P}(T_i^t, c_i, C_i^t|\tau_u=0)}\right) \\
&= \log\left(\frac{e^{\Lambda_{i\to u}^t}\frac{Kp}{d_1}(1-\alpha)+\frac{(n-K)q}{d_1}}{e^{\Lambda_{i\to u}^t}\frac{Kq}{d_0}(1-\alpha)+\frac{(n-K)q}{d_0}}\right) \\
&= \log\left(\frac{d_0}{d_1}\right)+\log\left(\frac{e^{\Lambda_{i\to u}^t}\frac{Kp}{(n-K)q}(1-\alpha)+1}{e^{\Lambda_{i\to u}^t}\frac{K}{(n-K)}(1-\alpha)+1}\right). \qquad \text{(A.7)}
\end{aligned}$$

We then use the substitution $\nu := \log((n-K)/K)$ in the above equation. Finally combining (A.3), (A.4) and (A.7) and replacing $\Lambda_u^t$ with $R_u^t$ and $\Lambda_{i\to u}^t$ with $R_{i\to u}^t$, we arrive at (5.8). The recursive equation (5.7) can be derived in exactly the same way by looking at the children of $i \in \delta u$.

## A.2 Proof of Proposition 5.1

Since the statistical properties of $R_u^t$ and $\Lambda_u^t$ are the same in the $n \to \infty$ limit, we analyze the distribution of $\Lambda_u^t$. Let us define the posterior likelihood for $\tau_u$ given by

$$\widetilde{\Lambda}_i^t = \log\left(\frac{\mathbb{P}(\tau_i=1|T_i^t, C_i^t, c_i=0)}{\mathbb{P}(\tau_i=0|T_i^t, C_i^t, c_i=0)}\right).$$

Note that $\mathbb{P}(\tau_i=1|c_i=0) = \kappa(1-\alpha)/(1-\kappa\alpha)$ and $\mathbb{P}(\tau_i=0|c_i=0) = (1-\kappa)/(1-\kappa\alpha)$ are the prior probabilities of the uncued vertices. For convenience we use an overline for the symbols of expectation $\overline{\mathbb{E}}$ and probability $\overline{\mathbb{P}}$ to denote conditioning w.r.t $\{c_i=0\}$.

By a slight abuse of notation, let $\xi_0^t$ and $\xi_1^t$ denote the rvs whose distributions are the same as the distributions of $\widetilde{\Lambda}_i^t$ given $\{c_i=0, \tau_i=0\}$ and $\{c_i=0, \tau_i=1\}$ respectively in the limit $n \to \infty$. We need a relationship between $P_0$ and $P_1$, the probability measures of $\xi_0^t$ and $\xi_1^t$ respectively, stated in the following lemma.

**Lemma A.2.**

$$\frac{dP_0}{dP_1}(\xi) = \frac{\kappa(1-\alpha)}{1-\kappa}\exp(-\xi).$$

*In other words for any integrable function $g(\cdot)$*

$$\overline{\mathbb{E}}[g(\widetilde{\Lambda}_u^t)|\tau_u=0] = \frac{\kappa(1-\alpha)}{1-\kappa}\overline{\mathbb{E}}[g(\widetilde{\Lambda}_u^t)e^{-\widetilde{\Lambda}_u^t}|\tau_u=1].$$

*Proof.* Following the logic in [Montanari 2015], we show this result for $g(\widetilde{\Lambda}_u^t) = \mathbf{1}(\widetilde{\Lambda}_u \in A)$, $A$ being some measurable set. The result for general $g$ then follows because any integrable function can be obtained as the limit of a sequence of such rvs [Billingsley 2008]. Let

$Y = (T_u^t, C_u^t)$, the observed rv. Therefore

$$\begin{aligned}
\overline{\mathbb{E}}\left(\mathbf{1}\left(\widetilde{\Lambda}_u^t \in A\right)|\tau_u = 0\right) &= \overline{\mathbb{P}}\left(\widetilde{\Lambda}_u^t \in A|\tau_u = 0\right) \\
&= \frac{\overline{\mathbb{P}}(\widetilde{\Lambda}_u^t \in A, \tau_u = 0)}{\overline{\mathbb{P}}(\tau_u = 0)} \\
&= \frac{\overline{\mathbb{E}}_Y\left(\overline{\mathbb{P}}(\widetilde{\Lambda}_u^t \in A, \tau_u = 0|Y)\right)}{\overline{\mathbb{P}}(\tau_u = 0)} \\
&= \overline{\mathbb{E}}_Y\left[\frac{\mathbf{1}(\widetilde{\Lambda}_u^t \in A)\overline{\mathbb{P}}(\tau_u = 0|Y)}{\overline{\mathbb{P}}(\tau_u = 0)}\right] \\
&\overset{(a)}{=} \overline{\mathbb{E}}_Y\left(\frac{\mathbf{1}(\widetilde{\Lambda}_u^t \in A)e^{-\widetilde{\Lambda}_u^t}\overline{\mathbb{P}}(\tau_u = 1|Y)}{\overline{\mathbb{P}}(\tau_u = 0)}\right) \\
&= \frac{\overline{\mathbb{P}}(\tau_u = 1)}{\overline{\mathbb{P}}(\tau_u = 0)}\overline{\mathbb{E}}_1(\mathbf{1}(\widetilde{\Lambda}_u^t \in A)e^{-\widetilde{\Lambda}_u^t}) \\
&= \frac{\kappa(1-\alpha)}{1-\kappa}\overline{\mathbb{E}}_1(\mathbf{1}(\widetilde{\Lambda}_u^t \in A)e^{-\widetilde{\Lambda}_u^t}),
\end{aligned}$$

where in (a) we used the fact that $\frac{\overline{\mathbb{P}}(\tau_u=0|Y)}{\overline{\mathbb{P}}(\tau_u=1|Y)} = \exp(-\widetilde{\Lambda}_u^t)$, and $\mathbb{E}_1$ denotes expectation conditioned on the event $\{\tau_u = 1\}$. □

*Proof.* Since $\lambda_\alpha$ and $\kappa$ are fixed and $b \to \infty$, from (5.12) we have

$$\rho := a/b = 1 + \sqrt{\frac{\lambda_\alpha(1-\kappa)}{(1-\alpha)^2\kappa^2 b}} = 1 + O(b^{-1/2}). \tag{A.8}$$

Following [Montanari 2015], we prove the result by induction on $t$. First let us verify the result holds when $t = 0$, for the initial condition that $\xi_0^0 = \xi_1^0 = -\upsilon$. We only do this for $\xi_0^t$ since the steps are similar for $\xi_1^t$. Observe that

$$\begin{aligned}
f(-\upsilon) &= \log\left(\frac{\frac{\kappa(1-\alpha)\rho}{(1-\kappa)} + 1}{\frac{\kappa(1-\alpha)}{(1-\kappa)} + 1}\right) \\
&= \log\left(1 + (\rho - 1)\frac{\kappa(1-\alpha)}{1-\kappa\alpha}\right) \\
&\overset{(a)}{=} (\rho - 1)\frac{\kappa(1-\alpha)}{1-\kappa\alpha} - \frac{(\rho-1)^2}{2}\frac{\kappa^2(1-\alpha)^2}{(1-\kappa\alpha)^2} + O(b^{-3/2}),
\end{aligned} \tag{A.9}$$

where $(a)$ follows from (A.8), and Taylor's expansion around $\rho = 1$. Similarly,

$$f^2(-\upsilon) = (\rho - 1)^2\frac{\kappa^2(1-\alpha)^2}{(1-\kappa\alpha)^2} + O(b^{-3/2}), \tag{A.10}$$

$$\begin{aligned}
\log(\rho) &= \log(1 + (\rho - 1)) \\
&= \sqrt{\frac{\lambda_\alpha(1-\kappa)}{(1-\alpha)^2\kappa^2 b}} - \frac{\lambda_\alpha(1-\kappa)}{2(1-\alpha)^2\kappa^2 b} + O(b^{-3/2}),
\end{aligned} \tag{A.11}$$

and

$$\log^2(\rho) = \frac{\lambda_\alpha(1-\kappa)}{(1-\alpha)^2\kappa^2 b} + O(b^{-3/2}). \tag{A.12}$$

Let us verify the induction result for $t = 0$. Using the recursion (5.9) with $\xi_0^0 = \log \frac{\kappa(1-\alpha)}{1-\kappa} = -\upsilon$, we can express $\mathbb{E}\xi_0^1$ as

$$\mathbb{E}\xi_0^1 = -\kappa b(\rho - 1) - \upsilon + \kappa b\alpha \log(\rho) + b(1-\kappa\alpha) f(-\upsilon).$$

Now using (A.9) and (A.11) we obtain

$$\begin{aligned}
\mathbb{E}\xi_0^1 &= -\kappa\sqrt{\frac{\lambda_\alpha b(1-\kappa)}{(1-\alpha)^2\kappa^2}} - \upsilon + \kappa\alpha\sqrt{\frac{\lambda_\alpha(1-\kappa)b}{(1-\alpha)^2\kappa^2}} - \frac{\lambda_\alpha(1-\kappa)\alpha}{2(1-\alpha)^2\kappa} \\
&\quad + \sqrt{\frac{\lambda_\alpha(1-\kappa)b}{(1-\alpha)^2\kappa^2}}\kappa(1-\alpha) - \frac{\lambda_\alpha(1-\kappa)}{2(1-\kappa\alpha)} + O(b^{-1/2}) \\
&= -\upsilon - \frac{\lambda_\alpha(1-\kappa)}{2(1-\alpha)^2\kappa}\alpha - \frac{\lambda_\alpha(1-\kappa)}{2(1-\kappa\alpha)} + O(b^{-1/2}).
\end{aligned} \tag{A.13}$$

We also obtain, using the formula for the variance of a Poisson random variable

$$\begin{aligned}
\mathrm{Var}\xi_0^1 &= \log^2(\rho)\kappa b\alpha + f^2(-\upsilon)(1-\kappa)b + f^2(-\upsilon)\kappa b(1-\alpha) \\
&\overset{(a)}{=} \frac{\lambda_\alpha\alpha(1-\kappa)}{(1-\alpha)^2\kappa} + \frac{(1-\kappa)\lambda_\alpha}{1-\kappa\alpha} + O(b^{-1/2}),
\end{aligned} \tag{A.14}$$

where in (a) we used (A.12) and (A.10). Comparing (A.13) and (A.14), after letting $b \to \infty$ with $\mu^{(1)}$ in (5.13) using $\mu^{(0)} = 0$, we can verify the mean and variance recursions. Next we use Lemma 5.2 to prove gaussianity. Note that we can express $\xi_0^1 - h$ as the Poisson sum of i.i.d. mixture random variables as follows

$$\xi_0^1 - h = \sum_{i=1}^{L_0} X_i,$$

where $L_0 \sim \mathrm{Poi}(b)$, and $\mathcal{L}(X_i) = \kappa\alpha\mathcal{L}(\log(\rho)) + (1-\kappa)\mathcal{L}(f(-\upsilon)) + (\kappa(1-\alpha))\mathcal{L}(f(-\upsilon))$, keeping in mind the independent splitting property of Poissons, where $\mathcal{L}$ denotes the law of a rv[1]. Next we calculate $\mathbb{E}(|X_i|^3)$. It is easy to show using (A.9) and (A.11) that

$$\mathbb{E}(|X_i|^3) = \kappa\alpha \log^3(b) + (1-\kappa\alpha)|f^3(-\upsilon)| = O(b^{-3/2}). \tag{A.15}$$

Therefore the upper bound of Lemma 5.2 with $\lambda = b$ becomes

$$\frac{C_{BE}\mathbb{E}(|X_i|^3)}{\sqrt{\gamma(\mu^2+\sigma^2)^3}} = \frac{O(b^{-3/2})}{\sqrt{b\Omega(b^{-3})}} = O(b^{-1/2}).$$

By Lemma 5.2, taking $b \to \infty$ we obtain the convergence to gaussian.

Having shown the induction hypothesis for $t = 0$, we now assume it holds for some $t > 0$. By using (5.11), (5.15) and Lebesgue's dominated convergence theorem [Billingsley 2008, Theorem 16.4] we obtain

$$\mathbb{E}f(\xi_1^t) = (\rho-1)\mathbb{E}\left(\frac{e^{\xi_1^t}}{1+e^{\xi_1^t}}\right) - \frac{(\rho-1)^2}{2}\mathbb{E}\left(\frac{e^{2\xi_1^t}}{(1+e^{\xi_1^t})^2}\right) + O(b^{-3/2}), \tag{A.16}$$

[1] Clearly $X_i$ are i.i.d. with mean $\mu = \kappa\alpha\log(\rho) + (1-\kappa\alpha)f(-\nu) = \Omega(1/\sqrt{b})$ and $\sigma^2 = \Omega(1/b)$, both of which are bounded (fixed $b$ and as $n \to \infty$). Also $\mu^2 + \sigma^2 = \Omega(1/b)$.

and by using Lemma A.2 in addition we obtain

$$\mathbb{E}f(\xi_0^t) = (\rho-1)\frac{\kappa(1-\alpha)}{1-\kappa}\mathbb{E}\left(\frac{1}{1+e^{\xi_1^t}}\right) - \frac{(\rho-1)^2\kappa(1-\alpha)}{2(1-\kappa)}\mathbb{E}\left(\frac{e^{\xi_1^t}}{(1+e^{\xi_1^t})^2}\right) + O(b^{-3/2}). \tag{A.17}$$

Now we take the expectation of both sides of (5.9) and (5.10). Using the fact that $\mathbb{E}\sum_{i=1}^{L} X_i = \mathbb{E}X_i\mathbb{E}L$ if $L \sim$ Poi and $X_i$ are independent and identically distributed (i.i.d.) rv, we obtain

$$\mathbb{E}(\xi_0^{t+1}) = h + \log\left(\frac{p}{q}\right)\kappa b\alpha + \mathbb{E}\left(f(\xi_0^t)\right)(1-\kappa)b + \mathbb{E}\left(f(\xi_1^t)\right)\kappa b(1-\alpha) \tag{A.18}$$

and

$$\mathbb{E}(\xi_1^{t+1}) = h + \log\left(\frac{p}{q}\right)\kappa a\alpha + \mathbb{E}\left(f(\xi_0^t)\right)(1-\kappa)b + \mathbb{E}\left(f(\xi_1^t)\right)\kappa a(1-\alpha). \tag{A.19}$$

We now substitute (A.17) and (A.16) in (A.18) to get:

$$\begin{aligned}\mathbb{E}(\xi_0^{t+1}) = h + \kappa b\alpha\log(\rho) + (1-\kappa)b\Bigg[&(\rho-1)\frac{\kappa(1-\alpha)}{1-\kappa}\mathbb{E}\left(\frac{1}{1+e^{\xi_1^t}}\right)\\ &- \frac{(\rho-1)^2\kappa(1-\alpha)}{2(1-\kappa)}\mathbb{E}\left(\frac{e^{\xi_1^t}}{(1+e^{\xi_1^t})^2}\right) + O(b^{-3/2})\Bigg]\\ &+ \kappa b(1-\alpha)\Bigg[(\rho-1)\mathbb{E}\left(\frac{e^{\xi_1^t}}{1+e^{\xi_1^t}}\right) - \frac{(\rho-1)^2}{2}\mathbb{E}\left(\frac{e^{2\xi_1^t}}{(1+e^{\xi_1^t})^2}\right) + O(b^{-3/2})\Bigg],\end{aligned}$$

which on simplifying and grouping like terms gives

$$\mathbb{E}(\xi_0^{t+1}) = h + \kappa b\alpha\log(\rho) + \kappa(a-b)(1-\alpha) - \frac{\lambda_\alpha(1-\kappa)}{2(1-\alpha)\kappa}\mathbb{E}\left(\frac{e^{\xi_1^t}}{1+e^{\xi_1^t}}\right) + O(b^{-1/2}).$$

Substituting $h = -\kappa(a-b) - \log\left(\frac{1-\kappa}{\kappa(1-\alpha)}\right)$, we get

$$\mathbb{E}(\xi_0^{t+1}) = -\log\left(\frac{1-\kappa}{\kappa(1-\alpha)}\right) - \alpha\kappa(a-b) + \kappa b\alpha\log(\rho) - \frac{\lambda_\alpha(1-\kappa)}{2\kappa(1-\alpha)}\mathbb{E}\left(\frac{e^{\xi_1^t}}{1+e^{\xi_1^t}}\right) + O(b^{-1/2}).$$

Using (A.11) we get

$$\begin{aligned}-\alpha\kappa(a-b) + \kappa b\alpha\log(\rho) &= \kappa b\alpha(\log(\rho) - (\rho-1))\\ &= \kappa b\alpha\left(-\frac{\lambda_\alpha(1-\kappa)}{2\kappa^2 b(1-\alpha)^2} + O(b^{-3/2})\right)\\ &= -\frac{\lambda_\alpha\alpha(1-\kappa)}{2(1-\alpha)^2\kappa} + O(b^{-1/2}).\end{aligned}$$

Finally we obtain

$$\mathbb{E}(\xi_0^{t+1}) = -\log\left(\frac{1-\kappa}{\kappa(1-\alpha)}\right) - \frac{\lambda_\alpha\alpha(1-\kappa)}{2(1-\alpha)^2\kappa} - \lambda_\alpha\frac{(1-\kappa)}{2(1-\alpha)\kappa}\mathbb{E}\left(\frac{e^{\xi_1^t}}{1+e^{\xi_1^t}}\right) + O(b^{-1/2}). \tag{A.20}$$

Using exactly the same simplifications we can get

$$\mathbb{E}(\xi_1^{t+1}) = -\log\left(\frac{1-\kappa}{\kappa(1-\alpha)}\right) + \frac{\alpha\lambda_\alpha(1-\kappa)}{2\kappa(1-\alpha)^2} + \frac{\lambda_\alpha(1-\kappa)}{2\kappa(1-\alpha)}\mathbb{E}\left(\frac{e^{\xi_1^t}}{1+e^{\xi_1^t}}\right) + O(b^{-1/2}). \quad \text{(A.21)}$$

Our next goals are to compute $\text{var}(\xi_0^{t+1})$ and $\text{var}(\xi_1^{t+1})$. Towards this, observe that $f^2(x) = (\rho-1)^2\left(\frac{e^x}{1+e^x}\right)^2 + O(b^{-3/2})$. Therefore

$$\mathbb{E}(f^2(\xi_0^t)) = (\rho-1)^2\mathbb{E}\left(\frac{e^{2\xi_0^t}}{(1+e^{\xi_0^t})^2}\right) + O(b^{-3/2}),$$

and using Lemma A.2 the above becomes

$$\mathbb{E}(f^2(\xi_0^t)) = (\rho-1)^2\frac{\kappa(1-\alpha)}{1-\kappa}\mathbb{E}\left(\frac{e^{\xi_1^t}}{(1+e^{\xi_1^t})^2}\right) + O(b^{-3/2}). \quad \text{(A.22)}$$

Similarly,

$$\mathbb{E}\left(f^2(\xi_1^t)\right) = (\rho-1)^2\mathbb{E}\left(\frac{e^{2\xi_1^t}}{(1+e^{\xi_1^t})^2}\right) + O(b^{-3/2}). \quad \text{(A.23)}$$

Now we use the formula for the variance of Poisson sums $\text{Var}\sum_{i=1}^{L} X_i = \mathbb{E}(X_i^2)\mathbb{E}(L)$ to get

$$\text{Var}(\xi_0^{t+1}) = \log^2(\rho)\kappa b\alpha + (1-\kappa)b\mathbb{E}(f^2(\xi_0^t)) + \kappa b(1-\alpha)\mathbb{E}(f^2(\xi_1^t))$$

$$\text{Var}(\xi_1^{t+1}) = \log^2(\rho)\kappa a\alpha + (1-\kappa)b\mathbb{E}(f^2(\xi_0^t)) + \kappa a(1-\alpha)\mathbb{E}(f^2(\xi_1^t)).$$

Substituting (A.22) and (A.23) into the above equations and letting $b \to \infty$, we get

$$\lim_{b\to\infty}\text{Var}(\xi_1^{t+1}) = \lim_{b\to\infty}\text{Var}(\xi_0^{t+1}) = \mu^{(t+1)},$$

where

$$\mu^{(t+1)} = \frac{\lambda_\alpha\alpha(1-\kappa)}{\kappa(1-\alpha)^2} + \frac{\lambda_\alpha(1-\kappa)}{\kappa(1-\alpha)}\mathbb{E}\left(\frac{\exp\xi_1^t}{1+\exp(\xi_1^t)}\right). \quad \text{(A.24)}$$

Using $\mu^{(t+1)}$ of (A.24) in (A.20) and (A.21) we get

$$\begin{aligned}\mathbb{E}(\xi_0^{t+1}) &= -\log\left(\frac{(1-\kappa)}{\kappa(1-\alpha)}\right) - \frac{1}{2}\mu^{(t+1)} + O(b^{-1/2})\\ \mathbb{E}(\xi_1^{t+1}) &= -\log\left(\frac{(1-\kappa)}{\kappa(1-\alpha)}\right) + \frac{1}{2}\mu^{(t+1)} + O(b^{-1/2}).\end{aligned} \quad \text{(A.25)}$$

Now we use the fact the induction assumption that $\xi_1^t \to \mathcal{N}(\mathbb{E}(\xi_1^t), \mu^{(t)})$. Since the function $e^{\xi_1^t}/(1+e^{\xi_1^t})$ is bounded, by Lebesgue's dominated convergence theorem [Billingsley 2008, Theorem 16.4] this means $\mathbb{E}(1/(1+e^{-\xi_1^t})) \to \mathbb{E}(1/(1+e^{-\mathcal{N}(\mathbb{E}(\xi_1^t),\mu^{(t)})}))$ as $b \to \infty$.We can write $\mathcal{N}(\mathbb{E}(\xi_1^t), \mu^{(t)}) = \sqrt{\mu^{(t)}}Z + \mathbb{E}(\xi_1^t)$, where $Z \sim \mathcal{N}(0,1)$. Therefore we obtain

$$\begin{aligned}\mathbb{E}\left(\frac{1}{1+e^{-\xi_1^t}}\right) &= \mathbb{E}\left(\frac{1}{1+e^{-\sqrt{\mu^{(t)}}Z}\frac{(1-\kappa)}{\kappa(1-\alpha)}e^{-\frac{\mu^{(t)}}{2}}}\right)\\ &= \mathbb{E}\left(\frac{\kappa(1-\alpha)}{\kappa(1-\alpha)+(1-\kappa)e^{(-\sqrt{\mu^t}Z-\frac{\mu^{(t)}}{2})}}\right).\end{aligned}$$

Substituting the above into (A.24) gives us the recursion for $\mu^{(t+1)}$ given in (5.13).

Next we prove gaussianity. Consider

$$\xi_0^{t+1} - \mathbb{E}(\xi_0^{t+1}) = \log\left(\frac{p}{q}\right)(L_{0c} - \mathbb{E}(L_{0c})) + \sum_{i=1}^{L_{00}}(f(\xi_{0,i}^t) - \mathbb{E}(f(\xi_0^t)))+ \sum_{i=1}^{L_{01}}(f(\xi_{1,i}^t) - \mathbb{E}(f(\xi_1^t))) + (L_{00} - \mathbb{E}(L_{00}))\mathbb{E}(f(\xi_0^t)) + (L_{01} - \mathbb{E}(L_{01}))\mathbb{E}(f(\xi_1^t)). \tag{A.26}$$

Let us look at the second term. Let $X_i = f(\xi_{0,i}^t) - \mathbb{E}f(\xi_{0,i}^t)$. Then it can be shown that $\mathbb{E}X_i^2 = O(1/b)$. Let $D := \sum_{i=1}^{L_{00}} X_i - \sum_{i=1}^{\mathbb{E}L_{00}} X_i$. In the second term the summation is taken up to $i \leq \mathbb{E}L_{00}$. Then $\mathbb{E}(D^2) = |\sum_{i=1}^{\delta} X_i|^2$, where $\delta \leq |L_{00} - \mathbb{E}L_{00}|+1$, where the extra 1 is because $\mathbb{E}L_{00}$ may not be an integer. Therefore $\mathbb{E}D^2 = \mathbb{E}\delta\mathbb{E}|X_1|^2 \leq (C/b)((1-\kappa)b+1)^{1/2} = O(1/\sqrt{b})$. Thus, we can replace the Poisson upper limits of the summations in the second and third terms of (A.26) by their means, leading to

$$\begin{aligned}\xi_0^{t+1} - \mathbb{E}(\xi_0^{t+1}) = \log\left(\frac{p}{q}\right)(L_{0c} - \mathbb{E}(L_{0c})) + \sum_{i=1}^{\mathbb{E}(L_{00})}(f(\xi_{0,i}^t) - \mathbb{E}(f(\xi_0^t))) \\ + \sum_{i=1}^{\mathbb{E}(L_{01})}(f(\xi_{1,i}^t) - \mathbb{E}(f(\xi_1^t))) + (L_{00} - \mathbb{E}(L_{00}))\mathbb{E}f(\xi_0^t) \\ + (L_{01} - \mathbb{E}(L_{01}))\mathbb{E}(f(\xi_1^t)) + o_p(1),\end{aligned} \tag{A.27}$$

where $o_p(1)$ indicates a rv that goes to zero in probability in the limit. The combined variance of all other terms approaches $\mu^{(t+1)}$, defined in (5.13), as $b \to \infty$ and it is finite for a fixed $t$. Now since we have an infinite sum of independent rvs as $a, b \to \infty$, with zero mean and finite variance, from the standard CLT, we can conclude that the distribution tends to $\mathcal{N}(0, \mu^{t+1})$. The argument for $\xi_1^{t+1}$ is identical. □

## A.3 Finishing the proof of Theorem 5.1

*Proof.* We bound $\mathbb{E}(|\overline{S}\Delta\widehat{S}_0|)/(K(1-\alpha))$ as follows:

$$\begin{aligned}\lim_{b\to\infty}\lim_{n\to\infty}\frac{\mathbb{E}(|\overline{S}\Delta\widehat{S}_0|)}{K(1-\alpha)} &= \lim_{b\to\infty}\lim_{n\to\infty}\left(\frac{\mathbb{E}\left(\sum_{i=1}^n \mathbf{1}_{\sigma_i\neq\widehat{\sigma}_i}\right)}{K - K\alpha}\right) \\ &\leq \lim_{b\to\infty}\left(\frac{(1-\kappa)}{\kappa(1-\alpha)}\mathbb{P}(\xi_0^t \geq 0) + \mathbb{P}(\xi_1^t \leq 0)\right),\end{aligned} \tag{A.28}$$

since

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^n \mathbf{1}_{\sigma_i\neq\widehat{\sigma}_i}\right) = n(\mathbb{P}(c_i = 0, \sigma_i = 0)\mathbb{P}(R_i^t > \upsilon | c_i = 0, \sigma_i = 0)+ \\ \mathbb{P}(c_i = 0, \sigma_i = 1)\mathbb{P}(R_i^t < \upsilon | c_i = 0, \sigma_i = 1)),\end{aligned} \tag{A.29}$$

and since $R_i^t - R_{i\to u}^t = O(b^{-1/2})$. Indeed, given the $b \to \infty$ limit in (A.28), the bound $O(b^{-1/2})$ allows us to replace $R_i^t$ in (A.29) by the distribution limit when $n \to \infty$, which is $\xi_0^t$ or $\xi_1^t$ when conditioned on $\{\sigma_i = 0\}$ or $\{\sigma_i = 1\}$ respectively, for an arbitrary $i$. We now analyze each term in (A.28) separately. By Proposition 5.1 we have

$$\lim_{b\to\infty}\mathbb{P}(\xi_1^t \leq 0) = Q\left(\frac{1}{\sqrt{\mu^{(t)}}}\left(\frac{\mu^{(t)}}{2} - \log\frac{(1-\kappa)}{\kappa(1-\alpha)}\right)\right)$$

where $Q(\cdot)$ denotes the standard $Q$ function. Notice that by (5.13) we have that $\mu^{(t)} \geq \lambda_\alpha \alpha(1-\kappa)/(\kappa(1-\alpha)^2)$, since $\mathbb{E}\left(\frac{1-\kappa}{\kappa(1-\alpha)+(1-\kappa)\exp(-\mu/2-\sqrt{\mu}Z)}\right) \geq 0$. In addition, by (A.24), $\mu^{(t)} \leq \frac{\lambda_\alpha(1-\kappa)}{\kappa(1-\alpha)^2}$. Note that the lower bound on $\mu^{(t)}$ is not useful when $\alpha = 0$. Therefore by using the Chernoff bound for the $Q$ function, $Q(x) \leq \frac{1}{2}e^{-x^2/2}$, we get

$$\begin{aligned}
\lim_{b\to\infty} \mathbb{P}(\xi_1^t \leq 0) &\leq \frac{1}{2} e^{-\frac{1}{2\mu^{(t)}}\left(\frac{\mu^{(t)}}{2} - \log\left(\frac{1-\kappa}{\kappa(1-\alpha)}\right)\right)^2} \\
&= \frac{1}{2} e^{-\frac{\mu^{(t)}}{8}\left(1 - \frac{2}{\mu^{(t)}}\log\left(\frac{1-\kappa}{\kappa(1-\alpha)}\right)\right)^2} \\
&\leq \frac{1}{2} e^{-\frac{\mu^{(t)}}{8}} e^{\frac{1}{2}\log\left(\frac{1-\kappa}{\kappa(1-\alpha)}\right)} \\
&= \frac{1}{2}\sqrt{\frac{1-\kappa}{\kappa(1-\alpha)}} e^{-\frac{\mu^{(t)}}{8}},
\end{aligned} \tag{A.30}$$

where we used the fact that $(1-x)^2 \geq 1-2x$ for any $x > 0$. By employing similar reductions, we can show

$$\lim_{b\to\infty} \left(\frac{(1-\kappa)}{\kappa(1-\alpha)}\right) \mathbb{P}(\xi_0^t \geq 0) \leq \frac{1}{2}\sqrt{\frac{1-\kappa}{\kappa(1-\alpha)}} e^{-\frac{\mu^{(t)}}{8}}. \tag{A.31}$$

Substituting (A.38) and (A.39) back in (A.28) and using the fact that $\mu^{(t)} \geq \lambda_\alpha \alpha(1-\kappa)/(\kappa(1-\alpha)^2)$, we get

$$\lim_{b\to\infty}\lim_{n\to\infty} \frac{\mathbb{E}(|\overline{S}\Delta\widehat{S}_0|)}{K(1-\alpha)} \leq \sqrt{\frac{1-\kappa}{\kappa(1-\alpha)}} e^{-\frac{\lambda_\alpha \alpha(1-\kappa)}{8\kappa(1-\alpha)^2}}$$

Then using (5.22) we get the desired result in (5.21) . □

## A.4 Proof of Proposition 5.3

*Proof.* We derive the conditional distributions of the messages $R_{u\to i}^t$ for a finite $t$ given $\{\sigma_u = 0\}$ and given $\{\sigma_u = 1\}$. In this limit the tree coupling of Lemma A.1 holds with a slightly modified construction of the tree to accomodate the difference in the generation of cued nodes. It is similar to the tree coupling in Lemma A.1, with the only difference being the generation of cues. At any level of the tree, a node $u$ is labelled a cue such that $\mathbb{P}(c_u = 1|\tau_u = 1) = \alpha\beta$ and $\mathbb{P}(c_u = 1|\tau_u = 0) = \kappa\alpha(1-\beta)/(1-\kappa)$, so that the equalities in (5.2) and (5.4) hold, where $c_u$ denotes the cue membership of node $u$ on the tree. Let $F_{u\to i}^t$ be such that $R_{u\to i}^t = F_{u\to i}^t + h_u$, for any two neighbouring nodes $i$ and $u$. Then, it can be seen from (5.25) that $F_{u\to i}^t$ satisfies the following recursion

$$F_{u\to i}^{t+1} = -\kappa(a-b) + \sum_{l\in\delta u, l\neq i} f_{\text{isi}}(F_l^t + h_l), \tag{A.32}$$

where $f_{\text{isi}}(x) := \log\left(\frac{e^{(x-\nu)}\rho+1}{e^{(x-\nu)}+1}\right)$. Let $\Psi_0^t, \Psi_1^t$ be the rvs that have the conditional asymptotic distribution of $F_{u\to i}^t$ given $\{\sigma_u = 0\}$ and $\{\sigma_u = 1\}$ respectively in the limit $n \to \infty$. Then, by studying the recursion (A.32) on the tree we can conclude that $\Psi_0^t, \Psi_1^t$ satisfy the following

recursive distributional equations

$$\Psi_0^{t+1} \overset{\mathrm{D}}{=} -\kappa(a-b) + \sum_{i=0}^{L_{01c}} f_{\mathrm{isi}}(\Psi_{1i}^t + B_c) + \sum_{i=0}^{L_{01n}} f_{\mathrm{isi}}(\Psi_{1i}^t + B_n) + \sum_{i=0}^{L_{00c}} f_{\mathrm{isi}}(\Psi_{0i}^t + B_c) + \sum_{i=0}^{L_{00n}} f_{\mathrm{isi}}(\Psi_{0i}^t + B_n), \tag{A.33}$$

$$\Psi_1^{t+1} \overset{\mathrm{D}}{=} -\kappa(a-b) + \sum_{i=0}^{L_{11c}} f_{\mathrm{isi}}(\Psi_{1i}^t + B_c) + \sum_{i=0}^{L_{11n}} f_{\mathrm{isi}}(\Psi_{1i}^t + B_n) + \sum_{i=0}^{L_{10c}} f_{\mathrm{isi}}(\Psi_{0i}^t + B_c) + \sum_{i=0}^{L_{10n}} f_{\mathrm{isi}}(\Psi_{0i}^t + B_n), \tag{A.34}$$

where $\overset{\mathrm{D}}{=}$ represents equality in distribution, and the random sums are such that $L_{01c} \sim \mathrm{Poi}(\kappa b\alpha\beta), L_{01n} \sim \mathrm{Poi}(\kappa b(1-\alpha\beta)), L_{00c} \sim \mathrm{Poi}(\kappa b\alpha(1-\beta)), L_{00n} \sim \mathrm{Poi}(b(1-\kappa-\kappa\alpha(1-\beta)), L_{11c} \sim \mathrm{Poi}(\kappa a\alpha\beta), L_{11n} \sim \mathrm{Poi}(\kappa a(1-\alpha\beta)), L_{10c} \sim \mathrm{Poi}(\kappa b\alpha(1-\beta))$, and $L_{10n} \sim \mathrm{Poi}(b(1-\kappa-\kappa\alpha(1-\beta)))$, $B_c = \log\left(\frac{\beta(1-\kappa)}{(1-\beta)\kappa}\right)$; $B_n = \log\left(\frac{(1-\alpha\beta)(1-\kappa)}{(1-\kappa-\alpha\kappa+\alpha\kappa\beta)}\right)$; and $\Psi_{0,i}^t$ and $\Psi_{1,i}^t$ are i.i.d. rvs with the same distribution as $\Psi_0^t$ and $\Psi_1^t$ respectively.

We now derive the asymptotic distributions $\Psi_0^{t+1}$ and $\Psi_1^{t+1}$ when $a, b \to \infty$ such that $\lambda = \frac{\kappa^2(a-b)^2}{(1-\kappa)b}$ and $\kappa$ are fixed. Observe that $\rho = 1 + \sqrt{\frac{\lambda(1-\kappa)}{\kappa^2 b}} = 1 + \sqrt{\frac{r}{b}}$, where $r := \frac{\lambda(1-\kappa)}{\kappa^2}$. Notice that if $P_0 \sim \mathcal{L}(\Psi_0^t)$ and $P_1 = \mathcal{L}(\Psi_1^t)$, we have, since $\Psi = \log\left(\frac{dP_1}{dP_0}\right)$, that $\frac{dP_0}{dP_1}(\Psi) = e^{-\Psi}$. Also

$$f_{\mathrm{isi}}(x) = \log\left(1 + (\rho-1)\frac{e^{x-\nu}}{1+e^{x-\nu}}\right) \tag{A.35}$$

$$= \sqrt{\frac{r}{b}}\left(\frac{e^{x-\nu}}{1+e^{x-\nu}}\right) - \frac{1}{2}\frac{r}{b}\left(\frac{e^{x-\nu}}{1+e^{x-\nu}}\right)^2 + O(b^{-3/2}), \tag{A.36}$$

and

$$f_{\mathrm{isi}}^2(x) = \frac{r}{b}\left(\frac{e^{x-\nu}}{1+e^{x-\nu}}\right)^2 + O(b^{-3/2}). \tag{A.37}$$

Now we can reformulate the recursions in (A.33) and (A.34) as a Poisson sum as follows:

$$\Psi_0^{t+1} \overset{\mathrm{D}}{=} -\kappa(a-b) + \sum_{l=1}^{L_0} X_l \tag{A.38}$$

$$\Psi_1^{t+1} \overset{\mathrm{D}}{=} -\kappa(a-b) + \sum_{l=1}^{L_1} Y_l, \tag{A.39}$$

where $L_0 = \mathrm{Poi}(b), L_1 = \mathrm{Poi}(\kappa a + (1-\kappa)b)$ and $X_l$ and $Y_l$ are mixture rvs with laws defined as follows:

$$\mathcal{L}(X_l) = \alpha\kappa(1-\beta)\mathcal{L}(f_{\mathrm{isi}}(\Psi_0^t + B_c)) + (1-\kappa)(1-\alpha(1-\beta)e^{-\nu})\mathcal{L}(f_{\mathrm{isi}}(\Psi_0^t + B_n)) + \alpha\kappa\beta\mathcal{L}(f_{\mathrm{isi}}(\Psi_1^t + B_c)) + \kappa(1-\alpha\beta)\mathcal{L}(f_{\mathrm{isi}}(\Psi_1^t + B_n)),$$

$$\mathcal{L}(Y_l) = \frac{\alpha\kappa b(1-\beta)}{\kappa a + (1-\kappa)b}\mathcal{L}(f_{\mathrm{isi}}(\Psi_0^t + B_c)) + \frac{(1-\kappa)b(1-\alpha(1-\beta)e^{-\nu})}{\kappa a + (1-\kappa)b}\mathcal{L}(f_{\mathrm{isi}}(\Psi_0^t + B_n)) + \frac{\kappa a\alpha\beta}{\kappa a + (1-\kappa)b}\mathcal{L}(f_{\mathrm{isi}}(\Psi_1^t + B_c)) + \frac{\kappa a(1-\alpha\beta)}{\kappa a + (1-\kappa)b}\mathcal{L}(f_{\mathrm{isi}}(\Psi_1^t + B_n)).$$

Observe that we have $B_c-\nu = \log(\frac{\beta}{1-\beta})$ and $B_n-\nu = \log(\frac{\kappa(1-\alpha\beta)}{(1-\kappa-\alpha\kappa(1-\beta))})$. We can calculate $\mathbb{E}(X_l)$ as

$$\mathbb{E}(X_l) = \alpha\kappa\beta\sqrt{\frac{r}{b}} + \kappa(1-\alpha\beta)\sqrt{\frac{r}{b}} - \alpha\kappa\beta\frac{r}{2b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) - \kappa(1-\alpha\beta)\frac{r}{2b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) + O(b^{-3/2}),$$

which gives,

$$\mathbb{E}(X_l) = \kappa\sqrt{\frac{r}{b}} - \alpha\kappa\beta\frac{r}{2b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) - \kappa(1-\alpha\beta)\frac{r}{2b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) + O(b^{-3/2}).$$

Similarly

$$\mathbb{E}(X_l^2) = \alpha\kappa\beta\frac{r}{b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) + \frac{r\kappa(1-\alpha\beta)}{b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) + O(b^{-3/2}),$$

and

$$\mathbb{E}(|X_l^3|) = \alpha\kappa\beta(\frac{r}{b})^{3/2}\mathbb{E}\left(\frac{e^{2(\Psi_1^t+B_c-\nu)}}{(1+e^{\Psi_1^t+B_c-\nu)^2}}\right) + \frac{\kappa(1-\alpha\beta)r^{3/2}}{b^{3/2}}\mathbb{E}\left(\frac{e^{2(\Psi_1^t+B_n-\nu)}}{(1+e^{\Psi_1^t+B_n-\nu)^2}}\right) + O(b^{-2}). \tag{A.40}$$

Similarly we can calculate the moments of $Y_l$ as follows:

$$\begin{aligned}\mathbb{E}(Y_l) = {} & \frac{\alpha\kappa b\beta}{\kappa a+(1-\kappa)b}\sqrt{\frac{r}{b}}\mathbb{E}\left(\frac{1+\rho e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) \\ & - \frac{r\alpha\kappa\beta}{2(\kappa a+(1-\kappa)b)}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}(1+\rho e^{\Psi_1^t+B_c-\nu})}{(1+e^{\Psi_1^t+B_c-\nu})^2}\right) \\ & + \frac{\kappa b(1-\alpha\beta)}{\kappa a+(1-\kappa)b}\sqrt{\frac{r}{b}}\mathbb{E}\left(\frac{1+\rho e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) \\ & - \frac{r\kappa(1-\alpha\beta)}{2(\kappa a+(1-\kappa b))}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}(1+\rho e^{\Psi_1^t+B_n-\nu})}{(1+e^{\Psi_1^t+B_n-\nu})^2}\right) + O(b^{-3/2}),\end{aligned}$$

giving

$$\begin{aligned}\mathbb{E}(Y_l) = {} & \kappa\sqrt{rb}\frac{1}{\kappa a+(1-\kappa)b} + \frac{r\alpha\beta\kappa}{2(\kappa a+(1-\kappa)b)}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) \\ & + \frac{r\kappa(1-\alpha\beta)}{2(\kappa a+(1-\kappa)b)}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) + O(b^{-3/2}).\end{aligned}$$

In addition,

$$\mathbb{E}(Y_l^2) = \frac{\alpha\kappa\beta r}{\kappa a+(1-\kappa)b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) + \frac{\kappa r(1-\alpha\beta)}{\kappa a+(1-\kappa)b}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) + O(b^{-2})$$

and

$$\mathbb{E}(|Y_l^3|) = \frac{\alpha\kappa\beta r^{3/2}}{(\kappa a + (1-\kappa)b)b^{1/2}}\mathbb{E}\left(\frac{e^{2\Psi_1^t+2B_c-2\nu}}{(1+e^{\Psi_1^t+B_c-\nu})^2}\right) \tag{A.41}$$

$$+ \frac{\kappa(1-\alpha\beta)r^{3/2}}{(\kappa a + (1-\kappa)b)b^{1/2}}\mathbb{E}\left(\frac{e^{2\Psi_1^t+2B_n-2\nu}}{(1+e^{\Psi_1^t+2B_n-\nu})^2}\right)$$

$$+ O(b^{-2}). \tag{A.42}$$

Let us define $\mu^{(t)}$ as

$$\mu^{(t+1)} = \alpha\beta\kappa r\mathbb{E}\left(\frac{1}{1+e^{-\Psi_1^t-B_c+\nu}}\right) + \kappa r(1-\alpha\beta)\mathbb{E}\left(\frac{1}{1+e^{-\Psi_1^t-B_n+\nu}}\right). \tag{A.43}$$

Finally we have

$$\begin{aligned}\mathbb{E}(\Psi_0^{t+1}) &= -\kappa(a-b) + b\mathbb{E}(X_l)\\ &= -\frac{\alpha\kappa\beta r}{2}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) - \frac{\kappa(1-\alpha\beta)r}{2}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) + O(b^{-1/2})\\ &= -\frac{\mu^{t+1}}{2} + O(b^{-1/2}),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(\Psi_1^{t+1}) &= -\kappa(a-b) + (\kappa a + (1-\kappa)b)\mathbb{E}(Y_l)\\ &= \frac{\alpha\kappa\beta r}{2}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) + \frac{\kappa(1-\alpha\beta)r}{2}\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_1^t+B_n-\nu}}\right) + O(b^{-1/2})\\ &= \frac{\mu^{(t+1)}}{2} + O(b^{-1/2}).\end{aligned}$$

In addition, for the variances of $\Psi_0^{t+1}$ and $\Psi_1^{t+1}$ we have

$$\begin{aligned}\mathrm{Var}(\Psi_0^{t+1}) &= b\mathbb{E}(X_l^2)\\ &= \alpha\beta\kappa r\mathbb{E}\left(\frac{e^{\Psi_1^t+B_c-\nu}}{1+e^{\Psi_1^t+B_c-\nu}}\right) + \kappa(1-\alpha\beta)r\mathbb{E}\left(\frac{e^{\Psi_1^t+B_n-\nu}}{1+e^{\Psi_i^t+B_n-\nu}}\right) + O(b^{-1/2})\\ &= \mu^{(t+1)} + O(b^{-1/2}),\end{aligned} \tag{A.44}$$

and similarly

$$\mathrm{Var}(\Psi_1^{t+1}) = (\kappa a + (1-\kappa)b)\mathbb{E}(Y_l^2) \tag{A.45}$$

$$= \mu^{(t+1)} + O(b^{-1/2}). \tag{A.46}$$

Now we need to show the gaussianity of the messages $\Psi_0^t$ and $\Psi_1^t$, which we show using Lemma 5.2. For (A.38) the upperbound in Lemma 5.2 becomes

$$\begin{aligned}\frac{C_{BE}\mathbb{E}(|X_i|^3)}{\sqrt{\gamma(\mu^2+\sigma^2)^3}} &= \frac{C_{BE}b\mathbb{E}(|X_i|^3)}{\sqrt{(b(\mu^2+\sigma^2))^3}}\\ &= \frac{C_{BE}b\mathbb{E}(|X_i|^3)}{\mathrm{Var}(\Psi_0^{t+1})^{3/2}}\end{aligned} \tag{A.47}$$

Similarly for (A.39) we get

$$\frac{C_{BE}\mathbb{E}(|Y_i|^3)}{\sqrt{\gamma(\mu^2+\sigma^2)^3}} = \frac{C_{BE}(\kappa a+(1-\kappa)b)\mathbb{E}(|Y_i|^3)}{\mathrm{Var}(\Psi_1^{t+1})^{3/2}}. \tag{A.48}$$

In Lemma A.4 stated and proved below, we show that $\mu^{(t+1)} \geq \alpha\beta^2\lambda\frac{1-\kappa}{\kappa}$. Therefore for any $\kappa < 1/2$, we have

$$\mathrm{Var}(\Psi_0^{t+1}) = \mathrm{Var}(\Psi_1^{t+1}) \geq \frac{\alpha\beta^2\lambda}{2} + O(b^{-1/2}) = \Theta(1),$$

under the assumptions of the proposition. In addition we have $b\mathbb{E}(|X_i|^3) = O(b^{-1/2})$ and $(\kappa a+(1-\kappa)b)\mathbb{E}(|Y_i|^3) = O(b^{-1/2})$ from (A.40) and (A.42). Thus the bounds given in (A.47) and (A.48) both tend to zero as $b \to \infty$.

Hence by Lemma 5.2, we obtain that $\Psi_1^t \to \mathcal{N}(\frac{\mu^{(t)}}{2}, \mu^{(t)})$ and $\Psi_0^t \to \mathcal{N}(-\frac{\mu^{(t)}}{2}, \mu^{(t)})$ as $b \to \infty$, where from (A.43), $\mu^{(t)}$ satisfies the following recursion with inital condition $\mu^{(0)} = 0$ :

$$\begin{aligned}\mu^{(t+1)} = \alpha\beta\lambda\mathbb{E}\left(\frac{(1-\kappa)}{\kappa+(1-\kappa)e^{-\sqrt{\mu^{(t)}}Z-\frac{\mu^{(t)}}{2}-B_c}}\right)\\ +(1-\alpha\beta)\lambda\mathbb{E}\left(\frac{(1-\kappa)}{\kappa+(1-\kappa)e^{-\sqrt{\mu^{(t)}}Z-\frac{\mu^{(t)}}{2}-B_n}}\right).\end{aligned} \tag{A.49}$$

Consequently, the distributions of the messages $R^t_{u\to i}$ in the limit of $n \to \infty$ converge to $\Gamma^t_j + h_u$, given $\{\sigma_u = j\}$, where $\Gamma^t_1 \sim \mathcal{N}(\frac{\mu^{(t)}}{2}, \mu^{(t)})$ and $\Gamma^t_0 \sim \mathcal{N}(-\frac{\mu^{(t)}}{2}, \mu^{(t)})$, in the large degree limit where $b \to \infty$. □

### A.4.1 Proving the bound on $\mu^{(t)}$

Let $F(\mu)$ be defined as

$$\begin{aligned}F(\mu) = \alpha\beta^2\lambda\mathbb{E}\left(\frac{(1-\kappa)/\kappa}{\beta+(1-\beta)\exp(-\mu/2-\sqrt{\mu}Z)}\right)\\ +(1-\alpha\beta)^2\lambda\mathbb{E}\left(\frac{(1-\kappa)}{\kappa(1-\alpha\beta)+(1-\kappa-\alpha\kappa+\alpha\kappa\beta)e^{(-\mu/2-\sqrt{\mu}Z)}}\right).\end{aligned}$$

Then $\mu^{(t)}$ satisfies the recursion $\mu^{(t+1)} = F(\mu^{(t)})$, by substituting for $B_c$ and $B_n$ in (A.49). Below we show a lower bound on $F(\mu)$. For its proof we need the following Lemma from [Alon & Spencer 2004].

**Lemma A.3.** *[Alon & Spencer 2004, Theorem 6.2.1] If $f, g : \mathbb{R} \to \mathbb{R}$ are two non-decreasing functions, then $\mathbb{E}(fg) \geq \mathbb{E}(f)\mathbb{E}(g)$.*

Now we state our result on $F(\mu)$.

**Lemma A.4.** *For $0 < \beta < 1$,*

$$F(\mu) \geq \alpha\beta^2\lambda\frac{1-\kappa}{\kappa}.$$

*Proof.* We show that

$$g_\beta = \mathbb{E}\left(\frac{1}{\beta+(1-\beta)\exp(-\mu/2-\sqrt{\mu}Z)}\right)$$

is nonincreasing for $0 \leq \beta \leq 1$ as shown below. Let $X = \exp(-\sqrt{\mu}Z)$. Then $\frac{d}{d\beta}(g_\beta) = \mathbb{E}\left(\frac{\exp(-\mu/2)X-1}{(\beta+(1-\beta)e^{-\mu/2}X)^2}\right)$. Now we show $\frac{d}{d\beta}(g_\beta) < 0$ using Lemma A.3. In Lemma A.3, let $f = \exp(-\mu/2)X$ and $g = \frac{-1}{(\beta+(1-\beta)e^{-\mu/2}X)^2}$. Clearly these are non-decreasing in $X$. Therefore $\mathbb{E}(fg) \geq \mathbb{E}(f)\mathbb{E}(g) = \mathbb{E}(g)$, since $\mathbb{E}(f) = \mathbb{E}e^{-\mu/2}e^{\sqrt{\mu}Z} = 1$. Therefore we have

$$\mathbb{E}\left(\frac{-e^{-\mu/2}X}{(\beta+(1-\beta)e^{-\mu/2}X)^2}\right) \geq \mathbb{E}\left(\frac{-1}{(\beta+(1-\beta)e^{-\mu/2}X)^2}\right),$$

hence $\frac{dg_\beta}{d\beta} < 0$ for all $\beta$. Therefore $1 = g_\beta(1) \leq g_\beta(\beta)$ for $\beta < 1$. The result then follows by substituting this lower bound in the definition of $F(\mu)$ and observing that the second term is strictly non-negative. □

### A.4.2 Proof of Theorem 5.2

*Proof.* Notice that when we set $\beta = 1$ the recursion (5.28) becomes the same as (5.13). Also, when $\beta = 0$ we can retrieve the recursion for standard BP without side-information, i.e., and from this it can be gleaned that the asymptotic error rate is zero only if $\lambda > 1/e$.

Let us now consider $0 < \beta < 1$. By Lemma A.4, we have

$$\alpha\beta^2\lambda\frac{1-\kappa}{\kappa} \leq \mu^{(t)} \leq \lambda\frac{(1-\kappa)}{\kappa}.$$

Hence $\mu^{(t)} = \Theta\left(\frac{1-\kappa}{\kappa}\right)$. The asymptotic distributions of the messages are as follows:

$$\begin{aligned}
\Gamma^t_{0,0} &\sim \mathcal{N}(-\mu^{(t)}/2, \mu^{(t)}) + \log\left(\frac{(1-\alpha\beta)(1-\kappa)}{(1-\kappa-\alpha\kappa+\alpha\kappa\beta)}\right)\\
\Gamma^t_{0,1} &\sim \mathcal{N}(-\mu^{(t)}/2, \mu^{(t)}) + \log\left(\frac{\beta(1-\kappa)}{\kappa(1-\beta)}\right)\\
\Gamma^t_{1,0} &\sim \mathcal{N}(\mu^{(t)}/2, \mu^{(t)}) + \log\left(\frac{(1-\alpha\beta)(1-\kappa)}{(1-\kappa-\alpha\kappa+\alpha\kappa\beta)}\right)\\
\Gamma^t_{1,1} &\sim \mathcal{N}(\mu^{(t)}/2, \mu^{(t)}) + \log\left(\frac{\beta(1-\kappa)}{\kappa(1-\beta)}\right),
\end{aligned}$$

where $\Gamma^t_{j,k}$ is the rv with the asymptotic distribution of the messages $R^t_{u\to i}$ in the limit of $n \to \infty$ and $b \to \infty$, given $\{\sigma_u = j, c_u = k\}$. We can now write the probability of error $p_e^\beta$ of the per-node MAP detector $\widehat{S}_0$ as

$$\begin{aligned}
p_e^\beta &= p_e^\beta(i|\sigma_i = 0, c_i = 0)P(\sigma_i = 0, c_i = 0) + p_e^\beta(i|\sigma_i = 0, c_i = 1)P(\sigma_i = 0, c_i = 1)\\
&\quad + p_e^\beta(i|\sigma_i = 1, c_i = 0)P(\sigma_i = 1, c_i = 0) + p_e^\beta(i|\sigma_i = 1, c_i = 1)P(\sigma_i = 1, c_i = 1)\\
&= P_{0,0}(R^t_i > \nu)\pi_{0,0} + P_{0,1}(R^t_i > \nu)\pi_{0,1} + P_{1,0}(R^t_i < \nu)\pi_{1,0} + P_{1,1}(R^t_i < \nu)\pi_{1,1},
\end{aligned}$$

is the error rate of Algorithm 3, where $p_e^\beta(i|\sigma_i = 0, c_i = 0)$ denotes the probability that node $i$ is misclassified, given $\{\sigma_i = 0, c_i = 0\}$ and $\pi_{0,1} = \mathbb{P}(\sigma_i = 0, c_i = 1)$ etc. Then the

expected fraction of mislabelled nodes $\frac{\mathbb{E}(|\hat{S}_0 \Delta S|)}{K}$ in the limit $n \to \infty, b \to \infty$ is

$$\lim_{b\to\infty}\lim_{n\to\infty}\frac{np_e^\beta}{K} = Q\left(\frac{\frac{\mu^{(t)}}{2}+\log(\frac{\beta}{(1-\beta)})}{\sqrt{\mu^{(t)}}}\right)\alpha\beta + (1-\alpha\beta)Q\left(\frac{\frac{\mu^{(t)}}{2}-\log\left(\frac{1-\kappa}{\kappa}\left(\frac{1-\frac{\alpha\kappa(1-\beta)}{1-\kappa}}{1-\alpha\beta}\right)\right)}{\sqrt{\mu^{(t)}}}\right)$$
$$+\alpha(1-\beta)Q\left(\frac{\frac{\mu^{(t)}}{2}-\log(\frac{\beta}{1-\beta})}{\sqrt{\mu^{(t)}}}\right)$$
$$+(\frac{1-\kappa}{\kappa}-\alpha(1-\beta))Q\left(\frac{\frac{\mu^{(t)}}{2}-\log\left(\frac{(1-\alpha\beta)\kappa}{(1-\kappa-\alpha\kappa+\alpha\kappa\beta)}\right)}{\sqrt{\mu^{(t)}}}\right).$$

We can show, by a calculation similar to the one followed in the proof of Theorem 5.1, that

$$\lim_{b\to\infty}\lim_{n\to\infty}\frac{np_e^\beta}{K} \leq \left(\alpha\sqrt{\beta(1-\beta)}+\sqrt{(1-\alpha\beta)(\frac{1-\kappa}{\kappa}-\alpha(1-\beta))}\right)e^{-\frac{\lambda\alpha\beta^2(1-\kappa)}{8\kappa}}.$$

Finally by a similar calculation to (5.22),

$$\lim_{b\to\infty}\lim_{n\to\infty}\frac{\mathbb{E}(|S\Delta\widehat{S}|)}{K} \leq 2\left(\alpha\sqrt{\beta(1-\beta)}+\sqrt{(1-\alpha\beta)(\frac{1-\kappa}{\kappa}-\alpha(1-\beta))}\right)e^{-\frac{\lambda\alpha\beta^2(1-\kappa)}{8\kappa}}.$$

□

Appendix B

# Appendix: Chapter 6

## B.1 Proof of Lemma 6.6

From Lemma 6.1, we have for Chung-Lu graphs that: $d_i = w_i(1+\varepsilon_i)$, where $\eta \equiv \max_i \varepsilon_i = o(1)$ with high probability. In the proof we assume explicitly that $v_i = 1/n$, but the results hold in the slightly more general case where $v_i = O(1/n)$ uniformly $\forall i$, i.e., $\exists K$ such that $\max_i nv_i \leq K$. It can be verified easily that all the bounds that follow hold in this more general setting. The event $\{\eta = o(1)\}$, holds w.h.p. asymptotically from Lemma 6.1. In this case, we have

$$\sum_j \left( \frac{A_{ij}}{\sqrt{d_i d_j}} - \frac{\sqrt{d_i d_j}}{\sum_i d_i} \right) \frac{v_j}{\sqrt{d_j}} = \sum_j \left( \frac{A_{ij}}{\sqrt{d_i d_j}} - \frac{\sqrt{d_i d_j}}{\sum_k d_k} \right) \frac{v_j}{\sqrt{w_j}} (1+\varepsilon_j)$$

where $\varepsilon_j$ is the error of convergence, and we have $\max_j \varepsilon_j = O(\eta)$. Therefore,

$$\begin{aligned} \left\| \widetilde{\mathbf{Q}}\mathbf{v}' \right\|_\infty &\leq \left\| \widetilde{\mathbf{Q}}\mathbf{q} \right\|_\infty + \max_i \varepsilon_i \left\| \widetilde{\mathbf{Q}}\mathbf{q} \right\|_\infty \\ &\leq \left\| \widetilde{\mathbf{Q}}\mathbf{q} \right\|_\infty (1+o(1)) \quad \text{w.h.p.}, \end{aligned} \tag{B.1}$$

where $\mathbf{q}$ is a vector such that $q_i = \frac{nv_i}{\sqrt{w_i}}$. Furthermore, we have w.h.p.

$$\begin{aligned} \frac{A_{ij}}{\sqrt{d_i d_j}} - \frac{\sqrt{d_i d_j}}{\sum_k d_k} &= \frac{A_{ij}}{\sqrt{w_i(1+\varepsilon_i)w_j(1+\varepsilon_j)}} - \frac{\sqrt{w_i(1+\varepsilon_i)w_j(1+\varepsilon_j)}}{\sum_k w_k(1+\varepsilon_k)} \\ &= \frac{A_{ij}}{\sqrt{w_i w_j}} (1 + O(\varepsilon_i) + O(\varepsilon_j)) - \frac{\sqrt{w_i w_j}}{\sum_k w_k} \left( \frac{1 + O(\varepsilon_i) + O(\varepsilon_j)}{1 + O(\eta)} \right) \\ &= \left( \frac{A_{ij}}{\sqrt{w_i w_j}} - \frac{\sqrt{w_i w_j}}{\sum_k w_k} \right) (1 + \delta_{ij}), \end{aligned}$$

where $\delta_{ij}$ is the error in the $ij^{\text{th}}$ term of the matrix and $\delta_{ij} = O(\eta)$ uniformly, so that $\max_{ij} \delta_{ij} = o(1)$ w.h.p. Consequently, defining $\widetilde{\overline{Q}}_{ij} = \frac{A_{ij}}{\sqrt{w_i w_j}} - \frac{\sqrt{w_i w_j}}{\sum_k w_k}$ we have:

$$\begin{aligned} \left\| \widetilde{\mathbf{Q}}\mathbf{q} \right\|_\infty &\leq \left\| \widetilde{\overline{\mathbf{Q}}}\mathbf{q} \right\|_\infty + \max_i \left| \sum_j \widetilde{\overline{Q}}_{ij} \delta_{ij} q_j \right| \\ &\leq \left\| \widetilde{\overline{\mathbf{Q}}}\mathbf{q} \right\|_\infty + O(\eta) \max_i \frac{1}{\sqrt{w_{\min}}} \sum_j |\widetilde{\overline{\mathbf{Q}}}_{ij}| \\ &\leq \left\| \widetilde{\overline{\mathbf{Q}}}\mathbf{q} \right\|_\infty + o(1) \frac{1}{\sqrt{w_{\min}}} \left( C \sqrt{\frac{w_{\max}}{w_{\min}}} + \frac{w_{\max}}{w_{\min}} \right) \qquad \text{(B.2)} \\ &\leq \left\| \widetilde{\overline{\mathbf{Q}}}\mathbf{q} \right\|_\infty + o(1/\sqrt{w_{\min}}) \qquad \text{(B.3)} \end{aligned}$$

where in (B.2) we used the fact the $O(\eta)$ is a uniform bound on the error and it is $o(1)$ w.h.p. and $\max_j q_j \leq \frac{1}{\sqrt{w_{\min}}}$. In (B.2) we also used the fact that

$$\begin{aligned}\max_i \sum_j |\widetilde{\overline{Q}}_{ij}| &\leq \max_i \sum_j \frac{A_{ij}}{\sqrt{w_i w_j}} + \sum_j \frac{\sqrt{w_i w_j}}{\sum_k w_k} \\ &\leq \max_i \frac{1}{\sqrt{w_{\min}}} \frac{d_i}{\sqrt{w_i}} + \max_i \frac{\sqrt{w_i w_{\max}}}{w_{\min}} \\ &\overset{(a)}{\leq} C\sqrt{\frac{w_i}{w_{\min}}} + \frac{w_{\max}}{w_{\min}} \\ &\leq C\sqrt{\frac{w_{\max}}{w_{\min}}} + \frac{w_{\max}}{w_{\min}},\end{aligned}$$

where $C$ is some constant. In (a) above we used the fact that w.h.p. $d_i = w_i(1+o(1))$, by Lemma 6.1, hence $\exists C$ such that $\forall n$ large enough $d_i \leq C w_i$.

Now we proceed to bound $\left\|\widetilde{\overline{\mathbf{Q}}}\mathbf{q}\right\|_\infty$. Substituting for $q_i = \frac{1}{\sqrt{w_i}}$, we get

$$\begin{aligned}\sum_j \frac{1}{\sqrt{w_j}} \left( \frac{A_{ij}}{\sqrt{w_i w_j}} - \frac{\sqrt{w_i w_j}}{\sum_k w_k} \right) &= \sum_j \frac{1}{w_j \sqrt{w_i}} \left( A_{ij} - \frac{w_i w_j}{\sum_i w_i} \right) \\ &\equiv \frac{1}{\sqrt{w_i}} X_i. \end{aligned} \tag{B.4}$$

We seek to bound $\max_i |X_i|$:

$$X_i = \sum_j \frac{1}{w_j} \left( A_{ij} - \frac{w_i w_j}{\sum_i w_i} \right).$$

Furthermore, $\mathbb{E}(X_i^2) = \sum_j \frac{1}{w_j^2} \mathbb{E}(A_{ij} - p_{ij})^2$, with $p_{ij} = \frac{w_i w_j}{\sum w_i}$. So, $\mathbb{E}(X_i^2) = \sum_j \frac{1}{w_j^2} p_{ij}(1 - p_{ij}) \leq \frac{w_i}{\sum_i w_i} \sum_j \frac{1}{w_j} \leq n \frac{p_i}{w_{\min}}$, where $p_i = \frac{w_i}{\sum_i w_i}$, and $\frac{A_{ij}}{w_j} \leq 1/w_{\min}$. Therefore using Bernstein Concentration Lemma for $\varepsilon < n \max_i p_i$:

$$\begin{aligned}\mathbb{P}\left( \max_i |\sum_j (A_{ij} - p_{ij})/w_j| \geq \varepsilon \right) &\leq n \max_i \exp\left(-\frac{\varepsilon^2}{2(p_i n/w_{\min}) + \varepsilon/w_{\min}}\right) \\ &\leq n \max_i \exp\left(-\frac{w_{\min}\varepsilon^2}{2(np_i + \varepsilon)}\right) \\ &\leq n \exp(-\varepsilon^2 w_{\min}/(4n \max_i p_i)) \\ &\leq n \exp\left(\frac{-\varepsilon^2 \mathrm{vol} w_{\min}}{4 w_{\max} n}\right), \end{aligned} \tag{B.5}$$

where $\frac{\mathrm{vol}}{n} = \frac{\sum_i w_i}{n} \geq w_{\min}$. It can be verified that when $\varepsilon = \frac{1}{(\overline{w})^\alpha}$, for some $\alpha > 0$, the RHS of (B.5) can be upper bounded by $n^{-(\gamma K - 1)}$, if $\overline{w} \geq (\gamma \log(n))^{\frac{1}{1-2\alpha}}$, for some large enough $\gamma$, which can be easily satisfied if $w_{\min} \gg O(\log^c(n))$, for some $c > 1$, where $K$ is a constant such that $w_{\max} \leq K w_{\min}$. Thus, finally, from (B.4) and (B.3) we have $\left\|\widetilde{\mathbf{Q}}\mathbf{q}\right\|_\infty = o(1/\sqrt{w_{\min}})$, w.h.p., and therefore from (B.1), we get $\left\|\widetilde{\mathbf{Q}}\mathbf{v}'\right\|_\infty = o(1/\sqrt{w_{\min}})$.

□

## B.2 Proof of Lemmas in Section 6.5

### B.2.1 Proof of Lemma 6.8

The proof is an application of Bernstein's Concentration Lemma. Note that for $1 \leq i \leq m$, $D_i = \sum_j A_{ij}$ . Here the mean degree $\mathbb{E}(D_i) = mp + (n-m)q = t_1$, and the variance $B_n^2 = mp(1-p)+(n-m)q(1-q) \leq t_1$ for $i \leq m$. Similarly for $i > m$, $\mathbb{E}(D_i) = mq+(n-m)p = t_2$ is and variance $\mathrm{Var}[D_i] \leq t_2$. Then, the minimum average degree $w_{\min} = \min(t_1, t_2)$. By Bernstein's Lemma, for $\varepsilon = C\sqrt{\frac{\log(n)}{w_{\min}}}$,

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq i\leq m}|D_i - t_1| \geq \varepsilon t_1\right) &\leq 2m\exp\left(\frac{-\varepsilon^2 t_1^2}{2(t_1\varepsilon/3 + t_1)}\right)\\ &= 2m\exp\left(\frac{-\varepsilon^2 t_1}{1+\varepsilon/3}\right)\\ &= O(n^{-c}),\end{aligned}$$

for some $c$. Hence $\max_{1\leq i\leq m}\left|\frac{D_i - t_1}{t_1}\right| \leq C\sqrt{\frac{\log(n)}{w_{\min}}}$ w.h.p. Similarly

$$\max_{1+m\leq i\leq n/2}\left|\frac{D_i - t_2}{t_2}\right| \leq C\sqrt{\frac{\log(n)}{w_{\min}}},\ \text{w.h.p.}$$

Combining the two bounds above we get,

$$\max_{1\leq i\leq n}\left|\frac{D_i}{\mathbb{E}(D_i)} - 1\right| \leq C\sqrt{\frac{\log(n)}{w_{\min}}},\ \text{w.h.p.} \tag{B.6}$$

□

### B.2.2 Proof of Lemma 6.9

To prove Lemma 6.9 we need the following lemma on the spectral norm of the difference between the adjacency matrix and its mean.

**Lemma B.1.** *For an SBM matrix $G(m, n-m, p, q)$ with adjacency matrix $\mathbf{A}$ and $\overline{\mathbf{A}} = \mathbb{E}(\mathbf{A})$, there exists a constant $K$ s.t.*

$$\|\mathbf{A} - \overline{\mathbf{A}}\|_2 \leq K\sqrt{\log(n)w_{max}},\ \textit{w.h.p.},$$

*where $w_{max} = \max(m, n-m)p + \min(m, n-m)q$ is the maximum average degree, if $w_{\max} = \omega(\log^3(n))$.*

To prove this Lemma we need the Matrix Bernstein Concentration result, which we state below for the sake of completeness:

**Lemma B.2.** *[Tropp 2012a, Theorem 1.4]. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots \mathbf{S}_t$ be independent random matrices with common dimension $d_1 \times d_2$. Assume that each matrix has bounded deviation from its mean, i.e.,*

$$\|\mathbf{S}_k - \mathbb{E}(\mathbf{S}_k)\| \leq R,\ \textit{for each } k = 1, \ldots n.$$

*Let $\mathbf{Z} = \sum_{k=1}^t \mathbf{S}_k$ and introduce a variance parameter*

$$\sigma_{\mathbf{Z}}^2 = \max\left\{\|\mathbb{E}\left((\mathbf{Z} - \mathbb{E}(\mathbf{Z}))(\mathbf{Z} - \mathbb{E}(\mathbf{Z}))^H\right)\|, \|\mathbb{E}\left((\mathbf{Z} - \mathbb{E}(\mathbf{Z}))^H(\mathbf{Z} - \mathbb{E}(\mathbf{Z}))\right)\|\right\}.$$

*Then*

$$\mathbb{P}\{\|\mathbf{Z}-\mathbb{E}(\mathbf{Z})\|> t\} \leq (d_1+d_2).\exp\left(\frac{-t^2/2}{\sigma_{\mathbf{Z}}^2+Rt/3}\right), \tag{B.7}$$

*for all* $t \geq 0$.

**Proof of Lemma B.1:** With $\mathbf{Z}=\mathbf{A}$, in Lemma B.2, we can decompose $\mathbf{Z}$ as sums of Hermitian matrices $\mathbf{S}_{i'j'}$, $\mathbf{Z}=\sum_{1\leq i'<j'\leq n}\mathbf{S}_{i'j'}$ such that:

$$(\mathbf{S}_{i'j'})_{ij}=\begin{cases}A_{i'j'} \text{ if } i=i^{'}, j=j^{'},\\ A_{i'j'} \text{ if } i=j^{'}, j=i^{'},\\ 0 \text{ otherwise.}\end{cases} \tag{B.8}$$

Notice that if $\mathbf{x}\neq 0$, $\|(\mathbf{S}_{i'j'}-\mathbb{E}(\mathbf{S}_{i'j'}))\mathbf{x}\|_2= |2x_{i'}x_{j'}(A_{i'j'}-\mathbb{E}(A_{i'j'}))|< |x_{i'}^2+x_{j'}^2|$. Consequently $\|\mathbf{S}_{i'j'}-\mathbb{E}(\mathbf{S}_{i'j'})\|_2< 1$, giving $R=1$ in the statement of Lemma B.7. Let $\mathbf{Y}=\mathbb{E}\left((\mathbf{Z}-\mathbb{E}\mathbf{Z})^H(\mathbf{Z}-\mathbb{E}\mathbf{Z})\right)$, then

$$Y_{ij}=\begin{cases}v_1 & \text{if } i=j, i\leq m,\\ v_2 & \text{if } i=j, i>m,\\ 0 & \text{otherwise,}\end{cases} \tag{B.9}$$

where $v_1=mp(1-p)+q(1-q)(n-m), v_2=(n-m)p(1-p)+mq(1-q)$. Therefore $\sigma_{\mathbf{Z}}^2=\max(v_1,v_2)=\max(n-m,m)p+\min(n-m,m)q=\sigma^2$. By our assumptions on the probabilities, $\sigma^2=\omega(\log^3(n))$. Thus it follows that

$$\begin{aligned}\mathbb{P}(\|\mathbf{A}-\overline{\mathbf{A}}\|\geq t\sigma) &\leq 2n\exp\left(\frac{-t^2\sigma^2}{2\sigma^2+t\sigma/3}\right)\\ &\leq 2n\exp(-t^2/3),\end{aligned}$$

if $\sigma>t$. The RHS is $O(n^{-c})$ if $t>\sqrt{r\log(n)}$, for some $r$. □Finally we are in a position to prove Lemma 6.9

**Proof of Lemma 6.9:** We prove this result in two steps. First we show that

$$\|\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}-\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}\|_2= C\sqrt{\frac{\log(n)}{w_{\min}}}=o(1). \tag{B.10}$$

Observe that

$$\begin{aligned}&\|\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}-\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}\|_2= \|\mathbf{Q}-\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\|\\ &=\|\mathbf{Q}-\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}+\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}-\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}\mathbf{W}^{-1/2}\|_2\\ &=\|(\mathbf{I}-\mathbf{W}^{-1/2}\mathbf{D}^{1/2})\mathbf{Q}+\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\mathbf{Q}(\mathbf{I}-\mathbf{D}^{1/2}\mathbf{W}^{-1/2})\|_2\\ &\leq \delta+(1+\delta)\delta,\end{aligned}$$

where $\delta=\max_i\left|\frac{d_i}{w_i}-1\right|$. In the last line we used the fact that $\|\mathbf{Q}\|_2= 1$, $\|\mathbf{I}-\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\|_2= \max_i\left|\sqrt{\frac{d_i}{w_i}}-1\right|\leq\max_i\left|\frac{d_i}{w_i}-1\right|$ and

$$\|\mathbf{W}^{-1/2}\mathbf{D}^{1/2}\|_2\leq \|\mathbf{W}^{-1/2}\mathbf{D}^{1/2}-\mathbf{I}\|_2+\|\mathbf{I}\|_2\leq \delta+1.$$

By Lemma 6.8, $\delta\leq C\sqrt{\frac{\log(n)}{w_{\min}}}=o(1)$ w.h.p. Next we show that

$$\|\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2} - \mathbf{W}^{-1/2}\overline{\mathbf{A}}\mathbf{W}^{-1/2}\|_2 \leq \frac{C\sqrt{\log(n)w_{\max}}}{w_{\min}} = o(1). \tag{B.11}$$

Now using Lemma B.1 we have

$$\begin{aligned}\|\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2} - \mathbf{W}^{-1/2}\overline{\mathbf{A}}\mathbf{W}^{-1/2}\| &\leq \frac{\|\mathbf{A} - \overline{\mathbf{A}}\|_2}{w_{\min}} \\ &\leq \frac{c\sqrt{\log(n)w_{\max}}}{w_{\min}} \\ &= o(1), \text{ w.h.p.},\end{aligned}$$

if $w_{\min} = \omega(\sqrt{\log(n)w_{\max}})$, which is satisfied when $w_{\max} \leq Cw_{\min}$ for some $C$, and $w_{\max} = \omega(\log^3(n))$. The result of Lemma 6.9 then follows from (B.10) and (B.11) by applying the triangular inequality. □

# Bibliography


[Abbe & Sandon 2015a] Emmanuel Abbe and Colin Sandon. *Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery.* In FOCS, 2015, pages 670–688. IEEE, 2015.

[Abbe & Sandon 2015b] Emmanuel Abbe and Colin Sandon. *Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap.* arXiv preprint arXiv:1512.09080, 2015.

[Abounadi *et al.* 2001] Jinane Abounadi, D Bertsekas and Vivek S Borkar. *Learning algorithms for Markov decision processes with average cost.* SIAM J. Control Optim., vol. 40, no. 3, pages 681–698, 2001.

[Abramowitz & Stegun 1964] Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions: with formulas, graphs, and mathematical tables, volume 55. Courier Corporation, 1964.

[Ahlswede & Winter 2002] Rudolf Ahlswede and Andreas Winter. *Strong converse for identification via quantum channels.* IEEE Transactions on Information Theory, vol. 48, no. 3, pages 569–579, 2002.

[Akoglu *et al.* 2015] Leman Akoglu, Hanghang Tong and Danai Koutra. *Graph based anomaly detection and description: a survey.* Data Mining and Knowledge Discovery, vol. 29, no. 3, pages 626–688, 2015.

[Albert *et al.* 1999] Réka Albert*et al.* *Emergence of scaling in random networks.* science, vol. 286, no. 5439, pages 509–512, 1999.

[Aldous & Fill 2002] David Aldous and James Allen Fill. *Reversible Markov Chains and Random Walks on Graphs.* Unfinished monograph, recompiled 2014, available at http://www.stat.berkeley.edu/~aldous/RWG/book.html, 2002.

[Allahverdyan *et al.* 2010] Armen E Allahverdyan, Greg Ver Steeg and Aram Galstyan. *Community detection with and without prior information.* EPL (Europhysics Letters), vol. 90, no. 1, page 18002, 2010.

[Alon & Spencer 2004] Noga Alon and Joel H Spencer. The probabilistic method. John Wiley & Sons, 2004.

[Alon *et al.* 1998] Noga Alon, Michael Krivelevich and Benny Sudakov. *Finding a large hidden clique in a random graph.* Random Structures and Algorithms, vol. 13, no. 3-4, pages 457–466, 1998.

[Alon *et al.* 2002] Noga Alon, Michael Krivelevich and Van H Vu. *On the concentration of eigenvalues of random symmetric matrices.* Israel Journal of Mathematics, vol. 131, no. 1, pages 259–267, 2002.

[Ames 2013] Brendan PW Ames. *Robust convex relaxation for the planted clique and densest k-subgraph problems.* arXiv preprint arXiv:1305.4891, 2013.

[Andersen & Chung 2007] Reid Andersen and Fan Chung. *Detecting sharp drops in PageRank and a simplified local partitioning algorithm.* Theory Appl. Model. Comput., vol. 4484/2007, no. 3, pages 1–12, 2007.

[Andersen *et al.* 2006] Reid Andersen, Fan Chung and Kevin Lang. *Local Graph Partitioning using PageRank Vectors*. In 2006 47th Annu. IEEE Symp. Found. Comput. Sci., pages 475–486. IEEE, oct 2006.

[Anderson *et al.* 2009] Greg W Anderson, Alice Guionnet, Ofer Zeitouni, Greg W. Anderson Alice Guionnet, Ofer Zeitouni, Greg W Anderson, Alice Guionnet and Ofer Zeitouni. An Introduction to Random Matrices, volume 118 of *Cambridge studies in advanced mathematics*. Cambridge University Press, 2009.

[Arias-Castro *et al.* 2014] Ery Arias-Castro, Nicolas Verzelen*et al.* *Community detection in dense random networks*. The Annals of Statistics, vol. 42, no. 3, pages 940–969, 2014.

[Athreya *et al.* 2013] Avanti Athreya, Vince Lyzinski, David J. Marchette, Carey E. Priebe, Daniel L. Sussman and Minh Tang. *A central limit theorem for scaled eigenvectors of random dot product graphs*. arXiv Prepr. arXiv1305.7388, no. 1983, pages 1–15, 2013.

[Avrachenkov & Lebedev 2006] Konstantin Avrachenkov and Dmitri Lebedev. *PageRank of scale-free growing networks*. Internet Mathematics, vol. 3, no. 2, pages 207–231, 2006.

[Avrachenkov *et al.* 2008] Konstantin Avrachenkov, Vladimir Dobrynin, Danil Nemirovsky, Son Kim Pham and Elena Smirnova. *Pagerank based clustering of hypertext document collections*. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 873–874. ACM, 2008.

[Avrachenkov *et al.* 2012] Konstantin Avrachenkov, Paulo Gonçalves, Alexey Mishenin and Marina Sokol. *Generalized Optimization Framework for Graph-based Semi-supervised Learning*. In Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012., pages 966–974, 2012.

[Avrachenkov *et al.* 2015] Konstantin Avrachenkov, Laura Cottatellucci and Arun Kadavankandy. *Spectral properties of random matrices for stochastic block model*. In Model. Optim. Mobile, Ad Hoc, Wirel. Networks (WiOpt), 2015 13th Int. Symp., pages 537–544. IEEE, 2015.

[Avrachenkov *et al.* 2016] Konstantin Avrachenkov, Bruno Ribeiro and Jithin K Sreedharan. *Inference in OSNs via Lightweight Partial Crawls*. In Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, pages 165–177. ACM, 2016.

[Bai & Pan 2012] Z. D. Bai and G. M. Pan. *Limiting behavior of eigenvectors of large Wigner matrices*. J. Stat. Phys., vol. 146, no. 3, pages 519–549, 2012.

[Bai & Silverstein 1998] Z. D. Bai and Jack W. Silverstein. *No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices*. Ann. Probab., vol. 26, no. 1, pages 316–345, 1998.

[Bai & Silverstein 2009] Zhidong Bai and Jack W Silverstein. Spectral analysis of large dimensional random matrices. Springer, 2009.

[Bai *et al.* 2007] Z. D. Bai, B. Q. Miao and G. M. Pan. *On asymptotics of eigenvectors of large sample covariance matrix.* Ann. Probab., vol. 35, no. 4, pages 1532–1572, 2007.

[Bai 1999] Zhidong D Bai. *Methodologies in Spectral Analysis of Large Dimensional Random Matrices, A Review.* Stat. Sin., vol. 9, no. 3, pages 611–677, jul 1999.

[Basu *et al.* 2006] Sugato Basu, Mikhail Bilenko, Arindam Banerjee and Raymond J Mooney. *Probabilistic semi-supervised clustering with constraints.* Semi-supervised learning, pages 71–98, 2006.

[Benaych-Georges 2011] Florent Benaych-Georges. *Eigenvectors of Wigner matrices: universality of global fluctuations.* arXiv Prepr. arXiv1104.1219, 2011.

[Benson *et al.* 2015] Austin R Benson, David F Gleich and Jure Leskovec. *Tensor spectral clustering for partitioning higher-order network structures.* In Proceedings of the 2015 SIAM International Conference on Data Mining, pages 118–126. SIAM, 2015.

[Beutel *et al.* 2013] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow and Christos Faloutsos. *Copycatch: stopping group attacks by spotting lockstep behavior in social networks.* In Proceedings of the 22nd WWW, pages 119–130. ACM, 2013.

[Bhatia 2013] Rajendra Bhatia. Matrix analysis, volume 169. Springer Science & Business Media, 2013.

[Billingsley 2008] Patrick Billingsley. Probability and measure. John Wiley & Sons, 2008.

[Bollobás 1998] Béla Bollobás. *Random graphs.* In Modern Graph Theory, pages 215–252. Springer, 1998.

[Bordenave & Guionnet 2013] Charles Bordenave and Alice Guionnet. *Localization and delocalization of eigenvectors for heavy-tailed random matrices.* Probab. Theory Relat. Fields, vol. 157, no. 3-4, pages 885–953, 2013.

[Bordenave & Lelarge 2010] Charles Bordenave and Marc Lelarge. *Resolvent of large random graphs.* Random Structures & Algorithms, vol. 37, no. 3, pages 332–352, 2010.

[Bordenave *et al.* 2010] Charles Bordenave, Pietro Caputo and Djalil Chafaï. *Spectrum of large random reversible markov chains: two examples.* Lat. Am. J. Probab. Math. Stat., vol. 7, no. March, pages 1–20, 2010.

[Borkar *et al.* 2014] Vivek S Borkar, Rahul Makhijani and Rajesh Sundaresan. *Asynchronous Gossip for Averaging and Spectral Ranking.* IEEE J. Sel. Areas Commun., vol. 8, no. 4, pages 703–716, 2014.

[Borkar 2009] Vivek S Borkar. *Reinforcement learning: a bridge between numerical methods and Monte Carlo.* Perspectives in Mathematical Science–I: Probability and Statistics, pages 71–91, 2009.

[Bose *et al.* 2013] Subhonmesh Bose, Elizabeth Bodine-Baron, Babak Hassibi and Adam Wierman. *The cost of an epidemic over a complex network: A random matrix approach.* arXiv preprint arXiv:1309.2236, 2013.

[Boudin 2013] Florian Boudin. *A comparison of centrality measures for graph-based keyphrase extraction.* In International Joint Conference on Natural Language Processing (IJCNLP), pages 834–838, 2013.

[Brémaud 2013] Pierre Brémaud. Markov chains: Gibbs fields, Monte Carlo simulation, and queues, volume 31. Springer Science & Business Media, 2013.

[Cai *et al.* 2016] T Tony Cai, Tengyuan Liang and Alexander Rakhlin. *Inference via Message Passing on Partially Labeled Stochastic Block Models.* arXiv preprint arXiv:1603.06923, 2016.

[Caltagirone *et al.* 2016] Francesco Caltagirone, Marc Lelarge and Léo Miolane. *Recovering asymmetric communities in the stochastic block model.* In Allerton 2016 54th Annual Allerton Conference on Communication, Control, and Computing, Monticello, United States, September 2016.

[Chau *et al.* 2006] Duen Horng Chau, Shashank Pandit and Christos Faloutsos. *Detecting fraudulent personalities in networks of online auctioneers.* In PKDD, pages 103–114. Springer, 2006.

[Chen & Saad 2012] Jie Chen and Yousef Saad. *Dense subgraph extraction with application to community detection.* IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 7, pages 1216–1230, 2012.

[Chen & Xu 2016] Yudong Chen and Jiaming Xu. *Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices.* Journal of Machine Learning Research, vol. 17, no. 27, pages 1–57, 2016.

[Chen *et al.* 2012] Zhengzhang Chen, William Hendrix and Nagiza F Samatova. *Community-based anomaly detection in evolutionary networks.* Journal of Intelligent Information Systems, vol. 39, no. 1, pages 59–85, 2012.

[Chen *et al.* 2014] Ningyuan Chen, Nelly Litvak and Mariana Olvera-cravioto. *PageRank in scale-free random graphs.* arXiv.math, no. 288956, 2014.

[Chen *et al.* 2015] Siheng Chen, Aliaksei Sandryhaila, José MF Moura and Jelena Kovačević. *Signal recovery on graphs: Variation minimization.* IEEE Transactions on Signal Processing, vol. 63, no. 17, pages 4609–4624, 2015.

[Chen *et al.* 2016] Ningyuan Chen, Nelly Litvak and Mariana Olvera-Cravioto. *Generalized PageRank on directed configuration networks.* Random Structures & Algorithms, 2016.

[Chung & Lu 2002a] Fan Chung and Linyuan Lu. *The average distances in random graphs with given expected degrees.* Proceedings of the National Academy of Sciences, vol. 99, no. 25, pages 15879–15882, 2002.

[Chung & Lu 2002b] Fan Chung and Linyuan Lu. *Connected components in random graphs with given expected degree sequences.* Annals of combinatorics, vol. 6, no. 2, pages 125–145, 2002.

[Chung & Radcliffe 2011] Fan Chung and Mary Radcliffe. *On the spectra of general random graphs.* the electronic journal of combinatorics, vol. 18, no. 1, page P215, 2011.

[Chung *et al.* 2003] Fan Chung, Linyuan Lu and Van Vu. *Spectra of random graphs with given expected degrees.* Proc. Natl. Acad. Sci. U. S. A., vol. 100, no. 11, pages 6313–6318, 2003.

[Chung 1997] Fan R K Chung. Spectral graph theory, volume 92. American Mathematical Soc., 1997.

[Chung 2009] Fan Chung. *A local graph partitioning algorithm using heat kernel pagerank.* Internet Mathematics, vol. 6, no. 3, pages 315–330, 2009.

[Condon & Karp 1999] Anne Condon and Richard M Karp. *Algorithms for graph partitioning on the planted partition model.* In Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques, pages 221–232. Springer, 1999.

[Cooper *et al.* 2013] Colin Cooper, Tomasz Radzik and Yiannis Siantos. *Fast Low-Cost Estimation of Network Properties Using Random Walks.* In Proc. Workshop on Algorithms and Models for the Web-Graph (WAW), Cambridge, MA, USA, 2013.

[Cvetković *et al.* 1980] Dragoš M Cvetković, Michael Doob and Horst Sachs. Spectra of graphs: theory and application, volume 87. Academic Pr, 1980.

[Dasgupta *et al.* 2014] Anirban Dasgupta, Ravi Kumar and Tamas Sarlos. *On estimating the average degree.* In Proc. 23rd Int. Conf. World wide web, pages 795–806. ACM, 2014.

[Decelle *et al.* 2011] Aurelien Decelle, Florent Krzakala, Cristopher Moore and Lenka Zdeborová. *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications.* Phys. Rev. E, vol. 84, no. 6, page 66106, 2011.

[Deshpande & Montanari 2015] Yash Deshpande and Andrea Montanari. *Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time.* Foundations of Computational Mathematics, vol. 15, no. 4, pages 1069–1128, 2015.

[Diaconis & Janson 2007] Persi Diaconis and Svante Janson. *Graph limits and exchangeable random graphs.* arXiv preprint arXiv:0712.2749, 2007.

[Ding *et al.* 2003] Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha and Horst Simon. *PageRank, HITS and a unified framework for link analysis.* In Proceedings of the 2003 SIAM International Conference on Data Mining, pages 249–253. SIAM, 2003.

[Ding *et al.* 2010] Xue Ding, Tiefeng Jiang*et al.* *Spectral distributions of adjacency and Laplacian matrices of random graphs.* The annals of applied probability, vol. 20, no. 6, pages 2086–2117, 2010.

[Erdős *et al.* 2009] László Erdős, Benjamin Schlein, Horng-Tzer Yau, László Erdős, Benjamin Schlein and Horng-Tzer Yau. *Local semicircle law and complete delocalization for Wigner random matrices.* Commun. Math. Phys., vol. 287, no. 2, pages 641–655, 2009.

[Erdős & Rényi 1959] Paul Erdős and Alfréd Rényi. *On random graphs, I.* Publicationes Mathematicae (Debrecen), vol. 6, pages 290–297, 1959.

[Erdős & Wilson 1977] Paul Erdős and Robin J Wilson. *On the chromatic index of almost all graphs.* Journal of combinatorial theory, series B, vol. 23, no. 2-3, pages 255–257, 1977.

[Erdős *et al.* 2013] László Erdős, Antti Knowles, Horng Tzer Yau and Jun Yin. *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law.* Ann. Probab., vol. 41, no. 3 B, pages 2279–2375, 2013.

[Erdős 2011] László Erdős. *Universality of Wigner random matrices: a survey of recent results.* Russian Mathematical Surveys, vol. 66, no. 3, page 507, 2011.

[Filippone *et al.* 2008] Maurizio Filippone, Francesco Camastra, Francesco Masulli and Stefano Rovetta. *A survey of kernel and spectral methods for clustering.* Pattern recognition, vol. 41, no. 1, pages 176–190, 2008.

[Firouzi *et al.* 2013] Hamed Firouzi, Bala Rajaratnam and Alfred O Hero III. *Predictive Correlation Screening: Application to Two-stage Predictor Design in High Dimension.* In AISTATS, pages 274–288, 2013.

[Fortunato & Barthélemy 2007] Santo Fortunato and Marc Barthélemy. *Resolution limit in community detection.* Proc. Natl. Acad. Sci. U. S. A., vol. 104, no. 1, pages 36–41, jan 2007.

[Fortunato *et al.* 2006] Santo Fortunato, Marián Boguñá, Alessandro Flammini and Filippo Menczer. *Approximating PageRank from in-degree.* In International Workshop on Algorithms and Models for the Web-Graph, pages 59–71. Springer, 2006.

[Fortunato 2010] Santo Fortunato. *Community detection in graphs.* Physics reports, vol. 486, no. 3, pages 75–174, 2010.

[Füredi & Komlós 1981] Zoltán Füredi and János Komlós. *The eigenvalues of random symmetric matrices.* Combinatorica, vol. 1, no. 3, pages 233–241, 1981.

[Ghoshal *et al.* 2009] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli and M. E J Newman. *Random hypergraphs and their applications.* Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 79, no. 6, pages 1–11, 2009.

[Girko *et al.* 1994] V. Girko, W. Kirsch and a. Kutzelnigg. *A necessary and sufficient conditions for the semicircle law.* Random Oper. Stoch. Equations, vol. 2, no. 2, pages 195–202, 1994.

[Girko 1990] Vjačeslav L Girko. Theory of Random Determinants. Mathematics and Its Applications. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[Girko 2001] Vyacheslav L Girko. Theory of Stochastic Canonical Equations, volume 1. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[Gjoka *et al.* 2010] Minas Gjoka, Maciej Kurant, Carter T Butts and Athina Markopoulou. *Walking in Facebook: A case study of unbiased sampling of OSNs.* In INFOCOM, 2010 Proc. IEEE, pages 1–9. IEEE, 2010.

[Gkorou *et al.* 2013] Dimitra Gkorou, Tamás Vinkó, Johan Pouwelse and Dick Epema. *Leveraging node properties in random walks for robust reputations in decentralized networks.* In Peer-to-Peer Computing (P2P), 2013 IEEE Thirteenth International Conference on, pages 1–10. IEEE, 2013.

[Gleich & Kloster 2016] DF Gleich and Kyle Kloster. *Seeded PageRank solution paths*. European Journal of Applied Mathematics, pages 1–34, 2016.

[Gleich 2015] David F Gleich. *PageRank beyond the Web*. SIAM Review, vol. 57, no. 3, pages 321–363, 2015.

[Goel & Salganik 2009] Sharad Goel and Matthew J Salganik. *Respondent-driven sampling as Markov chain Monte Carlo*. Stat. Med., vol. 28, no. 17, pages 2202–2229, 2009.

[Goldberg 1984] Andrew V Goldberg. *Finding a maximum density subgraph*. Technical report, University of California Berkeley, CA, 1984.

[Hajek *et al.* 2015a] Bruce Hajek, Yihong Wu and Jiaming Xu. *Recovering a Hidden Community Beyond the Spectral Limit in $O(|E|\log^*|V|)$ Time*. arXiv Prepr. arXiv1510.02786, 2015.

[Hajek *et al.* 2015b] Bruce E Hajek, Yihong Wu and Jiaming Xu. *Computational Lower Bounds for Community Detection on Random Graphs*. In COLT, pages 899–928, 2015.

[Hajek *et al.* 2016a] Bruce Hajek, Yihong Wu and Jiaming Xu. *Achieving exact cluster recovery threshold via semidefinite programming*. IEEE Transactions on Information Theory, vol. 62, no. 5, pages 2788–2797, 2016.

[Hajek *et al.* 2016b] Bruce Hajek, Yihong Wu and Jiaming Xu. *Information limits for recovering a hidden community*. In Information Theory (ISIT), 2016 IEEE International Symposium on, pages 1894–1898. IEEE, 2016.

[Haveliwala 2002] Taher H Haveliwala. *Topic-sensitive pagerank*. In Proceedings of the 11th international conference on World Wide Web, pages 517–526. ACM, 2002.

[Heard *et al.* 2010] Nicholas A. Heard, David J. Weston, Kiriaki Platanioti and David J. Hand. *Bayesian anomaly detection methods for social networks*. Ann. Appl. Stat., vol. 4, no. 2, pages 645–662, 2010.

[Heimlicher *et al.* 2012] Simon Heimlicher, Marc Lelarge and L Massoulié. *Community Detection in the Labelled Stochastic Block Model*. arXiv Prepr. arXiv1209.2910, vol. 2, no. 1, pages 1–9, 2012.

[Hofstad 2016] Remco van der Hofstad. Random graphs and complex networks, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2016.

[Holland *et al.* 1983] Paul W. Holland, Kathryn Blackmond Laskey and Samuel Leinhardt. *Stochastic blockmodels: First steps*. Soc. Networks, vol. 5, no. 2, pages 109–137, 1983.

[Horn & Johnson 2012] Roger A Horn and Charles R Johnson. Matrix analysis. Cambridge university press, 2012.

[Hou *et al.* 2016] Jack P Hou, Amin Emad, Gregory J Puleo, Jian Ma and Olgica Milenkovic. *A new correlation clustering method for cancer mutation analysis*. Bioinformatics, vol. 32, no. 24, pages 3717–3728, 2016.

[Jiang *et al.* 2006] Tiefeng Jiang*et al. How many entries of a typical orthogonal matrix can be approximated by independent normals?* The Annals of Probability, vol. 34, no. 4, pages 1497–1529, 2006.

[Kadavankandy *et al.* 2016] Arun Kadavankandy, Laura Cottatellucci and Konstantin Avrachenkov. *Characterization of $L^1$-norm statistic for Anomaly Detection in Erdös Rényi Graphs.* In CDC. IEEE, 2016.

[Kamvar *et al.* 2003] Sepandar D Kamvar, Mario T Schlosser and Hector Garcia-Molina. *The eigentrust algorithm for reputation management in p2p networks.* In Proceedings of the 12th international conference on World Wide Web, pages 640–651. ACM, 2003.

[Kang *et al.* 2011] U Kang, Duen Horng Chau and Christos Faloutsos. *Mining large graphs: Algorithms, inference, and discoveries.* In 2011 IEEE 27th International Conference on Data Engineering, pages 243–254. IEEE, 2011.

[Karp 1972] Richard M Karp. *Reducibility among combinatorial problems.* In Complexity of computer computations, pages 85–103. Springer, 1972.

[Karrer & Newman 2011] Brian Karrer and Mark EJ Newman. *Stochastic blockmodels and community structure in networks.* Physical Review E, vol. 83, no. 1, page 016107, 2011.

[Kemeny & Snell 1983] John G. Kemeny and James L. Snell. Finite markov chains. Springer, New York, USA, 1983.

[Kepner & Gilbert 2011] Jeremy Kepner and John Gilbert. Graph algorithms in the language of linear algebra. SIAM, 2011.

[Kitano 2002] Hiroaki Kitano. *Computational systems biology.* Nature, vol. 420, no. 6912, pages 206–210, 2002.

[Kloumann *et al.* 2016] Isabel M Kloumann, Johan Ugander and Jon Kleinberg. *Block models and personalized PageRank.* Proceedings of the National Academy of Sciences, page 201611275, 2016.

[Koutra *et al.* 2011] Danai Koutra, Tai-You Ke, U Kang, Duen Horng Polo Chau, Hsing-Kuo Kenneth Pao and Christos Faloutsos. *Unifying guilt-by-association approaches: Theorems and fast algorithms.* In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 245–260. Springer, 2011.

[Langville & Meyer 2004] Amy Langville and Carl Meyer. *Deeper Inside PageRank.* Internet Math., vol. 1, no. 3, pages 335–380, 2004.

[Latouche *et al.* 2009] Pierre Latouche, Etienne Birmelé and Christophe Ambroise. *Overlapping stochastic block models.* arXiv preprint arXiv:0910.2098, 2009.

[Lee *et al.* 2010] Victor E Lee, Ning Ruan, Ruoming Jin and Charu Aggarwal. *A survey of algorithms for dense subgraph discovery.* In Managing and Mining Graph Data, pages 303–336. Springer, 2010.

[Lee *et al.* 2012] Chul-Ho Lee, Xin Xu and Do Young Eun. *Beyond Random Walk and Metropolis-hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling.* In Proc. ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, London, UK, 2012.

[Levin *et al.* 2009] David Asher Levin, Yuval Peres and Elizabeth Lee Wilmer. Markov chains and mixing times. American Mathematical Soc., 2009.

[Lieb 1973] Elliott H Lieb. *Convex trace functions and the Wigner-Yanase-Dyson conjecture.* Advances in Mathematics, vol. 11, no. 3, pages 267–288, 1973.

[Litvak *et al.* 2007] Nelly Litvak, Werner RW Scheinhardt and Yana Volkovich. *In-degree and PageRank: why do they follow similar power laws?* Internet mathematics, vol. 4, no. 2-3, pages 175–198, 2007.

[Louis 2015] Anand Louis. *Hypergraph markov operators, eigenvalues and approximation algorithms.* In Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, pages 713–722. ACM, 2015.

[Lovász & Pelikán 1973] László Lovász and József Pelikán. *On the eigenvalues of trees.* Periodica Mathematica Hungarica, vol. 3, no. 1-2, pages 175–182, 1973.

[Lovász 1993] L Lovász. *Random walks on graphs: A survey.* Comb. Paul Erdos is Eighty, vol. 2, no. Volume 2, pages 1–46, 1993.

[Lu & Peng 2013] Linyuan Lu and Xing Peng. *High-Order Random Walks and Generalized Laplacians on Hypergraphs.* Internet Math., vol. 9, no. 1, pages 3–32, 2013.

[Martinsson 2013] Anders Martinsson. *Lovasz $\theta$ function , SVMs and Finding Dense Subgraphs.* J. Mach. Learn. Res., vol. 14, pages 3495–3536, 2013.

[Massoulié *et al.* 2006] Laurent Massoulié, Erwan Le Merrer, Anne-Marie Kermarrec and Ayalvadi Ganesh. *Peer Counting and Sampling in Overlay Networks: Random Walk Methods.* In Proc. ACM Annual Symposium on Principles of Distributed Computing (PODC), Denver, Colorado, USA, 2006.

[Massoulié 2014] Laurent Massoulié. *Community detection thresholds and the weak Ramanujan property.* In Proceedings of the 46th Annual ACM Symposium on Theory of Computing, pages 694–703. ACM, 2014.

[Mezard & Montanari 2009] Marc Mezard and Andrea Montanari. Information, physics, and computation. Oxford University Press, 2009.

[Mifflin *et al.* 2004] T.L. Mifflin, C. Boner, G.A. Godfrey and J. Skokan. *A random graph model for terrorist transactions.* In 2004 IEEE Aerosp. Conf. Proc., volume 5, pages 3258–3264. IEEE, 2004.

[Miller *et al.* 2010] Benjamin Miller, Nadya Bliss and Patrick J Wolfe. *Subgraph detection using eigenvector L1 norms.* In Advances in Neural Information Processing Systems, pages 1633–1641, 2010.

[Miller *et al.* 2015a] Benjamin A Miller, Michelle S Beard, Patrick J Wolfe and Nadya T Bliss. *A spectral framework for anomalous subgraph detection.* Signal Processing, IEEE Transactions on, vol. 63, no. 16, pages 4191–4206, 2015.

[Miller *et al.* 2015b] Benjamin A Miller, Stephen Kelley, Rajmonda S Caceres and Steven T Smith. *Residuals-based subgraph detection with cue vertices.* In 2015 49th Asilomar Conference on Signals, Systems and Computers, pages 1530–1534. IEEE, 2015.

[Mitra 2009] Pradipta Mitra. *Entrywise bounds for eigenvectors of random graphs.* Electron. J. Comb., vol. 16, no. 1, page R131, 2009.

[Mohar & Woess 1989] Bojan Mohar and Wolfgang Woess. *A survey on spectra of infinite graphs*. Bull. London Math. Soc, vol. 21, no. 3, pages 209–234, 1989.

[Montanari 2015] Andrea Montanari. *Finding one community in a sparse graph*. Journal of Statistical Physics, vol. 161, no. 2, pages 273–299, 2015.

[Mossel & Xu 2016] Elchanan Mossel and Jiaming Xu. *Local Algorithms for Block Models with Side Information*. In ITCS '16, pages 71–80, New York, New York, USA, jan 2016. ACM Press.

[Mossel *et al.* 2012] Elchanan Mossel, Joe Neeman and Allan Sly. *Stochastic block models and reconstruction*. arXiv preprint arXiv:1202.1499, 2012.

[Nadakuditi & Newman 2012] Raj Rao Nadakuditi and M. E J Newman. *Graph spectra and the detectability of community structure in networks*. Phys. Rev. Lett., vol. 108, no. 18, pages 1–5, 2012.

[Nazi *et al.* 2015] Azade Nazi, Zhuojie Zhou, Saravanan Thirumuruganathan, Nan Zhang and Gautam Das. *Walk, not wait: Faster sampling over online social networks*. Proc. VLDB Endow., vol. 8, no. 6, pages 678–689, 2015.

[Newman & Girvan 2004] Mark E J Newman and Michelle Girvan. *Finding and evaluating community structure in networks*. Phys. Rev. E, vol. 69, no. 2, page 26113, 2004.

[Newman 2003] Mark EJ Newman. *The structure and function of complex networks*. SIAM review, vol. 45, no. 2, pages 167–256, 2003.

[Newman 2006] Mark EJ Newman. *Modularity and community structure in networks*. Proceedings of the national academy of sciences, vol. 103, no. 23, pages 8577–8582, 2006.

[Newman 2013] Mark EJ Newman. *Spectral methods for community detection and graph partitioning*. Physical Review E, vol. 88, no. 4, page 042822, 2013.

[Nica & Speicher 2006] Alexandru Nica and Roland Speicher. Lectures on the combinatorics of free probability, volume 13. Cambridge University Press, 2006.

[Nummelin 2002] Esa Nummelin. *MC's for MCMC'ists*. Int. Stat. Rev., vol. 70, no. 2, pages 215–240, 2002.

[Olshevsky & Tsitsiklis 2009] Alex Olshevsky and John N Tsitsiklis. *Convergence speed in distributed consensus and averaging*. SIAM Journal on Control and Optimization, vol. 48, no. 1, pages 33–55, 2009.

[O'Rourke *et al.* 2016] Sean O'Rourke, Van Vu and Ke Wang. *Eigenvectors of random matrices: A survey*. Journal of Combinatorial Theory, Series A, vol. 144, pages 361–442, 2016.

[Page *et al.* 1997] Larry Page, S Brin, R Motwani and T Winograd. *PageRank: Bringing order to the web*. Technical report, Stanford Digital Libraries Working Paper, 1997.

[Page *et al.* 1999] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. *The PageRank citation ranking: Bringing order to the web*. Technical report, Stanford InfoLab, 1999.

[Pandurangan *et al.* 2002] Gopal Pandurangan, Prabhakar Raghavan and Eli Upfal. *Using pagerank to characterize web structure*. In International Computing and Combinatorics Conference, pages 330–339. Springer, 2002.

[Penrose 2003] Mathew Penrose. Random geometric graphs. Oxford University Press, 2003.

[Ribeiro & Towsley 2010] Bruno Ribeiro and Don Towsley. *Estimating and sampling graphs with multidimensional random walks*. In Proc. ACM SIGCOMM Internet Measurement Conference (IMC), Melbourne, Australia, November 2010.

[Robert & Casella 2013] Christian Robert and George Casella. Monte Carlo statistical methods. Springer Science & Business Media, 2013.

[Roberts & Rosenthal 2004] Gareth O Roberts and Jeffrey S Rosenthal. *General state space Markov chains and MCMC algorithms*. Probability Surveys, vol. 1, pages 20–71, 2004.

[Rohe *et al.* 2011] Karl Rohe, Sourav Chatterjee and Bin Yu. *Spectral clustering and the high-dimensional stochastic blockmodel*. Ann. Stat., pages 1878–1915, 2011.

[Ross 2013] Sheldon M Ross. Applied probability models with optimization applications. Courier Corporation, 2013.

[Rudelson & Vershynin 2015] Mark Rudelson and Roman Vershynin. *No-gaps delocalization for general random matrices*. arXiv Prepr. arXiv1506.04012, page 45, 2015.

[Rump 2006] Siegfried M Rump. *Eigenvalues, pseudospectrum and structured perturbations*. Linear algebra and its applications, vol. 413, no. 2-3, pages 567–593, 2006.

[Saad 1992] Youcef Saad. Numerical methods for large eigenvalue problems, volume 158. SIAM, 1992.

[Saade *et al.* 2015] Alaa Saade, Marc Lelarge, Florent Krzakala and Lenka Zdeborova. *Spectral detection in the censored block model*. In 2015 IEEE Int. Symp. Inf. Theory, pages 1184–1188. IEEE, jun 2015.

[Salganik & Heckathorn 2004] Matthew J Salganik and Douglas D Heckathorn. *Sampling and estimation in hidden populations using respondent-driven sampling*. Sociol. Methodol., vol. 34, no. 1, pages 193–240, 2004.

[Shuman *et al.* 2013] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega and Pierre Vandergheynst. *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains*. IEEE Signal Processing Magazine, vol. 30, no. 3, pages 83–98, 2013.

[Silva & Willett 2009] Jorge Silva and Rebecca Willett. *Hypergraph-based anomaly detection in very large networks*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

[Silverstein 1990] Jack W Silverstein. *Weak convergence of random functions defined by the eigenvectors of sample covariance matrices*. The Annals of Probability, pages 1174–1194, 1990.

[Smith *et al.* 2014] Steven Thomas Smith, Edward K Kao, Kenneth D Senne, Garrett Bernstein and Scott Philips. *Bayesian discovery of threat networks*. IEEE Transactions on Signal Processing, vol. 62, no. 20, pages 5324–5338, 2014.

[Spielman 2007] Daniel A Spielman. *Spectral graph theory and its applications.* In Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on, pages 29–38. IEEE, 2007.

[Sussman *et al.* 2012] Daniel L Sussman, Minh Tang, Donniell E Fishkind and Carey E Priebe. *A consistent adjacency spectral embedding for stochastic blockmodel graphs.* Journal of the American Statistical Association, vol. 107, no. 499, pages 1119–1128, 2012.

[Tao & Vu 2011] Terence Tao and Van Vu. *Random matrices: Universality of local eigenvalue statistics.* Acta Math., vol. 206, no. 1, pages 127–204, 2011.

[Tao & Vu 2012] Terence Tao and Van Vu. *Random matrices: Universal properties of eigenvectors.* Random Matrices Theory Appl., vol. 1, no. 01, page 1150001, 2012.

[Tao 2012] Terence Tao. Topics in random matrix theory, volume 132. American Mathematical Society Providence, RI, 2012.

[Tropp 2012a] Joel A Tropp. *User-friendly tail bounds for sums of random matrices.* Foundations of computational mathematics, vol. 12, no. 4, pages 389–434, 2012.

[Tropp 2012b] Joel A Tropp. *User-friendly tools for random matrices: An introduction.* Technical report, DTIC Document, 2012.

[Vadhan *et al.* 2012] Salil P Vadhan*et al.* *Pseudorandomness.* Foundations and Trends® in Theoretical Computer Science, vol. 7, no. 1–3, pages 1–336, 2012.

[Vershynin 2011] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices.* arXiv Prepr. arXiv1011.3027, pages 1–66, 2011.

[Volkovich & Litvak 2010] Yana Volkovich and Nelly Litvak. *Asymptotic analysis for personalized web search.* Advances in Applied Probability, vol. 42, no. 02, pages 577–604, 2010.

[Volz & Heckathorn 2008] Erik Volz and Douglas D Heckathorn. *Probability based estimation theory for respondent driven sampling.* J. Off. Stat., vol. 24, no. 1, page 79, 2008.

[Von Luxburg 2007] Ulrike Von Luxburg. *A tutorial on spectral clustering.* Stat. Comput., vol. 17, no. 4, pages 395–416, 2007.

[Vu 2007] VanH. H. Vu. *Spectral norm of random matrices.* Combinatorica, vol. 27, no. 6, pages 721–736, 2007.

[Wang *et al.* 2015] Xiaohan Wang, Pengfei Liu and Yuantao Gu. *Local-set-based graph signal reconstruction.* IEEE Transactions on Signal Processing, vol. 63, no. 9, pages 2432–2444, 2015.

[Wigner 1955] Eugene P. Wigner. *Characteristic Vectors of Bordered Matrices With Infinite Dimensions.* Annals of Mathematics, vol. 62, no. 3, pages 548–564, 1955.

[Wigner 1958] Eugene P. Wigner. *On the Distribution of the Roots of Certain Symmetric Matrices.* Annals of Mathematics, vol. 67, no. 2, pages 325–327, 1958.

[Wigner 1967] Eugene P Wigner. *Random matrices in physics.* SIAM review, vol. 9, no. 1, pages 1–23, 1967.

[Yang & Leskovec 2015] Jaewon Yang and Jure Leskovec. *Defining and evaluating network communities based on ground-truth*. Knowledge and Information Systems, vol. 42, no. 1, pages 181–213, 2015.

[Yeh *et al.* 2009] Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre and Aitor Soroa. *WikiWalk: random walks on Wikipedia for semantic relatedness*. In Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, pages 41–49. Association for Computational Linguistics, 2009.

[Zhao *et al.* 2012] Yunpeng Zhao, Elizaveta Levina, Ji Zhu*et al. Consistency of community detection in networks under degree-corrected stochastic block models*. The Annals of Statistics, vol. 40, no. 4, pages 2266–2292, 2012.

[Zhou *et al.* 2004] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston and Bernhard Schölkopf. *Learning with local and global consistency*. Advances in neural information processing systems, vol. 16, no. 16, pages 321–328, 2004.

[Zhou *et al.* 2007] Dengyong Zhou, Jiayuan Huang and Bernhard Schölkopf. *Learning with Hypergraphs: Clustering, Classification, and Embedding*. Adv. Neural Inf. Process. Syst. 19, vol. 19, no. Figure 1, pages 1601–1608, 2007.

[Zhu *et al.* 2003] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty*et al. Semi-supervised learning using gaussian fields and harmonic functions*. In ICML, volume 3, pages 912–919, 2003.