University of Nottingham

School of Mathematical Sciences

Some Topics in Topological Data Analysis

Yang DI

A thesis submitted to the University of Nottingham for the degree of
Doctor of Philosophy

September 2021

Abstract

In recent years there has been growing interest within Statistics in topological aspects of random objects, one important direction being Topological Data Analysis (TDA) and the associated concept of Persistent Homology. This research aims to investigate both theoretical and computational aspects of TDA. In the first strand of this research the aim is to generalize the central limit theorem (CLT) given by Kahle and Meckes (2013, 2015) for Betti numbers in Erdős-Rényi random graphs, to a CLT for Betti numbers in the stochastic block model. In addressing this problem, we have provided results on the spectral structure of the adjacency matrix and the normalized graph Laplacian in stochastic block models which appear to be new. The second strand of the research is to investigate numerically the relationship between the topological summaries computed under the full sample and under subsamples. Subsampling often needs to be considered because existing computational algorithms for TDA tend to break down for larger sample sizes as computational demands grow rapidly with sample size. One important finding is that subsampling which exploits existing structure in the data is likely to do much better than purely random subsampling. In this PhD thesis, numerical results are given for various types of simulated data through to real datasets.

Acknowledgements

It would not be possible for me to get through the four-year PhD study and complete this thesis without the guidance, help and support from many people.

First of all, I would like to express my most sincere gratitude to my supervisors, Prof. Andrew Wood, Prof. Huiling Le and Dr. Karthik Bharath for their guidance, patience and encouragement throughout the development of this research project. Without their extensive knowledge as a teacher, their scientific experience as a researcher and their continuous support, I would not have been able to find the right way of my PhD and enjoy this four-year study at University of Nottingham.

Secondly, I am grateful to my close friend, Dr. Jing Ji, who eased my anxiety and alleviated my stress when I was depressed and felt negative.

Last but not least, I would like to dedicate this thesis to my family for their unconditional love and emotional support throughout my time in the UK. They were always there whenever I needed them. It would not have been possible for me to finish this project without their understanding and encouragement.

CONTENTS

1

INTRODUCTION

This thesis is concerned with aspects of the exciting new approach known as Topological Data Analysis (TDA). In Section 1.1, some background to TDA is given, while in Section 1.2 the aims of the thesis are outlined. In Section 1.3 the contents of the thesis are indicated.

1.1 TOPOLOGICAL DATA ANALYSIS: BACKGROUND

TDA is a new approach to data analysis with roots in the area of Pure Mathematics known as Topology. Topology is concerned with concepts such continuity, connectedness and shape and focuses on properties of an object which remain invariant under continuous transformations. This subject has evolved since the early 1900s. The basic object of study in Topology, topological spaces, have certain precisely defined properties which are described later. Topology has two main branches: point-set Topology which has close connections to analysis; and Algebraic Topology, which uses Algebra, especially Group Theory, to study the structure of topological spaces. We will have more to say about relevant aspects of Algebraic Topology in Section 1.2.

Topology, especially Algebraic Topology, has a reputation for being one of the most abstract and difficult subjects in Pure Mathematics. Nevertheless, in recent years Computational Topology, especially TDA, has developed rapidly as a field and has aroused a high level of interest among mathematicians, statisticians and computational scientists. These developments have been made possible due to the rapid advances in computer technology in recent decades, plus the fact that it is feasible to implement some of the key tools in Algebraic Topology, such as the calculation of Betti numbers (see Section 1.2) in algorithmic form.

A central concept in TDA is Persistent Homology. It is difficult to explain this concept in non-technical terms; a precise mathematical definition of this concept is given in Chapter 2. Persistent Homology allows one to distinguish between "topological signal" and "topological noise", where typically our interest lies in identifying topological signal. There is a close analogy with distinguishing between signal and noise in statistical settings. Two key outputs from Persistent Homology are barcodes and persistence diagrams; these outputs are defined in Chapter 2.

There are some excellent books and papers on TDA. One of the books we have found most useful and accessible is by Edelsbrunner and Harer (2010). Although much of the book is quite mathematical in its presentation it is aimed at a general mathematical audience and therefore is not too specialised. Interesting applications are given in the book to Gene Expression Data, Protein Docking, Image Segmentation and Root Architectures.

The paper by Ghrist (2008) also gives a valuable summary of TDA with a focus on barcodes. Carlsson (2009) is another important contribution to the TDA literature which lays out key ideas with a focus on image analysis. Another useful source of information about TDA, especially useful for a probability and statistics audience, are the four short papers by Adler (2014a,b,c, 2015). These papers present TDA in a broad context and highlight barcodes as mathematical objects worth further study in the future.

There are several computing packages available for performing the calculations required for TDA. These packages have played a crucial role in making TDA accessible to a broad range of researchers in different fields and in popularising TDA. The packages include R package TDA written by Fasy et al. (2019) at `https://CRAN.R-project.org/package=TDA`; MATLAB package JavaPlex written by Tausz et al. (2014) at `http://appliedtopology.github.io/javaplex/`. These packages can take point clouds as input data while Javaplex can also take graph as a raw data. Moreover, all packages generate persistent diagrams and barcodes as the basic output. At the present time, TDA package in R can also provide persistent landscape as an additional output results.

1.2 AIMS OF THE THESIS

This thesis has the following goals. The first is to explore the possibility of extending the Kahle and Meckes (2013, 2015) Central Limit Theorem (CLT)

for Betti numbers in Erdős-Rényi random graphs to stochastic block model, where several types of vertex exist. Roughly speaking, Betti numbers count topological features of different types. Specifically, β_0, the first Betti number, counts the number of connected components in a topological space X. The second Betti number, β_1, counts the number of "1-dimensional holes" in X. The third Betti number, β_2, counts the number of "2-dimensional holes" in X, and so on. A formal definition of of the Betti numbers in terms of dimensions of Homology Groups is given in Chapter 2.

In the Kahle and Meckes (2013, 2015) CLT, the relevant asymptotic regime is such that the number of vertices, N, goes to infinity, the probability of two typical vertices being connected by an edge goes to 0 at a suitable rate as N goes to infinity. Further details are given in Chapters 2, 3 and 4.

Our work on the problem of extending Kahle and Meckes (2013, 2015), CLT for Erdős-Rényi graphs to stochastic block models indicates that, at best, there are serious difficulties in extending the proof to stochastic block models. At worst, the CLT may only extend in rather restricted circumstances. An important part of the Kahle and Meckes (2013, 2015) proof is an application of the so-called spectral gap theorem. The spectral gap theorem, in the form used by Kahle and Meckes (2013, 2015), states that the difference between the most extreme eigenvalue and the second most extreme eigenvalue of a certain matrix (the normalised graph Laplacian, defined in Chapter 2) becomes large in a suitable sense as N goes to infinity. It turns out that the spectral gap theorem does typically hold in a suitable form with stochastic block models. This is shown in Chapter 3, where we present results on the asymptotic spectral structure of the so-called adjacency matrix and normalised graph Laplacian. So far as we are aware, the results in Chapter 3 are new.

In Chapter 4, the focus is on proving as many of the results as possible that generalize from the Erdős-Rényi model to the stochastic block model. Most of the results for the Erdős-Rényi model, with the exception of the spectral gap theorem, go through to the stochastic block model.

Chapters 3 and 4 have quite a theoretical focus, though we believe that the problems addressed are of considerable interest as there is still a shortage of theoretical results relating to TDA. In Chapter 5, our focus is very different and much more computational. We have found in numerical examples that we have used TDA software on, the sample size does not need to be very large for the TDA algorithms to break down, in the sense that

the programmes do not finish in reasonable time and memory limits are exceeded. This raised the need for some kind of subsampling. However, subsampling can lead to problems because there is no guarantee that subsamples have the same topological and statistical structure as the original samples. In some situations, however, e.g. with the Brain Artery Tree Data considered in Chapter 5, there is already a long of structure in the data. The main purpose of this chapter is to investigate the usefulness of subsampling with using TDA with larger sample sizes. A key idea in this chapter is, when possible, to subsample from some kind of skeleton of the original data so that to some extent at least some of the structure is retained in subsamples. The numerical results in this chapter indicate that, when it can be applied, structured subsampling does a much better job than purely random subsampling.

1.3 CONTENTS OF THE THESIS

In this section we describe the contents of the thesis.

Chapter 2 focuses primarily on definitions and results from various areas of mathematics, probability and statistics which are used later in the thesis. The material selected does not form a coherent whole. The aim is to make the thesis as self-contained as possible, from a statistics and probability point of view. Results from statistics and probability include definitions of the Kolmogorov-Smirnov and Cramer-von Mises goodness-of-fit statistics, along with permutation tests, which all appear in Chapter 5 although the settings in which they are used are somewhat non-standard. Results and concepts from probability include Chernoff bounds, Bernstein's inequality and uniform integrability. In Chapter 3 we make use of a dominated convergence type result, based on uniform integrability and convergence in probability that is given in e.g. Williams (1991). Basic material on linear algebra and vector spaces is also included in Chapter 2, as is elementary material on Erdős-Rényi random graphs. Most of the remainder of Chapter 2 aims to provide an elementary and, as far as possible self-contained, account of Persistent Homology. Chapter 2 concludes with brief introductions to results in a few papers that have played an important role in motivating the work of this thesis, especially in Chapter 3 and Chapter 4, including Kahle and Meckes (2013, 2015) and some of the papers they reference.

In Chapter 3 the asymptotic spectral structure (i.e. the eigenvalues and eigenvectors) of the adjacency matrix and the normalised graph Laplacian, defined in Chapter 2, are derived. The results in this chapter appear to be new.

Chapter 4 extends many of the auxiliary results derived by Kahle and Meckes (2013, 2015) in the Erdős-Rényi model case, to the stochastic block model case, an important omission being the extension of the spectral gap theorem.

In Chapter 5, we investigate and compare purely random subsampling with structured subsampling. In the latter approach, the idea is that one should try to exploit structure in the dataset when designing the subsampling algorithm. In the Brain Artery Tree Data considered in Chapter 5, the structured subsampling approach worked well and proved to be far superior to purely random subsampling.

Finally, in Chapter 6, conclusions and possibilities of future work are described.

2

BACKGROUND KNOWLEDGE

2.1 INTRODUCTION

In this chapter, we present technical background which is relevant to later chapters in this thesis. In Section 2.2, some important topics in probability and statistics are covered. A review of linear algebra is given in Section 2.3. In Section 2.4, random graph models are reviewed. Section 2.5 covers the Persistent Homology. In Section 2.6, some key results from different papers are covered. Finally, a review of some more advanced literature on the Persistent Homology is given in Section 2.7.

2.2 BACKGROUND ON PROBABILITY AND STATISTICS

In this section, some relevant probability and statistics results are listed.

2.2.1 *Some Test Statistics*

KOLMOGOROV–SMIRNOV TEST

The Kolmogorov–Smirnov (KS) test was introduced by Kolmogorov (1933) and Smirnov (1948) for testing goodness-of-fit. Let $\{x_{1,\ldots,}x_n\}$ be independent identically distributed (iid) random variables from F, then the empirical cumulative distribution function (cdf) is defined as

$$F_n(x) = \frac{card\{i : x_i \leq t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{x_i \leq t\}, \tag{2.2.1}$$

where $\mathbb{I}$ is the indicator function.

The KS statistic for testing $H_0 : F = F_0$ where F_0 is a given continuous cdf $F(x)$ is

$$d_n = \sup_{-\infty \leq x \leq \infty} |F_n(x) - F_0(x)|.$$

If $\{x_i\}$ comes from the distribution with cdf $F_0(x)$, then $d_n \to 0$ as $n \to \infty$.

The KS test may also be extended to test whether two independent samples are from the same or different distributions. In this case, let $\{x_i : i = 1, \ldots, n\}$ be iid random variables from F and let $\{y_i : i = 1, ..., m\}$ be iid random variables from G, then the KS statistic is defined as

$$d_{nm} = \sup_{-\infty \leq x \leq \infty} |F_n(x) - G_m(x)|,$$

where F_n and G_m are the empirical cdf for two independent samples respectively and the null hypothesis is $H_0 : F = G$.

The null hypothesis test that $H_0 : F = G$ is rejected at level α if

$$d_{nm} > c(\alpha) \sqrt{\frac{n+m}{nm}},$$

where $c(\alpha) = \sqrt{-\frac{1}{2} \ln\left(\left(\frac{\alpha}{2}\right)\right)}$ is given by Knuth (1997).

CRAMER-VON MISES TEST

The Cramer-von Mises (CvM) test was presented by Cramér (1928) and Von Mises (1928). Let $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ be iid random variables from F arranged in an increasing order. Then the one-sample CvM test statistic for testing $H_0 : F = F_0$, where F_0 is specified is defined as

$$T = n \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 \, dF_0(x) = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F_0\left(x_{(i)}\right) \right]^2,$$

where F_n is the empirical cdf for the $x_{(i)}$ and F_0 is the theoretical distribution.

If T is larger than the critical value, then the null hypothesis $H_0 : F = F_0$ is rejected.

Similarly to the KS test, the CvM test can also be extended to a two-sample CvM test.

If there is a second sample with order statistics $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(m)}$, then let G_m denote the empirical cdf for $y_{(j)}$, and let H_{n+m} be the empirical cdf of combined sample

$$\{z_1, ..., z_n, z_{n+1}, ..., z_{n+m}\} = \{x_1, ..., x_n, y_1, ..., y_m\}.$$

By arranging the z_i in increasing order, write $z_{(r_i)} = x_{(i)}$ and $z_{(s_j)} = y_{(j)}$ where $i = 1, ..., n$ and $j = 1, ..., m$, where r_i is the rank of $x_{(i)}$ and s_j is the rank of $y_{(j)}$ in the pooled sample. Then the test statistic for two-sample CvM is defined as

$$T = \frac{nm}{n+m} \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 dH_{n+m}(x) = \frac{U}{nm(n+m)} - \frac{4nm-1}{6(n+m)},$$

where

$$U = n \sum_{i=1}^{n} (r_i - i)^2 + m \sum_{j=1}^{m} (s_j - j)^2.$$

If T is larger than the critical value, then the null hypothesis that $H_0 : F = G$ is rejected (Anderson, 1962).

PERMUTATION TEST

The permutation test was suggested by Fisher (1936). Assume there are two samples $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_m\}$ with cdf's F_n and G_m respectively. If there is a suitable test statistic T_0 calculated jointly from $\{x_i\}$ and $\{y_j\}$, then combine the two samples as follows:

$$\{z_1, ..., z_n, z_{n+1}, ..., z_{n+m}\} = \{x_1, ..., x_n, y_1, ..., y_m\}.$$

The combined group is randomly allocated into two groups $\left\{r_1^{(l)}, ..., r_n^{(l)}\right\}$ and $\left\{s_1^{(l)}, ..., s_m^{(l)}\right\}$, where $l = 1, ..., M$ and M is the number of permutation considered. Then for $l = 1, ..., M$, the new test statistic T_l is calculated using the same method as T_0. The resulting p-value for permutation test is defined as

$$p = \frac{1}{M} \sum_{l=1}^{M} \mathbb{I}\{T_l > T_0\},$$

where $\mathbb{I}$ is the indicator function.

The KS, CvM and permutation test statistics are used in Chapter 5. However, as pointed out in Chapter 5, these statistics are used in a non-standard way because the relevant sample are non-iid.

2.2.2 *Probability Approximation*

Some important results from probability theory that are used later are now presented.

CHERNOFF BOUND (CHERNOFF ET AL., 1952)
Suppose $X_i \overset{indep}{\sim} Bernoulli\,(p_i)$ and write $X = \sum_{i=1}^{n} X_i$ and $\mu = \sum_{i=1}^{n} p_i$. Then,

$$\begin{cases} P\left[X \geq (1+\delta)\,\mu\right] \leq \exp\left\{-\frac{\mu\delta^2}{2+\delta}\right\} & \delta > 0 \\ P\left[X \leq (1-\delta)\,\mu\right] \leq \exp\left\{-\frac{\mu\delta^2}{2}\right\} & 0 < \delta < 1. \end{cases}$$

Then for $\delta \in (0,1)$,

$$P\left(|X-\mu| > \delta\mu\right) \leq 2\exp\left\{-\frac{\mu\delta^2}{3}\right\}. \tag{2.2.2}$$

BERNSTEIN'S INEQUALITY (BERNSTEIN, 1924)
Suppose X_i are independent, zero-mean random variables such that $|X_i| \leq M$ almost surely (a.s.) for all $1 \leq i \leq n$. Then for any $t \geq 0$,

$$P\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n} E\left(X_i^2\right) + \frac{2}{3}Mt}\right\}. \tag{2.2.3}$$

The term a.s. is defined as follows. An event E is said to occur almost surely (a.s.) if $P(E) = 1$.

2.2.3 *Uniformly Integrability*

Definition 2.2.1. (Definition 6.7 by Williams (1991)) For $1 \leq t < \infty$, $X \in \mathscr{L}^t$ if

$$E\left(|X|^t\right) \leq \infty.$$

Definition 2.2.2. (Definition 13.2 by Williams (1991)) A class H of random variables is called uniformly integrable (UI) if given $\varepsilon > 0$, there exists $C \in [0,\infty)$ such that

$$E\left(|X|\,;|X| > C\right) = \int_{\{x:|x|>C\}} |X|\,dX < \varepsilon$$

for all X.

Proposition 2.2.3. *(Proposition 13.3(a) by Williams (1991)) Suppose that H is a class of random variables which is bounded in $\mathscr{L}^t$ for some $t > 1$; thus, for some $C \in [0, \infty)$,*

$$\sup_{X \in H} E\left(|X|^t\right) < C.$$

Then H is a uniformly integrable class of random variables.

Theorem 2.2.4. *(Theorem 13.7 by Williams (1991)) Let $(X_n)_{n \geq 1}$ be a sequence in $\mathscr{L}^1$ and let $X \in \mathscr{L}^1$. Then $E(|X_n - X|) \to 0$ if and only if the following two conditions hold*

1. *$X_n \to X$ in probability*
2. *the sequence $(X_n)_{n \geq 1}$ is uniformly integrable.*

2.3 BACKGROUND ON LINEAR ALGEBRA

Define $P = \left\{\mathbf{v}_i \in \mathbb{R}^d : i = 1, \ldots, N\right\}$. A point $x = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i$ is an affine combination of the $\mathbf{v}_i$ if $\sum_{i=1}^{N} \lambda_i = 1$. If in addition λ_i is non-negative for all i, then $x = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i$ is a convex combination. The convex hull is the set of all such convex combinations. Moreover, k points $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k$ are said to be affinely independent if and only if $\mathbf{v}_i - \mathbf{v}_1$, $2 \leq i \leq k$, are linearly independent.

Let $T : V \to W$ be a linear map where V and W are two vector spaces. Then the kernel and image are defined as $\text{Ker}(T) = \{v \in V | T(v) = 0\}$ and $\text{Im}(T) = \{T(v) | v \in V\}$. The dimension of a vector space V is the maximum number of linearly independent vectors in V. We write $\dim(V)$ for the dimension of V.

Then rank–nullity theorem states that

$$\dim(\text{Im}(T)) + \dim(\text{Ker}(T)) = \dim(V) \tag{2.3.1}$$

where $\text{rank}(T) = \dim(\text{Im}(T))$.

An $n \times m$ matrix A is defined as being in Smith normal form if and only if

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & \dots & 0 \\ \vdots & 0 & a_{kk} & 0 & \cdots & 0 \\ 0 & \dots & 0 & \dots & \dots & 0 \\ \vdots & \dots & \vdots & \dots & \dots & 0 \\ 0 & \cdots & 0 & \dots & \dots & 0 \end{pmatrix} = \begin{bmatrix} \text{diag}\{a_{11}, ..., a_{kk}\} & \mathbf{0}_{k,m-k} \\ \mathbf{0}_{n-k,k} & \mathbf{0}_{n-k,m-k} \end{bmatrix}$$

i.e. the only potentially non-zero elements are the diagonal elements a_{ii}, $i = 1, ..., k$.

2.4 BACKGROUND ON THE RANDOM GRAPH MODEL

A graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is a mathematical structure consisting of two sets, $\mathcal{V}$ and $\mathcal{E}$. $\mathcal{V}$ is a non-empty set, whose elements are called the vertices or nodes, and $\mathcal{E}$ is the edge set, where $\mathcal{E}$ is a subset of $\mathcal{V} \times \mathcal{V}$. The elements of $\mathcal{E}$ are called edges. If $e = (u, v) \in \mathcal{E}$ where $u, v \in \mathcal{V}$, then u and v are said to be adjacent. The graph is said to be undirected when $(v, u) \in \mathcal{E}$ if and only if $(u, v) \in \mathcal{E}$, for all $u, v \in \mathcal{V}$.

An edge which connects a vertex to itself is called a loop. If there is potentially more than one edge connecting two different vertices, the graph is said to be a multi-edge graph. A graph without loops or parallel edges is called a simple graph. In this thesis, we only consider simple undirected graphs.

Throughout this thesis, the number of elements of $\mathcal{V}$, $card(\mathcal{V})$, is denoted by N.

The degree of a vertex,

$$\deg(v) = card\{u : (u, v) \in \mathcal{E}\},$$

is the number of edges with an end-point in that vertex.

A graph is said to be connected if and only if there are no isolated vertices, i.e. there is no vertices with degree 0.

Two graphs $\mathcal{G}_1(\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2(\mathcal{V}_2, \mathcal{E}_2)$ are said to be isomorphic if there is a one-one relationship between $\mathcal{V}_1$ and $\mathcal{V}_2$

$$f : \mathcal{V}_1 \to \mathcal{V}_2$$

such that

$$(u_1, v_1) \in \mathcal{E}_1 \Leftrightarrow (f(u_1), f(v_1)) \in \mathcal{E}_2.$$

For a graph, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with block structure, there exists a partition of $\mathcal{V}$ into vertices of different types, i.e. there are blocks, $\mathcal{V}_1, ..., \mathcal{V}_\zeta$ where

$$\begin{gathered} \bigcup_{i=1}^{\zeta} \mathcal{V}_i = \mathcal{V} \\ \mathcal{V}_i \bigcap \mathcal{V}_j = \emptyset \end{gathered}$$

where $i \neq j$ and $\emptyset$ is the empty set. We called this a block model graph.

In graph theory, there are several different ways to represent a graph mathematically. Three of these representations are introduced here. Consider a simple undirected graph with N vertices and each vertex is labeled as $v_1, ..., v_N$.

The adjacency matrix, denoted $\mathbf{A} = \{a_{ij}\}_{1 \leq i \leq j \leq N}$ where a_{ij} is an indicator for edge (i, j)

$$a_{ij} = \begin{cases} 1 & \text{if there is an edge between } i \text{ and } j \\ 0 & \text{otherwise.} \end{cases}$$

For a simple undirected graph, the adjacency matrix is symmetric with $a_{ij} = a_{ji}$ for all $i \neq j$ and $a_{ii} = 0$ for all i.

Another matrix of interest is what is called the normalized graph Laplacian.

To define this, let $\mathbf{A}$ be the adjacency matrix and $\mathbf{D} = \{d_{ij}\}$ be the degree matrix of a random graph $\mathcal{G}$, where

$$d_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Then define

$$\mathbf{A}_{norm} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}.$$

Then the normalized graph Laplacian is defined as

$$\mathbf{L} = \mathbf{L}(\mathcal{G}) = \mathbf{I}_N - \mathbf{A}_{norm}$$

where

$$l_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & \text{if } i \neq j \text{ and } (v_i, v_j) \in \mathcal{E} \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbf{I}_N$ is the $N \times N$ identity matrix (Kahle, 2014).

In Chapter 3 and Chapter 4 of this thesis, random block models are subjects of primary interest. The most extensively studied one is called the Erdős-Rényi model, in which there is just one type of vertex.

The Erdős-Rényi model (ERM) was introduced by Erdős and Rényi (1959, 1960, 1961) and Gilbert (1959). The ERM, denoted as $\mathcal{G}(N, p)$, is an undirected graph with N vertices and each edge included independently of the others with probability $0 < p < 1$.

Stochastic block model (SBM) were discussed by Kolaczyk (2009). SBM, denoted as $\mathcal{G}((N_r), (p_{rs}), \zeta)$ is an undirected graph with N vertices and ζ blocks. Define

$$N = \sum_{r=1}^{\zeta} N_r$$

where $N_r = card(\mathcal{V}_r)$ and

$$p_{rs} = P[(u, v) \in \mathcal{E}]$$

where $u \in \mathcal{V}_r$ and $v \in \mathcal{V}_s$.

Without lost of generality, in this thesis, we often assume $N_1 \leq N_2 \leq ... \leq N_\zeta$.

In this thesis, the main results focus on the SBM with $\mathcal{G}((N_r), (p_{rs}), \zeta)$. However, there are some parts of the thesis where the focus is on the ERM, $\mathcal{G}(N, p)$ especially when discussing the CLT results of Kahle and Meckes (2013, 2015).

2.5 PERSISTENT HOMOLOGY

2.5.1 *Introduction*

In this section, we discuss relevant background and concepts of Persistent Homology. These concepts and ideas are applied later in the thesis. This section follows the book by Edelsbrunner and Harer (2010).

In general, Persistent Homology is composed of the following parts, of which a more detailed description will be given in the following discussions:

1. A set of points $P = \left\{ \mathbf{v}_i \in \mathbb{R}^d : i = 1, \ldots, N \right\}$, sometimes called a point cloud;

2. A sequence of mathematical objects, $(K_s)_{s>0}$ with a natural ordering is known as a filtration (see Section 2.5.2.2);

3. Each object K_s has topological features. As s increases, topological features are born and then die (see Section 2.5.2.2);

4. This collection of birth-and-death processes can be represented in terms of persistence diagrams and barcodes both of which are closely related to Betti numbers (see Section 2.5.4).

This section is organised as follows. In Section 2.5.2, we discuss some key concepts of Persistent Homology. Section 2.5.3 applies this theory to the simplicial complex, graphs and point cloud. In Section 2.5.4, we review different summaries of Persistent Homology.

2.5.2 *Key Concepts*

In this section, we first give the definitions for topological space and topology. Then we consider two different types of dataset: a point cloud dataset (Example A) and a dataset of points with no coordinate information, i.e. data points from a simple graph (Example B). These two examples are used to illustrate all definitions in Section 2.5.

Sutherland (2009) provides the definition for topological space and topology as follow.

A topological space $T = (X, \mathcal{T})$ consists of a non-empty set X together with a fixed family subsets of X satisfying

1. $X, \emptyset \in \mathcal{T}$, where $\emptyset$ is the empty set;
2. if $U_1, U_2 \in \mathcal{T}$, then $U_1 \bigcap U_2 \in \mathcal{T}$;
3. if $U_i \in \mathcal{T}$, then $\bigcup U_i \in \mathcal{T}$ where i is an index set which can be uncountable infinite.

X is defined as the topology of the family $\mathcal{T}$ and U_i is called the open sets of T. Therefore, '$U \in \mathcal{T}$' is equivalent to 'U is open in T'. However, in practise, as the topological space. In this thesis, we use X as the topological space in Chapter 1.

EXAMPLE A (PART 1)

Four points are generated randomly from $Uniform\left([0,10]^2\right)$ as shown in Figure 2.5.1. Then,

$$P = \left\{\mathbf{v}_i \in \mathbb{R}^2 : i = 1, ..., 4\right\}$$

is a set of data points. P is a simple example of a point cloud.

EXAMPLE B (PART 1)

In another case, there are 4 isolated points, or vertices, which we mark as 1 to 4. In this example, points can also be written as $P = \{v_1, v_{2,} v_3, v_4\}$.

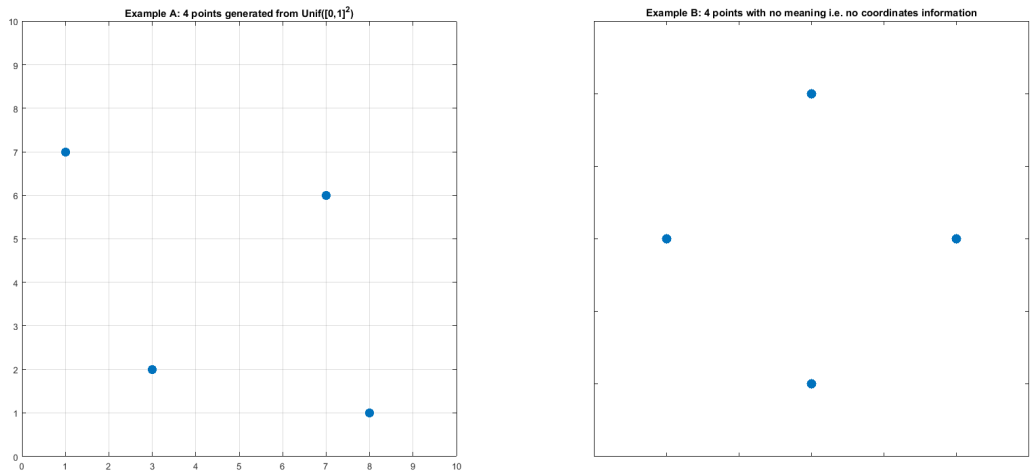

Figure 2.5.1: Example A (Part 1) (left) and Example B (Part 1) (right).

2.5.2.1 *Simplices, Simplicial Complexes and Chains*

The fundamental element of Persistent Homology is based on the idea of a simplex, σ.

A k-simplex σ_k is the convex hull of $k+1$ affinely independent points

$$\sigma_k = conv\{v_i : i = 1, \ldots, k+1\}$$

which is defined in Section 2.3. Its dimension is $\dim(\sigma_k) = k$. We use special names for the first few dimensions, i.e. a vertex is a 0-simplex σ_0, an edge is a 1-simplex σ_1, a triangle is a 2-simplex σ_2 and a tetrahedron is a 3-simplex σ_3, etc. A face of σ_k is the convex hull of a non-empty proper subset of $\{v_1, v_2, ..., v_{k+1}\}$. Here a proper subset is a subset not equal to the whole set. Figure 2.5.2 illustrates examples for the first 4 simplices, $\sigma_0, ..., \sigma_3$.

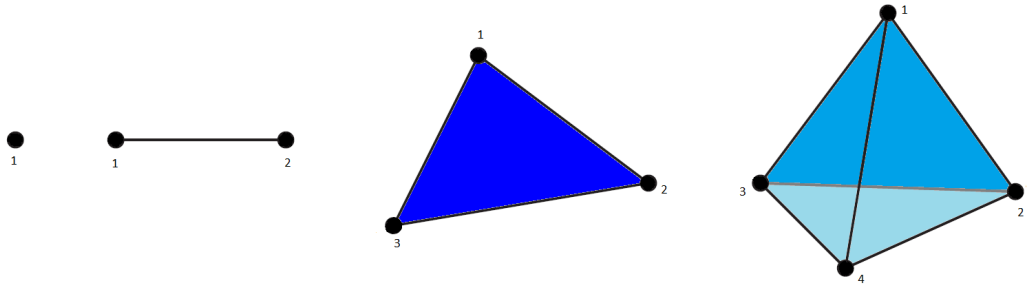

Figure 2.5.2: Simplices in $\mathbb{R}^3$: 0-simplex (vertex), 1-simplex (edge), 2-simplex (triangle) and 3-simplex (tetrahedron).

Definition 2.5.1. A simplicial complex, K, is a finite collection of simplices, $\sigma, \tau, ...$ such that $\sigma \in K$ and $\tau \leq \sigma$ implies that $\tau \in K$, and $\sigma, \tau \in K$ implies $\sigma \bigcap \tau$ is either empty or a face of both.

There are three types of subset of a complex which are called link, star and skeleton. For a simplicial complex, K, and a face σ_1 of K, the link, $lk_K(\sigma_1)$, is the set of faces σ_i such that faces $\sigma_1 \cap \sigma_i = \emptyset$ and $\sigma_1 \cup \sigma_i = K$. In addition, the star, $st(v)$, is the subcomplex of all faces in K containing v where v is a vertex of K.

Definition 2.5.2. The k-skeleton of a simplicial complex K, denoted $skel_k$, is the collection of all faces of all simplicies in K which have dimension at most k, i.e.

$$skel_k = \{\sigma \in K : \dim(\sigma) \leq k\}.$$

The three most commonly used complexes are the clique complex, Čech complex and Rips complex which are discussed later in Section 2.5.3. A clique complex is often used with abstract points whereas Čech and Rips complex are defined with point clouds.

A k-chain for k-simplices in a complex K is defined as

$$c_k = \sum_i a_{k,i}\sigma_{k,i}$$

where $a_{k,i} = 0$ or 1 indicates the exclusion or inclusion of the i-th k-simplex. There is a k-chain for each integer $0 \leq k \leq \dim(K)$.

Definition 2.5.3. For a k-chain, c_k, the support $supp(c_k)$ is the union of k-faces in c_k with non-zero coefficients. Similarly the vertex support $vsupp(c_k)$ is the underlying vertex set of $supp(c_k)$.

The group of k-chains, C_k, is the k-chain with normal addition operation under modulo 2. To relate these groups of chains to each other, the boundary of a k-simplex is defined as the sum of the k-simplex's $(k-1)$-simplices faces, which is

$$\partial_k \sigma_{k,i} = \sum_i a_{k-1,i}\sigma_{k-1,i}$$

where $a_{k-1,i} = 1$, if $\sigma_{k-1,i}$ is a face of $\sigma_{k,i}$ and $a_{k-1,i} = 0$ if $\sigma_{k-1,i}$ is not a face of $\sigma_{k,i}$.

Therefore, the boundary maps may be written in a sequence as

$$\ldots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \ldots \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

$(C_{-1} = 0)$.

Recall from the discussion as (2.3.1) that Ker denotes the kernel of a map (the set of elements mapped to the zero elements) and Im denotes the image of a map.

By letting $Z_k = \text{Ker}(\partial_k)$ and $B_k = \text{Im}(\partial_{k+1})$, we also define Z_k as the group of $k-$cycle and B_k as the group of the $p-$boundaries both with normal addition operation under modulo 2 . Since $C_{-1} = 0$, $Z_0 = \text{Ker}(\partial_0) = C_0$. The group of $k-$chains are illustrated in both Example B (Part 2) and B (Part 3).

EXAMPLE B (PART 2)

Following the setting from Example B (Part 1), Example B (Part 2) gives a basic example of chains, simplices and boundary.

As shown in Figure 2.5.3, we can connect the vertices and the yellow triangle indicates the presence of $2-$simplex.

The 0-simplices set is

$$\sigma_0 = \{v_1, v_2, v_3, v_4\} = \{\sigma_{0,1}, \sigma_{0,2}, \sigma_{0,3}, \sigma_{0,4}\},$$

and $c_1 = \sum_{i=1}^{6} a_i \sigma_{1,i}$ where $a_i = 1$ for $i = 1, ..., 5$ and $a_6 = 0$, since there should be a total number of $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ edges in the complete graph. Also,

$$\partial_2 \sigma_{2,1} = \sum a_{1,i} \sigma_{1,i} = \sigma_{1,1} + \sigma_{1,2} + \sigma_{1,3}.$$

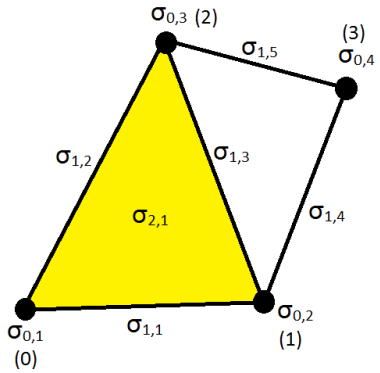

Figure 2.5.3: Example B (Part 2) and B (Part 3) illustrates how to calculate the different Betti numbers from matrix operations. See (2.5.4).

The fundamental property that makes homology work is that the boundary of a boundary is necessarily zero .

FUNDAMENTAL LEMMA OF HOMOLOGY

$$\partial_k \cdot \partial_{k+1} = \mathbf{0}_{n_{k+1}, n_{k-1}} \tag{2.5.1}$$

for all integers k and $\mathbf{0}_{n_{k+1}, n_{k-1}}$ is a zero matrix.

An example is given in Example B (Part 3).

2.5.2.2 *Betti Numbers and Filtrations*

Definition 2.5.4. The k-th homology group, H_k, is the quotient group $H_k = Z_k / B_k$, where Z_k and B_k are defined in the description of group of k-chains. The k-th Betti number is defined as

$$\begin{aligned}\beta_k &= \text{rank}\,(H_k) \\ &= \text{rank}\,(Z_k) - \text{rank}\,(B_k) \\ &= \text{rank}\,\{\text{Ker}\,(\partial_k)\} - \text{rank}\,\{\text{Im}\,(\partial_{k+1})\}\,.\end{aligned}$$

Therefore, if we rewrite $z_k = \text{rank}\,(Z_k)$, $b_k = \text{rank}\,(B_k)$, then

$$\beta_k = z_k - b_k. \tag{2.5.2}$$

Moreover, we also define $n_k = \text{rank}\,(C_k)$, then by rank–nullity theorem

$$n_k = z_k + b_{k-1}. \tag{2.5.3}$$

An example of a Betti number is given below in Example B (Part 3).

We have direct interpretations of Betti numbers for the first few dimensions, i.e. β_0 is the number of connected components, β_1 is the number of one-dimensional holes, β_2 is the number of two-dimensional voids.

MATRIX OPERATIONS

We can re-write the boundary map in matrix form

$$\begin{pmatrix} \sigma_{k,1} \\ \sigma_{k,2} \\ \vdots \\ \sigma_{k,m} \end{pmatrix} = \begin{pmatrix} a^1_{k-1,1} & a^1_{k-1,2} & \cdots & a^1_{k-1,m} \\ a^2_{k-1,1} & a^2_{k-1,2} & \cdots & a^2_{k-1,m} \\ \vdots & \vdots & \ddots & \vdots \\ a^m_{k-1,1} & a^m_{k-1,2} & \cdots & a^m_{k-1,m} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{k-1,1} \\ \sigma_{k-1,2} \\ \vdots \\ \sigma_{k-1,m} \end{pmatrix} \tag{2.5.4}$$

which can be simplified as $\boldsymbol{\sigma}_k = \mathbf{A}_{k,k-1}\boldsymbol{\sigma}_{k-1}$, where $\mathbf{A}_{k,k-1}$ is a matrix consisting of 0's and 1's with modulo 2 operations.

By doing both row and column operations, we can reduce $\mathbf{A}_{k,k-1}$ to Smith normal form. Therefore, we can present the information as

$$\begin{array}{cccccccc}
 & & \leftarrow & & n_{k-1} & & & \rightarrow \\
 & & \leftarrow & b_{k-1} & \rightarrow & & & \\
\uparrow & & 1 & 0 & \dots & \dots & \dots & 0 \\
 & & 0 & \ddots & 0 & \dots & \dots & 0 \\
n_k & & \vdots & 0 & 1 & 0 & \cdots & 0 \\
 & \uparrow & 0 & \dots & 0 & \dots & \dots & 0 \\
 & z_k & \vdots & \dots & \vdots & \dots & \dots & 0 \\
\downarrow & \downarrow & 0 & \cdots & 0 & \dots & \dots & 0
\end{array}$$

The matrix operations are presented in Example B (Part 3).

EXAMPLE B (PART 3)

Following the setting from Example B (Part 2), since

$$\sigma_{2,1} = \sigma_{1,1} + \sigma_{1,2} + \sigma_{1,3} = \sigma_{1,1} + \sigma_{1,2} + \sigma_{1,3} + 0 \cdot \sigma_{1,4} + 0 \cdot \sigma_{1,5},$$

the boundary matrix is the 1×1 matrix

$$(\sigma_{2,1}) = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma_{1,1} \\ \sigma_{1,2} \\ \sigma_{1,3} \\ \sigma_{1,4} \\ \sigma_{1,5} \end{pmatrix}.$$

Therefore, $\mathbf{A}_{2,1} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \end{pmatrix}$. By applying column operations to $\mathbf{A}_{2,1}$, it can be reduced to $\mathbf{A}_{2,1} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}$. As a result, $n_2 = 1$, $z_2 = 0$, $b_1 = 1$ and $n_1 = 5$.

Similarly, the boundary matrix for σ_1 is

$$\begin{pmatrix} \sigma_{1,1} \\ \sigma_{1,2} \\ \sigma_{1,3} \\ \sigma_{1,4} \\ \sigma_{1,5} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{0,1} \\ \sigma_{0,2} \\ \sigma_{0,3} \\ \sigma_{0,4} \end{pmatrix} = \mathbf{A}_{1,0}\sigma_0.$$

By applying row operations to $\mathbf{A}_{1,0}$, it can be reduced to

$$\mathbf{A}_{1,0} \rightarrow \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{2.5.5}$$

After applying column operations to the last column of (2.5.5) , it is reduced to

$$\mathbf{A}_{1,0} \rightarrow \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Consequently, $n_1 = 5$, $z_1 = 2$, $b_0 = 3$ and $n_0 = 4$.

Since $Z_0 = C_0$, $z_0 = \text{rank}\,(Z_0)$ =rank $(C_0) = n_0 = 4$, $n_3 = \text{rank}\,(C_3) = 0 = z_3 = b_2$

$$\begin{aligned} \beta_0 &= z_0 - b_0 = 4 - 3 = 1 \\ \beta_1 &= z_1 - b_1 = 2 - 1 = 1 \\ \beta_2 &= z_2 - b_2 = 0 - 0 = 0. \end{aligned}$$

As a result, $\beta_0 = 1$, $\beta_1 = 1$ and $\beta_2 = 0$ for Figure 2.5.3.

Furthermore,

$$\mathbf{A}_{2,1}\mathbf{A}_{1,0} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{0}_{n_2,n_0}$$

which satisfies the Fundamental Lemma of Homology stated in (2.5.1).

EULER-POINCARE THEOREM

For simplicial complexes, the Euler characteristic, χ, is defined as the alternating sum of the number of k-simplices. Moreover, using (2.5.3)

$$\begin{aligned}\chi &= \sum (-1)^k n_k \\ &= \sum (-1)^k (z_k + b_{k+1}),\end{aligned}$$

where n_k, z_k and b_k are defined in Section 2.5.2.2.

The Euler characteristic can also be defined as the alternating sum of the Betti numbers, combined with (2.5.2):

$$\begin{aligned}\chi &= \sum (-1)^k \beta_k \\ &= \sum (-1)^k (z_k - b_k),\end{aligned}$$

which implies

$$\begin{aligned}\chi &= \sum (-1)^k n_k \\ &= \sum (-1)^k (z_k + b_{k+1}) \\ &= \sum (-1)^k (z_k - b_k) \\ &= \sum (-1)^k \beta_k.\end{aligned}$$

An example of Euler characteristic is given in Example B (Part 4).

EXAMPLE B (PART 4)

Following the settings from B (Part 3), the alternating sum of the number of $k-$simplices

$$\begin{aligned}\chi &= \sum (-1)^k n_k \\ &= (-1)^0 n_0 + (-1)^1 n_1 + (-1)^2 n_2. \\ &= 4 - 5 + 1 = 0,\end{aligned}$$

while the alternating sum of the β_k numbers is

$$\begin{aligned}\chi &= \sum (-1)^k \beta_k \\ &= (-1)^0 \beta_0 + (-1)^1 \beta_1 + (-1)^2 \beta_2 \\ &= 1 - 1 + 0 = 0,\end{aligned}$$

which satisfies the results of the Euler-Poincare Theorem for the Euler characteristic.

FILTRATION

Let K be a simplicial complex with m simplices, then there are $n \leq m$ different sub-complexes, which can be arranged as an increasing sequence,

$$\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_m = K. \tag{2.5.6}$$

This nested sequence of complexes is called the filtration of K.

We then define a function $g : K \to \mathbb{R}$, where g is monotonic increasing along the increasing chains of the faces, i.e. if σ is a face of τ, then $g(\sigma) \leq g(\tau)$. Monotonicity implies that the sublevel set, $K_s = g^{-1}(-\infty, a_s]$, is a sub-complex of K for $a_0 \leq a_1 \leq ... \leq a_m$.

The sequence of inclusion maps from (2.5.6) induces maps on homology for each dimension k,

$$0 = H_k(K_0) \xrightarrow{f_k^{0,1}} H_k(K_1) \xrightarrow{f_k^{1,2}} \ldots \xrightarrow{f_k^{k-1,k}} H_k(K_m) = H_k(K). \tag{2.5.7}$$

In order to understand the changing space, we focus on where homology classes appear (are born) and disappear (i.e. die) in this sequence.

Let $f_k^{i,j} : H_k(K_i) \to H_k(K_j)$ be the map from (2.5.7). Then the k-th Persistent Homology group is defined as $H_k^{i,j} = \text{Im}\left(f_k^{i,j}\right)$ for $0 \leq i \leq j \leq m$. The corresponding k-th Betti number is defined as $\beta_k^{i,j} = \text{rank}\left(H_k^{i,j}\right)$.

Let γ be an element in $H_k(K_i)$, we define that γ is born at K_i if $\gamma \notin H_k^{i-1,i}$. Furthermore, the death time of γ is defined as K_j if $f_k^{i,j-1}(\gamma) \notin H_k^{i-1,j-1}$ but $f_k^{i,j-1}(\gamma) \in H_k^{i-1,j}$, as shown in Figure 2.5.4.

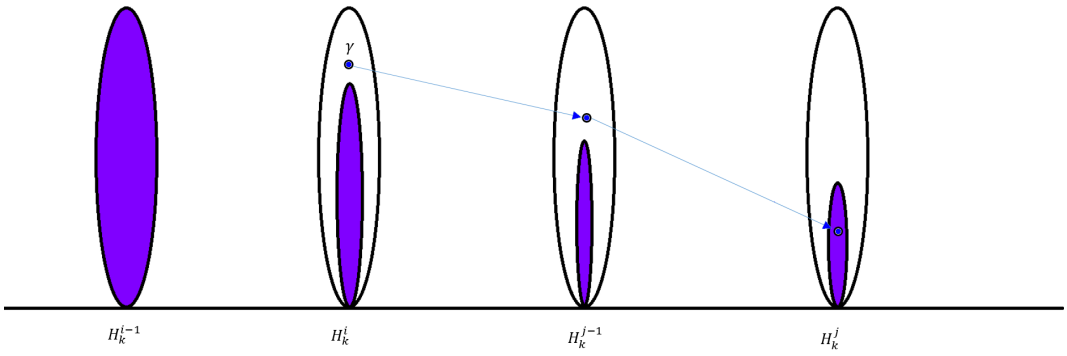

Figure 2.5.4: The class γ is born at K_i as it is not in the image of the $H_k^{i-1,i-1}$. It dies at K_j because it merges with the image of $H_k^{j-1,j-1}$.

If γ is born at K_i and dies at K_j, the difference in the function values is defined as the persistence, $\text{pers}(\gamma) = a_j - a_i$. If γ is born at K_i but never dies then we set its persistence to be $\text{pers}(\gamma) = \infty$.

In TDA, we often refer the birth time as b and death time as d, i.e. we write $b = a_i$ and $d = a_j$. This process is viewed as the birth-and-death process.

Figure 2.5.5 and 2.5.6 give examples of filtrations in the discrete case and the Čech complex, which is going to be defined in Section 2.5.3.

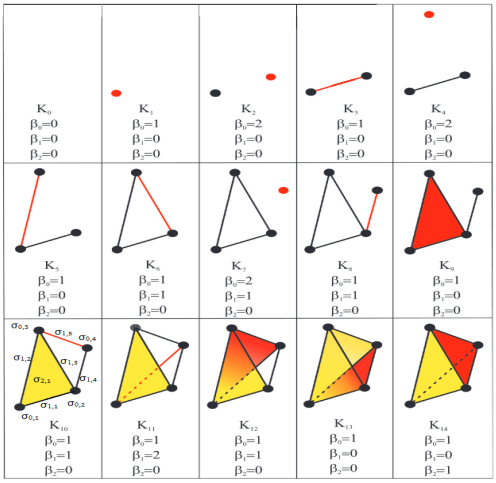

Figure 2.5.5: Filtration of a simplicial complex (tetrahedron) and its topological characterization. At each stage, a vertex, a line or a face which is newly added is presented in red.

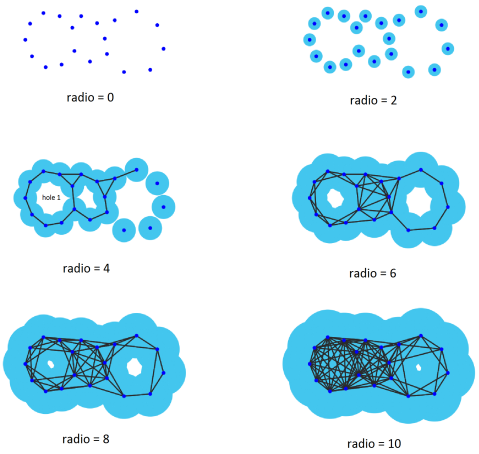

Figure 2.5.6: Filtration of a point cloud under Čech complex.

2.5.3 *Simplicial Complexes*

As mentioned in Section 2.5.2, the three most commonly used complexes are the clique complex, Čech complex and Rips complex.

CLIQUE COMPLEX

The clique complex $\mathcal{X}(H)$ is a simplical complex of a general graph H with N vertices, and it has a simplex $\sigma = \{v_1, \ldots, v_l\}$ if and only if all edges $(v_i, v_j) \in \mathcal{X}(H)$ for $1 \leq i \leq j \leq l \leq n$.

We also call the clique complex a random clique if the graph H is a random graph. The most commonly used random graph is the Erdős-Rényi model (ERM) which is defined in Section 2.4.

An ERM $\mathcal{G}(4, 0.4)$ is illustrated in Figure 2.5.7 in which only vertices labelled $(1,3)$, $(1,4)$ and $(3,4)$ are connected.

EXAMPLE B (PART 5)

Following the setting from Example B (Part 4), for random clique complex in Figure 2.5.7,

$$\sigma_0 = \{v_1, v_2, v_3, v_4\} = \{\sigma_{0,1}, \sigma_{0,2}, \sigma_{0,3}, \sigma_{0,4}\},$$

$\sigma_1 = \{\sigma_{1,1}, \sigma_{1,2}, \sigma_{1,3}\}$ and $\sigma_2 = \{v_1, v_3, v_4\}$ since all edges (v_i, v_j) are present.

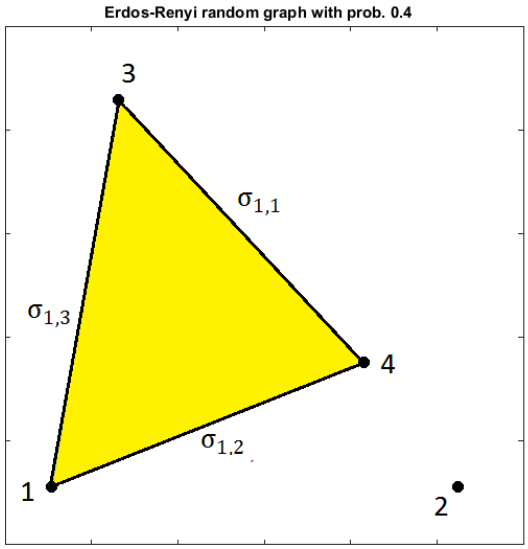

Figure 2.5.7: Example B (Part 5): Erdős-Rényi random graph $\mathcal{G}(4, 0.4)$. Only edges between vertices labelled $(1,3)$, $(1,4)$ and $(3,4)$ are connected.

Therefore, the β values are $\beta_0 = 2$, $\beta_1 = \beta_2 = 0$.

ČECH COMPLEX

Consider a point cloud $P = \left\{\mathbf{v}_i \in \mathbb{R}^d : i = 1, \ldots, n\right\}$. By placing a set of balls centered at the data points with common radius r, the Čech complex $C(r)$ is defined as follows. A simplex $\sigma = \left\{v_{i_1}, \ldots, v_{i_l}\right\}$ lies in $C(r)$ if and only if $\bigcap_{j=0}^{l} B_r(v_{i_j}) \neq \emptyset$ where $B_r\left(v_{i_j}\right) = \left\{v_{i_j} \in \sigma | d\left(v_{i_j}, v_{i_z}\right) \leq r, j \neq z\right\}$; $r > 0$ is the radius of the ball and $1 \leq j \leq l \leq n$.

A example of a Čech complex is given in Example A (Part 2).

EXAMPLE A (PART 2)

Following the setting from Example A (Part 1), as can be seen in Figure 2.5.8, each vertex is placing a circle with common radius from 2 to 4. The top line of Figure 2.5.8 illustrates the increase of the circle radius while the bottom line indicates the corresponding existence of the simplices. For $r = 2$, since no balls intersect, there is no edge present. For $r = 3$, only 2 balls intersect, as shown in the area marked as darker blue. Therefore, only edges are added. The empty triangle indicates that the three balls do not intersect with each other. For $r = 4$, the area where three balls intersect are marked as purple. This implies that two 2-simplices are presented which are marked as yellow in bottom line. In conclusion, for $r = 0$, $r = 1$ and $r = 2$, $\beta_0 = 4$, $\beta_1 = \beta_2 = 0$; for $r = 3$, $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 0$; for $r = 4$, $\beta_0 = 1$, $\beta_1 = \beta_2 = 0$. As a result, for persistent diagrams which are discussed in Section 2.5.4, Dgm_0 and Dgm_1 are $Dgm_0 = \{(0,3), (0,3), (0,3), (0,\infty)\}$, $Dgm_1 = \{(3,4)\}$ and Dgm_2 is null.

In general, from the computational point of view, Čech complexes are very expensive. In Example A (Part 2), establishing the existence of a 2-simplex involves searching all subsets of points of size 3. If we want to know whether or not there is a 8-simplex, we need to consider all the combinations of points of size 9.

Therefore, from a practical point of view, a computationally simpler type of complex is needed to replace the Čech complex.

RIPS COMPLEX

Consider the point cloud $P = \left\{\mathbf{v}_i \in \mathbb{R}^d : i = 1, \ldots, N\right\}$. By placing a set of balls centered at the data points with common radius r, the Vietoris-Rips complex, shortened to Rips complex and denoted $R(r)$, is defined as follows. A simplex $\sigma = \{\mathbf{v}_1, \ldots, \mathbf{v}_l\} \in R(r)$ if and only if $d\left(\mathbf{v}_i, \mathbf{v}_j\right) \leq 2r$

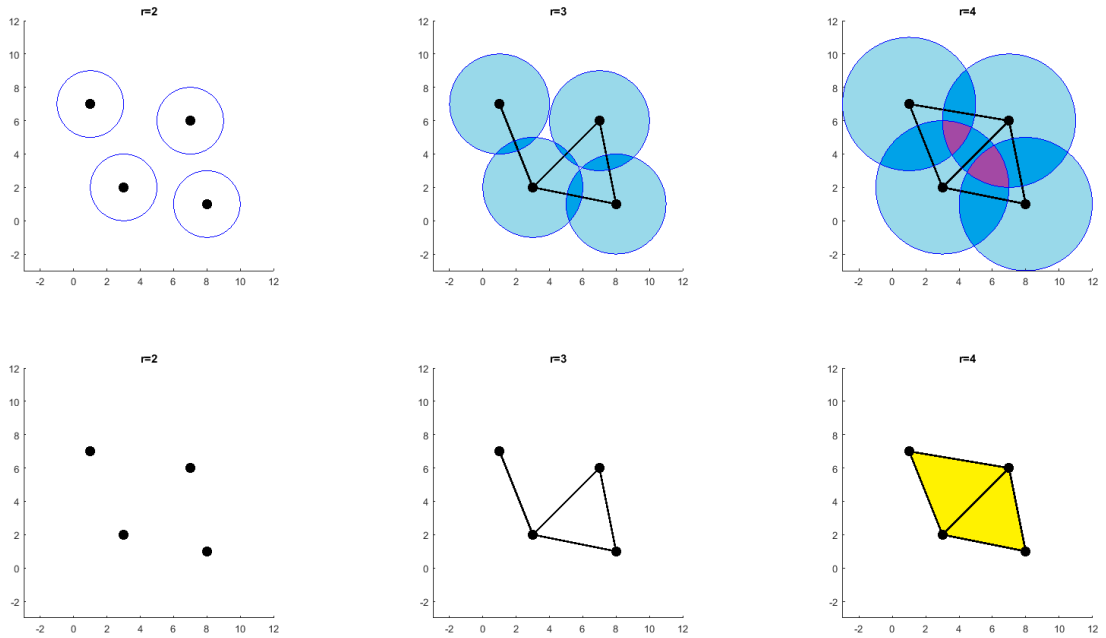

Figure 2.5.8: Example A (Part 2): A set of points with balls with common radius from 2 to 4 are shown in the top line. The darker blue indicates that two balls intersect with each other while the purple areas indicate three balls intersect with each other. The Čech complex under the filtration is given in the bottom line. The yellow areas indicate that two 2-simplices exist while the empty triangle at $r = 3$ indicates that β_1 is non-zero i.e. $\beta_1 = 1$.

where $d(\cdot,\cdot)$ is the Euclidean distance between two points, $r > 0$ is the radius of the ball and $1 \leq i < j \leq l \leq n$.

EXAMPLE A (PART 3)

Following the setting from Example A (Part 1), the top line of Figure 2.5.9 is the same as Example A (Part 2). However, the difference between Rips complex and Čech complex occurs when $r = 3$. For $r = 3$, since all 3 edges are included in the Rips complex, the triangle is also included in the Rips complex.

In summary, $Dgm_0 = \{(0,3), (0,3), (0,3), (0,\infty)\}$ and there are no Dgm_1 and Dgm_2.

The Rips complex is much easier to compute as it is only determined by the combinations of vertices and edges. This means that Rips complex is also a clique complex.

Furthermore, the Example A (Part 2 and 3) show that the Čech complex and Rips complex do not always have the same topological features, i.e. when $r = 3$, $\beta_1 > 0$ for the Čech but $\beta_1 = 0$ for the Rips complex as there is an empty hole formed by the triangle for Čech complex but not for the Rips complex. However, there is an inclusion relationship between Čech and Rips complex, which is

$$C(r) \subseteq R(r) \subseteq C\left(\sqrt{2}r\right) \subseteq R\left(\sqrt{2}r\right)$$

for any $r > 0$ (De Silva and Ghrist, 2007).

2.5.4 *Summaries of Persistent Homology*

PERSISTENCE DIAGRAM

In order to visualize the changing homology along a f, H notation, we draw persistence diagrams, denoted as Dgm_k for each dimension k. A persistence diagram is a set of points in the upper half plane $\{(b,d) \in \mathbb{R}^2 \mid d \geq b\}$ along with all the points on the diagonal $\{(b,b) \in \mathbb{R}^2\}$ where b is the birth time and d is the death time. Let $\mu_k^{i,j}$ be the number of k-dimensional classes born at K_i and dying at K_j. Then for each class γ that is born at K_i and dies at K_j, we draw point (b_γ, d_γ) with multiplicity $\mu_k^{i,j}$.

A point that is far away from the diagonal has a longer lifetime while the one close to diagonal indicates a shorter lifetime. In general, long persis-

Figure 2.5.9: Example A.2: A set of points with balls with common radius from 2 to 4 are shown in the top line. The darker blue indicates that two balls intersect with each other while the purple areas indicate three balls intersect with each other. The Rips complex under the filtration is given in bottom line. The yellow area indicates that a 2−simplex exists.

tence is likely to indicate the true signal and true topological feature while the topological features with short persistence are usually considered to be noise.

BARCODE

If we rewrite the coordinates (b_γ, d_γ) as intervals $[b_\gamma, d_\gamma)$ then a barcode is the collection of these intervals as horizontal line segments. Barcodes were introduced by Ghrist (2008).

Figure 2.5.10 gives an example for both persistence diagram (left) and barcode (right), where black ones indicate β_0 and red one means β_1. In this example, we take

$$Dgm_0 = \{(0,3), (0,3), (2,4), (1,5), (1,2)\}$$

and

$$Dgm_1 = \{(4.01, 4.03)\}.$$

For persistence diagram, the x-axis is the birth time and the $y-$axis is the death time of the topological features. For the barcode, the $x-$axis is the parameter of the filtration, i.e. the s value which is defined in the description of filtration in Section 2.5.2. If this is a barcode for a Čech or Rips complex which will be defined in Section 2.5.3, the x-axis in barcode is going to be r, the radius of the ball. Assume this barcode is for the general case, i.e. at K_s , then the Betti number at s, which is written as $\beta_0(s)$, is the number of line segments that intersect with the vertical line $x = s$. As shown in Figure 2.5.10 (right), $\beta_0(2.5) = 4$. Moreover, each of the symbols in the persistence diagram on the left is corresponding to at least one line segment with the same colour on the right in Figure 2.5.10. Taking $(0, 3)$ as an example, there are 2 line segments, however, only one black dot is shown on the persistence diagram. Since the red triangle, which is corresponding to β_1, is very close to the diagonal line, it is going to be considered as topological noise. In the persistence diagram, this noise can be easily visualised. However, for the barcode, the red line segment indicating β_1 is negligible and may be missed. As a result, a wrong conclusion that β_1 is zero may be made.

OTHER SUMMARIES

Apart from persistence diagrams and barcodes, other summaries have been introduced for Persistent Homology in recent years. One is due to

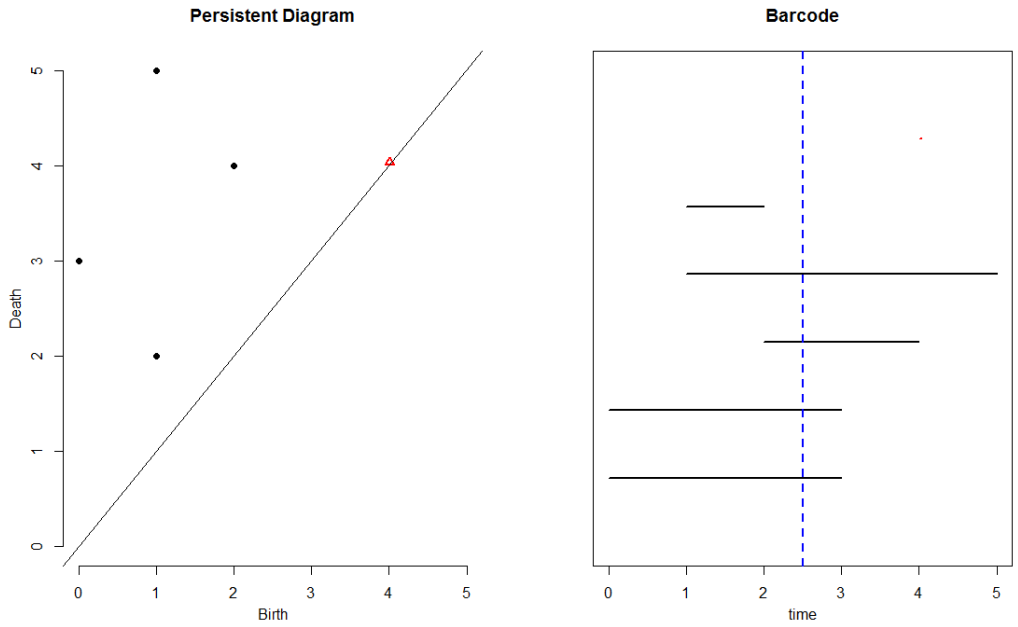

Figure 2.5.10: Persistent diagram (left) and barcode (right) corresponding to $Dgm_0 = \{(0,3),(0,3),(2,4),(1,5),(1,2)\}$ and $Dgm_1 = \{(4.01,4.03)\}$ where black corresponds to β_0 and red corresponds to β_1.

Bubenik (2015), who has introduced persistent landscapes. An example of a persistent landscape is shown in Figure 2.5.11.

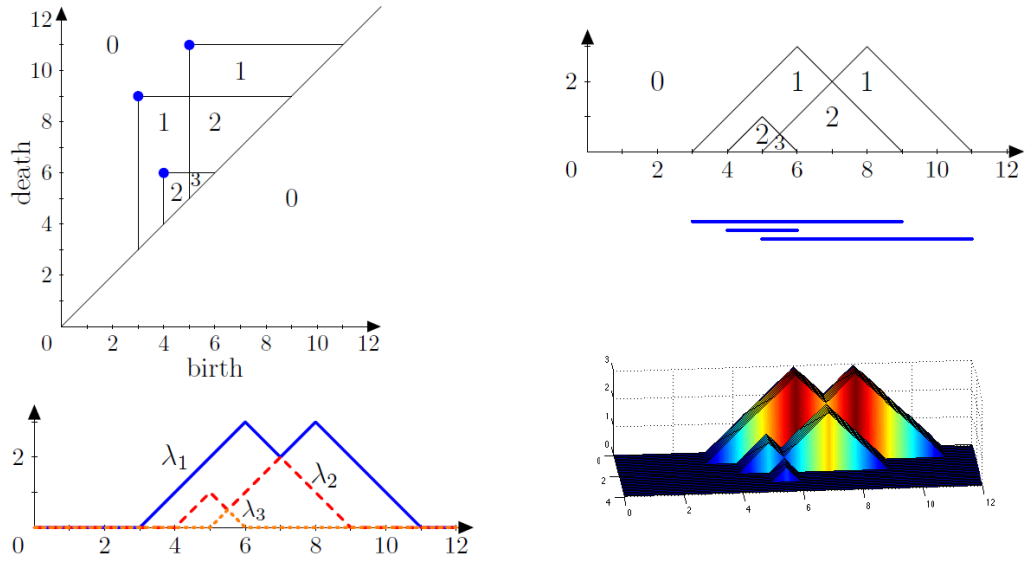

Figure 2.5.11: The top left is a standard persistent diagram for β_1 and the top right is the rescaled persistent diagram. Below the top right graph is the corresponding barcode. The bottom left is the corresponding persistence landscape and the bottom right is its 3-dimensional version.

A persistent landscape transforms the persistence diagram into a sequence of continuous functions. To define the landscape, we first introduce the triangle function

$$\Lambda(t) = \begin{cases} t - m + h & t \in [m-h, m] \\ m + h - t & t \in (m, m+h] \\ 0 & otherwise \end{cases} = \begin{cases} t - b & t \in \left[b, \frac{b+d}{2}\right] \\ d - t & t \in \left(\frac{b+d}{2}, d\right] \\ 0 & otherwise \end{cases}$$

where $m = \frac{b+d}{2}$ is the mean lifetime of a topological feature and $h = \frac{d-b}{2}$ is the half lifetime of a topological feature. By overlaying the graphs of the functions $\Lambda(t)$, we would construct an arrangement of curves which is shown in Figure 2.5.12. The persistence landscape is a summary of this arrangement. More precisely, persistence landscape is defined as the collection of the functions

$$\lambda_i(t) = i\max\Lambda(t)$$

where $i$max is the i-th largest value in the set, especially, 1max is the usual maximum function. One advantage of persistent landscapes is that one can make calculations directly on $\beta_k \geq i$. For example, λ_1 is the function for $\beta_k \geq 1$ and λ_2 is the function for $\beta_k \geq 2$, etc.

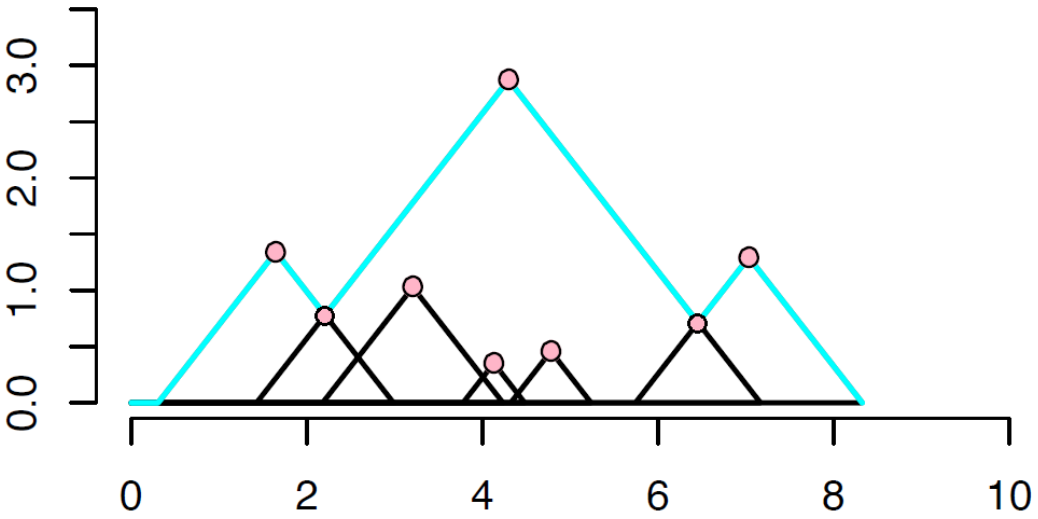

Figure 2.5.12: The pink circles are the points in a persistence diagram. The cyan curve is the persistence landscape λ_1.

Biscio and Møller (2019) have also recently introduced a new type of summary for TDA which is called the accumulative persistence function (APF) . The APF is defined, for topological features of dimension k, as

$$APF_k(m) = \sum c_i l_i \mathbb{I}(m_i \leq m), \ m \geq 0$$

where $m = \frac{b+d}{2}$ is the mean lifetime of a topological feature and c_i is the multiplicity. The β_0 and β_1 information from brain artery trees which were reviewed in Biscio and Møller (2019) are used to illustrate the APF which is shown in Figure 2.5.13.

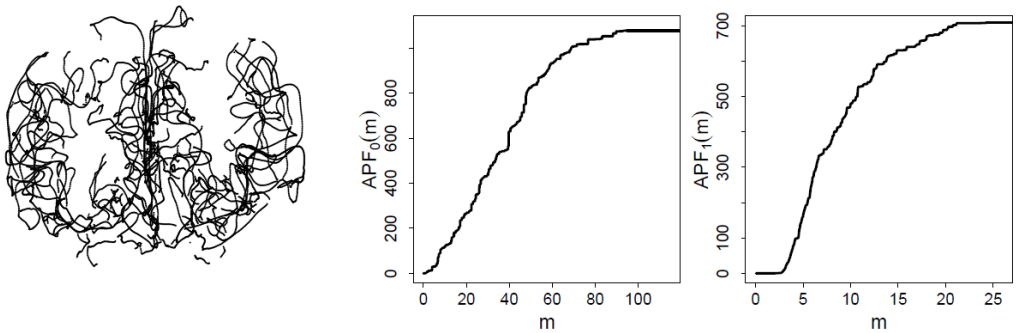

Figure 2.5.13: A brain artery tree (left), its corresponding APF_0 (middle) obtained from the sublevel set of the height function, and its corresponding APF_1 (right) obtained from the Rips complex.

2.6 KEY PAPERS

In this section, some of the relevant results from the literature are presented.

FAMILY OF BACHMANN-LANDAU NOTATIONS

For Family of Bachmann-Landau notations, for non-negative functions g and h,

- $g(n) = o(h(n))$ means for any $\epsilon > 0$, there exists an n_0 such that for $n > n_0$, $g(n) < \epsilon h(n)$ i.e. $\lim_{n\to\infty} \frac{g(n)}{h(n)} = 0$. (small-$o$)
- $g(n) = O(h(n))$ means for at least one $\epsilon > 0$, there exists an n_0 such that for $n > n_0$, $g(n) \leq \epsilon h(n)$ i.e. $\limsup_{n\to\infty} \frac{g(n)}{h(n)} < \infty$. (big-$O$)
- $g(n) = \omega(h(n))$ means for any $\epsilon > 0$, there exists an n_0 such that for $n > n_0$, $g(n) \geq \epsilon h(n)$ i.e. $\liminf_{n\to\infty} \frac{g(n)}{h(n)} \to \infty$. (small-$\omega$)

2.6.1 *Relevant Paper on CLT for Betti Numbers*

In Kahle and Meckes (2013, 2015), several limit theorems for Betti number have been proved for ERM. In particular, they prove that if the number of vertices N in the ERM tends to infinity, then the Betti number of the clique complex of an ERM tends to the normal distribution. To extend CLT result to a SBM, we need to extend results satisfied by the ERM to the SBM. Below, we state the results which are independent of the structure of the ERM. These results are from four different papers: Kahle and Meckes (2015); Feige and Ofek (2005); Kahle (2009); Hoffman et al. (2019).

Definition 2.6.1. (Kahle, 2009)

1. Let γ be a non-trivial k-cycle in a simiplicial complex K with minimal $vsupp$ where $vsupp$ is given in Definition 2.5.3, and write it as a linear combination of faces

$$\gamma_k = \sum_{f_k \in supp(\gamma)} \lambda_f f_k$$

with $\lambda_f \in \mathbb{Z}$.

2. Suppose X is the full induced subcomplex on $vsupp(\gamma)$. For $v \in vsupp(\gamma)$ define the k-chain

$$\gamma \bigcap st(v) = \sum_{f \in st(v)} \lambda_f f_k$$

and the $(k-1)$-chain

$$\gamma \bigcap lk(v) = \sum_{f \in st(v)} \lambda_f (f_k - \{v\}).$$

Order the vertices with v last and let this induce an orientation on every face. Then

$$\gamma \bigcap lk(v) = \partial_k \left(\gamma \bigcap st(v) \right)$$

and $\partial_{k-1} \cdot \partial_k = 0$ by (2.5.1) this gives that $\gamma \bigcap lk(v)$ is a $(k-1)$-cycle.

Definition 2.6.1 is required for Lemma 2.6.2.

Lemma 2.6.2. *(Lemma 5.2 in Kahle (2009)) If γ is a non-trivial $k-$cycle and $v \in vsupp(\gamma)$, then $\gamma \bigcap lk(v)$ is a non-trivial $(k-1)$-cycle in $lk(v)$.*

Definition 2.6.3. (Kahle, 2009)

1. A simiplicial complex X is said to be pure $k-$dimensional if every face of X is contained in a k-dimensional face.

2. A pure k-dimensional subcomplex X is said to be strongly connected if every pair of $k-$faces $\sigma, \tau \in X$ can be connected by a sequence of facets which is $(k-1)$-faces $\sigma = \{\sigma_0, ..., \sigma_n\} = \tau$ such that $\dim(\sigma_i \bigcap \sigma_{i+1}) = d-1$. for $0 \leq i \leq n-1$.

3. Every k-cycle is a $\mathbb{Z}-$linear combination of k-cycle with strongly connected support.

Lemma 2.6.4. *(Lemma 5.3 in Kahle (2009))Let $\mathcal{G}$ be a graph and $X(\mathcal{G})$ be its Clique complex. If γ is a non-trivial k-cycle in $X(\mathcal{G})$, then $|vsupp(\gamma)| \geq 2k+2$.*

Lemma 2.6.2 and 2.6.4 are going to be applied in Section 4.6.

Lemma 2.6.5. *(Lemma 4.1 in Hoffman et al. 2019) Let $\mathcal{G}$ be a graph with N vertices. For some positive constants C_1, C_2, C_3 and M, assume that G satisfies the following conditions:*

1. b.d.c.: every vertex has degree at most $C_1 d$;

2.

$$\sup_{\substack{\|\mathbf{x}\| = 1 \\ \mathbf{x}^T \mathbf{1}_N = 0 \\ \|\mathbf{y}\| = 1}} \left| \mathbf{x}^T \mathbf{A} \mathbf{y} \right| \leq C_2 \sqrt{d}$$

where $\mathbf{A}$ is the adjacency matrix of $\mathcal{G}$ and $\mathbf{1}_N$ is a vector whose elements are 1;

3. there are no edges between vertices of η_M, $|\eta_M| < \frac{N}{2}$ *and*

$$\max_{u \in \eta_M^c} \mathcal{E}(u, \eta_M) \leq 1$$

where $\eta_M = \left\{ v : \deg(v) \leq \frac{d}{M} \right\}$ *is a set of vertices of small degree and* $d \geq 1$ *is a function of* N. $\mathcal{E}(u, \eta_M)$ *is the number of edges between point* u *and set* η_M*;*

4.

$$\sup_{\substack{\|\mathbf{x}\| = 1 \\ \mathbf{x}^T \mathbf{D}^{\frac{1}{2}} \mathbb{I}_W = 0}} \left| \mathbf{x}^T \mathbf{D}^{-\frac{1}{2}} \mathbb{I}_{\eta_M^c} \right| \leq C_3 \frac{\sqrt{N}}{d};$$

Then there is a constant $C = C(C_1, C_2, C_3, M)$ *such that* $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n \leq 2$ *from normalized graph Laplacian defined in Section 2.4 satisfies*

$$\max_{\lambda_i \neq 0} |1 - \lambda_i| < \frac{C}{\sqrt{d}}.$$

Lemma 2.6.5 is going to be used in the proof of spectral gap theorem in Chapter 4.

Lemma 2.6.6. *(Lemma 2.3 in Feige and Ofek (2005)) Let*

$$S = \left\{ \sum_i \mathbf{v}_i = 0 : \|\mathbf{v}\| \leq 1, \mathbf{v} = (v_1, ..., v_n) \right\},$$

and define a grid which approximates S *as*

$$T = \left\{ \mathbf{x} \in \left(\frac{\epsilon}{\sqrt{n}} \mathbb{Z} \right)^n : \sum_i x_i = 0, \ \|\mathbf{x}\| \leq 1 \right\}$$

where $\mathbb{Z}$ *denotes the set of the integer values,* $0 < \epsilon < 1$ *and* ϵ *can be considered as the constant* $\frac{1}{2}$*. Then every vector* $\mathbf{v} \in S$ *whose norm is less than* $1 - \epsilon$ *is a convex combination of vertices from* T*.*

Lemma 2.6.7. *(Lemma 2.4 in Feige and Ofek (2005)) Let* $c \in \mathbb{R}$ *be an arbitrary constant. If for every* $\mathbf{x}, \mathbf{y} \in T$, $\left| \mathbf{x}^T \mathbf{A} \mathbf{y} \right| \leq c$*, then for every* $\mathbf{x} \in S$, $\left| \mathbf{x}^T \mathbf{A} \mathbf{x} \right| \leq \frac{c}{(1-\epsilon)^2}$ *where* T *and* S *are defined in Lemma 2.6.6 and* A *is an adjacency matrix.*

Claim 2.6.8. (Claim 2.9 in Feige and Ofek (2005)) If a set T is defined as

$$T = \left\{ \mathbf{x} \in \left(\frac{1}{2\sqrt{N}} \mathbb{Z} \right)^N : \sum_i x_i = 0, \ \|\mathbf{x}\| \leq 1 \right\}$$

where N is the vertices number in a graph $\mathcal{G}$ and $\mathbb{Z}$ is the integer set, then

$$card\,(T) \leq \exp\{N \log(18)\}.$$

Claim 2.6.8, Lemma 2.6.6 and 2.6.7 are going to be used in the proof of condition 2 in Lemma 2.6.5.

Theorem 2.6.9. *(Theorem 2.5 in Ballmann and Świątkowski (1997)) Let* $0 < k < n = \dim(K)$ *where K is a simplicial complex. Assume that* K_τ *is connected and that there is an* $\varepsilon > 0$ *such that*

$$\kappa_\tau \geq \frac{k(N-k)}{k+1} + \varepsilon$$

for all $(k-1)$*-simplices* τ *of K where* κ_τ *is the smallest positive eigenvalue of* $\mathbf{A}_{1,0}$ *which is defined in Section 2.5.2.2. Then* $H^k(K, \mathbb{Q}) = 0$.

Theorem 2.6.10. *(Cohomology Vanishing Theorem in Kahle (2014))Let K be a pure* $k-$*dimensional finite simplicial complex such that for every* $(k-2)$*-dimensional face* σ*, the link* $lk_K(\sigma)$ *is connected and has spectral gap*

$$\lambda_2\left[lk_K(\sigma)\right] > 1 - \frac{1}{k}.$$

Then $H^{k-1}(K, \mathbb{Q}) = 0$.

Theorem 2.6.11. *(Theorem 1 in Barbour et al. (1989)) Let* $\left\{X_{\mathbf{j}} : \mathbf{j} = (j_1, ..., j_r) \in J\right\}$ *be a dissociated set of random variables, such that* $E\left(X_{\mathbf{j}}\right) = 0$ *for all j. Let* $W = \sum_{\mathbf{j} \in J} X_{\mathbf{j}}$ *and suppose that the* $X_{\mathbf{j}}$ *are normalized such that* $E\left(W^2\right) = 1$*. Then*

$$d_1(W, Z) \leq K \left(\sum_{\mathbf{j} \in J}\right) \left(\sum_{\mathbf{k},\mathbf{l} \in L_j}\right) E\left[\left|X_j X_k X_l\right|\right] + E\left[\left|X_j X_k\right|\right] E\left[\left|X_l\right|\right] \quad (2.6.1)$$

where Z is a standard normal random variables, $d_1(\cdot, \cdot)$ *is the 1-norm distance and*

$$L_j = \left\{\mathbf{k} \in J : \{k_1, ..., k_r\} \bigcap \{j_1, ..., j_r\} \neq \emptyset\right\}.$$

In Theorem 2.6.11, the term dissociated is defined as follows. Let $J = \{\mathbf{j} = (j_1, ..., j_r)\}$ be a set of ordered list of elements. We define the set $\left\{X_{\mathbf{j}} : \mathbf{j} = (j_1, ..., j_r) \in J\right\}$ for J to be a dissociated set if two sub-collections of

the random variables $\{X_{\mathbf{j}} : \mathbf{j} \in K\}$ and $\{X_{\mathbf{j}} : \mathbf{j} \in L\}$ are independent whenever $\left(\bigcup_{\mathbf{j}\in K} \{j_1, ..., j_r\}\right) \bigcap \left(\bigcup_{\mathbf{j}\in L} \{j_1, ..., j_r\}\right) = \emptyset$.

Theorem 2.6.11 is going to be employed in the proof of CLT for Betti number for SBM in Chapter 4.

Theorem 2.6.12. *(Theorem 1.1 in Kahle and Meckes (2015)) Consider Clique complex $\mathcal{X}(\mathcal{G})$ with Erdős-Rényi model $\mathcal{G}(N, p)$. Assume that $N^{-\frac{1}{k}} < p < N^{-\frac{1}{k+1}}$, then*

$$\beta_k(\mathcal{X}) - E\{\beta_k(\mathcal{X})\} \to Normal(0, Var\{\beta_k(\mathcal{X})\}) \tag{2.6.2}$$

for each k.

Theorem 2.6.12 is the CLT for Betti numbers for ERM. In Chapter 4, we are going to extend Theorem 2.6.12 to Theorem 4.7.2 which is the CLT for Betti numbers for SBM.

Remark 2.6.13. In these four papers, there are two terminologies which have very similar definitions. They are asymptotic almost surely (a.a.s) and with high probability (w.h.p.). For a.a.s., an event E depending on x is said to occur a.a.s. if $P(E_x) \to 1$ as $x \to \infty$. Meanwhile for w.h.p., if there exist a graph $\mathcal{G}$ and a graph property $\mathcal{P}$, it is be said that $\mathcal{G} \in \mathcal{P}$ w.h.p. if $P(\mathcal{G} \in \mathcal{P}) \to 1$ as the number of vertices $n \to \infty$.

Since a.a.s. and w.h.p. have nearly identical definitions, throughout this thesis, we are going to use only the terminology a.a.s.

Currently, Theorem 2.6.12 has not been widely used in practise. The only relevant paper is given by Carstens and Horadam (2013), who have used Theorem 2.6.12 to to study four weighted collaboration networks. In Carstens and Horadam (2013), β_0 and β_1 formed by Theorem 2.6.12 are used to determine the difference between a collaboration network and a random network. Moreover, the weights do not make any contribution when identifying the difference between the two models using the first two Betti numbers.

2.6.2 *Relevant Paper for Analysis of Brain Tree Data*

In Bendich et al. (2016), Persistent Homology has been introduced to study the human brain. The dataset has been constructed by using images from a 3-dimensional Magnetic Resonance Angiography (MRA). A tube-tracking

vessel segmentation algorithm was applied to the data collected and this information was combined into trees by a combination of automatic and manual techniques. In this way, each of the 98 such trees, taken from people between 18 and 72 years old, represents the tree of arteries in a person's brain.

While applying Persistent Homology, β_0 and β_1 are formed using different methods. By using mean-difference as test statistic, a two-sample permutation test has been performed. The resulting p-value is 0.03 for β_1 which suggests a sex effect is associated with the loops. They first used a sub-level set function for β_0. The sub-level set function is illustrated in Figure 2.6.1, the graph on the left is named as K and let $f(a)$ be the height of vertex a measured in the vertical direction. Extend f to a function on the edge set by setting $f(a,b) = \max(f(a), f(b))$ for each edge (a,b) of K. The persistence diagram $Dgm_0(f)$ takes K and f as input and outputs a multi-scale summary of the component evolution of the threshold sets of K. Each person's data then implies a persistent diagram, Dgm_0. For each of the 98 Dgm_0, the persistence of each dot is computed, which is the death time minus birth time, i.e. $d - b$. Then these lengths of barcodes are sorted in the descending order and picked the first 100 to produce a vector $(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{100})$ for each brain in 0-dimension.

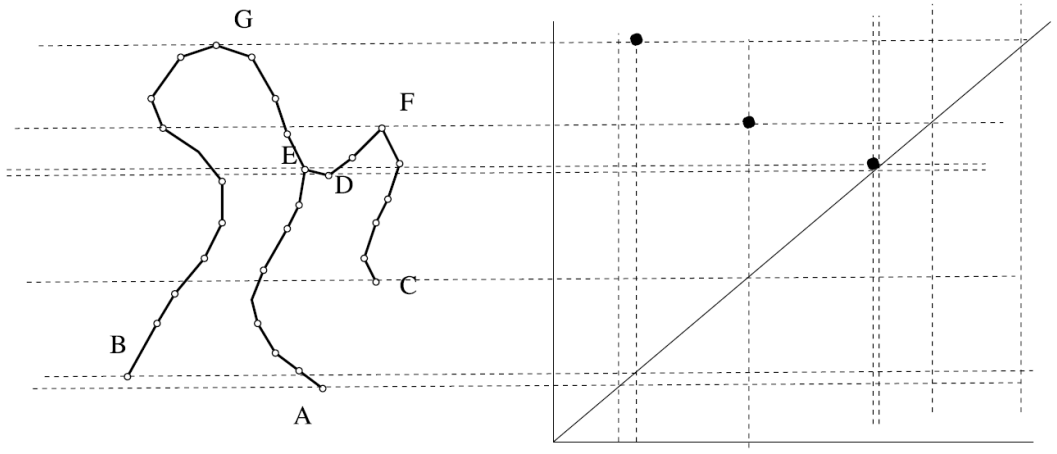

Figure 2.6.1: (left) graph K; (right) persistence diagram $Dgm_0(f)$ with function f measuring the height in the vertical direction. The coordinates of the dots are $(f(A), \infty)$, $(f(B), f(G))$, $(f(C), f(F))$ and $(f(D), f(E))$ respectively from left to right.

Secondly, the standard Rips complex is used for calculating β_1. The same procedure on the Dgm_1 leads to the vector $(\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_{100})$, in which the number q_j represents the size of the j-th most persistent loop in the brain.

Furthermore, they also have studied the sex effect by considering the arithmetic mean of the vector $(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{100})$ for both male and female subjects. The Euclidean distance between the two means are computed in $\mathbb{R}^{100}$ for β_0 , by a simple permutation test on the mean-difference statistic, which randomly assigns the 98 vectors into two groups of equal size, computes the difference between the means of two groups, and repeats this procedure 1000 times. In the test, 98 of the reassignments leads to a larger mean-difference than the original men-female split, giving an p-value of 0.1. However, by repeating the permutation test procedure for β_1, $(\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_{100})$ gives a lower p-value 0.03.

2.7 FURTHER TOPICS IN PERSISTENT HOMOLOGY

In Cohen-Steiner et al. (2007, 2010), they have discussed the stability theorem of persistence diagram. According this theorem, if Y' is a subsample of Y, then

$$W_k\left(Dgm_1\left(Y\right), Dgm_1\left(Y'\right)\right) \leq K \cdot d_H\left(Y, Y'\right)$$

where

$$W_k\left(D, D'\right) = \inf_{f: D \to D'} \left(\sum_{d \in D} \left\| d - f\left(d\right) \right\|_k \right)^{1/k}$$

is the k-th Wasserstein distance ; D and D' is short for $Dgm_1\left(Y\right), Dgm_1\left(Y'\right)$ respectively; f is an arbitrary bijection function; $d_H\left(Y, Y'\right)$ is the Hausdorff distance between Y and Y' (Edelsbrunner and Harer, 2010; Bendich et al., 2016). In the present setting, the Hausdorff distance between two sets $A \subset \mathbb{R}^p$ and $B \subset \mathbb{R}^p$ is defined as

$$d_H\left(A, B\right) = \max\left\{ \sup_{a \in A} \inf_{b \in B} d\left(a, b\right), \sup_{b \in B} \inf_{a \in A} d\left(a, b\right) \right\} \tag{2.7.1}$$

where $d\left(\cdot, \cdot\right)$ is the Euclidean distance between two points. This ensures that $Dgm_1\left(Y\right)$ is well approximated by $Dgm_1\left(Y'\right)$ if Y and Y' are close under the Hausdorff metric.

Otter et al. (2017) provided an overview paper for computational TDA. They have introduced 7 different packages written in Java or C++, which

are JavaPlex, JHoles, Perseus, Dionysus, PHAT, DIPHA, Gudhi SimpPers, Ripser. 5 different simplicial complexes including Rips and Čech complex have been computed using these 7 packages. Otter et al. (2017) suggested that the best suited library for the Rips complex is Gudhi while Dionysus is the most suitable packages for the Čech complex. Besides, Otter et al. (2017) summarised various fast algorithms to simplify the boundary matrix which is the most time-consuming part for working out the barcodes for different packages.

Most of the approaches to reducing sample size in Persistent Homology are based on reselecting subsamples from the original dataset. Then one or more test statistics can be produced from the calculated persistence diagrams, such as the mean of the lifetime which is presented in Bendich et al. (2016). However, Adler et al. (2017) have introduced a new approach which is first generating a persistence diagram from one sample of data. Secondly, by fitting a parametric model on this persistence diagram, a sequence of persistence diagrams can be generated by a Monte Carlo Markov chain method from this parametric model. This method has very distinct advantages and disadvantages. As the most time consuming process in Persistent Homology is getting the persistence diagram from the original dataset, therefore, this method only needs one persistence diagram to generate the rest of the diagrams. However, the drawback of this approach is that in general, as the first persistence diagram can only be generated using a portion of the full data. Therefore, generating the other persistence diagrams based on this one may cause loss of original information.

Other simplical complexes such as the alpha complex of Edelsbrunner and Harer (2010) have also been introduced by different researchers in other areas.

van de Weygaert et al. (2010); Van de Weygaert et al. (2011); van de Weygaert et al. (2011) have applied Persistent Homology to study the universe which shows a weblike network called the Cosmic Web. The Betti numbers and Euler characteristic formed by the alpha complex have been used to identify different models of the Cosmic Web.

Kovacev-Nikolic et al. (2016) apply Persistent Homology on the maltose-binding protein which is found in Escherichia coli where its primary function is to bind and transport sugar molecules across cell membranes. The protein can be either open or closed conformation. The closed confirmation occurs when ligand attaches to the protein molecule. The aim of this

paper is to distinguish between the state of a protein. By computing β_0 and using integrated distance between persistent landscape Bubenik (2015) as a test statistic, a two-sample permutation test has been performed. The resulting p-value is 5.83×10^{-4} which suggests a difference between open and closed conformation. In addition, they also suggest that the active sites are associated with loops i.e. β_1 in the protein.

Other research works have been done using Persistent Homology including sensor coverage area Ghrist and Muhammad (2005), image compression and segmentation Carlsson et al. (2008), shape classification Richardson and Werman (2014), and more.

3

SPECTRAL PROPERTIES OF STOCHASTIC BLOCK MODELS

3.1 INTRODUCTION

In this section, the main focus is on Stochastic Block Models (SBM) with a finite number of blocks, ζ. The aim is to determine, as far as possible, the spectral structure of the adjacency matrix and the normalized graph Laplacian defined in Section 2.1. As will be seen later, in Section 4.3, these results on spectral structure are relevant to the proof of CLT for Betti numbers in the SBM. Specifically, these results show that the method of proof of the CLT used in the Erdős-Rényi case breaks down in the SBM setting. The breakdown occurs due to the lack of separation of the larger eigenvalues.

The outline of this chapter is as follows. In Section 3.2, the spectral structure of the adjacency matrix in the 2-block model is determined under the asymptotic limit considered in (3.2.8). See in particular Proposition 3.2.3 and Proposition 3.2.7. It turns out that in this setting there are two eigenvalues which dominate in magnitude, with associated eigenvectors given by (3.2.9). In Section 3.3, the results are extended to the spectral structure of the adjacency matrix of the ζ-block model. The key results are Proposition 3.3.1 and Proposition 3.3.2. In this case, there are ζ eigenvalues which dominate in magnitude with associated eigenvectors as specified in (3.3.3) and (3.3.4). The derivation of the spectral structure of the normalised graph Laplacian in the ζ-block model is considered in Section 3.4. In the case of the normalised graph Laplacian, the situation is more complex than it is in the case of the adjacency matrix due to the dependencies which are introduced by the factor $\left(d_i d_j\right)^{-\frac{1}{2}}$ in (3.4.4). In this case, Proposition 3.4.1 is the relevant analogue of Proposition 3.3.1, where the latter result applies to the adjacency matrix. However, we do not yet have an analogue of Proposition

3.3.2, due to the added complexity which arises due to the dependencies mentioned above, but we believe that a result similar to Proposition 3.3.2 does hold. In Section 3.5, the relevance of the results in Chapter 3 to the CLT for Betti numbers in the SBM is explained; more details are given in Chapter 4.

In Table 3.1.1, some notations for this chapter are stated for convenience.

$\mathcal{V}_r$	set of vertices of type r
$\mathcal{V}$	$\bigcup_{r=1} \mathcal{V}_r$
N_r	$card(\mathcal{V}_r)$ Without lost of generality, let $N_1 \leq N_2 \leq ... \leq N_\zeta$
N	$\sum_{r=1}^{\zeta} N_r$ =Total number of vertices in the graph
r, s	letter used for vertex type $1 \leq r, s \leq \zeta$
i, j	letter used for vertex label $1 \leq i, j \leq N$
p_{rs}	probability $u \in \mathcal{V}_r$ and $v \in \mathcal{V}_s$ are connected by an edge
$p_{\min}$	$\min(p_{rs} : 1 \leq r, s \leq \zeta)$
$p_{\max}$	$\max(p_{rs} : 1 \leq r, s \leq \zeta)$
$\mathcal{G}((N_r), (p_{rs}), \zeta)$	SBM with ζ blocks, and N_r, p_{rs} are defined as above
$\mathbf{1}_N$	$N \times 1$ vector of ones
$\mathbf{I}_N$	$N \times N$ identity matrix
$f(N, d; p)$	$\binom{N}{d} p^d (1-p)^{N-d}$ binomial probability

Table 3.1.1: Some notations defined for Chapter 3.

3.2 ADJACENCY MATRIX: THE 2−BLOCK MODEL

In this section, the standard 2−block model is considered where $N_1 = card(\mathcal{V}_1)$, $N_2 = card(\mathcal{V}_2)$, $N = N_1 + N_2$ and $\mathcal{V} = \mathcal{V}_1 \bigcup \mathcal{V}_2$. Define the adjacency matrix $\mathbf{A} = (a_{ij})_{i,j=1}^{N}$ where $a_{ii} = 0$ for $i \in \mathcal{V}$ and

$$a_{ij} = \begin{cases} 0 & \text{no edge between } i \text{ and } j \\ 1 & \text{edge present between } i \text{ and } j. \end{cases}$$

In the stochastic 2−block model, the a_{ij} are independent random variables and for $i \neq j$,

$$P\left(a_{ij}=1\right)=\begin{cases} p_{11} & 1 \leq i,j \leq N_1 \\ p_{22} & N_1 \leq i,j \leq N \\ p_{12} & 1 \leq i \leq N_1 < j \leq N \\ p_{12} & 1 \leq j \leq N_1 < i \leq N. \end{cases} \tag{3.2.1}$$

The main goal in Section 3.2 is to determine the spectral structure of $\mathbf{A}$ as far as possible, where

$$\begin{aligned} &N \to \infty \\ &\frac{N_1}{N_2} \to \omega \in (0,\infty). \end{aligned} \tag{3.2.2}$$

It is also be assumed that $p_{rs} \to 0$ as $N \to \infty$, where $r,s \in \{1,2\}$.

Define $\bar{\mathbf{A}} = E(\mathbf{A})$ where the expectation is taken under the 2−block model with probabilities given by (3.2.1). Let $\mathbf{x}$ denote an $N \times 1$ unit vector, i.e. $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = 1$. Write

$$\mathcal{P}_2 = \{p_{11}, p_{22}, p_{12}\} \tag{3.2.3}$$

and define

$$p_2^* = \arg\min_{p \in \mathcal{P}_2} \left| p - \frac{1}{2} \right|. \tag{3.2.4}$$

Proposition 3.2.1. *For any given $N \times 1$ unit vector $\mathbf{x}$,*

$$Var\left[\mathbf{x}^T\left(\mathbf{A}-\bar{\mathbf{A}}\right)\mathbf{x}\right] \leq 2p_2^*\left(1-p_2^*\right). \tag{3.2.5}$$

Proof. Since $\mathbf{A}$ is the adjacency matrix for a 2−block model, the final sum below,

$$\begin{aligned} \mathbf{x}^T\left(\mathbf{A}-\bar{\mathbf{A}}\right)\mathbf{x} &= \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j\left(a_{ij}-\bar{a}_{ij}\right) \\ &= 2\sum_{1\leq i<j\leq N} x_i x_j\left(a_{ij}-\bar{a}_{ij}\right) \end{aligned}$$

consists of a double sum of independent random variables.

Consequently, as $Var(X+Y) = Var(X) + Var(Y)$ for independent random variables, it follows that

$$Var\left[\mathbf{x}^T\left(\mathbf{A}-\bar{\mathbf{A}}\right)\mathbf{x}\right] = 4\sum_{1\leq i<j\leq N} x_i^2 x_j^2 Var\left(a_{ij}\right). \tag{3.2.6}$$

The function $f(p) = p(1-p)$ has a maximum at $p = 0.5$, is monotonic increasing for $p < 0.5$, is monotonic decreasing for $p > 0.5$ and is symmetric about $p = 0.5$. Therefore,

$$\max_{i,j=1,...,N} Var\left(a_{ij}\right) \leq \max_{p\in\mathcal{P}_2} p(1-p) = p_2^*(1-p_2^*), \tag{3.2.7}$$

where $\mathcal{P}_2$ is defined in (3.2.3) and p_2^* is defined in (3.2.4). Continuing from the RHS of (3.2.6), it is found that

$$\begin{aligned}
4\sum_{1\leq i<j\leq N} x_i^2 x_j^2 Var\left(a_{ij}\right) &= 2\sum_{i=1}^{N}\sum_{j=1}^{N} x_i^2 x_j^2 Var\left(a_{ij}\right) \\
&\leq 2\sum_{i=1}^{N}\sum_{j=1}^{N} x_i^2 x_j^2 p_2^*(1-p_2^*) \\
&= 2p_2^*(1-p_2^*)\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{j=1}^{N} x_j^2\right) \\
&= 2p_2^*(1-p_2^*),
\end{aligned}$$

as $\mathbf{x}$ is a unit vector, so $\sum_{i=1}^{N} x_i^2 = 1$. □

Before moving on, the general version of Proposition 3.2.1 is given which applies to a general $\zeta-$block model.

For $r, s = 1, ..., \zeta$, let p_{rs} denote the probability of an edge being present between a vertex in $\mathcal{V}_r$ and a vertex in $\mathcal{V}_s$. Following (3.2.3) and (3.2.4), define

$$\mathcal{P}_\zeta = \{p_{rs} : 1 \leq r \leq s \leq \zeta\}$$

and

$$p_\zeta^* = \arg\min_{p\in\mathcal{P}_\zeta} \left|p - \frac{1}{2}\right|.$$

The general version of Proposition 3.2.1 is as follows.

Proposition 3.2.2. *For any given $N \times 1$ unit vector $\mathbf{x}$,*

$$Var\left[\mathbf{x}^T(\mathbf{A} - \bar{\mathbf{A}})\mathbf{x}\right] \leq 2p_\zeta^*\left(1 - p_\zeta^*\right).$$

Proof. The proof is almost identical to that of Proposition 3.2.1. All that changes is that $\mathcal{P}_2$ is replaced by $\mathcal{P}_\zeta$ and p_2^* is replaced by p_ζ^*. □

A full description of the spectral structure of $\bar{\mathbf{A}} = E(\mathbf{A})$ is now given where the expectation is taken under model (3.2.1).

Proposition 3.2.3. *The $N \times N$ matrix $\bar{\mathbf{A}}$ has the following spectral structure.*

1. *Any vector of the form $\left(\mathbf{v}_{N_1}^T, \mathbf{0}_{N_2}^T\right)^T$ where $\mathbf{v}_{N_1}$ is an $N_1 \times 1$ vector with $\mathbf{1}_{N_1}^T \mathbf{v}_{N_1} = 0$, is an eigenvector of $\bar{\mathbf{A}}$ with corresponding eigenvalue $\lambda = -p_{11}$.*
2. *Any vector of the form $\left(\mathbf{0}_{N_1}^T, \mathbf{v}_{N_2}^T\right)^T$ where $\mathbf{v}_{N_2}$ is an $N_2 \times 1$ vector with $\mathbf{1}_{N_2}^T \mathbf{v}_{N_2} = 0$, is an eigenvector of $\bar{\mathbf{A}}$ with corresponding eigenvalue $\lambda = -p_{22}$.*
3. *Suppose that the 2×2 matrix*

$$\mathbf{B}_2 = \begin{bmatrix} (N_1 - 1)\, p_{11} & N_2 p_{12} \\ N_1 p_{12} & (N_2 - 1)\, p_{22} \end{bmatrix} \tag{3.2.8}$$

has eigenvalues λ_1 and λ_2 with corresponding eigenvectors $(\gamma_{11}, \gamma_{12})^T$ and $(\gamma_{21}, \gamma_{22})^T$ respectively. Then $\bar{\mathbf{A}}$ has eigenvalues λ_1 and λ_2 with corresponding unit eigenvectors

$$\frac{1}{\delta_1} \begin{pmatrix} \gamma_{11} \mathbf{1}_{N_1} \\ \gamma_{12} \mathbf{1}_{N_2} \end{pmatrix} \text{ and } \frac{1}{\delta_2} \begin{pmatrix} \gamma_{21} \mathbf{1}_{N_1} \\ \gamma_{22} \mathbf{1}_{N_2} \end{pmatrix} \tag{3.2.9}$$

where

$$\delta_r = \left(\gamma_{r1}^2 N_1 + \gamma_{r2}^2 N_2\right)^{\frac{1}{2}},\ r = 1, 2. \tag{3.2.10}$$

Before proving Proposition 3.2.3, some remarks are presented.

Remark 3.2.4. In part 1 of Proposition 3.2.3, the relevant eigenspace has dimension $N_1 - 1$; the $'1'$ is subtracted because of the constraint $\mathbf{1}_{N_1}^T \mathbf{v}_{N_1} = 0$. Similarly, in part 2 of Proposition 3.2.3, the relevant eigenspace has dimension $N_2 - 1$; the $'1'$ is subtracted because of the constraint $\mathbf{1}_{N_2}^T \mathbf{v}_{N_1} = 0$. Finally, two eigenvalue/eigenvector combinations are identified in part 3 of Proposition 3.2.3 which are different to those identified in parts 1 and 2. Therefore, the total dimension of all the eigenspaces is

$$N_1 - 1 + N_2 - 1 + 1 + 1 = N_1 + N_2 = N$$

which establishes that all eigenvalue/eigenvector combinations have been identified.

Remark 3.2.5. From the eigenvalue equation

$$\det\left(\mathbf{B}_2 - \lambda \mathbf{I}_2\right) = 0$$

it is found that the eigenvalues in part 3 of Proposition 3.2.3 satisfy

$$\left[(N_1 - 1)\, p_{11} - \lambda\right]\left[(N_2 - 1)\, p_{22} - \lambda\right] - N_1 N_2 p_{12}^2 = 0$$

or, equivalently,

$$\lambda^2 - \lambda\left[(N_1-1)\,p_{11} + (N_2-1)\,p_{22}\right] + (N_1-1)\,(N_2-1)\,p_{11}p_{22} - N_1N_2p_{12}^2 = 0.$$

This implies that

$$\lambda = \frac{\left[(N_1-1)\,p_{11} + (N_2-1)\,p_{22}\right] \pm \sqrt{\left[(N_1-1)\,p_{11} - (N_2-1)\,p_{22}\right]^2 + 4N_1N_2p_{12}^2}}{2}. \tag{3.2.11}$$

Now write λ_1 and λ_2 for the eigenvalues with positive and negative square root terms respectively. Then, using standared arguments, eigenvectors corresponding to the eigenvalues λ_1 and λ_2 are given by

$$\begin{pmatrix} N_2p_{12} \\ \lambda_1 - (N_1-1)\,p_{11} \end{pmatrix} \text{ and } \begin{pmatrix} \lambda_2 - (N_2-1)\,p_{22} \\ N_1p_{12} \end{pmatrix} \tag{3.2.12}$$

respectively.

Remark 3.2.6. A specific and convenient choice for a set of orthonormal eigenvectors in parts 1 and 2 of Proposition 3.2.3 is given by the columns of the transpose Helment submatrix of appropriate dimension by Dryden and Mardia (2016).

Proof of Proposition 3.2.3. For a 2−block model arranged as indicated by (3.2.1), $\bar{\mathbf{A}}$, the expectation of $\mathbf{A}$ under model (3.2.1), may be written in block form as

$$\bar{\mathbf{A}} = \begin{bmatrix} p_{11}\left(\mathbf{1}_{N_1}\mathbf{1}_{N_1}^T - \mathbf{I}_{N_1}\right) & p_{12}\mathbf{1}_{N_1}\mathbf{1}_{N_2}^T \\ p_{12}\mathbf{1}_{N_2}\mathbf{1}_{N_1}^T & p_{22}\left(\mathbf{1}_{N_2}\mathbf{1}_{N_2}^T - \mathbf{I}_{N_2}\right) \end{bmatrix}, \tag{3.2.13}$$

where $\mathbf{1}_N$ is the $N \times 1$ vector of ones and $\mathbf{I}_N$ is the $N \times N$ identity matrix.

Proof of Part 1 of Proposition 3.2.3. Using (3.2.13), it is seen that

$$\bar{\mathbf{A}}\begin{pmatrix} \mathbf{v}_1 \\ \mathbf{0}_{N_2} \end{pmatrix} = \begin{pmatrix} p_{11}\mathbf{1}_{N_1}\left(\mathbf{1}_{N_1}^T\mathbf{v}_{N_1}\right) - p_{11}\mathbf{v}_{N_1} \\ p_{12}\mathbf{1}_{N_2}\mathbf{1}_{N_1}^T\mathbf{v}_{N_1} \end{pmatrix}.$$

For $\left(\mathbf{v}_{N_1}^T, \mathbf{0}_{N_2}^T\right)^T$ to be an eigenvector of $\bar{\mathbf{A}}$, it must have for some scalar λ,

$$\begin{pmatrix} p_{11}\mathbf{1}_{N_1}\left(\mathbf{1}_{N_1}^T\mathbf{v}_{N_1}\right) - p_{11}\mathbf{v}_{N_1} \\ p_{12}\mathbf{1}_{N_2}\left(\mathbf{1}_{N_1}^T\mathbf{v}_{N_1}\right) \end{pmatrix} = \lambda\begin{pmatrix} \mathbf{v}_{N_1} \\ \mathbf{0}_{N_2} \end{pmatrix},$$

or, equivalently,

$$p_{11}\mathbf{1}_{N_1}\left(\mathbf{1}_{N_1}^T\mathbf{v}_{N_1}\right) - p_{11}\mathbf{v}_{N_1} = \lambda\mathbf{v}_{N_1}, \tag{3.2.14}$$

and

$$p_{12}\mathbf{1}_{N_2}\left(\mathbf{1}_{N_1}^T\mathbf{v}_{N_1}\right) = \mathbf{0}_{N_2}. \tag{3.2.15}$$

Since $p_{12} \neq 0$, (3.2.15) can only be satisfied if $\mathbf{1}_{N_1}^T\mathbf{v}_{N_1} = 0$. Moreover, if $\mathbf{1}_{N_1}^T\mathbf{v}_{N_1} = 0$ then (3.2.14) is satisfied if $\lambda = -p_{11}$. Therefore, for any $\mathbf{v}_{N_1}$ such that $\mathbf{1}_{N_1}^T\mathbf{v}_{N_1} = 0$, $\left(\mathbf{v}_{N_1}^T, \mathbf{0}_{N_2}^T\right)^T$ will be an eigenvector of $\bar{\mathbf{A}}$ with corresponding eigenvalue $\lambda = -p_{11}$ as required. Note that the corresponding eigenspace has dimension $N_1 - 1$.

Proof of Part 2 of Proposition 3.2.3. The proof of part 2 is very similar to that of part 1. In this case, any vector of the form $\left(\mathbf{0}_{N_1}^T, \mathbf{v}_{N_2}^T\right)^T$ is an eigenvector of $\bar{\mathbf{A}}$ provided that $\mathbf{1}_{N_2}^T\mathbf{v}_{N_2} = 0$, with corresponding eigenvalue $\lambda = -p_{22}$, and associated eigenspace of dimension $N_2 - 1$.

Proof of Part 3 of Proposition 3.2.3. It will be checked directly that, with suitable choice of γ_1 and γ_2, $\left(\gamma_1\mathbf{1}_{N_1}^T, \gamma_2\mathbf{1}_{N_2}^T\right)^T$ is an eigenvector of $\bar{\mathbf{A}}$. In particular, using the block structure of $\bar{\mathbf{A}}$ in (3.2.13),

$$\begin{aligned}\bar{\mathbf{A}}\begin{pmatrix}\gamma_1\mathbf{1}_{N_1}\\ \gamma_2\mathbf{1}_{N_2}\end{pmatrix} &= \begin{bmatrix} p_{11}\left(\mathbf{1}_{N_1}\mathbf{1}_{N_1}^T - \mathbf{I}_{N_1}\right) & p_{12}\mathbf{1}_{N_1}\mathbf{1}_{N_2}^T \\ p_{12}\mathbf{1}_{N_2}\mathbf{1}_{N_1}^T & p_{22}\left(\mathbf{1}_{N_2}\mathbf{1}_{N_2}^T - \mathbf{I}_{N_2}\right)\end{bmatrix}\begin{pmatrix}\gamma_1\mathbf{1}_{N_1}\\ \gamma_2\mathbf{1}_{N_2}\end{pmatrix}\\ &= \begin{pmatrix}\left[p_{11}\left(N_1-1\right)\gamma_1 + p_{12}N_2\gamma_2\right]\mathbf{1}_{N_1}\\ \left[p_{12}N_1\gamma_1 + p_{22}\left(N_2-1\right)\gamma_2\right]\mathbf{1}_{N_2}\end{pmatrix}.\end{aligned} \tag{3.2.16}$$

So for $\left(\gamma_1\mathbf{1}_{N_1}^T, \gamma_2\mathbf{1}_{N_2}^T\right)^T$ to be an eigenvector of $\bar{\mathbf{A}}$, (3.2.16) must be a scalar multiple of $\left(\gamma_1\mathbf{1}_{N_1}^T, \gamma_2\mathbf{1}_{N_2}^T\right)^T$, in which case

$$\begin{cases} p_{11}\left(N_1-1\right)\gamma_1 + p_{12}N_2\gamma_2 = \lambda\gamma_1 \\ p_{12}N_1\gamma_1 + p_{22}\left(N_2-1\right)\gamma_2 = \lambda\gamma_2 \end{cases}$$

for some scalar λ, which in turn is equivalent to

$$\mathbf{B}_2\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} = \begin{bmatrix}\left(N_1-1\right)p_{11} & N_2p_{12}\\ N_1p_{12} & \left(N_2-1\right)p_{22}\end{bmatrix}\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} = \lambda\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix},$$

where $\mathbf{B}_2$ is defined in (3.2.8). Therefore λ_1 and λ_2, the required eigenvalues of $\bar{\mathbf{A}}$, are also eigenvalues of the 2×2 matrix $\mathbf{B}_2$. The corresponding eigenvectors of $\bar{\mathbf{A}}$ may be written as

$$\begin{pmatrix} \gamma_{11}\mathbf{1}_{N_1} \\ \gamma_{12}\mathbf{1}_{N_2} \end{pmatrix} \text{ and } \begin{pmatrix} \gamma_{21}\mathbf{1}_{N_1} \\ \gamma_{22}\mathbf{1}_{N_2} \end{pmatrix},$$

where

$$\begin{pmatrix} \gamma_{11} \\ \gamma_{12} \end{pmatrix} \text{ and } \begin{pmatrix} \gamma_{21} \\ \gamma_{22} \end{pmatrix}$$

are the eigenvectors of $\mathbf{B}_2$ corresponding to λ_1 and λ_2. The corresponding unit eigenvectors of $\bar{\mathbf{A}}$ are given by

$$\frac{1}{\delta_1}\begin{pmatrix} \gamma_{11}\mathbf{1}_{N_1} \\ \gamma_{12}\mathbf{1}_{N_2} \end{pmatrix} \text{ and } \frac{1}{\delta_2}\begin{pmatrix} \gamma_{21}\mathbf{1}_{N_1} \\ \gamma_{22}\mathbf{1}_{N_2} \end{pmatrix},$$

where

$$\delta_r = \left\| \begin{pmatrix} \gamma_{r1}\mathbf{1}_{N_1} \\ \gamma_{r2}\mathbf{1}_{N_2} \end{pmatrix} \right\| = \left(\gamma_{r1}^2 N_1 + \gamma_{r2} N_2\right)^{\frac{1}{2}},\ r = 1, 2$$

□

The next step is to identify a class of cases in which the largest eigenvalues in absolute value of $\mathbf{A}$ are determined by the largest eigenvalues of $\bar{\mathbf{A}}$. Here, the largest eigenvalues are the eigenvalues λ_1 and λ_2 obtained in part 3 of Proposition 3.2.3. The following asymptotic regime is considered. Write $p = p_{11} + p_{22} + p_{12}$. Since $N = N_1 + N_2$ and suppose that $p \to 0$, $N \to \infty$ and $Np \to \infty$;

$$\begin{aligned} &\frac{p_{11}(N_1 - 1)}{Np} \to \psi_{11}; \qquad \frac{p_{22}(N_2 - 1)}{Np} \to \psi_{22}; \\ &\frac{p_{12}N_1}{Np} \to \psi_{21}; \qquad \frac{p_{12}N_2}{Np} \to \psi_{12}; \\ &\psi_{11}, \psi_{12}, \psi_{21}, \psi_{22} \in (0, \infty). \end{aligned} \tag{3.2.17}$$

Note that (3.2.17) implies that

$$\frac{1}{Np}\mathbf{B}_2 \to \bar{\mathbf{\Psi}}_2 = (\psi_{rs})_{r,s=1,2}.$$

Proposition 3.2.7. *Suppose that $\bar{\mathbf{\Psi}}_2$ has full rank. Then*

$$\frac{1}{\delta_r^2}\left(\gamma_{r1}\mathbf{1}_{N_1}^T, \gamma_{r2}\mathbf{1}_{N_2}^T\right)\mathbf{A}\begin{pmatrix}\gamma_{r1}\mathbf{1}_{N_1}\\ \gamma_{r2}\mathbf{1}_{N_2}\end{pmatrix} \sim \lambda_r = Np\bar{\lambda}_r$$

in probability, where

$$\begin{cases}\bar{\lambda}_1 = \frac{\psi_{11}+\psi_{22}+\sqrt{(\psi_{11}-\psi_{22})^2+4\psi_{12}\psi_{21}}}{2}\\ \bar{\lambda}_2 = \frac{\psi_{11}+\psi_{22}-\sqrt{(\psi_{11}-\psi_{22})^2+4\psi_{12}\psi_{21}}}{2} \neq 0\end{cases} \tag{3.2.18}$$

are the eigenvalues of $\bar{\mathbf{\Psi}}_2$.

Proof. The assumption that $\bar{\mathbf{\Psi}}_2$ has full rank implies that $\bar{\lambda}_1$ and $\bar{\lambda}_2$ are non-zero (this is immediate for $\bar{\lambda}_1$). Write

$$\mathbf{x}_r = \frac{1}{\delta_r}\begin{pmatrix}\gamma_{r1}\mathbf{1}_{N_1}\\ \gamma_{r2}\mathbf{1}_{N_2}\end{pmatrix},\ r = 1, 2.$$

Then

$$\mathbf{x}_r^T\mathbf{A}\mathbf{x}_r = \mathbf{x}_r^T\bar{\mathbf{A}}\mathbf{x}_r + \mathbf{x}_r^T(\mathbf{A}-\bar{\mathbf{A}})\mathbf{x}_r.$$

Note that $\mathbf{x}_r^T\bar{\mathbf{A}}\mathbf{x}_r = \lambda_r = Np\bar{\lambda}_r$ as $N \to \infty$, where $\bar{\lambda}_1$ and $\bar{\lambda}_2$ are defined in (3.2.18). For $\mathbf{x}_r^T(\mathbf{A}-\bar{\mathbf{A}})\mathbf{x}_r$, combining Proposition 3.2.1 with Chebychev's inequality,

$$\begin{aligned}P\left[\left|\mathbf{x}_i^T(\mathbf{A}-\bar{\mathbf{A}})\mathbf{x}_i\right| > (p_2^*)^{\frac{1}{4}}\right] &\leq \frac{Var\left\{\mathbf{x}_i^T(\mathbf{A}-\bar{\mathbf{A}})\mathbf{x}_i\right\}}{\left(p_2^*\right)^{\frac{2}{4}}}\\ &\leq \frac{2p_2^*\left(1-p_2^*\right)}{\left(p_2^*\right)^{\frac{1}{2}}}\\ &\leq 2\left(p_2^*\right)^{\frac{1}{2}} \to 0\end{aligned}$$

as $N \to \infty$, since p_2^*, defined in (3.2.4) converges to 0 in the asymptotic regime considered here. □

3.3 ADJACENCY MATRIX: THE $\zeta-$BLOCK MODEL

The purpose of this section is to generalize Proposition 3.2.3 and 3.2.7 from the 2-block case to the $\zeta-$block case. In the ζ-block case, there are ζ types of vertex as opposed to just 2. It is supposed that there are N_r vertices of

type r $(r = 1, ..., \zeta)$ and that the vertices have been labelled so that vertex i is of type r if

$$\begin{cases} i \leq N_1 & r = 1 \\ N_1 + ... + N_{r-1} < i \leq N_1 + ... + N_r & r = 2, ..., \zeta. \end{cases}$$

It is assumed also that the probability of a link between a type r vertex and type s vertex is given by p_{rs} $(r, s = 1, ..., \zeta)$. As before define $\mathbf{A}$ to be the adjacency matrix and let $\bar{\mathbf{A}}$ be the expectation of $\mathbf{A}$ under the ζ-block model with probabilities $(p_{rs})_{r,s=1}^{\zeta}$. The $N \times N$ matrix $\bar{\mathbf{A}}$, where $N = \sum_{t=1}^{\zeta} N_t$ may be written in block form as $\bar{\mathbf{A}} = (\bar{\mathbf{A}}_{rs})_{r,s=1}^{\zeta}$ where

$$\bar{\mathbf{A}}_{rs} = \begin{cases} p_{rr}\left(\mathbf{1}_{N_r}\mathbf{1}_{N_r}^T - \mathbf{I}_{N_r}\right) & r = s \\ p_{rs}\mathbf{1}_{N_r}\mathbf{1}_{N_s}^T & r \neq s. \end{cases} \tag{3.3.1}$$

The ζ-block analogue of Proposition 3.2.3 is now stated.

Proposition 3.3.1. *The $N \times N$ matrix $\bar{\mathbf{A}}$ has the following spectral structure.*

1. Any vector of the form

$$\left(\mathbf{v}_{N_1}^T, \mathbf{0}_{N_2}^T, ..., \mathbf{0}_{N_\zeta}^T\right)^T, \left(\mathbf{0}_{N_1}^T, ..., \mathbf{0}_{N_{r-1}}^T, \mathbf{v}_{N_r}^T, \mathbf{0}_{N_{r+1}}^T, ..., \mathbf{0}_{N_\zeta}^T\right)^T \text{ or } \left(\mathbf{0}_{N_1}^T, ..., \mathbf{0}_{N_{\zeta-1}}^T, \mathbf{v}_{N_\zeta}^T\right)^T$$

where $\mathbf{v}_{N_r}$ is an $N_r \times 1$ vector with $\mathbf{1}_{N_r}^T\mathbf{v}_{N_r} = 0$, is an eigenvector of $\bar{\mathbf{A}}$ with corresponding eigenvalue $\lambda = -p_{rr}$.

2. Define $\mathbf{B}_\zeta = (b_{rs})_{r,s=1}^{\zeta}$ where

$$b_{rs} = \begin{cases} p_{rr}(N_r - 1) & r = s \\ p_{rs}N_s & r \neq s, \end{cases} \tag{3.3.2}$$

and suppose that $\mathbf{B}_\zeta$ has eigenvalues $\lambda_1, ..., \lambda_\zeta$ with corresponding eigenvectors

$$\begin{pmatrix} \gamma_{11} \\ \vdots \\ \gamma_{1\zeta} \end{pmatrix} ... \begin{pmatrix} \gamma_{\zeta 1} \\ \vdots \\ \gamma_{\zeta\zeta} \end{pmatrix}. \tag{3.3.3}$$

Then $\bar{\mathbf{A}}$ *has eigenvalues* $\lambda_1, ..., \lambda_\zeta$ *with corresponding eigenvectors*

$$\begin{pmatrix} \gamma_{11}\mathbf{1}_{N_1} \\ \vdots \\ \gamma_{1\zeta}\mathbf{1}_{N_\zeta} \end{pmatrix} ... \begin{pmatrix} \gamma_{\zeta 1}\mathbf{1}_{N_1} \\ \vdots \\ \gamma_{\zeta\zeta}\mathbf{1}_{N_\zeta} \end{pmatrix} \tag{3.3.4}$$

Proof of Part 1 of Theorem 3.3.1. Using the block structure of $\bar{\mathbf{A}}$ indicated in (3.3.1), it is seen that

$$\begin{aligned} \bar{\mathbf{A}} \begin{pmatrix} \mathbf{0}_{N_1} \\ \vdots \\ \mathbf{0}_{N_{r-1}} \\ \mathbf{v}_{N_r} \\ \mathbf{0}_{N_{r+1}} \\ \vdots \\ \mathbf{0}_{N_\zeta} \end{pmatrix} &= \begin{bmatrix} \bar{\mathbf{A}}_{11} & \bar{\mathbf{A}}_{12} & \cdots & \bar{\mathbf{A}}_{1\zeta} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{22} & \cdots & \bar{\mathbf{A}}_{2\zeta} \\ \vdots & \vdots & \cdots & \vdots \\ \bar{\mathbf{A}}_{\zeta 1} & \bar{\mathbf{A}}_{\zeta 2} & \cdots & \bar{\mathbf{A}}_{\zeta\zeta} \end{bmatrix} \begin{pmatrix} \mathbf{0}_{N_1} \\ \vdots \\ \mathbf{0}_{N_{r-1}} \\ \mathbf{v}_{N_r} \\ \mathbf{0}_{N_{r+1}} \\ \vdots \\ \mathbf{0}_{N_\zeta} \end{pmatrix} \\ &= \begin{pmatrix} \bar{\mathbf{A}}_{1r}\mathbf{v}_{N_r} \\ \bar{\mathbf{A}}_{2r}\mathbf{v}_{N_r} \\ \vdots \\ \bar{\mathbf{A}}_{\zeta r}\mathbf{v}_{N_r} \end{pmatrix}. \end{aligned} \tag{3.3.5}$$

The vector

$$\left(\mathbf{0}_{N_1}^T, ..., \mathbf{0}_{N_{r-1}}^T, \mathbf{v}_{N_r}^T, \mathbf{0}_{N_{r+1}}^T, ..., \mathbf{0}_{N_\zeta}^T\right)^T$$

is an eigenvector of $\bar{\mathbf{A}}$ if and only if it has (3.3.5) as a scalar multiple. A necessary and sufficient condition for this vector to have (3.3.5) as a scalar multiple is that $\mathbf{1}_{N_r}^T\mathbf{v}_{N_r} = 0$, in which case the corresponding eigenvalue is $\lambda = -p_{rr}$.

Similarly,

$$\left(\mathbf{v}_{N_1}^T, \mathbf{0}_{N_2}^T, ..., \mathbf{0}_{N_\zeta}^T\right)^T$$

is an eigenvector of $\bar{\mathbf{A}}$ if and only if $\mathbf{1}_{N_1}\mathbf{v}_{N_1} = 0$, in which case the corresponding eigenvalue is $\lambda = -p_{11}$; and

$$\left(\mathbf{0}_{N_1}^T, ..., \mathbf{0}_{N_{\zeta-1}}^T, \mathbf{v}_{N_\zeta}^T\right)^T$$

is an eigenvector of $\bar{\mathbf{A}}$ if and only if $\mathbf{1}_{N_\zeta}^T\mathbf{v}_{N_\zeta} = 0$, in which case the corresponding eigenvalue is $\lambda = -p_{\zeta\zeta}$.

Proof of Part 2 of Proposition 3.3.1. Since for $r, s = 1, ..., \zeta$, using (3.3.1) and (3.3.2),

$$\bar{\mathbf{A}}_{rs}\mathbf{1}_{N_s} = \begin{cases} p_{rr}\left(N_r - 1\right)\mathbf{1}_{N_r} & r = s \\ p_{rs}N_s\mathbf{1}_{N_r} & r \neq s. \end{cases}$$

It follows that, for $r = 1, ..., \zeta$,

$$\sum_{s=1}^{\zeta} \bar{\mathbf{A}}_{rs}\mathbf{1}_{N_s} = \left\{ p_{rr}\left(N_r - 1\right)\gamma_r + \sum_{s\neq r} p_{rs}N_s\gamma_s \right\} \mathbf{1}_{N_r}. \tag{3.3.6}$$

Therefore, for general $\gamma_1, ..., \gamma_\zeta$, and using (3.3.6),

$$\begin{aligned} \bar{\mathbf{A}} \begin{pmatrix} \gamma_1\mathbf{1}_{N_1} \\ \vdots \\ \gamma_\zeta\mathbf{1}_{N_\zeta} \end{pmatrix} &= \begin{bmatrix} \sum_{s=1}^{\zeta} \bar{\mathbf{A}}_{1s}\gamma_s\mathbf{1}_{N_s} \\ \vdots \\ \sum_{s=1}^{\zeta} \bar{\mathbf{A}}_{\zeta s}\gamma_s\mathbf{1}_{N_s} \end{bmatrix} \\ &= \begin{bmatrix} \left\{ p_{11}\left(N_1 - 1\right)\gamma_1 + \sum_{s=2}^{\zeta} p_{1s}N_s\gamma_s \right\} \mathbf{1}_{N_1} \\ \vdots \\ \left\{ p_{\zeta\zeta}\left(N_\zeta - 1\right)\gamma_\zeta + \sum_{s=1}^{\zeta-1} p_{\zeta s}N_s\gamma_s \right\} \mathbf{1}_{N_\zeta} \end{bmatrix}. \end{aligned} \tag{3.3.7}$$

It follows from (3.3.7) that the condition for

$$\left(\gamma_{r1}\mathbf{1}_{N_1}^T, ..., \gamma_{r\zeta}\mathbf{1}_{N_\zeta}^T\right)^T$$

is an eigenvector of $\bar{\mathbf{A}}$ if and only if for some scalar λ,

$$\begin{cases} p_{11}\left(N_1 - 1\right)\gamma_1 + \sum_{s=2}^{\zeta} p_{1s}N_s\gamma_s = \lambda\gamma_1 \\ \vdots \\ p_{\zeta\zeta}\left(N_\zeta - 1\right)\gamma_\zeta + \sum_{s=1}^{\zeta-1} p_{\zeta s}N_s\gamma_s = \lambda\gamma_1 \end{cases}$$

which in turn is equivalent to

$$\mathbf{B}_\zeta \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_\zeta \end{pmatrix} = \lambda \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_\zeta \end{pmatrix}, \tag{3.3.8}$$

where

$$\begin{aligned} \mathbf{B}_\zeta &= (b_{rs})_{r,s=1}^{\zeta} \\ &= \begin{bmatrix} (N_1-1)\,p_{11} & N_2 p_{12} & N_3 p_{13} & ... & N_\zeta p_{1\zeta} \\ N_1 p_{12} & (N_2-1)\,p_{22} & N_3 p_{23} & \dots & N_\zeta p_{2\zeta} \\ N_3 p_{13} & N_3 p_{23} & (N_3-1)\,p_{33} & \dots & N_\zeta p_{3\zeta} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ N_1 p_{1\zeta} & N_\zeta p_{2\zeta} & N_\zeta p_{3\zeta} & \dots & (N_\zeta-1)\,p_{\zeta\zeta} \end{bmatrix} \end{aligned}$$

is defined in (3.3.2). Consequently, from (3.3.8), and writing $\lambda_1, ..., \lambda_\zeta$ for the eigenvalues of $\mathbf{B}_\zeta$ with corresponding eigenvectors

$$\begin{pmatrix} \gamma_{11} \\ \vdots \\ \gamma_{1\zeta} \end{pmatrix} ... \begin{pmatrix} \gamma_{\zeta 1} \\ \vdots \\ \gamma_{\zeta\zeta} \end{pmatrix}$$

respectively, it follows from (3.3.7) that the vector

$$\left(\gamma_{r1}\mathbf{1}_{N_1}^T, ..., \gamma_{r\zeta}\mathbf{1}_{N_\zeta}^T\right)^T$$

is an eigenvector of $\bar{\mathbf{A}}$ with corresponding eigenvalue λ_r, for $r = 1, ..., \zeta$. □

Now, a generalization of Proposition 3.2.7 is considered to the case of a ζ-block model. Assume the same notation and setup as was considered in Proposition 3.3.1. Define $N = \sum_{r=1}^{\zeta} N_r$ and $p = \sum_{1\le r\le s\le\zeta} p_{rs}$. As previously, it is assumed that $N \to \infty$, $p \to 0$ and $Np \to \infty$ and that the following limits exist: for $r = 1, ..., \zeta$,

$$\frac{p_{rr}\,(N_r - 1)}{Np} \to \psi_{rr} \tag{3.3.9}$$

and for $1 \leq r \neq s \leq \zeta$,

$$\frac{p_{sr}N_r}{Np} \to \psi_{sr}, \ \frac{p_{rs}N_s}{Np} \to \psi_{rs} \tag{3.3.10}$$

where $\psi_{rs} \in (0, \infty)$ for all $1 \leq r, s \leq \zeta$.

Write

$$\bar{\mathbf{\Psi}}_\zeta = (\psi_{rs})_{r,s=1,...,\zeta}.$$

The next result is a generalization of Proposition 3.2.7 to the $\zeta-$block model.

Proposition 3.3.2. *Suppose that $\bar{\mathbf{\Psi}}_\zeta$ has full rank. Then for $r = 1, .., \zeta$,*

$$\frac{1}{\delta_r^2}\left(\gamma_{r1}\mathbf{1}_{N_1}^T, ..., \gamma_{r\zeta}\mathbf{1}_{N_\zeta}^T\right)\mathbf{A}\begin{pmatrix}\gamma_{r1}\mathbf{1}_{N_1}\\ \vdots \\ \gamma_{r\zeta}\mathbf{1}_{N_\zeta}\end{pmatrix} \sim \lambda_r = Np\bar{\lambda}_r$$

where $\delta_r = \left(\sum_{s=1}^{\zeta}\gamma_{rs}^2 N_s\right)^{\frac{1}{2}}$ *and* $\bar{\lambda}_1, ..., \bar{\lambda}_\zeta$ *are the non-zero eigenvalues of* $\bar{\mathbf{\Psi}}_\zeta$.

Proof. The proof is the same as that for Proposition 3.2.7. □

3.4 THE NORMALIZED GRAPH LAPLACIAN FOR $\zeta-$BLOCK MODELS

The goal of this section is to determine the asymptotic spectral structure of the expectation of the normalized graph Laplacian under the $\zeta-$block model.

Define for any vertex i, $1 \leq i \leq N$,

$$\tilde{d}_i = card\left\{j \in \{1, .., N\} : a_{ij} = 1\right\}, \tag{3.4.1}$$

where $a_{ij} = 1$ if there is an edge between vertex i and vertex j, and $a_{ij} = 0$ otherwise. Then define

$$d_i = \max\left\{\tilde{d}_i, 1\right\}. \tag{3.4.2}$$

Note that $\tilde{d}_i$ is the degree of vertex i and $d_i = \tilde{d}_i$ unless $\tilde{d}_i = 0$, i.e. unless i is an isolated vertex. In the asymptotic framework which we consider, it will be seen later that the graph will be connected with probability approaching to 1 as $N \to \infty$, so that the difference between and $\tilde{d}_i$ and d_i does not concern us.

The normalized graph Laplacian is defined here by

$$\mathbf{L} = \mathbf{I}_N - \mathbf{D}_N^{-\frac{1}{2}} \mathbf{A} \mathbf{D}_N^{-\frac{1}{2}} \tag{3.4.3}$$

where $\mathbf{D}_N = \text{diag}\{d_1, ..., d_N\}$ and d_i is defined in (3.4.2). The focus of interest is the spectral structure (i.e. eigenvalues and eigenvectors) of $\mathbf{L}$ in (3.4.3). However, it is slightly easier to work with $\mathbf{J} = \mathbf{D}_N^{-\frac{1}{2}} \mathbf{A} \mathbf{D}_N^{-\frac{1}{2}}$. Then λ is an eigenvalue of $\mathbf{J}$ if and only if $1 - \lambda$ is an eigenvalue of $\mathbf{L}$. Moreover, the eigenvectors of $\mathbf{J}$ and $\mathbf{L}$ are the same.

Our main goal now is to find an asymptotic expression for $\bar{\mathbf{J}} = E(\mathbf{J})$, where the expectation is taken under the $\zeta-$block model. Note that $\bar{\mathbf{J}}$ has the same type of block structure as $\bar{\mathbf{A}}$ written in (3.3.1):

$$\bar{\mathbf{J}} = (\bar{\mathbf{J}}_{rs})_{r,s=1,...,\zeta} = \begin{cases} q_{rr} \left(\mathbf{1}_{N_r} \mathbf{1}_{N_r}^T - \mathbf{I}_{N_r} \right) & r = s \\ q_{rs} \mathbf{1}_{N_r} \mathbf{1}_{N_s}^T & r \neq s \end{cases}$$

where

$$q_{rs} = E\left[\frac{a_{ij}}{\sqrt{d_i d_j}} \right], \; i \neq j, \tag{3.4.4}$$

with i and j vertices of type r and type s, respectively. In Proposition 3.4.1 below, we find an asymptotic expression for the q_{rs} in (3.4.4).

As before we define $N = N_1 + ... + N_\zeta$ where $N_1 \leq ... \leq N_\zeta$ and $p = \sum_{1 \leq r \leq s \leq \zeta} p_{rs}$, and assume $N \to \infty$, $p \to 0$ and $Np \to \infty$. The asymptotic regime indicated in (3.3.9) and (3.3.10) in Section 3.3 is also assumed.

Proposition 3.4.1. *Suppose that for some constants $C > 0$ and $\varepsilon > 0$,*

$$N_1 p_{\min} \geq C N^{\varepsilon} \tag{3.4.5}$$

where $p_{\min} = \min(p_{rs} : 1 \leq r, s \leq \zeta)$. Then as $N \to \infty$, for $r, s = 1, ..., \zeta$,

$$N_s q_{rs} \to \frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma} \right) \left(\sum_{\gamma=1}^{\zeta} \psi_{s\gamma} \right)}}$$

or, equivalently, we can rewrite as

$$q_{rs} \sim \frac{1}{N_s} \cdot \frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{s\gamma}\right)}}.$$

Proof. Let ρ_{is} denote the number of edges between vertex i $(i = 1, .., N)$ and vertices of type s $(s = 1, ..., \zeta)$. Write

$$\boldsymbol{\rho}_i = \left(\rho_{i1}, ..., \rho_{i\zeta}\right)^T.$$

Note that $\rho_{is} \sim Binomial\left(N_{is}, p_{rs}\right)$ where vertex i is of type r and

$$\tilde{d}_i = \sum_{s=1}^{\zeta} \rho_{is},$$

where $\tilde{d}_i$ is defined in 3.4.1. The quantity

$$q_{rs} = E\left[\frac{a_{ij}}{\sqrt{d_i d_j}}\right],$$

where i is of type r and j is of type s, is difficult to calculate directly. The plan here is to perform an asypmtotic calculation in the following steps.

Step 1 Calculate the conditional expectation

$$E\left[a_{ij} | \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right].$$

Step 2 Find an asymptotic expression for the expectation over $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_j$ of

$$\frac{1}{\sqrt{d_i d_j}} E\left[a_{ij} | \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right].$$

Define the binomial probability

$$f\left(N, d; p\right) = \binom{N}{d} p^d \left(1 - p\right)^{N-d}.$$

Also for $r = 1, ..., \zeta$, write

$$N_{jr} = \begin{cases} N_r - 1 & \text{if } j \text{ is of type } r = s \\ N_r & \text{if } j \text{ is of type } r \neq s. \end{cases}$$

Step 1 Using Bayes Theorem we write the conditional expectation as

$$\begin{aligned} E\left[a_{ij}|\boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] &= P\left[a_{ij} = 1|\boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] \\ &= \frac{P\left[a_{ij} = 1, \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right]}{P\left[a_{ij} = 1, \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] + P\left[a_{ij} = 0, \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right]} \\ &= \frac{1}{1 + T_{ij}} \end{aligned} \tag{3.4.6}$$

where

$$T_{ij} = \frac{P\left[a_{ij} = 0, \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right]}{P\left[a_{ij} = 1, \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right]}. \tag{3.4.7}$$

From elementary considerations,

$$\begin{aligned} P\left[a_{ij} = 1, \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] =& p_{rs} \cdot f\left(N_{is} - 1, \rho_{is} - 1; p_{rs}\right) f\left(N_{jr} - 1, \rho_{jr} - 1; p_{rs}\right) \\ &\times \left[\prod_{t \neq s} f\left(N_{it}, \rho_{it}; p_{rt}\right)\right] \left[\prod_{t \neq r} f\left(N_{jt}, \rho_{jt}; p_{ts}\right)\right] \end{aligned} \tag{3.4.8}$$

and

$$\begin{aligned} P\left[a_{ij} = 0, \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] =& \left(1 - p_{rs}\right) f\left(N_{is} - 1, \rho_{is}; p_{rs}\right) f\left(N_{jr} - 1, \rho_{jr}; p_{rs}\right) \\ &\times \left[\prod_{t \neq s} f\left(N_{it}, \rho_{it}; p_{rt}\right)\right] \left[\prod_{t \neq r} f\left(N_{jt}, \rho_{jt}; p_{ts}\right)\right]. \end{aligned} \tag{3.4.9}$$

The most important points here are that the product over $t \neq s$ on the RHS of (3.4.8) is the same as the product over $t \neq s$ on the RHS of (3.4.9), and the product over $t \neq r$ on the RHS of (3.4.8) is the same as the product

over $t \neq r$ on the RHS of (3.4.9). Therefore, substituting (3.4.8) and (3.4.9) into (3.4.7) and cancelling, it is seen that

$$\begin{aligned}
T_{ij} &= \frac{(1-p_{rs})}{p_{rs}} \cdot \frac{f\left(N_{is}-1, \rho_{is}; p_{rs}\right) f\left(N_{jr}-1, \rho_{jr}; p_{rs}\right)}{f\left(N_{is}-1, \rho_{is}-1; p_{rs}\right) f\left(N_{jr}-1, \rho_{jr}-1; p_{rs}\right)} \\
&= \frac{(1-p_{rs})}{p_{rs}} \cdot \frac{\binom{N_{is}-1}{\rho_{is}} p_{rs}^{\rho_{is}} \left(1-p_{rs}\right)^{(N_{is}-1)-\rho_{is}} \cdot \binom{N_{jr}-1}{\rho_{jr}} p_{rs}^{\rho_{jr}} \left(1-p_{rs}\right)^{\left(N_{jr}-1\right)-\rho_{jr}}}{\binom{N_{is}-1}{\rho_{is}-1} p_{rs}^{\rho_{is}-1} \left(1-p_{rs}\right)^{N_{is}-\rho_{is}} \cdot \binom{N_{jr}-1}{\rho_{jr}-1} p_{rs}^{\rho_{jr}-1} \left(1-p_{rs}\right)^{N_{jr}-\rho_{jr}}} \\
&= \frac{p_{rs}^2 \left(1-p_{rs}\right)}{p_{rs} \left(1-p_{rs}\right)^2} \cdot \frac{\frac{(N_{is}-1)!}{\rho_{is}!(N_{is}-1-\rho_{is})!} \cdot \frac{\left(N_{jr}-1\right)!}{\rho_{jr}!\left(N_{jr}-1-\rho_{jr}\right)!}}{\frac{(N_{is}-1)!}{(\rho_{is}-1)!(N_{is}-\rho_{is})!} \cdot \frac{\left(N_{jr}-1\right)!}{\left(\rho_{jr}-1\right)!\left(N_{jr}-\rho_{jr}\right)!}} \\
&= \frac{p_{rs}}{1-p_{rs}} \cdot \frac{\left(N_{is}-\rho_{is}\right)\left(N_{jr}-\rho_{jr}\right)}{\rho_{is}\rho_{jr}} \\
&= \frac{p_{rs}}{1-p_{rs}} \cdot \frac{\left(1-\frac{\rho_{is}}{N_{is}}\right)\left(1-\frac{\rho_{jr}}{N_{jr}}\right)}{\frac{\rho_{is}}{N_{is}} \cdot \frac{\rho_{jr}}{N_{jr}}}.
\end{aligned} \tag{3.4.10}$$

Therefore, substituting (3.4.10) into (3.4.6) and rearranging,

$$\begin{aligned}
E\left[a_{ij} | \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] &= \frac{1}{1+T_{ij}} \\
&= \frac{\left(1-p_{rs}\right) \frac{\rho_{is}}{N_{is}} \cdot \frac{\rho_{jr}}{N_{jr}}}{\left(1-p_{rs}\right) \frac{\rho_{is}}{N_{is}} \cdot \frac{\rho_{jr}}{N_{jr}} + p_{rs}\left(1-\frac{\rho_{is}}{N_{is}}\right)\left(1-\frac{\rho_{jr}}{N_{jr}}\right)}.
\end{aligned}$$

This completes Step 1 of the proof.

Step 2 Define

$$\begin{aligned}
\tilde{e}_{ij}^{(N)} &= \tilde{e}_{ij} \\
&= \frac{1}{\sqrt{d_i d_j}} E\left[a_{ij} | \boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] \\
&= \frac{1}{\sqrt{d_i d_j}} \cdot \frac{\left(1-p_{rs}\right) \frac{\rho_{is}}{N_{is}} \cdot \frac{\rho_{jr}}{N_{jr}}}{\left(1-p_{rs}\right) \frac{\rho_{is}}{N_{is}} \cdot \frac{\rho_{jr}}{N_{jr}} + p_{rs}\left(1-\frac{\rho_{is}}{N_{is}}\right)\left(1-\frac{\rho_{jr}}{N_{jr}}\right)}
\end{aligned} \tag{3.4.11}$$

and write

$$e_{ij}^{(N)} = e_{ij} = N_{is} \tilde{e}_{ij}.$$

Recall that the definition of d_i in (3.4.2), and the fact that

$$0 \leq E\left[a_{ij}|\boldsymbol{\rho}_i, \boldsymbol{\rho}_j\right] \leq 1,$$

it follows that $0 \leq \tilde{e}_{ij} \leq 1$ and so $0 \leq e_{ij} \leq N$. For $\delta \in \left(\frac{1}{2}, 1\right)$ define the event

$$D_{N,s,\delta,r} = \left\{|\rho_{is} - N_{is}p_{rs}| \leq \left(N_{is}p_{rs}\right)^{\delta}\right\},$$

where vertex i is of type r.

Now define

$$D_{N,\delta,r} = \bigcap_{s=1}^{\zeta} D_{N,s,\delta,r}.$$

Under the assumptions of Proposition 3.4.1, on the event $D_{N,\delta,r}$ as $N \to \infty$,

$$\begin{aligned} |\rho_{is} - N_{is}p_{rs}| &= \left|\frac{\rho_{is}}{N_{is}p_{rs}} - 1\right| \\ &\leq \frac{\left(N_{is}p_{rs}\right)^{\delta}}{N_{is}p_{rs}} \\ &\leq \left(N_{is}p_{rs}\right)^{-(1-\delta)} \to 0, \end{aligned} \tag{3.4.12}$$

where vertex i is of type r. Similarly,

$$\left|\frac{\rho_{jr}}{N_{jr}p_{rs}} - 1\right| \leq \left(N_{jr}p_{rs}\right)^{-(1-\delta)} \to 0, \tag{3.4.13}$$

where vertex j is of type s, and

$$\begin{aligned}\left|\frac{d_i}{Np}-\sum_{s=1}^{\zeta}\psi_{rs}\right| &= \left|\frac{\sum_{s=1}^{\zeta}\rho_{is}}{Np}-\sum_{s=1}^{\zeta}\psi_{rs}\right| \\ &\leq \sum_{s=1}^{\zeta}\left|\frac{\rho_{is}}{Np}-\psi_{rs}\right| \\ &= \sum_{s=1}^{\zeta}\left|\frac{\rho_{is}}{Np}\cdot\frac{1}{\psi_{rs}}-1\right| \\ &= \sum_{s=1}^{\zeta}\left|\frac{\rho_{is}}{Np}\cdot\frac{Np}{p_{rs}N_s}-1\right| \\ &\leq \sum_{s=1}^{\zeta}\left(N_{is}p_{rs}\right)^{-(1-\delta)} \\ &\leq CN^{-(1-\delta)}\to 0.\end{aligned} \tag{3.4.14}$$

It follows that on $D_{N,\delta,r}$, when N is sufficiently large, for any $\tau>0$, (3.4.11) implies that

$$\begin{aligned}e_{ij} =& N_s\cdot\frac{1}{\sqrt{d_id_j}}\cdot\frac{(1-p_{rs})\,\frac{\rho_{is}}{N_{is}}\cdot\frac{\rho_{jr}}{N_{jr}}}{(1-p_{rs})\,\frac{\rho_{is}}{N_{is}}\cdot\frac{\rho_{jr}}{N_{jr}}+p_{rs}\left(1-\frac{\rho_{is}}{N_{is}}\right)\left(1-\frac{\rho_{jr}}{N_{jr}}\right)} \\ =& N_s\cdot\frac{Np}{\sqrt{d_id_j}}\cdot\frac{\frac{1}{Np}\cdot\frac{1}{p_{rs}}\cdot(1-p_{rs})\cdot\frac{\rho_{is}}{N_{is}}\cdot\frac{\rho_{jr}}{N_{jr}}}{\frac{1}{p_{rs}}\cdot(1-p_{rs})\,\frac{\rho_{is}}{N_{is}}\cdot\frac{\rho_{jr}}{N_{jr}}+\left(1-\frac{\rho_{is}}{N_{is}}\right)\left(1-\frac{\rho_{jr}}{N_{jr}}\right)} \\ =& N_s\cdot\frac{1}{\sqrt{\frac{d_i}{Np}\cdot\frac{d_j}{Np}}}\cdot\frac{\frac{\rho_{is}}{Np}\cdot\frac{\rho_{jr}}{N_{jr}p_{rs}}\cdot\frac{1}{N_{is}}(1-p_{rs})}{(1-p_{rs})\,\frac{\rho_{is}}{N_{is}p_{rs}}\cdot\frac{\rho_{jr}}{N_{jr}}+\left(1-\frac{\rho_{is}}{N_{is}}\right)\left(1-\frac{\rho_{jr}}{N_{jr}}\right)} \\ \to& \frac{1}{\sqrt{\left(\sum_{\gamma=1}^{\zeta}\psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta}\psi_{\gamma s}\right)}}\cdot\frac{\psi_{rs}\cdot 1\cdot 1}{1\cdot 1\cdot p_{rs}+1\cdot 1} \\ =& \frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta}\psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta}\psi_{\gamma s}\right)}} \\ \leq& (1+\tau)\,\frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta}\psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta}\psi_{\gamma s}\right)}}\end{aligned}$$

in probability.

From Bernstein's inequality and using (3.4.5),

$$\begin{aligned}
P\left(D_{N,\delta,r}^{c}\right) &= P\left[\left(\bigcap_{s=1}^{\zeta} D_{N,s,\delta,r}\right)^{c}\right] \\
&= P\left(\bigcup_{s=1}^{\zeta} D_{N,s,\delta,r}^{c}\right) \\
&\leq \sum_{s=1}^{\zeta} P\left(D_{N,s,\delta,r}^{c}\right) \\
&= \sum_{s=1}^{\zeta} P\left[\left|\frac{\rho_{is}}{N_{is}p_{rs}} - 1\right| \geq \left(N_{is}p_{rs}\right)^{-(1-\delta)}\right] \\
&\leq C_0 \exp\left\{-C_1\left(N_{is}p_{rs}\right)^{-2(1-\delta)}\right\} \\
&\leq C_0 \exp\left\{-C_1 N^{\varepsilon}\right\},
\end{aligned}$$

for some $C_0 > 0$, $C_1 > 0$ and $\varepsilon > 0$. Therefore, using

$$E\left(|X|^{\alpha}\right) \leq P\left(A\right) \max_{x\in A} |x|^{\alpha} + P\left(A^{c}\right) \max_{x\in A^{c}} |x|^{\alpha},$$

for $\alpha > 1$ and $\tau > 0$ and N sufficiently large,

$$\begin{aligned}
E\left(e_{ij}^{\alpha}\right) &\leq P\left(D_{N,\delta,r}^{c}\right) \cdot N^{\alpha} + P\left(D_{N,\delta,r}\right) \cdot (1+\tau)^{\alpha} \frac{\psi_{rs}^{\alpha}}{\left[\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{\gamma s}\right)}\right]^{\alpha}} \\
&\leq C_0 \exp\left\{-C_1 N^{\varepsilon}\right\} N^{\alpha} \\
&\quad + \left(1 - C_0 \exp\left\{-C_1 N^{\varepsilon}\right\}\right) \cdot (1+\tau)^{\alpha} \frac{\psi_{rs}^{\alpha}}{\left[\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{\gamma s}\right)}\right]^{\alpha}} \\
&\to 0 + (1+\tau)^{\alpha} \frac{\psi_{rs}^{\alpha}}{\left[\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{\gamma s}\right)}\right]^{\alpha}}.
\end{aligned}$$

Therefore, using Proposition 2.2.3, it follows that the family of random variables

$$\left\{e_{ij} = e_{ij}^{(N)} : N \geq 1\right\}$$

is uniformly integrable. Moreover, (3.4.12), (3.4.13) and (3.4.14) imply that

$$e_{ij} \xrightarrow{P} \frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{\gamma s}\right)}}.$$

Consequently, we may use Theorem 2.2.4 to conclude that

$$E\left(e_{ij}\right) \to \frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{\gamma s}\right)}}$$

as $N \to \infty$. Therefore,

$$N_s q_{rs} \to \frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{\gamma s}\right)}}$$

which is equivalent to

$$q_{rs} \sim \frac{1}{N_s} \cdot \frac{\psi_{rs}}{\sqrt{\left(\sum_{\gamma=1}^{\zeta} \psi_{r\gamma}\right)\left(\sum_{\gamma=1}^{\zeta} \psi_{\gamma s}\right)}}$$

as required. □

3.5 DISCUSSION

The relevance of the results in this chapter to the CLT for Betti numbers in the SBM is now explained. For simplicity we focus on the following cases in the 2-block model:

$$p_{11} = p_{22} = p_0;\ p_{12} = p_{21} = \theta p_0;\ N_1 = N_2 = N_0, \tag{3.5.1}$$

where $\theta > 0$ is fixed. Note that if $\theta \neq 1$ then the model is SBM but not an ERM.

From (3.2.11), and using (3.5.1),

$$\begin{aligned}\lambda =& \frac{(N_0-1)p_0 + (N_0-1)p_0 \pm 2N_0\theta p_0}{2} \\ =& N_0 p_0 \bar{\lambda},\end{aligned}$$

where

$$\bar{\lambda} = 1 - \frac{1}{N_0} \pm \theta.$$

Consequently, the eigenvalues on this scale are given by

$$\bar{\lambda}_1 = 1 - \frac{1}{N_0} + \theta,\ \bar{\lambda}_2 = 1 - \frac{1}{N_0} - \theta. \tag{3.5.2}$$

By choosing the constant θ sufficiently close to 0 we can make the ratio $\frac{\bar{\lambda}_2}{\bar{\lambda}_1}$ arbitrarily close to 1. Therefore the two largest eigenvalues of $\bar{\mathbf{A}}$, the expectation of the adjacency matrix, are arbitrarily close together on the scale for which $\bar{\lambda}_1$ and $\bar{\lambda}_2$ are bounded away from 0 and bounded above.

The conclusion to be drawn is that the SGT fails to hold in this case when θ is small. The SGT and its role in the proof of the CLT for Betti numbers in the ERM is explained in Chapter 4.

The above comments apply to the adjacency matrix. We believe the same conclusions hold for the normalized graph Laplacian but our results in this direction are still incomplete. Specifically, we do not yet have an analogue of Proposition 3.3.2 for the normalized graph Laplacian. These findings lead to the following questions.

1. In those cases of the SBM where there is sufficiently large separation between the largest and second largest eigenvalue, can the CLT for Betti numbers be proved using the same method of proof as in the ERM case, as given in Kahle and Meckes (2013, 2015)?

2. Does the CLT for Betti numbers in the SBM still hold in general, even though the method of proof breaks down in a broad range of cases? Or, alternatively, does the proof break down because the CLT does not hold in general in the SBM? In Chapter 4 it is proved that the answer to Question 1 is affirmative; see Theorem 4.7.2. Question 2 is an open question and we do not know the answer.

4

TOWARDS THE CLT FOR BETTI NUMBERS IN THE STOCHASTIC BLOCK MODEL

4.1 INTRODUCTION

In this chapter, the aim is to go as far as possible in proving the CLT for Betti numbers in the stochastic block model (SBM). This CLT has been proved by Kahle and Meckes (2013, 2015) in the special case of the Erdős-Rényi model (ERM).

The outline of this chapter is as follows. Since the structure of the proof is rather complex even in the ERM case, our first goal is to study the structure of the proof given by Kahle and Meckes (2013, 2015) in detail. This is done in Section 4.2. It turns out that some parts of the proof extend to the SBM case without difficulty while in other parts of the proof there are serious difficulties in extending the proof. The most serious difficulties arise in proving a suitable form of the spectral gap theorem (SGT) for SBM. Our results in Chapter 3 show that suitable versions of the SGT do not hold for the SBM in wide generality. These difficulties are discussed in detail in Section 4.3. We also discuss and where possible prove component results which do generalise to the SBM case including the lower vanishing threshold (Section 4.5) and the upper vanishing threshold (Section 4.6). In Section 4.7, it is proved that the CLT for the SBM does hold for the subclass of SBMs which satisfy a sufficiently strong version of the spectral gap theorem. In Section 4.8, some simulation results for CLT for the SBM are presented.

An important question is the following: is the failure to extend the method of proof of the CLT due to Kahle and Meckes (2013, 2015) to the SBM a question of the method of proof breaking down but the CLT still holding; or does the CLT in fact fail to hold in generality in the SBM? We do not know the

answer to this question. It will be interesting to see if it can be resolved in future work.

In Table 4.1.1, some notations for this chapter are stated for convenience.

$\mathcal{V}_r$	vertices set for type r
$\mathcal{V}$	$\bigcup_{i=1} \mathcal{V}_r$
N_r	$card\left(\mathcal{V}_r\right)$ Without lost of generality, let $N_1 \leq N_2 \leq ... \leq N_\zeta$
N	$\sum_{r=1}^{\zeta} N_r$ =Total number of vertices in the graph
p_{rs}	probability $u \in \mathcal{V}_r$ and $v \in \mathcal{V}_s$ are connected by an edge
$p_{\min}$	$\min\left(\{p_{rs}\}_{1\leq r\leq s\leq \zeta}\right)$
$p_{\max}$	$\max\left(\{p_{rs}\}_{1\leq r\leq s\leq \zeta}\right)$
$\mathcal{G}\left(\left(N_r\right),\left(p_{rs}\right),\zeta\right)$	SBM with ζ blocks, and N_r, p_{rs} are defined as above
$\mathcal{X}$	$\mathcal{X} \sim \mathcal{X}\left(\mathcal{G}\right)$ where $\mathcal{G} \sim \mathcal{G}\left(\left(N_r\right),\left(p_{rs}\right),\zeta\right)$
$\mathbf{1}_N$	$N \times 1$ vector of ones
$\mathbf{I}_N$	$N \times N$ identity matrix
ξ_R	simplices formed by $\lvert n_R\rvert$ numbers of vertices
ξ_S	simplices formed by $\lvert n_S\rvert$ numbers of vertices
ξ_T	simplices formed by $\lvert n_T\rvert$ numbers of vertices (Without lost of generality, $\lvert n_R\rvert \geq \lvert n_S\rvert \geq \lvert n_T\rvert$)
a_i; α_i; γ_i	$card\left(\{i : i \in \xi_R\}\right)$; $card\left(\{i : i \in \xi_S\}\right)$; $card\left(\{i : i \in \xi_T\}\right)$
η	$card\left\{i : i \in \xi_R \bigcup \xi_S \bigcup \xi_T\right\}$
η_i	$card\left\{i : i \in \{\xi_R \bigcap \xi_S\} \bigcap \mathcal{V}_i\right\}$
$\eta_{S\cap T\setminus R,i}$	$card\left\{i : i \in \{\xi_S \bigcap \xi_T \setminus \xi_R\} \bigcap \mathcal{V}_i\right\}$
$\eta_{R\cap S\cap T,i}$	$card\left\{i : i \in \{\xi_S \bigcap \xi_T \bigcap \xi_R\} \bigcap \mathcal{V}_i\right\}$
$T\left(n,a\right)$	$\frac{n!}{a!(n-a)!}$ Binomial coefficient
$\tau\left(a,b\right)$	$T\left(a+b,2\right) - T\left(a,2\right) - T\left(b,2\right)$
$f\left(a_r, k+1, \zeta\right)$	$\sum_{a_1=0}^{k+1} \sum_{a_2=0}^{k+1} ... \sum_{a_\zeta=0}^{k+1}$ $\sum_{r=1}^{\zeta} a_r = k+1$
$\mathbb{I}_R$	the vector that is 1 in every coordinate corresponding to set R and 0 elsewhere

Table 4.1.1: Some notation defined for Chapter 4.

4.2 STRUCTURE OF PROOF OF CLT IN ERM

[1]	Kahle and Meckes (2013)
[2]	Kahle and Meckes (2015)
[3]	Kahle (2014)
[4]	Hoffman et al. (2019)
[5]	Kahle (2009)
[6]	Ballmann and Światkowski (1997)

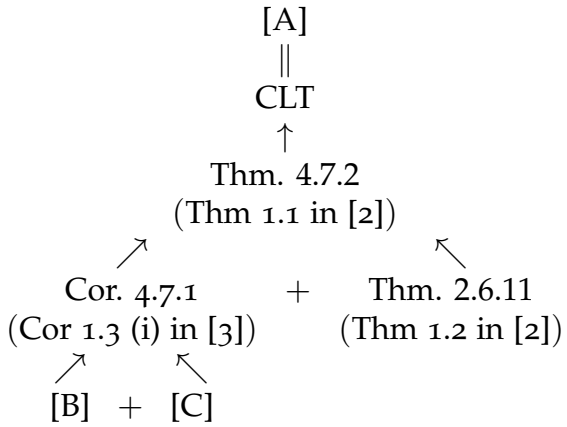

Figure 4.2.1: Proof structure of statement A, the Central Limit Theorem for Betti Numbers.

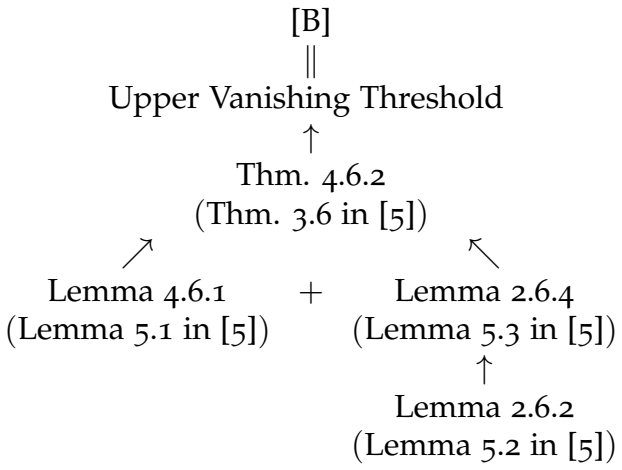

Figure 4.2.2: Proof structure of statement B, the Upper Vanishing Threshold

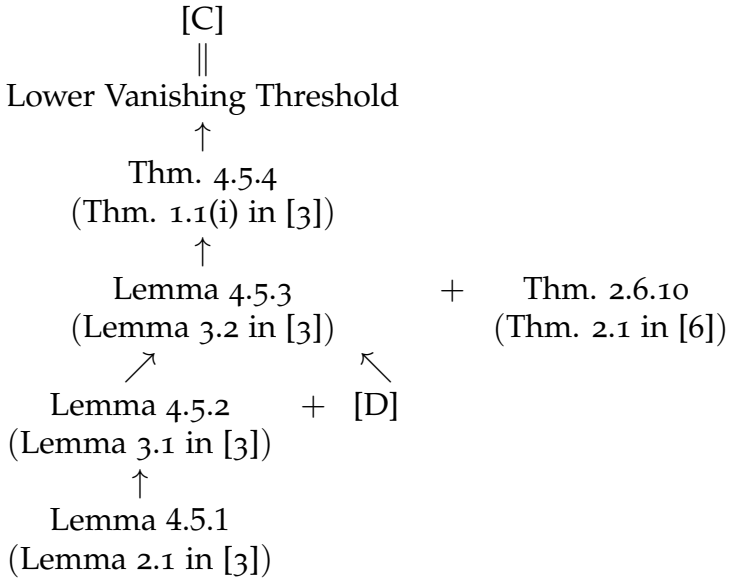

Figure 4.2.3: Proof of statement C, the Lower Vanishing Threshold.

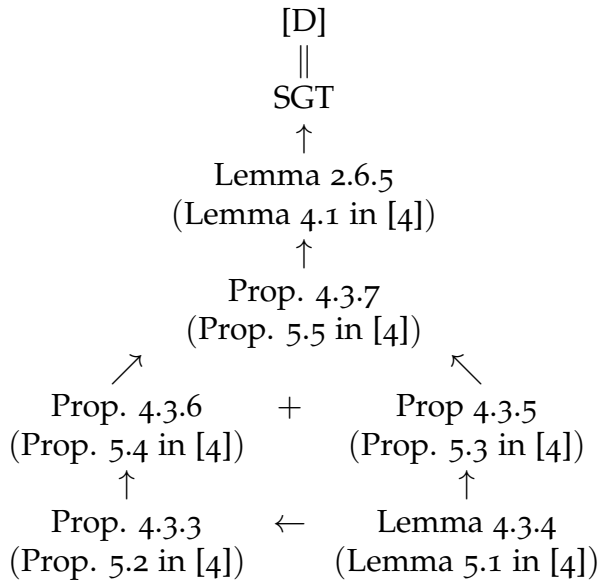

Figure 4.2.4: Proof structure of statement D, the Spectral Gap Theorem (SGT)

Thm./Lemma/Prop.	Origin	Result
Cor. 4.7.1	Cor 1.3 (i) in [3]	Solved under assumption
Thm. 2.6.11	Thm 1.2 in [2]	Unchanged
Thm. 4.7.2	Thm 1.1 in [2]	Solved under assumption

Table 4.2.1: Table of proof structure of statement A, the Central Limit Theorem for Betti Numbers.

Thm./Lemma/Prop.	Origin	Result
Lemma 4.6.1	Lemma 5.1 in [5]	Solved
Lemma 2.6.2	Lemma 5.2 in [5]	Unchanged
Lemma 2.6.4	Lemma 5.3 in [5]	Unchanged
Thm. 4.6.2	Thm. 3.6 in [5]	Solved

Table 4.2.2: Table of proof structure of statement B, the Upper Vanishing Threshold.

Thm./Lemma/Prop.	Origin	Result
Lemma 4.5.1	Lemma 2.1 in [3]	Solved
Lemma 4.5.2	Lemma 3.1 in [3]	Solved
Lemma 4.5.3	Lemma 3.2 in [3]	Proof under assumption
Thm. 4.5.3	Thm 2.1 in [6]	Unchanged
Thm. 4.5.4	Thm. 1.1(i) in [3]	Proof under assumption

Table 4.2.3: Table of proof of statement C, the Lower Vanishing Threshold.

Thm./Lemma/Prop.	Origin	Results
Lemma 4.3.4	Lemma 5.1 in [4]	Solved
Prop. 4.3.3	Prop. 5.2 in [4]	Unsolved
Prop 4.3.5	Prop. 5.3 in [4]	Solved
Prop. 4.3.6	Prop. 5.4 in [4]	Solved
Prop. 4.3.7	Prop. 5.5 in [4]	Unsolved
Lemma 2.6.5	Lemma 4.1 in [4]	Unsolved

Table 4.2.4: Table of proof structure of statement D, the Spectral Gap Theorem (SGT)

To extend CLT from ERM to SBM, a similar pattern of proof is followed as for SBM. The main theorem CLT for ζ-block model $\mathcal{G}\left((N_r), (p_{rs}), \zeta\right)$ is given as Theorem 4.7.2. Theorem 4.7.2 states that for each k and certain range of $P = \{p_{\min} \leq p_{rs} \leq p_{\max} : 1 \leq r \leq s \leq \zeta\}$, $\beta_k \geq 0$ a.a.s and in this regime β_k follows a normal distribution with mean $E(\beta_k)$ and variance $Var(\beta_k)$. The

range of P for β_k is presented in Corollary 4.7.1 where the proof is given by induction for both $i < k$ and $i > k$.

The next step is to prove the upper vanishing threshold for the range of P, which is given in Theorem 4.6.2. Theorem 4.6.2 states that for each k, with $p_{\max} \leq p$, $\beta_k = 0$ a.a.s. Theorem 4.6.2 is proved by 3 lemmas where Lemma 2.6.4 and 2.6.2 are counting the non-trivial k-cycles in the sample graph which are irrelevant to ERM. Lemma 4.6.1 for Theorem 4.6.2 proves that there is no k-complex for subgraph with size $M + k + 1$ where M is a function of k for $p_{\max} \leq p$. As a result, Theorem 4.6.2 proves that $\mathcal{G}$ is formed by the subgraph with zero k-cycle, i.e. $\beta_k = 0$.

For the lower vanishing threshold for the CLT regime, the SGT is introduced, which we have not been able to prove for the SBM in this thesis. In Section 4.3, we show that SGT breaks down in 4 different places. However, for CLT for SBM, the full force of the SGT is not required. The only requirement is that $\lambda_2 > 1 - \frac{1}{k+1}$ as N tends to infinity. This implies that as k increases, more edges are required for β_k. We proved in Section 3.5 that for a 2-block model, λ_2 is close to 1. The simulation results in Section 4.4 suggest that the λ_2 for SBM is always bounded by the λ_2 value from ERM. As a result, we assume SGT is true for the lower vanishing threshold for SBM.

For the rest of the lower vanishing threshold, we first prove Lemma 4.5.1 and 4.5.2 which are not related to SGT. Lemma 4.5.1 states that there are no $(k+1)$-simplices in any graph if probability $p_{\min} \geq p$. Whereas Lemma 4.5.2 indicates that a subgraph with $k+1$ vertices of the ζ-block model is not a $(k+1)$-simplex with probability $p_{\min} \geq p$. Lemma 4.5.2 is proved by Lemma 4.5.1. Furthermore, Lemma 4.5.3 proves that the conditions for using SGT on a subgraph with $k+1$ vertices are satisfied. Finally, the main theorem, Theorem 4.5.4 states that for each k, with $p_{\min} \geq p$, $\beta_k = 0$ a.a.s. We prove Theorem 4.5.4 by checking every subgraph with size $k+1$ in $\mathcal{G}((N_r), (p_{rs}), \zeta)$ satisfies the conditions on SGT, if SGT holds, then Theorem 4.5.3 suggests that $\beta_k = 0$.

4.3 SPECTRAL GAP THEOREM FOR SBMS

In this section, we are going to show how Lemma 2.6.5 breaks down in four different places when it is applied to the general SBM.

First, two addition lemmas are proved below.

Lemma 4.3.1. *Let* $X_i \overset{indep}{\sim} Bernoulli\,(p_i)$ *where* $i = 1, ..., N$ *and* $X = \sum_i X_i$, $\mu = \sum_i p_i$. *Then for any* $t < \mu$

$$P\,(X \leq t) \leq \exp\left\{-\mu + t\left(1 + \log\frac{\mu}{t}\right)\right\}.$$

Proof. For any $\lambda \in \mathbb{R}$, the MGF for $X_i \overset{indep}{\sim} Bernoulli\,(p_i)$ can be bounded by

$$\begin{aligned}
E\left(e^{\lambda X}\right) &= E\left(e^{\lambda \sum X_i}\right) \\
&= \prod E\left(e^{\lambda X_i}\right) \\
&= \prod\left[1 + p_i\left(e^{\lambda} - 1\right)\right] \\
&\leq \prod \exp\left\{p_i\left(e^{\lambda} - 1\right)\right\} \quad (1 + x \leq e^x,\ x > 0) \\
&= \exp\left\{\sum p_i\left(e^{\lambda} - 1\right)\right\} \\
&= \exp\left\{\mu\left(e^{\lambda} - 1\right)\right\}.
\end{aligned}$$

If $\lambda < 0$, then by Markov's inequality,

$$\begin{aligned}
P\,(X \leq t) &= P\left(e^{\lambda X} \geq e^{\lambda t}\right) \\
&\leq \frac{E\left(e^{\lambda X}\right)}{e^{\lambda t}} \\
&\leq \exp\left\{\mu\left(e^{\lambda} - 1\right) - \lambda t\right\}.
\end{aligned}$$

Assuming that $t < \mu$, let $\lambda = \log\left(\frac{t}{\mu}\right)$, which gives

$$\begin{aligned}
P\,(X \leq t) &\leq \exp\left\{\mu\left(e^{\log\left(\frac{t}{\mu}\right)} - 1\right) - t\log\left(\frac{t}{\mu}\right)\right\} \\
&= \exp\left\{-\mu + t\left(1 + \log\frac{\mu}{t}\right)\right\}.
\end{aligned}$$

□

Lemma 4.3.2. *Let* $X_i \overset{indep}{\sim} Bernoulli\,(p_i)$ *where* $i = 1, ..., N$ *and* $X = \sum_i X_i$ *and* $\mu = \sum_i p_i$. *Then for any* $t > 4$

$$P\,(X \geq t\mu) \leq \exp\left\{-\frac{t\mu\log t}{3}\right\}.$$

Proof. For any $\lambda \in \mathbb{R}$, the MGF for $X_i \overset{indep}{\sim} Bernoulli\,(p_i)$ can be bounded by

$$\begin{aligned}
E\left(e^{\lambda X}\right) &= E\left(e^{\lambda \sum X_i}\right) \\
&= \prod E\left(e^{\lambda X_i}\right) \\
&= \prod \left[1 + p_i\left(e^{\lambda} - 1\right)\right] \\
&\leq \prod \exp\left\{p_i\left(e^{\lambda} - 1\right)\right\} \quad (1 + x \leq e^{x},\ x > 0) \\
&= \exp\left\{\sum p_i\left(e^{\lambda} - 1\right)\right\} \\
&= \exp\left\{\mu\left(e^{\lambda} - 1\right)\right\}.
\end{aligned}$$

If $\lambda > 0$, then by Markov's inequality,

$$\begin{aligned}
P\left(X \geq t\mu\right) &= P\left(e^{\lambda X} \geq e^{\lambda t \mu}\right) \\
&\leq \frac{E\left(e^{\lambda X}\right)}{e^{\lambda t \mu}} \\
&\leq \exp\left\{\mu\left(e^{\lambda} - 1\right) - \lambda t \mu\right\}.
\end{aligned}$$

For $t > 1$, let $\lambda = \log(t)$, which gives

$$\begin{aligned}
P\left(X \geq t\mu\right) &\leq \exp\left\{\mu\left(e^{\log(t)} - 1\right) - t\mu\log(t)\right\} \\
&= \exp\left\{\mu\left(t - t\log t - 1\right)\right\}.
\end{aligned}$$

As $\mu\left(t-1\right) - \mu t \log t \leq -\dfrac{t\mu\log t}{3}$ is required, we need to show $(t-1) \leq \frac{2}{3} t \log t$ for $t > 4$. Since $\frac{t}{t-1}\log t$ is a increasing function for $t > 1$, $\frac{4}{3}\log 4 \geq \frac{3}{2}\log 3 \geq \frac{3}{2}$. This implies $3 \leq \frac{2}{3} \cdot 4 \log 4$. Therefore,

$$\begin{aligned}
P\left(X \geq t\mu\right) &\leq \exp\left\{\mu\left(\frac{2}{3}t\log t - t\log t\right)\right\} \\
&= \exp\left\{-\frac{\mu t \log t}{3}\right\}.
\end{aligned}$$

□

4.3.1 *Difficulty 1 of Proof of SGT*

Lemma 4.3.3 below, which is condition 2 of Lemma 2.6.5 is the first place where the proof of the SBM fails.

In the proof of SGT, since only the upper bound is required, the ideal scenario is that $d = (N-1)p$ for ERM in Lemma 2.6.5 can be replaced directly by $d_{\max} = (N-1)p_{\max}$. However, this is not true for Proposition 4.3.3. The original updated version of Condition 2 of Lemma 2.6.5 is given as Lemma 4.3.3.

Lemma 4.3.3. *For each $\delta > 0$ and $m \geq 0$, there is a constant $C = C(\delta, m)$ sufficiently large so that if $p_{\min} \geq \frac{\delta \log N}{N}$ then*

$$\sup_{\substack{\|\mathbf{x}\| = 1 \\ \mathbf{x}^T \mathbf{1}_N = 0 \\ \|\mathbf{y}\| = 1}} \left|\mathbf{x}^T \mathbf{A} \mathbf{y}\right| \leq C\sqrt{d_{\max}}$$

with probability at least $1 - C\exp\left\{-md_{\max}^2\right\}$

Proof. Define

$$T = \left\{\mathbf{x} \in \left(\frac{1}{2\sqrt{N}}\mathbb{Z}\right)^N : \|\mathbf{x}\| \leq 1\right\} \text{ and } U = \left\{\mathbf{x} \in T : \sum_i x_i = 0\right\}.$$

By Lemma 2.6.6, $U = \left\{\mathbf{x} : \|\mathbf{x}\| = 1, x^T \mathbf{1} = 0\right\}$ is in the convex hull of T. Let $Q = \{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$. For any vector $\mathbf{x} \in Q$, we can find a hypercube C with length of side $\frac{\epsilon}{\sqrt{N}}$ as described below.

Fix any weight vector α where $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$. Choose $\mathbf{x} \in Q$ such that $\sum \alpha_i x_i$ attains its maximum value. Therefore, if $i_1, ..., i_l$ are the indices of non-integer coordinates, choose a non-integer coordinates i_j by adding γ to x_{i_j}, objective function $\sum \alpha_i x_i$ changes by $\gamma \alpha_{i_j}$. If the sign of γ is chosen to be the sign of α_{i_j}, the objective function does not decrease. Increase $|\gamma| > 0$ until x_{i_j} becomes integer. Repeat this process for each i_j until x_{i_j} is an integral value.

By Lemma 2.6.7, let $\mathbf{z_x} = \frac{1}{2}\mathbf{x}$ where $\mathbf{x} \in S$ and $\mathbf{z_y} = \frac{1}{2}\mathbf{y}$ where $\mathbf{y} \in Q$. Then $\mathbf{z_x} = \sum_i \alpha_i \mathbf{v}_i$, $\mathbf{v}_i \in T$ and $\mathbf{z_y} = \sum_j \alpha_j \mathbf{v}_j$, $\mathbf{v}_j \in U$. The result of Lemma 2.6.7 follows and $\frac{1}{(1-\epsilon)^2} = \left(\frac{1}{2}\right)^{-2} = 4$. Thus,

$$\sup_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T\mathbf{1}_N=0 \\ \|\mathbf{y}\|=1}} \left|\mathbf{x}^T\mathbf{A}\mathbf{y}\right| \leq 4 \sup_{\substack{\mathbf{x}\in U \\ \mathbf{y}\in T}} \left|\mathbf{x}^T\mathbf{A}\mathbf{y}\right|.$$

For a fixed pair of vectors $(\mathbf{x},\mathbf{y}) \in U \times T$, define the light couples $L = L(\mathbf{x},\mathbf{y})$ to be all ordered pairs $(u,v) \in N \times N$ such that $|x_u y_v| \leq \frac{\sqrt{d_{\max}}}{N}$ and let heavy couples $H = H(x,y)$ be all those pairs that are not light. The notation which will be used is the following

$$Y = l(\mathbf{x},\mathbf{y}) = \sum_{(u,v)\in L} x_u A_{uv} y_v \tag{4.3.1}$$

$$h(\mathbf{x},\mathbf{y}) = \sum_{(u,v)\in H} x_u A_{uv} y_v \ . \tag{4.3.2}$$

For the light couples $l(\mathbf{x},\mathbf{y})$, let

$$X_i = x_u A_{uv} y_v \mathbb{I}\{(u,v) \in L\} + x_v A_{uv} y_u \mathbb{I}\{(v,u) \in L\},$$

where $i = 1,...,\binom{N}{2}$ is corresponding to (u,v) and assume M is the number of the light couples i.e. $M = card(L)$.

Then

$$X_i = x_u y_v + x_v y_u \quad (\mathbf{A} \text{ is adjency matrix})$$

which implies that

$$|X_i| \leq 2\frac{\sqrt{d_{\max}}}{N}.$$

Moreover,

$$\begin{aligned} X_i^2 &\leq \begin{cases} (x_u y_v + x_v y_u)^2 & \operatorname{sgn}(x_u y_v) = \operatorname{sgn}(x_v y_u) \\ (x_u y_v)^2 + (x_v y_u)^2 & \operatorname{sgn}(x_u y_v) \neq \operatorname{sgn}(x_v y_u) \end{cases} \\ &\leq \left[(x_u y_v)^2 + 2|x_u y_v x_v y_u| + (x_v y_u)^2\right]. \end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=1}^{M} X_i^2 &\leq \frac{1}{2}\left[\sum_{(u,v)} (x_u y_v)^2 + 2\sum_{(u,v)} |x_u y_v x_v y_u| + \sum_{(u,v)} (x_v y_u)^2\right] \\ &\leq \frac{1}{2}\left[\sum_u x_u^2 \sum_v y_v^2 + 2\langle x, y\rangle + \sum_v x_v^2 \sum_u y_u^2\right] \quad (\langle x, y\rangle \text{ is dot product}) \\ &\leq 2\|x\| \cdot \|y\| \\ &\leq 2.\end{aligned}$$

This implies that $\sum_{i=1}^{N} E\left(X_i^2\right) \leq 2p_{\max}$.

Moreover, to control the expectation

$$\begin{aligned}&E\left[l\left(\mathbf{x}, \mathbf{y}\right)\right] + E\left[h\left(\mathbf{x}, \mathbf{y}\right)\right] \\ =&E\left[\sum_{(u,v)\in L} x_u A_{uv} y_v\right] + E\left[\sum_{(u,v)\in H} x_u A_{uv} y_v\right] \\ =&E\left(\mathbf{x}^T \mathbf{A} \mathbf{y}\right) \\ =&\mathbf{x}^T E\left(\mathbf{A}\right) \mathbf{y} \\ \leq&\mathbf{x}^T p_{\max} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \mathbf{y} \quad \left(x^T \overrightarrow{1} = 0\right) \\ \neq&0.\end{aligned}$$

Unlike Proposition 5.2 in Hoffman et al. (2019), the condition that

$$E\left[l\left(\mathbf{x}, \mathbf{y}\right)\right] + E\left[h\left(\mathbf{x}, \mathbf{y}\right)\right] = 0$$

fails in block model. This is because $\mathbf{x}$ and $\mathbf{y}$ are unit vectors, but the sign of each component in the unit vectors is unkonwn. So it is impossible to get the conclusion that $p_{\max} \mathbf{x}^T \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \mathbf{y} = 0$. □

Apply Proposition 3.4.1, instead of $\left|\mathbf{x}^T \mathbf{A} \mathbf{y}\right| \leq C\sqrt{d_{\max}}$, the new upper bound is now assumed to be $\left|\mathbf{x}^T \mathbf{A} \mathbf{y}\right| \leq C\sqrt{d_{\max}^2}$.

4.3.2 *Difficulty 2 of Proof of SGT*

However, since the conditions in Lemma 2.6.5 are connected to each other, by changing one of the conditions of the lemma 2.6.5 from d to $d_{\max}^2$, all 4 conditions of the Lemma 2.6.5 also need to be modified in the same manner. Moreover, an additional condition is added for the original setup of the lemma which is $\eta_M = \left\{v \in V : \deg(v) \leq \frac{d_{\max}^2}{M}\right\}$ as $M = d_{\max}^2$. In this case, the condition 1 and 3 of Lemma 2.6.5 are still held as below.

Lemma 4.3.4. *For $\delta > 0$ and $m \geq 0$, there is a constant $C = C(\delta, m)$ such that every vertex has degree at most $Cd_{\max}^2$ with probability at least $1 - C\exp\{-md_{\max}^2\}$. This is called the bounded degree condition (b.d.c.).*

Proof. In SBM, for any vertex v_i, $d_{\min} \leq E[\deg(v_i)] \leq d_{\max} < d_{\max}^2$. Since $p_{\min} \geq \frac{\delta \log N}{N}$, $d_{\min} > \delta \log N$ for large N.

By Lemma 4.3.2, $P(X \geq t\mu) \leq \exp\left\{-\frac{t\mu \log t}{3}\right\}$, assume $\alpha = E[\deg(v_i)]$,

$$\begin{aligned} P\left[\deg(v_i) > c_0 d_{\max}^2 \cdot \alpha\right] &\leq \exp\left\{-\frac{c_0 d_{\max}^2 \cdot \alpha \cdot \log c_0 d_3}{3}\right\} \\ &\leq \exp\left\{-\frac{c_0 d_3 \cdot d_{\min} \cdot \log c_0 d_3}{3}\right\}, \end{aligned}$$

where $c_0 > 4$. Using the fact that $\log N < \frac{d_{\max}^2}{\delta}$,

$$\begin{aligned} P(\text{b.d.c. fails}) =& P(\text{b.d.c does not hold for at least one } v_i) \\ =& P\left[\bigcup_{i=1}^{N} \deg(v_i) > c_0 d_{\max}^2 \cdot \alpha\right] \\ \leq& \sum_{i=1}^{N} P\left[\deg(v_i) > c_0 d_{\max}^2 \cdot \alpha\right] \\ \leq& \sum_{i=1}^{N} \exp\left\{-\frac{c_0 d_{\max}^2 \cdot d_{\min} \cdot \log c_0 d_{\max}^2}{3}\right\} \\ \leq& N \exp\left\{-\frac{c_0 d_3 \cdot d_{\min} \cdot \log c_0 d_{\max}^2}{3}\right\} \\ \leq& \exp\left\{d_3\left[\frac{1}{\delta} - \frac{1}{3} d_{\min} c_0 \log c_0 d_{\max}^2\right]\right\}. \end{aligned}$$

By choosing c_0 large enough, we may take

$$\frac{1}{\delta} - \frac{1}{3} d_{\min} c_0 \log c_0 d_{\max}^2 \leq -m.$$

Thus

$$P\left(\text{b.d.c. fails}\right) = O\left(\exp\left\{-md_{\max}^2\right\}\right).$$

□

Proposition 4.3.5. *For each $\delta > 0$ and each $\epsilon > 0$, if $p_{\min} \geq \frac{\delta \log N}{N}$ there is an $M = M(\delta, \epsilon) > 1$, such that*

a). $|\eta_M| < \frac{N}{100 d_{\max}^2}$ *where* $\eta_M = \left\{v \in V : \deg(v) \leq \frac{d_{\max}^2}{M}\right\}$

b). $\max_{u \in \eta_M^c} Edge(u, \eta_m) \leq 1$

with probability at least $1 - CN\exp\left\{-(2-\epsilon) d_{\max}^2\right\} - C\exp\left\{-cN\right\}$.

Proof. 1): Let $s = \frac{N}{100 d_{\max}^2}$, since in this case only upper bound is required

$$P\left(|\eta_M| \geq s\right) \leq \binom{N}{s} P\left[\deg(u_i) \leq \frac{d_{\max}^2}{M}, 1 \leq i \leq s\right].$$

Let $U = \eta_M^c = \{u_1, .., u_s\}$ and $S = \eta_M = U^c = \{u_{s+1}, ..., u_N\}$

$$\begin{aligned} &P\left[\deg(u_i) \leq \frac{d_{\max}^2}{M}, 1 \leq i \leq s\right] \\ \leq &P\left[Edge(u_i, S) \leq \frac{d_{\max}^2}{M}, 1 \leq i \leq s\right]. \end{aligned}$$

Since $Edge(u, \eta_M)$ are independent Bernoulli r.v., by Lemma 4.3.1,

$$\begin{aligned} &P\left[Edge(u_i, S) \leq \frac{d_{\max}^2}{M}\right] \\ \leq &\exp\left\{-(N-s) p_{\min} + \frac{d_{\max}^2}{M}\left(1 + \log\frac{(N-s) p_{\max}}{\frac{d_{\max}^2}{M}}\right)\right\}. \end{aligned}$$

Thus,

$$
\begin{aligned}
&\log P\left(|\eta_M| \geq s\right) \\
\leq &\log \binom{N}{s} + s\left\{-\left(N-s\right)p_{\min} + \frac{d_{\max}^2}{M}\left(1+\log\frac{M\left(N-s\right)p_{\max}}{d_{\max}^2}\right)\right\} \qquad (4.3.3)\\
\leq &\log\left(\frac{e^s N^s}{s^s}\right) + s\left\{-\left(d_{\min}-sp_{\min}\right) + \frac{d_{\max}^2}{M}\left[\left(1+\log M\right)+\log\frac{\left(N-k\right)}{\left(N-1\right)}\cdot\frac{1}{d_{\max}^2}\right]\right\} \\
\leq &s + s\log\left(\frac{N}{s}\right) + s\left\{-\left(d_{\min}-sp_{\min}\right) + \frac{d_{\max}^2}{M}\left(1+\log M\right)\right\} \qquad (4.3.4)\\
\leq &s\left\{1+\log\left(\frac{N}{s}\right) - \left(d_{\min}-sp_{\min}\right) + \frac{d_{\max}^2}{M}\left(1+\log M\right)\right\} \\
\leq &s\cdot f\left(M\right).
\end{aligned}
$$

For (4.3.3),

$$
\begin{aligned}
-\left(N-s\right)p_{\min} &= -Np_{\min} + sp_{\min} \\
&\leq -Np_{\min} + p_{\min} + sp_{\min} \\
&= -d_{\min} + sp_{\min}.
\end{aligned}
$$

As $s = \frac{N}{100d_{\max}^2}$ and let $M = d_{\max}^2$, using the fact that $\frac{d_{\min}}{100d_{\max}^2} < 1$,

$$
\begin{aligned}
f\left(M\right) \leq &1+\log\left(\frac{N}{s}\right) - \left(d_{\min}-sp_{\min}\right) + \frac{d_{\max}^2}{M}\left(1+\log M\right) \\
\leq &1+\log\left(\frac{N}{\frac{N}{100d_{\max}^2}}\right) - \left(d_{\min}-sp_{\min}\right) + \frac{d_{\max}^2}{M} + C \\
\leq &1+\log\left(100d_{\max}^2\right) - \left(d_{\min}-\frac{d_{\min}}{100d_{\max}^2}\right) + \frac{d_{\max}^2}{M} + C \qquad (4.3.5)\\
\leq &-d_{\min} + \log\left(100d_{\max}^2\right) + 2 + \frac{d_{\max}^2}{M} + C \\
\leq &-d_{\min} + \log\left(100d_{\max}^2\right) + C,
\end{aligned}
$$

where C is a positive constant. As a result,

$$\begin{aligned}
&\log P\left(|\eta_M| \geq s\right) \\
\leq & s \cdot f\left(M\right) \\
\leq & \frac{N}{100 d_{\max}^2}\left\{-d_{\min}+\log\left(100 d_{\max}^2\right)+C\right\} \\
\leq & -\frac{1}{100}\cdot\frac{N}{\left(N-1\right)^2}\cdot\frac{\left(N-1\right)p_{\min}}{p_{\max}^2}+C \\
\leq & -\frac{1}{100}\cdot\frac{p_{\min}}{p_{\max}^2}+C.
\end{aligned}$$

As a result, if $\frac{p_{\min}}{p_{\max}^2} \to \infty$, then $P\left(|\eta_M| \geq s\right) \to C\exp\left\{-CN\right\}$.

Hence we have that $|\eta_M| < \frac{N}{100 d_{\max}^2}$ with probability at least

$$1 - O\left(\exp\left\{-CN\right\}\right)$$

for $C > 0$.

2): We try to bound the probability that there are at least two edges between η_M and η_M^c. And we require that the degree of η_M^c is bounded by $Cd_{\max}$ given by Lemma 4.3.4 1).

$$\begin{aligned}
&P\left[\exists u \in \eta_M^c:\ \mathcal{E}\left(u, \eta_M\right) \geq 2 \bigcap \text{b.d.c}\right] \\
\leq & N^3 P\left(u \in \eta_M^c,\ v \in \eta_M,\ w \in \eta_M,\ u \leftrightarrow v,\ u \leftrightarrow w \bigcap \text{b.d.c.}\right),
\end{aligned}$$

where b.d.c. is the bounded degree condition from Lemma 4.3.4, $Edge\left(u, \eta_M\right)$ is the edge number between vertex u and set η_M and $u \leftrightarrow v$ indicates an

edge is connected between vertices u and v. By conditioning on $\deg(u) = d_u$, $\deg(v) = d_v$, $\deg(w) = d_w$ and

$$
\begin{aligned}
&P\left(u \leftrightarrow v,\ u \leftrightarrow w | d_u, d_v, d_w\right) \\
\leq &\frac{P\left(u \leftrightarrow v,\ u \leftrightarrow w | d_u, d_v, d_w\right)}{P\left(u \leftrightarrow v,\ v \leftrightarrow w,\ u \leftrightarrow w | d_u, d_v, d_w\right)} \\
\leq &\frac{p_{\max}^2 \left(1-p_{\min}\right) \cdot T\left(N-3, d_u-2\right) p_{\max}^{d_u-2} \left(1-p_{\min}\right)^{(N-3)-(d_u-2)}}{\left(1-p_{\max}\right)^3 T\left(N-3, d_u\right) p_{\max}^{d_u} \left(1-p_{\min}\right)^{(N-3)-d_u}} \\
&\quad \times \frac{T\left(N-3, d_v-1\right) p_{\max}^{d_v-1} \left(1-p_{\min}\right)^{(N-3)-(d_v-1)}}{T\left(N-3, d_v\right) p_{\max}^{d_v} \left(1-p_{\min}\right)^{(N-3)-d_v}} \\
&\quad \times \frac{T\left(N-3, d_w-1\right) p_{\max}^{d_w-1} \left(1-p_{\min}\right)^{(N-3)-(d_w-1)}}{T\left(N-3, d_w\right) p_{\max}^{d_w} \left(1-p_{\min}\right)^{(N-3)-d_w}} \\
&\quad + \frac{p_{\max}^3 \cdot T\left(N-3, d_1-2\right) p_{\max}^{d_u-2} \left(1-p_{\min}\right)^{(N-3)-(d_u-2)}}{\left(1-p_{\min}\right)^3 T\left(N-3, d_1\right) p_{\max}^{d_u} \left(1-p_{\min}\right)^{(N-3)-d_u}} \\
&\quad \times \frac{T\left(N-3, d_2-2\right) p_{\max}^{d_v-2} \left(1-p_{\min}\right)^{(N-3)-(d_v-2)}}{T\left(N-3, d_2\right) p_{\max}^{d_v} \left(1-p_{\min}\right)^{(N-3)-d_v}} \\
&\quad \times \frac{T\left(N-3, d_3-2\right) p_{\max}^{d_w-2} \left(1-p_{\min}\right)^{(N-3)-(d_w-2)}}{T\left(N-3, d_3\right) p_{\max}^{d_w} \left(1-p_{\min}\right)^{(n-3)-d_w}} \\
\leq &\frac{\left(1-p_{\min}\right)^2}{p_{\max}^2} \cdot \frac{T\left(N-3, d_u-2\right) T\left(N-3, d_v-1\right) T\left(N-3, d_w-1\right)}{T\left(N-3, d_u\right) T\left(N-3, d_v\right) T\left(N-3, d_w\right)} \\
&\quad + \frac{\left(1-p_{\min}\right)^3}{p_{\max}^3} \cdot \frac{T\left(N-3, d_u-2\right) T\left(N-3, d_v-2\right) T\left(N-3, d_w-2\right)}{T\left(N-3, d_u\right) T\left(N-3, d_v\right) T\left(N-3, d_w\right)}. \\
\leq &\frac{\left(1-p_{\min}\right)^2}{p_{\max}^2} \cdot \frac{N^{d_u+d_v+d_w-4}}{N^{d_u+d_v+d_w}} + \frac{\left(1-p_{\min}\right)^3}{p_{\max}^3} \cdot \frac{N^{d_u+d_v+d_w-6}}{N^{d_u+d_v+d_w}} \\
\leq &\frac{\left(1-p_{\min}\right)^2}{p_{\max}^2} \cdot \frac{1}{N^4} + \frac{\left(1-p_{\min}\right)^3}{p_{\max}^3} \cdot \frac{1}{N^6} \\
\leq &\frac{\left(1-p_{\min}\right)^2}{d_{\max}^2} \cdot \frac{1}{N^2} + o\left(1\right) \\
\leq &C \frac{d_{\max}^4}{N^2}.
\end{aligned}
$$

Then it remains to estimate the probability for both $v, w \in \eta_M$,

$$
P\left[\deg(v) \leq \frac{d_{\max}^2}{M},\ \deg(w) \leq \frac{d_{\max}^2}{M}\right] = \left[P\left(X \leq \frac{d_{\max}^2}{M}\right)\right]^2.
$$

By (4.3.4), putting $s = 2$

$$\begin{aligned}\left[P\left(X \leq \frac{d_{\max}^2}{M}\right)\right]^2 &\leq \exp\left\{2\left[-\left(d_{\min} - 2p_{\min}\right) + \frac{d_{\max}^2}{M}\left(1 + \log M\right)\right]\right\} \\ &\leq \exp\left\{-2d_{\min} + 4 + \frac{2d_{\max}^2}{M} + C\right\} \\ &\leq \exp\left\{-2d_{\min} + C\right\} \quad \left(M = d_{\max}^2\right) \\ &= O\left(\exp\left\{-d_{\max}^2\left(2 - \frac{\epsilon}{2}\right)\right\}\right),\end{aligned}$$

using the fact that

$$d_{\min} = \frac{1}{2}d_{\max}^2\left(\left[2 - \frac{\epsilon}{2}\right]\right).$$

Thus,

$$\begin{aligned}P\left[\max_{u \in \eta_M^c} \mathcal{E}\left(u, \eta_m\right) > 1\right] \leq & C \cdot N^3 \cdot \frac{\left(1 - p_{\min}\right)^2}{d_{\max}^2} \cdot \frac{1}{N^2} \cdot \exp\left\{-d_{\max}^2\left(2 - \frac{\epsilon}{2}\right)\right\} \\ = & O\left(N \exp\left\{-d_{\max}^2\left(2 - \epsilon\right)\right\}\right).\end{aligned}$$

□

Proposition 4.3.5 is the condition 3 in Lemma 2.6.5.

Proposition 4.3.6. *For fixed $\delta > 0$ and $m \geq 0$, there is a constant $C = C(\delta, m)$ sufficiently large so that if $p_{\min} \geq \frac{\delta \log N}{N}$ then*

$$\sum_{v \in V}\left[\deg\left(v\right) - d_{\max}^2\right]^2 \leq CNd_{\max}^2$$

with probability at least $1 - C\exp\left\{-md_{\max}^2\right\}$.

Proof. Note that

$$\sum_{v \in V}\left[\deg\left(v\right) - d_{\max}^2\right]^2 = \left\|\left(\mathbf{A} - d_{\max}^2\mathbf{I}_N\right)\overrightarrow{1}_N\right\|^2$$

where $\mathbf{A}$ is the adjacency matrix of $\mathcal{G}$, $\mathbf{I}_N$ is the identity matrix and $\mathbf{1}_N$ is the vector whose elements are all 1. Thus,

$$\left\|\mathbf{A} - d_{\max}^2\mathbf{1}_N\right\| = \sup_{\|\mathbf{x}\|=1}\left|\mathbf{x}^T\left(\mathbf{A} - d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N\right|.$$

For a fixed vector $\mathbf{x}$, by orthogonal decomposition,

$$\begin{aligned}\tilde{\mathbf{x}} &= \frac{\mathbf{x}\cdot\mathbf{1}_N}{\|\mathbf{1}_N\|^2}\cdot\mathbf{1}_N \\ &\leq \frac{\|\mathbf{x}\|\cdot\|\mathbf{1}_N\|}{\|\mathbf{1}_N\|^2}\cdot\mathbf{1}_N \quad (|\mathbf{x}\cdot\mathbf{1}_N| \leq \|x\|\cdot\|\mathbf{1}_N\|) \\ &= \frac{1}{\|\mathbf{1}_N\|}\cdot\mathbf{1}_N \quad (\|\mathbf{x}\|=1) \\ &\leq \frac{1}{\sqrt{N}}\cdot\mathbf{1}_N.\end{aligned}$$

Thus $\mathbf{x} = \mathbf{y} + c\cdot\mathbf{1}_N$ where $c \leq \frac{1}{\sqrt{N}}$. This implies that

$$\begin{aligned}\left|\mathbf{x}^T\left(\mathbf{A}-d_{\max}^2 I_N\right)\mathbf{1}_N\right| &= \left|(\mathbf{y}+c\mathbf{1}_N)^T\left(\mathbf{A}-d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N\right| \\ &\leq \left|\mathbf{y}^T\left(\mathbf{A}-d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N\right| + \left|c\mathbf{1}_N^T\left(\mathbf{A}-d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N\right|.\end{aligned}$$

For $\left|\mathbf{y}^T\left(\mathbf{A}-d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N\right|$,

$$\mathbf{y}^T\left(\mathbf{A}-d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N = \mathbf{y}^T\mathbf{A}\mathbf{1}_N$$

as $\mathbf{y}^T\mathbf{1}_N = 0$. From Proposition 4.3.3,

$$\sup_{\substack{\|\mathbf{x}\|=1\\ \mathbf{x}^T\mathbf{1}_N=0\\ \|\mathbf{y}\|=1}} \left|\mathbf{x}^T\mathbf{A}\mathbf{y}\right| \leq C\sqrt{d_{\max}^2}.$$

Therefore,

$$\begin{aligned}\sup_{\substack{\|\mathbf{y}\|=1\\ \mathbf{y}^T\mathbf{1}_N=0}} \left|\mathbf{y}^T\mathbf{A}\mathbf{1}_N\right| &= \sup\left|\mathbf{y}^T\mathbf{A}\frac{\mathbf{1}_N}{\|\mathbf{1}_N\|}\right|\cdot\|\mathbf{1}_N\| \\ &\leq C\sqrt{d_{\max}^2}\cdot\sqrt{N}\end{aligned}$$

with probability $1 - O\left(\exp\left\{-md_{\max}^2\right\}\right)$.

Note that

$$\mathbf{1}_N^T\left(\mathbf{A}-d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N = \sum_{v\in V}\deg(v) - Nd_{\max}^2.$$

Since $\sum_{v\in V} \deg(v) = 2\sum_i X_i$ where X_i are independent but not identically distributed Bernoulli random variables, by Chernoff Bounds which is given in (2.2.2),

$$P(|X-\mu| > \delta\mu) \leq 2\exp\left\{-\frac{\mu\delta^2}{3}\right\}.$$

Write $X = \sum_{v\in V} \deg(v)$, let $\mu = E(X)$ then

$$\begin{aligned}
&\mu \leq \sum_{v\in V} (N-1)\, p_{\max} \\
\Rightarrow &\mu \leq N d_{\max} \\
\Rightarrow &\mu < N d_{\max}^2 \\
\Rightarrow &\exp\left\{-\frac{1}{\mu}\right\} < \exp\left\{-\frac{1}{N d_{\max}^2}\right\}.
\end{aligned}$$

Let $t = mN\sqrt{d_{\max}^2}$, then $\delta = \frac{mN\sqrt{d_{\max}^2}}{\mu}$. This implies that by Chernoff Bound,

$$\begin{aligned}
P(|X-\mu| > t) \leq &2\exp\left\{-\frac{m^2N^2d_{\max}^2}{3\mu}\right\} \\
\leq &2\exp\left\{-\frac{t^2}{3\mu}\right\} \quad \left(t^2 = m^2 N d_{\max}^2\right) \\
\leq &C\exp\left\{-\frac{t^2}{CNd_{\max}^2}\right\}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&P\left[\left|\mathbf{1}_N^T\left(\mathbf{A} - d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N\right| \leq t\right] \geq 1 - C\exp\left\{-\frac{t^2}{CNd_{\max}^2}\right\} \\
\Rightarrow &\left|c\mathbf{1}_N^T\left(\mathbf{A} - d_{\max}^2\mathbf{I}_N\right)\mathbf{1}_N\right| \leq \frac{1}{\sqrt{N}}\cdot mN\sqrt{d_{\max}^2} \\
\Rightarrow &\left\|\mathbf{A} - d_{\max}^2\mathbf{1}_N\right\| \leq C\sqrt{Nd_{\max}^2} + m\sqrt{Nd_{\max}^2} \\
\Rightarrow &\sum_{v\in V}\left[\deg(v) - d_{\max}^2\right]^2 = \left\|\mathbf{A} - d_{\max}^2\mathbf{1}_N\right\|^2 \leq CNd_{\max}^2.
\end{aligned}$$

□

Proposition 4.3.6 is not a condition for lemma 2.6.5, however, it is required for the proof of the condition 4 of the lemma 2.6.5 which is the next proposition, Proposition 4.3.7.

4.3.3 *Difficulty 3 of Proof of SGT*

For condition 4 of Lemma 2.6.5, it is unclear that the conditions for the size of the $|\eta_M|$ if $|\eta_M| = 0$ is acceptable, then condition 4 holds. Otherwise, condition 4 fails completely.

Proposition 4.3.7. *Let $W = \{v \in V : \deg(v) > 0\}$ and let η_m be defined in Proposition 4.3.5. For each $\delta > 0$ and $m \geq 0$, there is a constant $C = C(\delta, m)$ sufficiently large so that if $p_{\min} \geq \frac{\delta \log N}{N}$ then*

$$\sup_{\substack{\|\mathbf{x}\| = 1 \\ \mathbf{x}^T \mathbf{D}^{\frac{1}{2}} \mathbb{I}_W = 0}} \left| \mathbf{x}^T \mathbf{D}^{-\frac{1}{2}} \mathbb{I}_{\eta_M^c} \right| \leq C \frac{\sqrt{N}}{d_{\max}^2}$$

with probability at least $1 - C \exp\left\{-m d_{\max}^2\right\}$ where $\mathbf{D}$ is the degree matrix defined in Section 2.4.

Proof. Since $|\eta_M| < \frac{N}{100 d_{\max}^2}$ from Proposition 4.3.5 1), therefore,

$$\left| \mathbf{x}^T \mathbf{D}^{-\frac{1}{2}} \mathbb{I}_{\eta_M^c} \right| \leq \|\mathbf{x}\| \cdot \left\| \mathbf{D}^{\frac{1}{2}} \mathbb{I}_{\eta_M} \right\|$$

as $\|\mathbf{x}\| = 1$. Moreover, $\eta_M = \left\{ v : \deg(v) \leq \frac{d_{\max}^2}{M} \right\}$, this implies $\left\| \mathbf{D}^{\frac{1}{2}} \mathbb{I}_{\eta_M} \right\|$ has value if and only if at $\deg(v) \leq \frac{d_{\max}^2}{M}$. Thus,

$$\begin{aligned}
\left\| \mathbf{D}^{\frac{1}{2}} \mathbb{I}_{\eta_M} \right\| &\leq \left[\sum_{i=1}^{|\eta_M|} \left(\sqrt{\frac{d_{\max}^2}{M} - \epsilon_i} \right)^2 \right]^{\frac{1}{2}} \quad (\epsilon_i \geq 0) \\
&\leq \left(\sum_{i=1}^{|\eta_M|} \frac{d_{\max}^2}{M} \right)^{\frac{1}{2}} \\
&\leq \left(\sum_{i=1}^{|\eta_M|} d_{\max}^2 \right)^{\frac{1}{2}} \\
&\leq \sqrt{d_{\max}^2 \cdot |\eta_M|} \\
&< \sqrt{d_{\max}^2 \cdot \frac{N}{100 d_{\max}^2}} \\
&= O\left(\sqrt{N}\right).
\end{aligned}$$

Furthermore, we need to show that $\mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\left(\mathbb{I}_{\eta_M}+\mathbb{I}_{\eta_M^c}\right)=0$. We can write $\mathbf{D}$ as

$$\mathbf{D}=\begin{bmatrix}\mathbf{D}_W & 0\\ 0 & \mathbf{D}_{W^c}\end{bmatrix}=\begin{pmatrix}\mathbf{D}_W & 0\\ 0 & 0\end{pmatrix}$$

as $W=\{v\in V:\ \deg(v)>0\}$ and

$$\mathbf{D}_W=\operatorname{diag}\left\{\frac{d}{M}+\delta_1,...,\frac{d}{M}+\delta_i,\frac{d}{M}-\epsilon_{i+1},...,\frac{d}{M}-\epsilon_j\right\}\quad(\delta_i>0).$$

Thus,

$$\begin{aligned}&\mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_W=0\\ \Rightarrow&\mathbf{x}^T\mathbf{D}_W^{\frac{1}{2}}\mathbb{I}_W=0\\ \Rightarrow&\mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbf{1}_N=0\quad(\mathbb{I}_W+\mathbb{I}_{W^c}=0)\end{aligned}$$

where $\mathbb{I}_W$ is the indicator vector. As a result,

$$\begin{aligned}&\left(\mathbb{I}_{\eta_M}+\mathbb{I}_{\eta_M^c}\right)=\mathbf{1}_N\\ \Rightarrow&\mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\left(\mathbb{I}_{\eta_M}+\mathbb{I}_{\eta_M^c}\right)=0\\ \Rightarrow&\mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_{\eta_M}=-\mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_{\eta_M^c}.\end{aligned}$$

This implies that

$$\begin{aligned}\sup_{\substack{\|\mathbf{x}\|=1\\ \mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_W=0}}\left|\mathbf{x}^T\mathbf{D}^{-\frac{1}{2}}\mathbb{I}_{\eta_M^c}\right|&\le\sup_{\substack{\|\mathbf{x}\|=1\\ \mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_W=0}}\left|\mathbf{x}^T\mathbf{D}^{-\frac{1}{2}}\mathbb{I}_{\eta_M^c}+O\left(\sqrt{N}\right)\right|\\ &\le\sup_{\substack{\|\mathbf{x}\|=1\\ \mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_W=0}}\left|\mathbf{x}^T\mathbf{D}^{-\frac{1}{2}}\mathbb{I}_{\eta_M^c}+\mathbf{x}^T\frac{\mathbf{D}^{\frac{1}{2}}}{d_{\max}^2}\mathbb{I}_{\eta_M}\right|\\ &\le\sup_{\substack{\|\mathbf{x}\|=1\\ \mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_W=0}}\left|\mathbf{x}^T\mathbf{D}^{-\frac{1}{2}}\mathbb{I}_{\eta_M^c}-\mathbf{x}^T\frac{\mathbf{D}^{\frac{1}{2}}}{d_{\max}^2}\mathbb{I}_{\eta_M^c}\right|.\end{aligned}$$

Taking norms,

$$\left|\mathbf{x}^T\left(\mathbf{D}^{-\frac{1}{2}}-\frac{\mathbf{D}^{\frac{1}{2}}}{d_{\max}^2}\right)\mathbb{I}_{\eta_M^c}\right|\le\left\|\left(\mathbf{D}^{-\frac{1}{2}}-\frac{\mathbf{D}^{\frac{1}{2}}}{d_{\max}^2}\right)\mathbb{I}_{\eta_M^c}\right\|.$$

Squaring this norm,

$$\begin{aligned}
\left\| \left(\mathbf{D}^{-\frac{1}{2}} - \frac{\mathbf{D}^{\frac{1}{2}}}{d_{\max}^2} \right) \mathbb{I}_{\eta_M^c} \right\|^2 &\leq \sum_{v \in \eta_M^c} \left[\frac{1}{\sqrt{\deg(v)}} - \frac{\sqrt{\deg(v)}}{d_{\max}^2} \right]^2 \\
&\leq \sum_{v \in \eta_M^c} \frac{1}{\deg(v)} \left[1 - \frac{\deg(v)}{d_{\max}^2} \right]^2 \\
&\leq \sum_{v \in \eta_M^c} \frac{1}{d_{\max}^4 \deg(v)} \left[d_{\max}^2 - \deg(v) \right]^2 \\
&\leq \sum_{v \in \eta_M^c} \frac{M}{d_{\max}^6} \left[d_{\max}^2 - \deg(v) \right]^2 \quad \left(\deg(v) \geq \frac{d_{\max}^2}{M} \right) \\
&\leq \sum_{v \in V} \frac{M}{d_{\max}^6} \left[d_{\max}^2 - \deg(v) \right]^2 \quad \text{(Lemma 4.3.6)} \\
&\leq \frac{CNM}{d_{\max}^4}.
\end{aligned}$$

Therefore, by letting $C = C(C, M)$,

$$\sup_{\substack{\|\mathbf{x}\| = 1 \\ \mathbf{x}^T \mathbf{D}^{\frac{1}{2}} \mathbb{I}_W = 0}} \left| \mathbf{x}^T \mathbf{D}^{-\frac{1}{2}} \mathbb{I}_{\eta_M^c} \right| \leq \sqrt{\frac{CN}{d_{\max}^4}} \leq C \frac{\sqrt{N}}{d_{\max}^2}.$$

□

Although the proof of the Proposition 4.3.7 does not contain any probability calculations, we have expanded it and given the details on the calculation.

Theorem 4.3.8. *Fix $\delta > 0$ and let $p_{\min} \geq \frac{\left(\frac{1}{2}+\delta\right) \log N}{N}$. Let $d_{\max} = (N-1) p_{\max}$ denote the expected degree of vertex. Let $\mathcal{G}$ be the random graph with block structure. For every fixed $\epsilon > 0$, there is a constant $C = C(\delta, \epsilon)$, so that*

$$\max_{i \neq 1} |1 - \lambda_i| < \frac{C}{\sqrt{d_{\max}^2}}$$

with probability at least $1 - CN \exp\left\{ -(2-\epsilon) d_{\max}^2 \right\} - C \exp\left\{ -d_{\max}^{\frac{1}{2}} \log N \right\}$ where λ_n is the eigenvalue from normalized graph Laplacian which is defined in Section 2.4.

Proof. Proof follows by an extension of Lemma 2.6.5,

1.Lemma 4.3.4 satisfies condition 1 which is every vertex has degree at most $C_1 d_{\max}^2$;

2.Proposition 4.3.3 satisfies condition 2 which is

$$\sup_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T\mathbf{1}_N=0 \\ \|\mathbf{y}\|=1}} \left|\mathbf{x}^T\mathbf{A}\mathbf{y}\right| \leq C_2\sqrt{d_{\max}^2}$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{1}_N$ is a vector whose components are all 1;

3.Proposition 4.3.5 satisfies condition 3 which is $\max_{u\in\eta_M^c}\mathcal{E}(u,\eta_m) \leq 1$

4.Lemma 4.3.7 satisfies condition 4 which is

$$\sup_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T\mathbf{D}^{\frac{1}{2}}\mathbb{I}_W=0}} \left|\mathbf{x}^T\mathbf{D}^{-\frac{1}{2}}\mathbb{I}_{\eta_M^c}\right| \leq C_3\frac{\sqrt{N}}{d_{\max}^2}$$

As a result, Lemma 2.6.5 concludes that

$$\max_{\lambda_i\neq 0}|1-\lambda_i| < \frac{C}{\sqrt{d_{\max}^2}}$$

where $C = C(C_1, C_2, C_3, M)$. □

4.3.4 *Difficulty 4 of Proof of SGT*

Although the original proof of Lemma 2.6.5 is deterministic, given the additional conditions that $M = d_{\max}^2$, it break at 2 different parts.

As in the original proof in ERM,

$$|1-\lambda_i| \leq f_1(x) + f_2(x) + f_3(x) + f_4(x)$$

where $f_i(x) \leq \frac{C}{d} \leq \frac{C}{\sqrt{d}}$ for all $i = 1, ..., 4$.

However, two of the $f_i(x)$ have upper bounds which contain the M value in general, which are

$$f_2(x) \leq \frac{CM}{\sqrt{d}}$$

and

$$f_4(x) \leq \frac{\sqrt{M}}{\sqrt{d}}.$$

As a result, by setting $M = d_{\max}^2$,

$$f_2(x) \leq \frac{C d_{\max}^2}{\sqrt{d_{\max}^2}} = C\sqrt{d_{\max}^2}$$

and

$$f_4(x) \leq \frac{\sqrt{d_{\max}^2}}{\sqrt{d_{\max}^2}} = 1.$$

This implies that the original upper bound $\frac{C}{\sqrt{d}}$ given in Lemma 2.6.5 for the SGT does not hold.

4.4 NEW SIMULATION EVIDENCE FOR SGT

From Section 4.3, the original SGT fails in 4 different ways when extended from ERM to SBM. However, when we prove the CLT for Betti numbers for SBM, the full force of SGT is not needed. From Theorem 2.6.10, the only requirement for β_k is

$$P\left[\lambda_2 > 1 - \frac{1}{k+1}\right] = 1 - o\left(N^{-\alpha}\right)$$

as $N \to \infty$ where $\alpha \geq 0$, k is β_k the degree of the Betti number.

Furthermore, in ERM, the regime for β_k for CLT is

$$N^{-\frac{1}{k}} < N^{-x} < N^{-\frac{1}{k+1}}.$$

As a result, the corresponding theoretical results are shown below.

k	$\frac{1}{k}$	$\frac{1}{k+1}$	λ_2
1	1	0.5	0.5
2	0.5	0.333...	0.666...
3	0.333...	0.25	0.75
4	0.25	0.2	0.8
5	0.2	0.166...	0.833...

We choose N from 500 to 3000 and x from 0.2 to 1.0 to check whether or not λ_2 satisfies the results that whether or not p is outside the range then λ_2 is no longer greater than $1 - \frac{1}{k+1}$.

For each N and x, the sample size $n_s = 100$. One of the scatter plot for λ_N for $N = 2000$ and $x = 0.6$ is shown in Figure 4.4.1. In this case, $\lambda_1 = 2.77 \times 10^{-15}$ and $\lambda_2 = 0.5778$ which satisfied the condition of SGT.

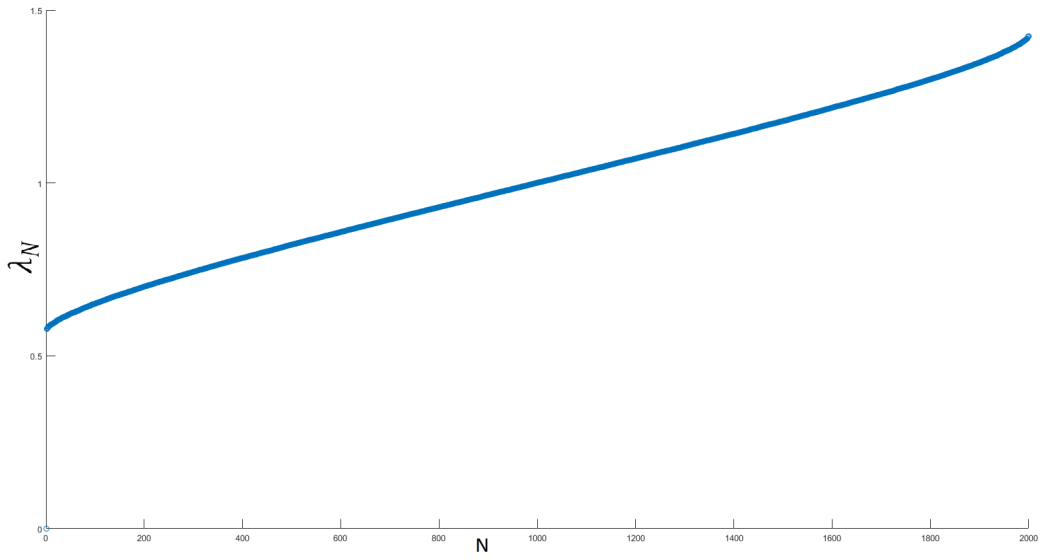

Figure 4.4.1: The λ values for SGT by setting $N = 2000$, $x = 0.6$. The simulation results show that $\lambda_1 = 2.77 \times 10^{-15}$, $\lambda_2 = 0.5778$ and other λ values are gradually increase as required.

For a fixed $N = 2000$, the λ_2 values for x between 0.2 and 0.7 are given in Figure 4.4.2. As can be seen, for $x = 0.2$ to 0.6, λ_2 is greater than 0.5 as required.

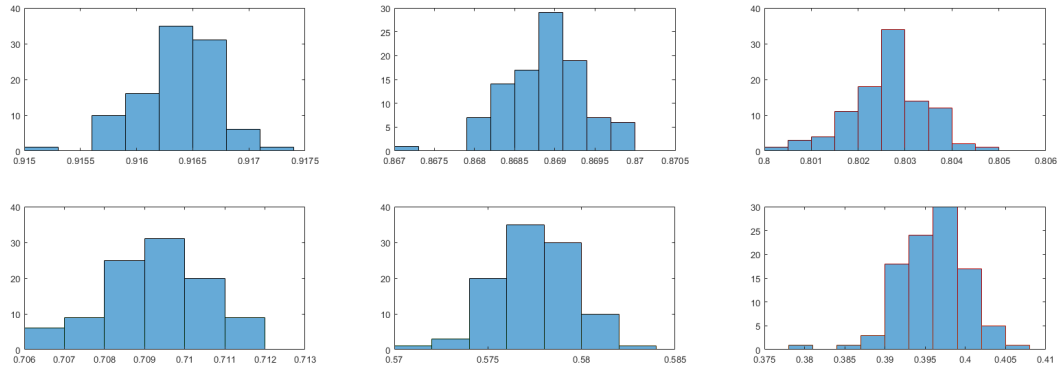

Figure 4.4.2: 100 λ_2 values for $N = 2000$ and $x = 0.2$ to 0.7

Moreover, in Figure 4.4.3, the graph displays values of λ_2 for different x values as N increases. In general, λ_2 increases as N increases for each x values which is results that required for Theorem 2.6.10. Although for $x = 0.7,\ 0.8$, the λ_2 are small, the trends are increasing. It suggests that for $x = 0.7, 0.8$ and 0.9, N needs to be far larger than 5000 to achieve the limiting results. While for $x = 1.0$ as it is outside the maximum range for x so 0 is expected.

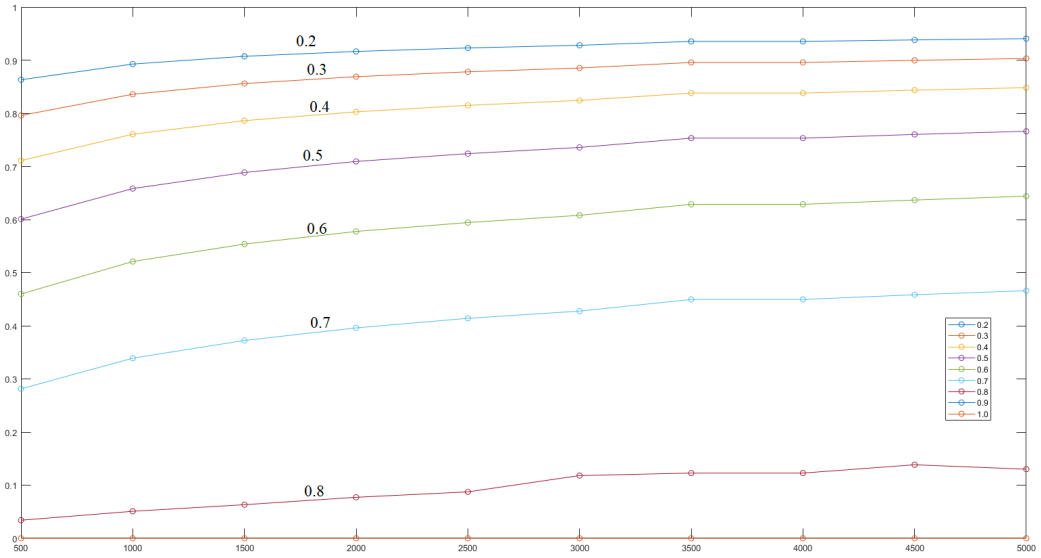

Figure 4.4.3: $N = 500$ to 5000 and $x = 0.2$ to 1.0 for $p = N^{-x}$

In Figure 4.4.3, the line chats show the values for different x values for ERM. Since CVM is a unique theorem for both ERM and SBM, the corresponding theoretical results for k and λ_2 are remains unchanged. However, the regime for β_k for SBM for CLT becomes

$$N_1^{-\frac{1}{k}} = N^{-\phi} < N^{-x} < N^{-\frac{1}{k+1}}.$$

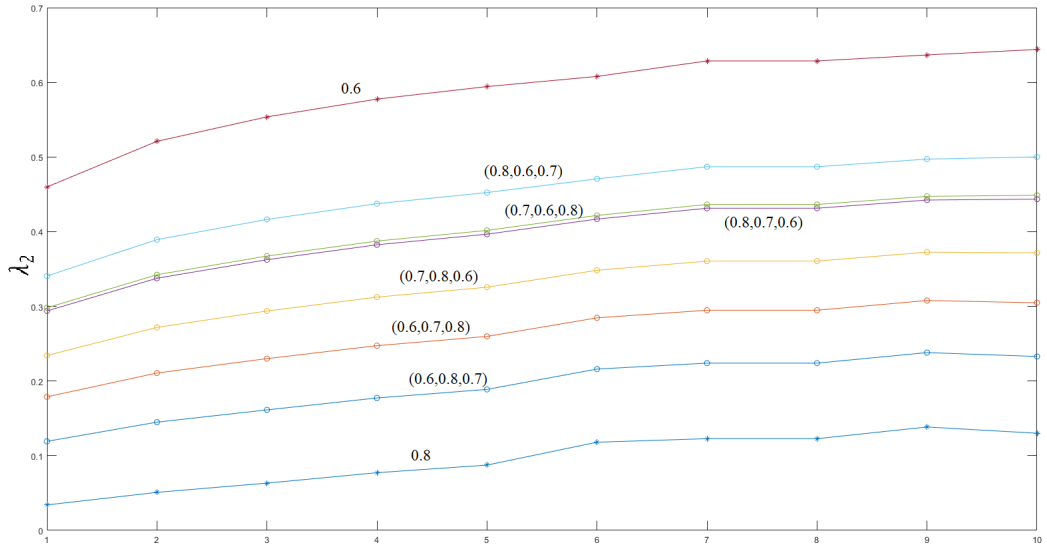

Figure 4.4.4: λ_2 values for $N = 500$ to 5000 for SBM and ERM. The number above lines indicate (x_1, x_2, x_3) where $p_{\min} < N^{-x_i} < p_{\max}$ where $i = 1, ..., 3$.

As can be seen in Figure 4.4.4, the λ_2 for $\mathcal{G}(N, \zeta)$ with $p_{\min}$, $p_{\max}$ is always between λ_2 for $\mathcal{G}(N, p_{\min})$ and $\mathcal{G}(N, p_{\max})$. Since the limiting behaviour forλ_2 for ERM is always true, we make the assumption that the λ_2 for $\mathcal{G}(N, \zeta)$ with $p_{\min}$, $p_{\max}$ is also true.

In conclusion, we assume that the SGT for SBM is true and the following result is stated.

Conjecture 4.4.1. *Consider a clique complex $\mathcal{X}(\mathcal{G})$. Assume that for each k, $N_1^{-\frac{1}{k}} < p_{\min} \leq p_{\max} < N^{-\frac{1}{k+1}}$. Then λ_2 from Normalized graph Laplacian satisfies that*

$$P\left(\lambda_2 > 1 - \frac{1}{k+1}\right) \to 1$$

as $N \to \infty$.

4.5 LOWER VANISHING THRESHOLD FOR BETTI NUMBERS IN SBMS

In this section, we are going to prove the lower vanishing threshold for CLT for Betti numbers

Lemma 4.5.1. *For ζ blocks defined as before, if*

$$p_{\min} \geq \left(\frac{\left[\frac{k}{2}+1\right] \log N + \left[\frac{k}{2}\right] \log\log N + \omega(1)}{N} \right)^{\frac{1}{k+1}}$$

then asymptotic almost surely (a.a.s.) $M_{k+1} = \sum_{i \in \binom{n}{k+1}} Y_i = 0$ *where* Y_i *is the indicator function for* ξ_R *with* $|n_R| = k+1$.

Proof. For a particular Y_i, there are a_r vertices chosen from $\mathcal{V}_r$, then

$$E(Y_i|a_r) = \prod_{i=1}^{\zeta} p_{rr}^{T(a_r,2)} \prod_{1 \leq r < s < \zeta} p_{rs}^{\tau(a_r,a_s)} \prod_{r=1}^{\zeta} \left(1 - \prod_{1 \leq r \leq s \leq \zeta} p_{rs}^{a_s}\right)^{N_r - a_r},$$

and

$$E(M_{k+1}) = f(a_r, k+1, \zeta) \left[\prod_{r=1}^{\zeta} T(N_r, a_r) E(Y_i|a_r) \right].$$

Since we want to prove $E(M_{k+1}) \to 0$, we need to prove the upper bound of $E(M_{k+1})$ tends to zero. Therefore,

$$E(Y_i|a) \leq p_{\max}^{T(k+1,2)} \left(1 - p_{\min}^{k+1}\right)^{N_1 - k - 1}$$

and

$$
\begin{aligned}
E\left(M_{k+1}\right) \leq & \binom{N}{k+1} p_{\max}^{\binom{k+1}{2}} \left(1-p_{\min}^{k+1}\right)^{N_1-k-1} \\
\leq & \frac{N^{k+1}}{(k+1)!} p_{\max}^{\binom{k+1}{2}} \left(1-\frac{N_1 p_{\min}^{k+1}}{N_1}\right)^{N_1-k-1} \\
\leq & \frac{N^{k+1}}{(k+1)!} p_{\max}^{\binom{k+1}{2}} e^{-N_1 p_{\min}^{k+1}} \\
= & \frac{N^{k+1}}{(k+1)!} \left\{\left(\frac{1}{N}\right)^{\frac{1}{k+1}}\right\}^{\frac{(k+1)k}{2}} \\
& \times \exp\left\{-N_1\left(\frac{\left[\frac{k}{2}+1\right]\log N+\left[\frac{k}{2}\right]\log\log N+(c-c_1)}{N}\right)\right\} \\
= & \frac{N^{k+1}}{(k+1)!} N^{-\frac{k}{2}} \cdot \exp\left\{\frac{N_1}{N}\right\} \\
& \times \exp\left\{-\left[\frac{k}{2}+1\right]\log N-\left[\frac{k}{2}\right]\log\log N-c+c_1\right\} \\
= & \frac{N^{k+1}}{(k+1)!} N^{-\frac{k}{2}} \cdot \exp\left\{\frac{N_1}{N}\right\} \cdot N^{-\frac{k}{2}-1} \left(\log N\right)^{-\frac{k}{2}} e^{-c+c_1} \\
= & \frac{1}{(k+1)!} \left(\log N\right)^{-\frac{k}{2}} e^{\frac{N_1}{N}} e^{-c+c_1} \\
= & \frac{e^{-c+c_1}}{(k+1)!} e^{\frac{N_1}{N}} \left[\frac{1}{\log N}\right]^{\frac{k}{2}}.
\end{aligned}
$$

As $N_1 \to \infty$, $N_1-k-1 \to \infty$, $N \to \infty$, $\frac{1}{\log N} \to 0$. As $c \to \infty$, c_1 is a fixed constant, $e^{-c+c_1} \to 0$. then $E\left(M_{k+1}\right) \to 0$. Since

$$
N = N_1 + \sum_{s=2}^{\zeta} N_s = N_1 + \sum_{s=2}^{\zeta} \theta_{1s} N_1
$$

where $\theta_{1s} = \frac{N_s}{N_1} \geq 1$. Then $\frac{N_1}{N} = \frac{N_1}{N_1\left(1+\sum_{s=2}^{\zeta}\theta_{1s}\right)} = \frac{1}{\left(1+\sum_{s=2}^{\zeta}\theta_{1s}\right)}$, then $\exp\left\{-1-\sum_{s=2}^{\zeta}\theta_{1s}\right\}$ is either a constant or tends to e^0.

Since upper bound of $E\left(M_{k+1}\right) \to 0$, the resulting $E\left(M_{k+1}\right)$ tends to zero. □

Moreover, we need to use CVT introduced at Theorem 2.6.10 to prove Theorem 4.5.4. We state Lemma 4.5.2 and 4.5.3 to show that the requirement for CVM is satisfied.

Lemma 4.5.2. *Set*

$$\bar{p} = \left(\frac{\left[\frac{k}{2}+1\right] \log N + C_k \sqrt{\log N} \log\log N}{N} \right)^{\frac{1}{k+1}}$$

where C_k is a constant depending on k. If $p_{\min} \geq \bar{p}$, then a.a.s. the $skel_{k+1}(\mathcal{G})$ where $skel_{k+1}$ is defined in Definition 2.5.2 and $\mathcal{G}$ is the graph with ζ-blocks is pure $(k+1)$-dimensional; in other words, every face is contained in the boundary of a $(k+1)-$face.

Proof. Since a k-face which is not in a $(k+1)$-face would correspond to a maximal $(k+1)$-clique. As

$$\bar{p} \geq \left(\frac{\left[\frac{k}{2}+1\right] \log N + \left[\frac{k}{2}\right] \log\log N + \omega(1)}{N} \right)^{\frac{1}{k+1}}$$

by Lemma 4.5.1, $P(M_{k+1}) \to 0$ as $N \to \infty$, where M_{k+1} is the number of maximal $(k+1)$-cliques in $\mathcal{G}$ as defined before.

For $0 \leq i < k$, let $p_i = \left(\frac{\left(\frac{i-1}{2}+1\right) \log N + \left[\frac{i-1}{2}\right] \log\log N + \omega(1)}{N} \right)^{\frac{1}{i}}$. Since $\bar{p} \geq \left(\frac{\left[\frac{k}{2}+1\right] \log N + \left[\frac{k}{2}\right] \log\log N + \omega(1)}{N} \right)^{\frac{1}{k+1}} > p_i$ for all $0 \leq i < k$. Therefore, by Lemma 4.5.1, $P(M_i) \to 0$ as $N \to \infty$. □

The proof of Lemma 4.5.2 is nearly identical to the proof given in (Kahle, 2014). We only replace the notation small n with big N and replace the p value to the $p_{\min}$.

Lemma 4.5.3. *Let $\mathcal{X} \sim \mathcal{X}(\mathcal{G})$. Set*

$$\bar{p} = \left(\frac{\left[\frac{k}{2}+1\right] \log N + C_k \log\log N}{N} \right)^{\frac{1}{k+1}}$$

where C_k is a constant depending on k. If $p_{\min} \geq \bar{p}$ then a.a.s.

$$p_{\min} \geq \frac{(\alpha+1)\log N_\sigma + C_\alpha \sqrt{\log N_\sigma}\log\log N_\sigma}{N_\sigma}$$

for every $(k-1)$-dimensional face $\sigma \in \mathcal{X}$, where M_σ is the number of vertices in $lk_{\mathcal{X}}(\sigma)$, $\alpha = \frac{k(k+3)}{2}$ and C_α is a constant depending only on α.

Proof. Assume there are a_r vertices chosen from $\mathcal{V}_r$ for σ then

$$E(M_\sigma) = f(a_r, k, \zeta)\left\{\prod_{r=1}^{\zeta}(N_r - a_r)\, p_{rr}^{a_r} \prod_{1\leq r<s\leq\zeta} p_{rs}^{a_r}\right\}$$

Set

$$g(x) = \frac{(\alpha+1)\log x + C_\alpha\sqrt{\log x}\log\log x}{x}$$

then the first derivative $g'(x)$ is

$$g'(x) = g_1(x) + g_2(x)$$

where

$$\begin{cases} g_1(x) = (\alpha+1)\left[\frac{1}{x^2} - \frac{\log x}{x^2}\right] \\ g_2(x) = C_\alpha\left[\frac{2+\log\log x}{2x^2\sqrt{\log x}} - \frac{2\log x\log\log x}{2x^2\sqrt{\log x}}\right]. \end{cases}$$

Since $g_1(x) < 0$ for $x \geq 4$ and $g_2(x) < 0$ for $x > 16$, so $g(x)$ is a decreasing function for $x > 16$. Note: $\log[\log(16)] > 1$ and x need to be an integer.

As $g(x)$ is a decreasing function, we want to get the lower bound of M_σ. Then

$$E(M_\sigma) \geq (N_1 - k)\, p_{\min}^k \approx N_1 p_{\min}^k.$$

Let $\mu = N_1 p_{\min}^k$, and $\delta = \mu^{-\frac{2}{5}}$, by Chernoff bound which is defined as (2.2.2),

$$\begin{aligned} P(|X-\mu| > \delta\mu) =& P\left(|M_\sigma - \mu| > \mu^{\frac{3}{5}}\right) \\ \leq& 2\exp\left\{-\frac{\mu^{\frac{1}{5}}}{3}\right\}. \end{aligned}$$

Since for $p_{\min} \geq N_1^{-\frac{1}{k+1}}$, $N_1 p_{\min}^k = N_1 \cdot N_1^{-\frac{k}{k+1}} = N_1^{\frac{1}{k+1}}$. Therefore, $2\exp\left\{-\frac{\mu^{\frac{1}{5}}}{3}\right\} = 2\exp\left\{-\frac{1}{3}N_1^{\frac{1}{k+1}}\right\} \to 0$ as $N_1 \to \infty$. Therefore,

$$\begin{aligned} P\left(|M_\sigma - \mu| \leq \mu^{\frac{3}{5}}\right) =& 1 - P\left(|M_\sigma - \mu| > \mu^{\frac{3}{5}}\right) \\ \to& 1. \end{aligned}$$

Chernoff Bound defined as (2.2.2) can be updated as

$$\mu - \mu^{\frac{3}{5}} \leq M_\sigma \leq \mu + \mu^{\frac{3}{5}}$$

Thus, let $N' = \mu - \mu^{\frac{3}{5}}$ and if a.a.s.

$$p_{\min} \geq \frac{(\alpha+1)\log N' + C_\alpha\sqrt{\log N'}\log\log N'}{N'} = g\left(N'\right)$$

then $p_{\min} \geq g\left(N'\right) \geq g\left(M_\sigma\right)$ as $N' \leq M_\sigma$ and $g\left(x\right)$ is a decreasing function.

Write

$$f\left(p_{\min}\right) = N' p_{\min} - (\alpha+1)\log N' + C_\alpha\sqrt{\log N'}\log\log N'. \tag{4.5.1}$$

Then we can prove that $f\left(p_{\min}\right) > 0$ for $p_{\min} \geq \bar{p}$. □

In Lemma 4.5.3, the updated proof of Chernoff bound which is not given in the original paper is included. Moreover, since the original proof of Lemma 4.5.3 is split into two parts by (4.5.1), the second part of the proof is only based on simple calculation which does not include any probability. Thus, it is a general results for SBM.

Combining Lemma 4.5.2, 4.5.3 and with the assumption of SGT, Conjecture 4.4.1 is true, Theorem 4.5.4 is now ready to be proved.

Theorem 4.5.4. *Let $k \geq 1$ and $\epsilon > 0$ be fixed and let graph $\mathcal{G}$ be a SBM. If*

$$p_{\min} \geq \left(\frac{\left[\frac{k}{2} + 1 + \epsilon\right]\log N}{N}\right)^{\frac{1}{k}}$$

then a.a.s.

$$H^k\left(\mathcal{X}, \mathbb{Q}\right) = 0.$$

Proof. Suppose $p \geq \left(\frac{\left[\frac{k}{2}+1\right]\log n + C_k \log\log n}{n}\right)^{\frac{1}{k+1}}$, $\mathcal{X} \sim \mathcal{X}(\mathcal{G})$ and f_{k-1} is the number of $(k-1)$-dimensional faces of $\mathcal{X}$. From Section 4.7.1,

$$\begin{aligned} E(f_{k-1}) &\leq T(N,k)\, p_{\max}^{T(k,2)} \\ Var(f_{k-1}) &\leq c_k N^{2(k-1)} p_{\max}^{2T(k,2)-1}. \end{aligned}$$

Then with standard Chebyshev's inequality,

$$\begin{aligned} P[|X - E(X)| \geq a] =& P[|f_{k-1} - \mu| \geq o(1)\,\mu] \\ \leq& \frac{c_k N^{2(k-1)} p_{\max}^{2T(k,2)-1}}{\epsilon \mu^2} \\ \approx& \frac{c_k N^{2(k-1)} p_{\max}^{2T(k,2)-1}}{\epsilon N^{2k} p_{\max}^{T(k,2)}} \\ =& \frac{c_k}{\epsilon N^2 p_{\max}}, \end{aligned}$$

where c_k is a constant depending on k and ϵ is a small constant.

For $p_{\max} \geq p_{\min} > N_1^{-\frac{1}{k+1}} > N^{-\frac{1}{k+1}}$, $N^2 p_{\min} = N^2 N_1^{-\frac{1}{k+1}} > N^2 N^{-\frac{1}{k+1}} = N^{\frac{2k+1}{2(k+1)}} \to \infty$ as $N \to \infty$. So

$$\begin{aligned} P[|f_{k-1} - \mu| \leq o(1)\,\mu] \geq& 1 - \frac{c_k}{\epsilon N^2 p_{\max}} \\ \to& 1. \end{aligned}$$

Lemma 4.5.3 shows that a.a.s.

$$p_{\min} \geq \frac{(\alpha+1)\log N_\sigma + C_\alpha \sqrt{\log N_\sigma}\log\log N_\sigma}{N_\sigma}$$

for every $(k-1)$-dimensional face σ where $\alpha = \frac{k(k+3)}{2}$.

Let $A = \{f_{k-1} \leq \gamma\}$ where $\gamma = (1 + o(1))\, T(N,k)\, p^{T(k,2)}$. By SGT, since $|1-\lambda_2| \leq \max_{i \neq 1} |1-\lambda_n|$, $\lambda_2 > 1 - \frac{1}{k}$. Let $R_\sigma = \lambda_2[\mathcal{G}] < \frac{1}{k-1}$ and $P_\sigma = P(R_\sigma) = o\left(N_\sigma^{-\alpha}\right)$ then

$$
\begin{aligned}
P_f =& P\left(\bigcup_\sigma R_\sigma\right) \\
=& P\left(\bigcup_\sigma R_\sigma | A\right) P(A) + P\left(\bigcup_\sigma R_\sigma | A^C\right) P\left(A^C\right) \\
=& P\left[\bigcup_\sigma \left(R_\sigma \bigcap A\right)\right] + o(1) \\
=& E\left[\sum_\sigma \mathbb{I}_{R_\sigma} \cdot \mathbb{I}_A\right] + o(1) \\
\leq& \sum_{i=1}^{\gamma} E\left[\mathbb{I}_{R_\sigma} \cdot \mathbb{I}_A\right] + o(1) \\
=& \gamma P(R_\sigma) + o(1) \\
\leq& \gamma o\left(M_\sigma^{-\alpha}\right) + o(1) \\
\leq& \gamma o\left(\mu^{-\alpha}\right) + o(1) \quad \left(\mu\left[1 - \mu^{-\frac{2}{5}}\right] \leq M_\sigma\right) \\
=& o\left\{T(N,k)\, p_{\min}^{T(k,2)} \left(N p_{\min}^k\right)^{-\frac{k(k+3)}{2}}\right\} \\
=& o\left\{N^{\frac{2k-k(k+3)}{2}} p_{\min}^{\frac{k(k-1)-k^2(k+3)}{2}}\right\} \\
=& o\left\{\left[N p_{\min}^{k+1}\right]^{-\frac{k(k+1)}{2}}\right\} \\
=& o(1),
\end{aligned}
$$

since $N p_{\min}^{k+1} \to \infty$ as $N \to \infty$ for $p_{\min} \geq N^{-\frac{1}{k+1}}$. □

4.6 UPPER VANISHING THRESHOLD FOR BETTI NUMBERS IN SBMS

In this section, the extension to the SBM of Theorem 4.6.2 is proved for CLT for Betti numbers. This section follows the paper by Hoffman et al. (2019). Firstly, we need to introduce the following Lemma in SBM.

Lemma 4.6.1. *Let $N^{-x} < N^{-\frac{1}{k}}$ and $0 < M^{-1} < -\frac{1}{k} - x$. Then a.a.s. there are no strongly connected pure k-dimensional subcomplexes of $\mathcal{X}(\mathcal{G})$ with vsupp of more than $M + k + 1$ vertices where vsupp is defined in Definition 2.5.3.*

Proof. The vertices in the support of a strongly connected subcomplex can be ordered $v_1, v_2, ..., v_n$ such that $\{v_1, ..., v_{k+1}\}$ spans a k-face and v_i is connected to at least k vertices v_j with $j < i$. One way to see this is to order the k-faces $f_1, f_2, ...$ so that each has $(k-1)$-dimensional intersection with the union of the previous faces. That this is possible is guaranteed by the assumption of strongly connected. Then let this ordering induce an ordering on vertices, since at most one vertex gets added at a time in the sequence $f_1,\ f_1 \bigcup f_2,\ f_1 \bigcup f_2 \bigcup f_3, ...$

Suppose complex $\mathcal{X}$ has $M + k + 1$ vertices, as f_1 has $\binom{k+1}{2}$ edges, $\mathcal{X}$ has at least $\binom{k+1}{2} + Mk$ edges. If the underlying graph of $\mathcal{X}$ is not a subgraph of $\mathcal{G}$ then $\mathcal{X}$ is not a subcomplex. Choose ϵ and M such that $M^{-1} < \epsilon < x - \frac{1}{k}$, assume $p_{\max} = N^{-x} < N^{-\frac{1}{k}-\epsilon}$ and $k < \epsilon M k$, thus

$$
\begin{aligned}
&P\left(\exists \text{subcomplex } \mathcal{X}\right) \\
=&P\left(\exists \text{subcomplex } \mathcal{X} | \mathcal{X} \text{ is a subgraph}\right) P\left(\mathcal{X} \text{ is a subgraph}\right) \\
&\quad + P\left(\exists \text{subcomplex } \mathcal{X} | \mathcal{X} \text{ is not a subgraph}\right) P\left(\mathcal{X} \text{ is not a subgraph}\right) \\
\leq&P\left(\exists \text{subcomplex } \mathcal{X} | \mathcal{X} \text{ is a subgraph}\right) \\
\leq&(M+k+1)! \cdot T\left(N, M+k+1\right) p_{\max}^{T(k+1,2)+Mk} \\
\leq&N^{M+k+1} N^{-\frac{1}{k}[T(k+1,2)+Mk]} N^{-\epsilon[T(k+1,2)+Mk]} \\
=&N^{M+k+1-\frac{k+1}{2}-M-\epsilon T(k+1,2)-\epsilon Mk} \\
=&N^{1-\frac{k+1}{2}-\epsilon T(k+1,2)} \\
=&O\left(N^{-\epsilon}\right).
\end{aligned}
$$

The proof of Lemma 4.6.1 is separated into two parts, the first part is based only on TDA, which is identical for both ERM and SBM. The second part is a modified version of the proof in Hoffman et al. (2019), obtained by replacing n with N and p with $p_{\max}$.

Moreover, since Lemma 2.6.2 and 2.6.4 stated in Section 2.6 are independent to the structure of ERM, these two lemmas are assumed to be true for SBM. The results from Lemma 4.6.1, 2.6.2 and 2.6.4 are applied to prove Theorem 4.6.2. □

Theorem 4.6.2. *If $k \geq 1$ and $\epsilon > 0$ are fixed and suppose graph $\mathcal{G}$ with N vertices has clique complex $\mathcal{X}(\mathcal{G})$. If*

$$p_{\max} \leq \frac{1}{N^{\frac{1}{k}+\epsilon}}$$

then a.a.s. $H^k(\mathcal{X}, \mathbb{Q}) = 0$ where $H^k(\mathcal{X}, \mathbb{Q})$ is defined in Section 2.5.

Proof. Any non-trivial k-cycle with minimal $vsupp$ must have minimum vertex degree at least $2k$ in its supporting subgraph. Since by Lemma 2.6.2, each vertex link is a non-trivial $(k-1)$-cycle, hence by Lemma 2.6.4 it contains at least $2(k-1)+2 = 2k$ vertices.

Let H be any fixed graph with minimal vertex degree $2k$. Let m =number of vertices in H and $\mathcal{E}(H) \geq \frac{m \cdot 2k}{2}$ as the edges get double counting. Then if $-x < -\frac{1}{k}$ and $p_{\max} = N^{-x}$, H is almost always not a subgraph of $\mathcal{G}$. This is because by Lemma 4.6.1,

$$\begin{aligned} & m! T(N, m)\, p_{\max}^{mk} \\ \leq & N^m N^{-xmk} \\ = & o(1). \quad (-xk < -1) \end{aligned}$$

There are only finite many isomorphism types of graphs of minimal degree $2k$ on $m = N + k$ vertices. Each has at least km edges. Applying this argument to each of them, it can be concluded that $\mathcal{X}$ a.s has no vertex minimal non-trivial $k-$cycles, so a.s $\tilde{H}_k(\mathcal{X}, \mathbb{Z}) = 0$. □

4.7 CLT FOR BETTI NUMBERS AND VANISHING THRESHOLDS

In this section, we are going to extend CLT for clique complex of ERM results by Kahle and Meckes (2015) to SBM.

4.7.1 *Moments and Simplices*

Since Theorem 2.6.11 does not include any probability, it can be assumed to be true and use it directly when proving Theorem 4.7.2. However, we are going to include the proof of Conjecture 4.4.1, as it is not given in any paper.

Proof of Theorem 2.6.11. Under present setup, let

$$X_R = \frac{1}{\sigma}(-1)^{card(R)+k+1}\left[\xi_R - E(\xi_R)\right]$$

therefore,

$$\begin{aligned}
&E\{|X_R X_S X_T|\}\\
=&\frac{1}{\sigma^3}E\{[\xi_R - E(\xi_R)][\xi_S - E(\xi_S)][\xi_T - E(\xi_T)]\}\\
=&\frac{1}{\sigma^3}E\{|\xi_R\xi_S\xi_T - \xi_R E(\xi_S)\xi_T - E(\xi_R)\xi_S\xi_T\\
&\quad + E(\xi_R)E(\xi_S)\xi_T - \xi_R\xi_S E(\xi_T) - \xi_R E(\xi_S)E(\xi_T)\\
&\quad - E(\xi_R)\xi_S E(\xi_T) + E(\xi_R)E(\xi_S)E(\xi_T)|\}\\
=&\frac{1}{\sigma^3}\{E(\xi_R\xi_S\xi_T) + E(\xi_R\xi_T)E(\xi_S) + E(\xi_R)E(\xi_S\xi_T)\\
&\quad + E(\xi_R)E(\xi_S)E(\xi_T) + E(\xi_R\xi_S)E(\xi_T) + E(\xi_R)E(\xi_S)E(\xi_T)\\
&\quad + E(\xi_R)E(\xi_S)E(\xi_T) + E(\xi_R)E(\xi_S)E(\xi_T)\}\\
\leq&\frac{8}{\sigma^3}E(\xi_R\xi_S\xi_T),
\end{aligned} \tag{4.7.1}$$

and

$$\begin{aligned}
&E\{|X_R X_S|\}E\{|X_T|\}\\
=&\frac{1}{\sigma^3}E\{|[\xi_R - E(\xi_R)][\xi_S - E(\xi_S)]|\}E\{|[\xi_T - E(\xi_T)]|\}\\
=&\frac{1}{\sigma^3}\{E[|\xi_R\xi_S - \xi_R E(\xi_S) - E(\xi_R)\xi_S + E(\xi_R)E(\xi_S)|]\times 2E(\xi_T)\}\\
=&\frac{1}{\sigma^3}\{2[E(\xi_R\xi_S) + E(\xi_R)E(\xi_S)]\times 2E(\xi_T)\}\\
\leq&\frac{8}{\sigma^3}E(\xi_R\xi_S\xi_T).
\end{aligned} \tag{4.7.2}$$

Since $E(\xi_R)$ is counting the number of edges that exist in the graph, the mean value should be the sum of non-negative number times the probability, i.e. $E(|\xi|) = E(\xi)$. From the results above, for both ERM and SBM, the power of probability is determined by the number of intersection points ,η, between the simplices.

For both

$$E(\xi_R\xi_S)E(\xi_T) \leq E(\xi_R\xi_S\xi_T)$$

and

$$E\left(\xi_R\right) E\left(\xi_S\right) E\left(\xi_T\right) \leq E\left(\xi_R \xi_S \xi_T\right),$$

the LHS of the inequality can not have the case when three simplices are intersect together, i.e. take triangle as an example, the extreme case for $E\left(\xi_R \xi_S \xi_T\right)$ is that all three simplices are sharing 3 vertices then the power for the probability is p^3. While for $E\left(\xi_R \xi_S\right) E\left(\xi_T\right)$, the extreme case can only be ξ_R and ξ_S are sharing the same vertices, therefore, the least possible power is p^6. Similarly, $E\left(\xi_R\right) E\left(\xi_S\right) E\left(\xi_T\right)$ can not have any points of intersection, the only possible power is p^9. Since $0 < p < 1$, $p^3 > p^6 > p^9$, $E\left(\xi_R \xi_S\right) E\left(\xi_T\right) \leq E\left(\xi_R \xi_S \xi_T\right)$ and $E\left(\xi_R\right) E\left(\xi_S\right) E\left(\xi_T\right) \leq E\left(\xi_R \xi_S \xi_T\right)$.

As a result, from Theorem 2.6.11,

$$d_1\left(W, Z\right) \leq \frac{16}{\sigma^3} \left(\sum_{R \subseteq \mathcal{V}}\right) \left(\sum_{S,T \in L_R}\right) E\left(\xi_R \xi_S \xi_T\right) \tag{4.7.3}$$

where $\mathcal{V}$ is the collection of the vertices set and for $R \subseteq \mathcal{V}$, L_R is the collection of subsets of $\mathcal{V}$ where at least two vertices is chosen from simplicies ξ_R. □

Therefore, if we want to extend this proof for SBM, we need to calculate the first three moments.

For the first moment, we can consider a_1 vertices first chosen from $\mathcal{V}_1$. There are $T\left(N_1, a_1\right)$ possibilities and in $\mathcal{V}_1$ each edge has probability $p_{11}^{T(N_1, a_1)}$. Then a_2 vertices are chosen from $\mathcal{V}_2$. There are $T\left(N_2, a_2\right)$ such choices and the edge probabilities are $T\left(N_2, a_2\right)$ combinations with probability $p_{22}^{T(a_2, 2)}$, etc. If an edge is connected between block r and s, the probability of this edge is

$$p_{rs}^{T(a_r + a_s, 2) - T(a_r, 2) - T(a_s, 2)},$$

and

$$E\left(f_k\right) = f\left(a_r, k+1, \zeta\right) \left[\prod_{r=1}^{\zeta} T\left(N_r, a_r\right) p_{rr}^{T(a_r, 2)} \prod_{1 \leq r < s < \zeta} p_{rs}^{\tau(a_r, a_s)}\right] . \tag{4.7.4}$$

For the second moment, suppose two simplices both contain $k+1$ vertices. The first simplicies same way as first moment, then the common vertices are chosen from each block r from 1 to ζ,

$$\begin{aligned} E\left(f_k^2\right) =& f\left(a_r, k+1, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r, a_r\right) \sum_{\eta=0}^{k+1} f\left(\eta_r, \eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(a_r, \eta_r\right) \\ & \times f\left(a_r, k+1-\eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r - a_r, \alpha_r - \eta_r\right) \\ & \times \prod_{r=1}^{\zeta} p_{rr}^{T(a_r,2)+T(\alpha_r,2)-T(\eta_r,2)} \prod_{1\le r<s<\zeta} p_{ij}^{\tau(a_r,a_s)+\tau(\alpha_r,\alpha_s)-\tau(\eta_r,\eta_s)}. \end{aligned} \tag{4.7.5}$$

For the second moment, assume two simpilicies have diffident sizes. Then without loss of generality, assume the first simplex is the one with larger the one with larger number of vertices. So

$$\begin{aligned} E\left(f_k f_{k+j}\right) =& f\left(a_r, k+j+1, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r, a_r\right) \sum_{\eta=0}^{k+1} f\left(\eta_r, \eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(a_r, \eta_r\right) \\ & \times f\left(\alpha_r, k+1-\eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r - a_r, \alpha_r - \eta_r\right) \\ & \times \prod_{r=1}^{\zeta} p_{rr}^{T(a_r,2)+T(\alpha_r,2)-T(\eta_r,2)} \prod_{1\le r<s\le\zeta} p_{rs}^{\tau(a_r,a_s)+\tau(\alpha_r,\alpha_s)-\tau(\eta_r,\eta_s)} \end{aligned}$$

where $j \geq 0$.

For SBM, by induction from first and second moment, the third moment is

$$\begin{aligned}
E\left(\xi_R\xi_S\xi_T\right) =& f\left(a_r, n_R, \zeta\right) f\left(\alpha_r, n_S, \zeta\right) f\left(\gamma_r, n_T, \zeta\right) \\
&\times \left(\sum_{\eta_{R\cap S}=0}^{n_S}\right) f\left(\eta_{R\cap S,r}, \eta_{R\cap S}, \zeta\right) \binom{\min(n_T,\eta_{R\cap S})}{\sum_{\eta_{R\cap S\cap T}=0}} \left(\sum_{\eta_{R\cap T\setminus S}}^{n_T}\right) \left(\sum_{\eta_{S\cap T\setminus R}}^{n_T}\right) \\
&\times f\left(\eta_{R\cap S\cap T,r}, \eta_{R\cap S\cap T}, \zeta\right) f\left(\eta_{R\cap T\setminus S,r}, \eta_{R\cap T\setminus S}, \zeta\right) \\
&\times f\left(\eta_{S\cap T\setminus R,i}, \eta_{S\cap T\setminus R}, \zeta\right) \\
&\prod_{r=1}^{\zeta} T\left(N_r, a_r\right) T\left(a_r, \eta_{R\cap S,r}\right) T\left(N_r - a_r, \alpha_r - \eta_{R\cap S,r}\right) \\
&\quad\times T\left(\eta_{R\cap S,r}\eta_{R\cap S\cap T,r}\right) T\left(a_r - \eta_{R\cap S,r}, \eta_{R\cap T\setminus S,r}\right) \\
&\quad\times T\left(a_r - \eta_{R\cap S,r}, \eta_{S\cap T\setminus R,r}\right) \\
&\quad\times T\left(N_r - a_r - \alpha_r + \eta_{R\cap S,r}, \gamma_r - \eta_{R\cap T\setminus S,r} - \eta_{S\cap T\setminus R,r} - \eta_{R\cap S\cap T,r}\right) \\
&\prod_{r=1}^{\zeta} p_{rr} g^{rr} \prod_{1\le r<s\le\zeta} p_{rs} g^{rs},
\end{aligned} \tag{4.7.6}$$

where

$$\begin{aligned}
g_{rr} =& T\left(a_r, 2\right) + T\left(\alpha_r, 2\right) - T\left(\eta_{R\cap S,r}, 2\right) \\
&+ T\left(\gamma_r, 2\right) - T\left(\eta_{R\cap T\setminus S,r} + \eta_{R\cap S\cap T,r}, 2\right) \\
&- T\left(\eta_{S\cap T\setminus R,r} + \eta_{R\cap S\cap T,r}, 2\right) + T\left(\eta_{R\cap S\cap T,r}, 2\right),
\end{aligned}$$

and

$$\begin{aligned}
g_{rs} =& \tau\left(a_r, a_s\right) + \tau\left(\alpha_r, \alpha_s\right) - \tau\left(\eta_{R\cap S,r}, \eta_{R\cap S,s}\right) \\
&+ \tau\left(\gamma_r, \gamma_s\right) - \tau\left(\eta_{R\cap T\setminus S,r} + \eta_{R\cap S\cap T,r}, \eta_{R\cap T\setminus S,s} + \eta_{R\cap S\cap T,s}\right) \\
&- \tau\left(\eta_{S\cap T\setminus R,r} + \eta_{R\cap S\cap T,r}, \eta_{S\cap T\setminus R,s} + \eta_{R\cap S\cap T,s}\right) + \tau\left(\eta_{R\cap S\cap T,r}, \eta_{R\cap S\cap T,s}\right).
\end{aligned}$$

The main calculation for getting these moments has been put in Section 4.9. In this section, we only give the final results for each moment.

4.7.2 *CLT for Betti Numbers in SBMs*

In this section, we are going to prove the CLT for Betti numbers in SBM.

Corollary 4.7.1. *Let $k \geq 1$ and $\epsilon > 0$ be fixed. If*

$$\left(\frac{\left[\frac{k}{2}+1+\epsilon\right]\log N}{N}\right)^{\frac{1}{k}} \leq p_{\min} \leq p_{\max} \leq \frac{1}{N^{\frac{1}{k+1}+\epsilon}}$$

then a.a.s.

$$\widetilde{H}_i(\mathcal{X},\mathbb{Q}) = 0.$$

Proof. Under the present setup, we need to prove for both $i < k$ and $i > k$ by induction.

For $i = k+1$, let $p_{\max} = N^{-\frac{1}{k+1}-\epsilon}$.

From Theorem 4.6.2, if $p_{k+1} \leq \frac{1}{N^{\frac{1}{k+1}+\epsilon}}$, then $H^{k+1}(\mathcal{X},\mathbb{Q}) = 0$. Since $p_{\max} = N^{-\frac{1}{k+1}-\epsilon} \leq p_{k+1}$ is the upper bound in Corollary 4.7.1, then the lower upper bound is less than p_{k+1}.

For $i = k+2$, $p_{k+2} \leq \frac{1}{n^{\frac{1}{k+2}+\epsilon}}$

$$\begin{aligned} &p_{\max} \leq p_{k+1} < p_{k+2} \\ \Rightarrow &H^{k+2}(\mathcal{X},\mathbb{Q}) = 0 \end{aligned}$$

By induction,

$$\begin{aligned} &p_{k+1} < p_{k+j}\ \forall j > 1 \\ \Rightarrow &H^{k+j}(\mathcal{X},\mathbb{Q}) = 0\ \forall j > 1 \\ \Rightarrow &H^{i}(\mathcal{X},\mathbb{Q}) = 0\ \forall i > k \end{aligned}$$

For $i = k-1$, let $p_{\min} = \left(\frac{\left[\frac{k}{2}+1+\epsilon\right]\log N}{N}\right)^{\frac{1}{k}}$.

From Theorem 4.5.4, if $p_{k-1} \geq \left(\frac{\left[\frac{k-1}{2}+1+\epsilon\right]\log N}{N}\right)^{\frac{1}{k}}$, as $N \to \infty$, $\frac{k-1}{2} \approx \frac{k}{2}$, then $p_{\min} = \left(\frac{\left[\frac{k}{2}+1+\epsilon\right]\log N}{N}\right)^{\frac{1}{k}} \geq p_{k-1}$, therefore, $H^{k-1}(\mathcal{X},\mathbb{Q}) = 0$. Since $p_{\min}$ is the lower bound in Corollary 4.7.1, then the upper bound is greater than p_{k-1}.

For $i = k-2$, $p_{k-2} = \left(\frac{\left[\frac{k-2}{2}+1+\epsilon\right]\log n}{n}\right)^{\frac{1}{k-1}}$. Similarly, as $n \to \infty$, $\frac{k-2}{2} \approx \frac{k}{2}$, $p_{k-2} \approx \left(\frac{\left[\frac{k}{2}+1+\epsilon\right]\log n}{n}\right)^{\frac{1}{k-1}}$

$$\begin{aligned} & p_{\min \geq} p_{k-1} > p_{k-2} \\ \Rightarrow & H^{k-2}(X, \mathbb{Q}) = 0 \end{aligned}$$

By induction,

$$\begin{aligned} & p_{k-1} > p_{k-j} \ \forall j > 1 \\ \Rightarrow & H^{k-j}(X, \mathbb{Q}) = 0 \ \forall j > 1 \\ \Rightarrow & H^{i}(X, \mathbb{Q}) = 0 \ \forall i < k \end{aligned}$$

□

We now going to prove the CLT for Betti numbers in SBM. However, we need to note that the SGT has not yet been proved to hold for all values of the range of the p values. We only make the assumption that SGT is true for the CLT in SBM which is given as Conjecture 4.4.1. But this proof may not be true for all p values.

Theorem 4.7.2. *Consider Clique complex $\mathcal{X}(\mathcal{G})$ with ζ-blocks $\mathcal{G}((N_r),(p_{rs}),\zeta)$. Assume that $N_1^{-\frac{1}{k}} < p_{\min} \leq p_{\max} < N^{-\frac{1}{k+1}}$, then*

$$\beta_k(\mathcal{X}) - E\{\beta_k(\mathcal{X})\} \to N(0, Var\{\beta_k(\mathcal{X})\}) \tag{4.7.7}$$

for each k.

If some of $p_{ii}, p_{ij} = 0$, then if the complete graph can degenerate into two or more SBM, then each lower degree graph follows normal criterion, and since sum of normal is normal, the complete graph satisfies Theorem 4.7.2.

Moreover, from Corollary 4.7.1, for p_{ij} in the given regime, all the Betti numbers are zero expect for β_k a.a.s. Therefore, using

$$\sum (-1)^i \beta_i = \sum (-1)^i f_i,$$

which is the Euler characteristic defined in 2.5.2.2, it follows that

$$\tilde{\beta}_k = f_k - f_{k+1} - f_{k-1} + f_{k+2} + f_{k-2} - \ldots$$

To prove Theorem 4.7.2, we need to prove that the result which satisfies ERM also satisfies the SBM under the assumption that SGT is true.

Proof. Since in ERM,

$$N_i^{-\frac{1}{k}} < p_{ii} < N_i^{-\frac{1}{k+1}},$$

$$\left(N_i + N_j\right)^{-\frac{1}{k}} < p_{ij} < \left(N_i + N_j\right)^{-\frac{1}{k+1}}$$

and as $N_1 \leq N_i < N$ for $i \neq 1$ we can get that

$$N^{-\frac{1}{k}} < ... \leq N_1^{-\frac{1}{k}} < p_{\min} \leq p_{\max} < N^{-\frac{1}{k+1}} < ... \leq N_1^{-\frac{1}{k+1}}$$

Thus we want to show that if

$$N_1^{-\frac{1}{k}} < p_{\min} \leq p_{\max} < N^{-\frac{1}{k+1}}$$

then $\beta_k(\mathcal{X})$ tends to Normal as $N \to \infty$.

Let $\sigma^2 = Var\left(\tilde{\beta}_k\right)$, and define

$$W = \frac{\tilde{\beta}_k - E\left(\tilde{\beta}_k\right)}{\sqrt{Var\left(\tilde{\beta}_k\right)}} = \frac{1}{\sigma} \sum_{R \subseteq \mathcal{V}} (-1)^{card(R)+k+1} \left[\xi_R - E\left(\xi_R\right)\right];$$

where $\mathcal{V}$ is the collection of the vertices. Then from Section 4.7.1, we define

$$X_R = (-1)^{card(R)+k+1} \left[\xi_R - E\left(\xi_R\right)\right].$$

Since only the upper bound is required for Theorem 2.6.11, all N_i can be replaced by N. This implies that from (4.7.6)

$$E\left(\xi_R \xi_S \xi_T\right) = f_2\left(c_k, N\right) \times p_{\max}^{g_m}$$

where

$$f_2\left(c_k, N\right) \leq c_k N^{a+\alpha+\gamma-\eta_{R\cap S}-\eta_{R\cap T\setminus S}-\eta_{S\cap T\setminus R}+\eta_{R\cap T\cap S}}$$

with c_k is a constant depending only on k and

$$\prod_{r=1}^{\zeta} p_{rr}^{g_{rr}} \prod_{1\leq r<s\leq\zeta} p_{rs}^{g_{rs}} \leq p_{\max}^{\sum_{r=1}^{\zeta} g_{rr} + \sum_{1\leq r<s\leq\zeta} g_{rs}}.$$

Therefore,

$$\begin{aligned} g_m =& \sum_{r=1}^{\zeta} g_{rr} + \sum_{1\le r<s\le\zeta} g_{rs} \\ =& T(n_R,2) + T(n_S,2) + T(n_T,2) - T(\eta_{R\cap S},2) \\ & - T\left(\eta_{R\cap T\setminus S} + \eta_{R\cap S\cap T},2\right) - T\left(\eta_{S\cap T\setminus R} + \eta_{R\cap S\cap T},2\right) + T(\eta_{R\cap S\cap T},2) \end{aligned}$$

Then following the similar steps as Kahle and Meckes (2015), if we fix n_R, n_S, $\eta_{R\cap S}$ and ignore the factors which depend on these parameters, factors corresponding to n_T are left to sum up which is

$$\begin{aligned} &\frac{1}{\sigma^3} N^{n_T-\eta_{R\cap T\setminus S}-\eta_{S\cap T\setminus R}-\eta_{R\cap S\cap T}} \\ &\quad\times p_{\max}^{T(n_T,2)-T\left(\eta_{R\cap T\setminus S}+\eta_{R\cap S\cap T},2\right)-T\left(\eta_{S\cap T\setminus R}+\eta_{R\cap S\cap T},2\right)+T(\eta_{R\cap S\cap T},2)} \end{aligned} \tag{4.7.8}$$

If n_T increased by one and the new element of T is in $R\cap T\setminus S$, then the power of N in (4.7.8) does not change as $(n_T+1) - \left(\eta_{R\cap T\setminus S}+1\right)$ but the power of $p_{\max}$ does; the ratio of the new term to the old is

$$\begin{aligned} &\frac{p_{\max}^{T(n_T+1,2)-T\left(\eta_{R\cap T\setminus S}+1+\eta_{R\cap S\cap T},2\right)-T\left(\eta_{S\cap T\setminus R}+\eta_{R\cap S\cap T},2\right)+T(\eta_{R\cap S\cap T},2)}}{p_{\max}^{T(n_T,2)-T\left(\eta_{R\cap T\setminus S}+\eta_{R\cap S\cap T},2\right)-T\left(\eta_{S\cap T\setminus R}+\eta_{R\cap S\cap T},2\right)+T(\eta_{R\cap S\cap T},2)}} \\ =& p_{\max}^{\left[\frac{(n_T+1)n_T}{2}-\frac{n_T(n_T-1)}{2}\right]} \\ & \quad - \left[\frac{\left(\eta_{R\cap T\setminus S}+\eta_{R\cap S\cap T}+1\right)\left(\eta_{R\cap T\setminus S}+\eta_{R\cap S\cap T}\right)}{2}-\frac{\left(\eta_{R\cap T\setminus S}+\eta_{R\cap S\cap T}\right)\left(\eta_{R\cap T\setminus S}+\eta_{R\cap S\cap T}-1\right)}{2}\right] \\ & \quad\times p_{\max} \\ =& p_{\max}^{n_T-\eta_{R\cap T\setminus S}-\eta_{R\cap S\cap T}} \end{aligned}$$

Similarly, if n_T increased by one and the new element of T is in $R\cap S\cap T$, then the ratio of the new term to the old is

$$p_{\max}^{n_T-\eta_{R\cap T\setminus S}-\eta_{S\cap T\setminus R}-\eta_{R\cap S\cap T}}$$

Since in both cases, the power on $p_{\max}$ is non-negative, adding a new vertex to T which is already in $R\cup S$ can only make the summand smaller. On the other hand, if the new element is not in R or S, then the ratio of the new term to the old is $Np_{\max}^{n_T}$. In the regime that given in Theorem 4.7.2, this tends to infinity for $n_T < k$ and tends to zero for $n_T > k+1$. Thus, the largest

possible order for 4.7.8 is achieved when $n_T = k+1$, $\eta_{R\cap T\setminus S} + \eta_{R\cap S\cap T} = 2$ and $\eta_{S\cap T\setminus R} = 0$. Using these values in 4.7.8

$$\frac{1}{\sigma^3} N^{k-1} p_{\max}^{T(k+1,2)-1} \tag{4.7.9}$$

If only n_R is fixed, then similarly, the only sums over n_S is

$$\frac{1}{\sigma^3} N^{n_S-\eta_{R\cap S}+k-1} p_{\max}^{T(n_S,2)-T(\eta_{R\cap S},2)+T(k+1,2)-1} \tag{4.7.10}$$

Then once again, the largest possible order is $n_S = k+1$ and $\eta_{R\cap S} = 2$. Using these values in 4.7.10

$$\frac{1}{\sigma^3} N^{2k-2} p_{\max}^{2T(k+1,2)-2} \tag{4.7.11}$$

Finally, considering the full term, the upper bound is

$$\frac{1}{\sigma^3} N^{n_R+2k-2} p_{\max}^{T(n_R,2)+2T(k+1,2)-2} \tag{4.7.12}$$

by the same argument, 4.7.12 is maximized when $n_R = k+1$. Then using

$$\frac{1}{\sigma^3} N^{3k-1} p_{\max}^{3T(k+1,2)-2}$$

Thus, Theorem 2.6.11 implies that

$$d_1(W,Z) \leq \frac{C}{\sigma^3} N^{3k-1} p_{\max}^{3T(k+1,2)-2}$$

where $W = \frac{\tilde{\beta}_k - E(\tilde{\beta}_k)}{\sqrt{Var(\tilde{\beta}_k)}}$.

Furthermore, using the fact that $\sigma^2 \leq c_k N^{2k} p_{\max}^{2T(k+1,2)-1}$,

$$\begin{aligned} d_1(W,Z) \leq & C \frac{N^{3k-1} p_{\max}^{3T(k+1,2)-2}}{N^{3k} p_{\max}^{3T(k+1,2)-\frac{3}{2}}} \\ \leq & \frac{C}{N\sqrt{p_{\max}}}, \end{aligned}$$

which tends to 0 as N tends to infinity. □

4.8 SIMULATION RESULTS FOR CLT FOR SBM

In this section, some simulation results for CLT for SBM are implemented for both β_1 and β_2 using MATLAB package 'JavaPlex'. Consider the standard 2-block model where $N_1 = card(\mathcal{V}_1)$, $N_2 = card(\mathcal{V}_2)$, $N = N_1 + N_2$ and $\mathcal{V} = \mathcal{V}_1 \bigcup \mathcal{V}_2$. The regime for β_k for CLT for SBM is

$$N_1^{-\frac{1}{k}} = N^{-\phi} < N^{-x} < N^{-\frac{1}{k+1}}.$$

For simplicity, set the probabilities as (p_1, p_2, p_{12}) where

$$\begin{cases} p_{gap} = \frac{1}{4}\left(N^{-\frac{1}{k+1}} - N_1^{-\frac{1}{k}}\right) \\ p_1 = p_{gap} + N_1^{-\frac{1}{k}} \\ p_2 = 2p_{gap} + N_1^{-\frac{1}{k}} \\ p_{12} = 3p_{gap} + N_1^{-\frac{1}{k}}. \end{cases} \tag{4.8.1}$$

By choosing sample size $n_s = 10$ to 500 and 1000 for β_1, set $N = 1000$, the resulting qqplots are shown in Figure 4.8.1 and 4.8.2. It can be seen that for $N = 1000$, qqplots with $n_s \geq 50$ follow normal distribution while qqplots with $n_s < 50$ seem to have a heavy tail.

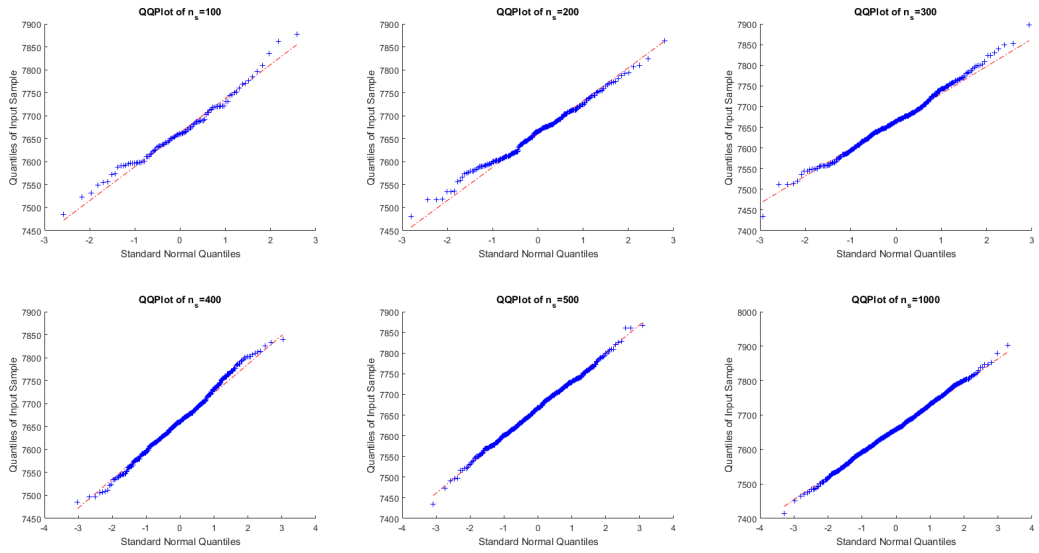

Figure 4.8.2: $\beta_1$for $n_s = 100$ to 500 and 1000 with $N = 1000$ and probability (p_1, p_2, p_{12})

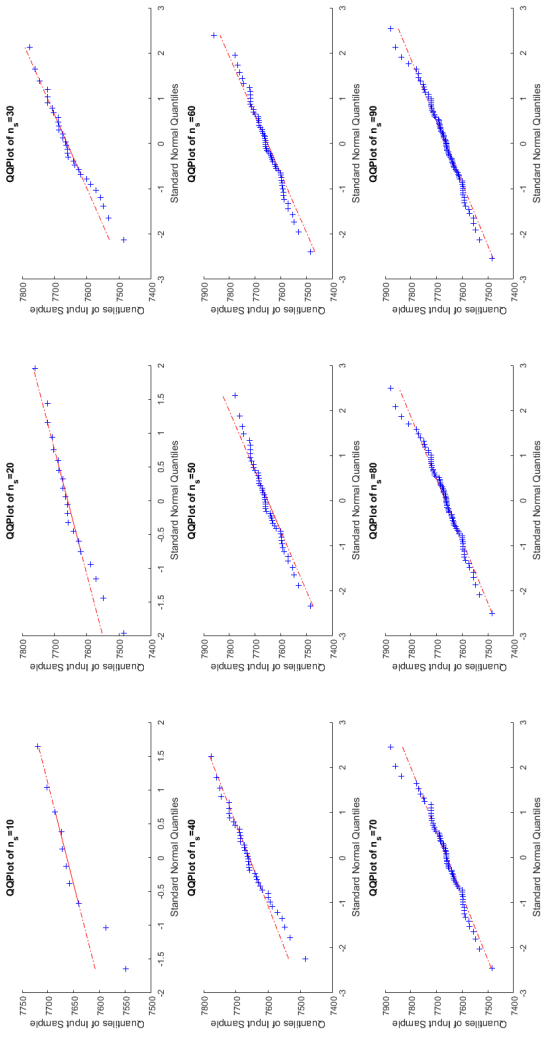

Figure 4.8.1: β_1 for $n_s = 10$ to 90 with $N = 1000$ and probability (p_1, p_2, p_{12})

Moreover, the line chart for sample mean $E(\beta_1)$ for $n_s = 10$ to 1000 is displayed in Figure 4.8.3. As can be seen, $E(\beta_1)$ is slowly steady to around 7660. Therefore, Figure 4.8.2 suggest that $n_s = 500$ is a good candidate for the rest of the example for β_1.

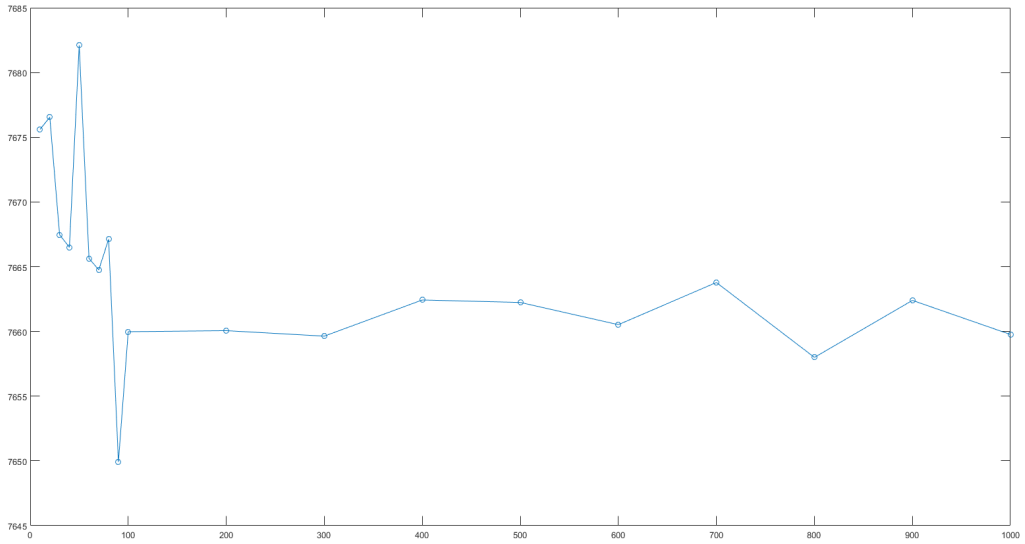

Figure 4.8.3: Sample mean $E(\beta_1)$ for $n_s = 10$ to 1000 with $N = 1000$ and probability (p_1, p_2, p_{12})

By setting $n_s = 500$, we choose $N = 10$ to 1000 for β_1 with probability $(p_1, p_{2,} p_3)$ given as (4.8.1). The resulting qqplots are given in Figure 4.8.4 and 4.8.5. It can be seen that for the small number $N \leq 50$, there are horizontal line segments on qqplots represent the repeated value for β_1. Take $N = 10$ as an example, β_1 only have values between 0 and 6, therefore, there are 5 horizontal lines in the qqplot represent the values between 0 and 5. Moreover, for $N \geq 70$, β_1 is normally distributed whereas for $N = 60$ qqplot shows a heavy tail.

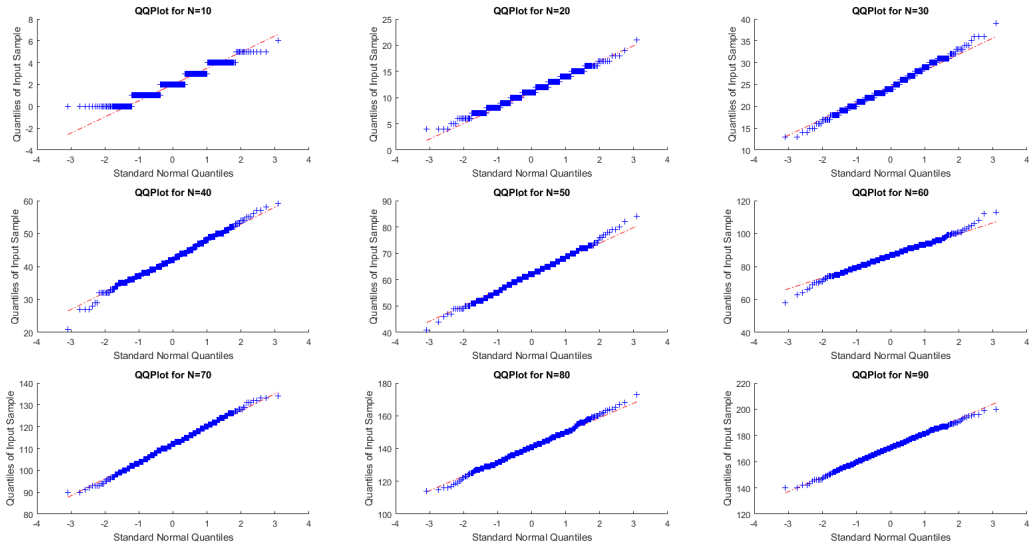

Figure 4.8.4: β_1 for $N = 10$ to 90 with $n_s = 500$ and probability (p_1, p_2, p_{12})

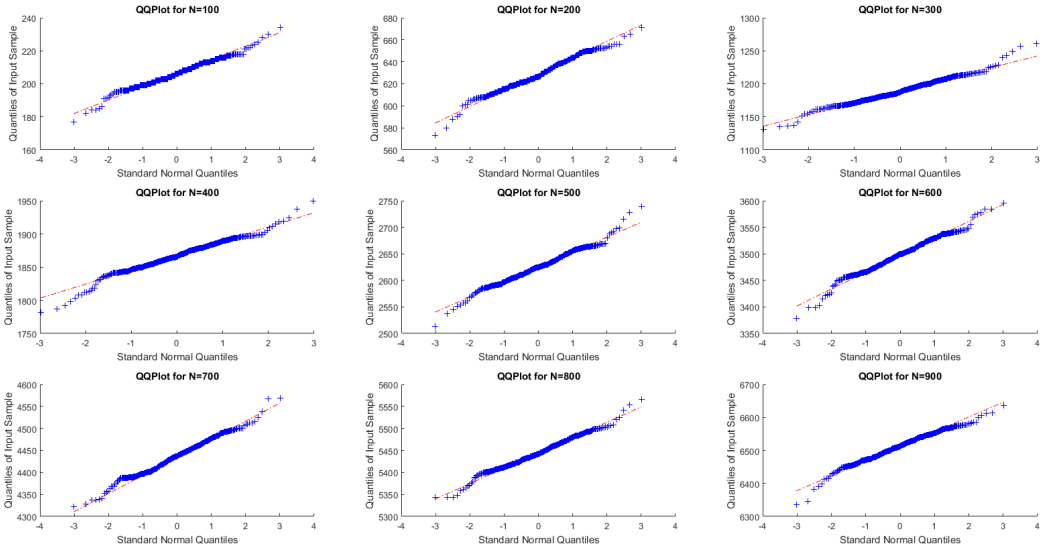

Figure 4.8.5: β_1 for $N = 100$ to 900 with $n_s = 500$ and probability (p_1, p_2, p_{12})

Moreover, we have also implemented β_2 for SBM with $N = 50$ to 100 and $n_s = 100$. We used smaller values because tetrahedrons are required for calculating β_2 which is very time consuming. As can be seen, in Figure 4.8.6 the qqplots indicate that β_2 generally follows a normal distribution although there are some outliers in the qqplots for $N \leq 70$.

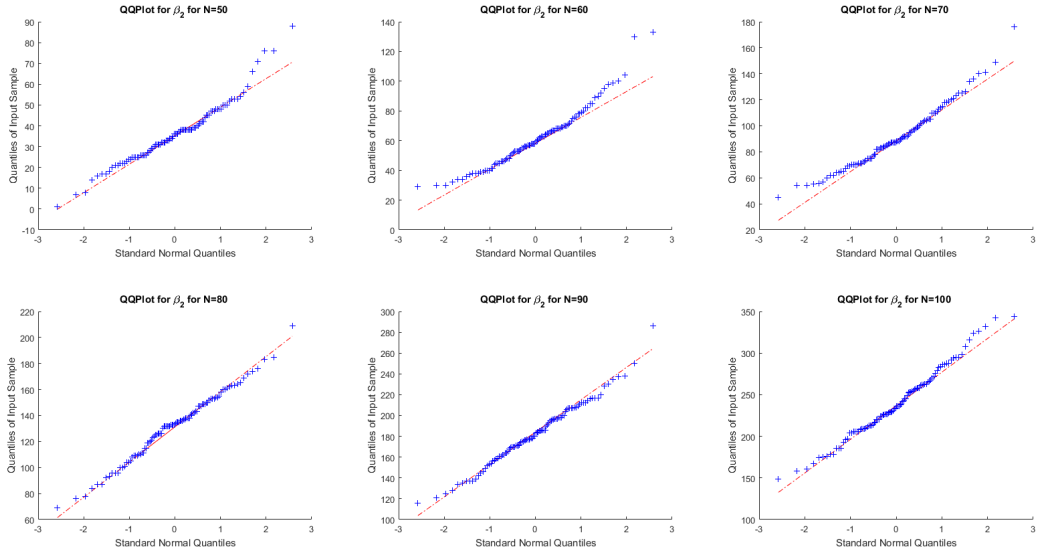

Figure 4.8.6: β_2 for $N = 50$ to 100 with $n_s = 100$ and probability (p_1, p_2, p_{12})

In addition, we have also recorded the processing time in seconds for both β_1 for β_2 with different N and the number of edges, triangles and tetrahedrons in Table 4.8.1. Besides, a comparison between including tetrahedrons or not is given for β_1 in row 5 (without tetrahedrons) and 6 (with tetrahedrons). This allows a direct insight into the computing time for the tetrahedrons.

As can be seen in Table 4.8.1, if only β_1 is considered, the number of edges and triangles increase as the numbers of vertices increase. The regime of β_1 ensures there are very few tetrahedron in the graph for β_1, i.e. 4 for 50 vertices and 8 for 100 vertices. Moreover, working with tetrahedron is very time-consuming as seen by row 6 being greater than row 5. Take $N = 300$ as an example, the processing time for including tetrahedrons is 22mins which is approximately 160 times larger than not including tetrahedrons for β_1. In addition, the number of simplexes is generally larger for β_2 than β_1, and the processing time for β_2 is about twice longer than β_1. Unfortunately, we are not able to work out anything for $N \geq 500$ for tetrahedrons.

In conclusion , one of the most time-consuming steps in calculating β_k for random graphs is working out the $(k+1)$-simplex, i.e. triangles for β_1, tetrahedrons for β_2, etc. Since this is not directly spotted by the MATLAB package 'JavaPlex', instead, we need to manually work out every simplex and add every simplex separately from the $0-$simplex to the $(k+1)$-simplex.

N	10	30	50	100	300	500	1000
edge	13	62	153	377	1794	3651	9924
triangle	2	9	42	65	311	507	1351
tetrahedron	0	0	4	0	8	N/A	N/A
β_1 w/o tet (seconds)	6.58	1.26	0.89	1.72	8.12	26.26	434.77
β_1 with tet (seconds)	6.91	1.66	3.80	33.08	1315.09	N/A	N/A
edge	21	128	340	1081	6170		
triangle	11	94	411	1612	11662		
tetrahedron	0	23	198	605	4642		
β_2 (seconds)	6.32	3.90	7.68	98.44	4534.42		

Table 4.8.1: Table of time in seconds needed for calculation of β_1 and β_2 for 2-block model using MATALB package 'JavaPlex'.

4.9 APPENDIX

In this appendix, the details are given of the derivation of how to get the first three moment of the simiplices for ζ blocks, obtained by generalising from 2 blocks to ζ blocks. See Table 4.1.1 for notations. However, in this part, two blocks is applied as an illustrative example to show how the ζ blocks results are obtained from two to ζ.

For ERM model, the mean, variance and covariance can be rewritten from Kahle and Meckes (2013), where $p = p_{rs}$ for all $1 \leq r \leq s \leq \zeta$. We have

$$E\left(f_k\right) = T\left(\sum_{r=1}^{\zeta} N_r, k+1\right) p^{T(k+1,2)}, \tag{4.9.1}$$

$$\begin{aligned} E\left(f_k^2\right) =& T\left(\sum_{r=1}^{\zeta} N_r, k+1\right) \sum_{\eta=0}^{k+1} T\left(k+1, \eta\right) T\left(\sum_{r=1}^{\zeta} N_r - \left(k+1\right), k+1-\eta\right) \\ & \times p^{T(k+1,2)} p^{T(k+1,2)-T(\eta,2)} \end{aligned} \tag{4.9.2}$$

and

$$\begin{aligned} E\left(f_k f_{k+j}\right) =& T\left(\sum_{r=1}^{\zeta} N_r, k+j+1\right) \sum_{\eta=0}^{k+1} T\left(k+j+1, r\right) \\ & \times T\left(\sum_{r=1}^{\zeta} N_r - (k+j+1), k+1-\eta\right) p^{T(k+j+1,2)} p^{T(k+1,2)-T(\eta,2)} \end{aligned} \tag{4.9.3}$$

where without loss of generality $j > 0$.

$e(f_2)$ for a 2-block model

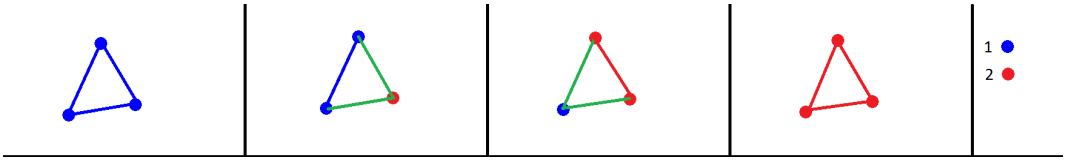

Figure 4.9.1: Take triangle as an example, $E(f_2)$ can have above 4 possibilities where a vertex marked blue indicates it belongs to $\mathcal{V}_1$ while the red one belongs to $\mathcal{V}_2$. An edge marked blue indicates it has probability p_{11} as two vertices are both in $\mathcal{V}_1$, and similar for $\mathcal{V}_2$. Meanwhile an edge is marked green if it is connected between $\mathcal{V}_1$ and $\mathcal{V}_2$, i.e. one node is $\mathcal{V}_1$ and the other one is from $\mathcal{V}_2$.

Take the triangle as an example, for $E(f_2)$ in 2-blocks model, there are four possibilities as shown in Figure 4.9.1

$$\begin{aligned} E(f_2) =& T(N_1,3)\,T(N_2,0)\,p_{11}^3 \\ &+ T(N_2,2)\,T(N_2,1)\,p_{11}^2 p_{12} \\ &+ T(N_1,1)\,T(N_2,2)\,p_{12}p_{22}^2 \\ &+ T(N_1,0)\,T(N_2,3)\,p_{22}^3. \end{aligned} \tag{4.9.4}$$

By letting $p_{11} = p_{12} = p_{12} = p$, (4.9.1) for f_2 should equal to (4.9.4), i.e.

$$\begin{aligned} T(N_1+N_2,3) =& T(N_1,3)\,T(N_2,0) + T(N_1,2)\,T(N_2,1) \\ &+ T(N_1,1)\,T(N_2,2) + T(N_1,0)\,T(N_2,3)\,. \end{aligned}$$

Then

$$\begin{aligned} 6\times LHS =& N_1^3 - 3N_1^2 + 2N_1 \\ &+ 3N_1^2N_2 + 3N_1N_2^2 \\ &- 6N_1N_2 \\ &+ N_2^3 - 3N_2^2 + 2N_2, \end{aligned}$$

and

$$\begin{aligned} 6\times RHS =& N_1^3 - 3N_1^2 + 2N_1 \\ &+ 3N_1^2N_2 - 3N_1N_2 \\ &+ 3N_1N_2^2 - 3N_1N_2 \\ &+ N_2^3 - 3N_2^2 + 2N_2. \end{aligned}$$

Therefore, $LHS = RHS$.

Moreover, (4.9.4) can be summarized as

$$E(f_2) = \sum_{a_1=0}^{3} \sum_{a_2=0}^{3} T(N_1, a_1)\, T(N_2, a_2) \sum_{r=1}^{2} a_r = 3 \times p_1^{T(a_1,2)} p_2^{T(a_2,2)} p_{12}^{T(3,2)-T(a_1,2)-T(a_2,2)},$$

which implies the general formula for the k-simplex is

$$\begin{aligned} E(f_k) =& T(N_1, a_1)\, T(N_2, a_2) \\ & \times p_{11}^{T(a_1,2)} p_{22}^{T(a_2,2)} p_{12}^{T(k+1,2)-T(a_1,2)-T(a_2,2)} \\ =& \sum_{a_1=0}^{k+1} \sum_{a_2=0}^{k+1} \prod_{r=1}^{2} T(N_r, a_r)\, p_{rr}^{T(a_r,2)} \prod_{1 \le r < s \le 2} p_{rs}^{\tau(a_1,a_2)}, \end{aligned}$$
$$\sum_{r=1}^{2} a_r = k+1$$

where $T(N, a) = \frac{N!}{a!(N-a)!}$ and $\tau(a_1, a_2) = T(a_1 + a_2, 2) - T(a_1, 2) - T(a_2, 2)$.

Therefore, if we extend from 2-blocks to ζ-blocks then

$$\begin{aligned} E(f_k) =& \sum_{a_1=0}^{k+1} \sum_{a_2=0}^{k+1} \cdots \sum_{a_\zeta=0}^{k+1} \prod_{r=1}^{\zeta} T(N_r, a_r)\, p_{rr}^{T(a_r,2)} \prod_{1 \le r < s \le \zeta} p_{rs}^{\tau(a_r,a_s)} \\ & \sum_{r=1}^{\zeta} a_r = k+1 \\ =& f(a_r k + 1, \zeta) \prod_{r=1}^{\zeta} T(N_r, a_r)\, p_r^{T(a_r,2)} \prod_{1 \le r < s \le \zeta} p_{rs}^{\tau(a_r,a_s)} \end{aligned}$$

where $f(a_r, k+1, \zeta) = \sum_{a_1=0}^{k+1} \sum_{a_2=0}^{k+1} \cdots \sum_{a_\zeta=0}^{k+1}$.
$$\sum_{r=1}^{\zeta} a_r = k+1$$

$e\left(f_2^2\right)$ for a 2-block model

For variance, as shown as first moment, we start with the 2-block model for triangles. Then two triangles can have the following possibilities as shown in the table.

(N_1, N_2)	$(3,0)$	$(2,1)$	$(1,2)$	$(0,3)$
$(3,0)$	$\eta = 0$ $\eta = 1$ $\eta = 2$ $\eta = 3$	$\eta = 0$ $\eta = 1$ $\eta = 2$	$\eta = 0$ $\eta = 1$	$\eta = 0$
$(2,1)$	$\eta = 0$ $\eta = 1$ $\eta = 2$	$r = 0$ $r = 1$ $r = 2$ $r = 3$	$\eta = 0$ $\eta = 1$ $\eta = 2$	$\eta = 0$ $\eta = 1$
$(1,2)$	$\eta = 0$ $\eta = 1$	$r = 0$ $r = 1$ $r = 2$	$\eta = 0$ $\eta = 1$ $\eta = 2$ $\eta = 3$	$\eta = 0$ $\eta = 1$ $\eta = 2$
$(0,3)$	$\eta = 0$	$\eta = 0$ $\eta = 1$	$\eta = 0$ $\eta = 1$ $\eta = 2$	$\eta = 0$ $\eta = 1$ $\eta = 2$ $\eta = 3$

Table 4.9.1: Possibilities for two triangles, where η =number of intersection points between two triangles. (N_1, N_2) =number of vertices chosen from each block.

We are going to perform some of the calculations in Table 4.9.1 as examples in this thesis. First of all, $(3,0) \to (3,0)$ or $(0,3) \to (0,3)$ as these two only contain $\mathcal{V}_1$ and $\mathcal{V}_2$, therefore, it will have a similar behaviour as the (4.9.2) which is the general case.

$$
\begin{aligned}
E\left(f_2^2|\mathcal{V}_1\right) &= T\left(N_1,3\right) \sum_{\eta=0}^{3} T\left(3,\eta\right) T\left(N_1-3,3-\eta\right) p_{11}^3 \cdot p_{11}^{3-T(\eta,2)} \\
&= T\left(N_1,3\right) T\left(3,0\right) T\left(N_1-3,3-0\right) p_{11}^6 \quad (\eta=0) \\
&\quad + T\left(N_1,3\right) T\left(3,1\right) T\left(N_1-3,3-1\right) p_{11}^6 \quad (\eta=1) \\
&\quad + T\left(N_1,3\right) T\left(3,2\right) T\left(N_1-3,3-2\right) p_{11}^5 \quad (\eta=2) \\
&\quad + T\left(N_1,3\right) T\left(3,3\right) T\left(N_1-3,3-3\right) p_{11}^3. \quad (\eta=3)
\end{aligned}
\tag{4.9.5}
$$

(4.9.5) is the 2nd column and 2nd row of the Table 4.9.1. If we change N_1 to N_2 in equation (4.9.5), then we will get the second moment results for 5th row and 5th column for vertices only in $\mathcal{V}_2$.

Secondly, we want to get the results for vertices only in $\mathcal{V}_1$ to only in $\mathcal{V}_2$ vice versa, which is the 5th row, 2nd column and 2nd row, 5th column in Table 4.9.1. In this case, it only have $\eta = 0$,

$$\begin{aligned} E\left(f_2^2|\mathcal{V}_1,\mathcal{V}_2\right) =& T\left(N_1,3\right)T\left(N_2,0\right)T\left(N_1-3,0\right)T\left(N_2,3\right)p_{11}^3p_{22}^3 \\ &+ T\left(N_1,0\right)T\left(N_2,3\right)T\left(N_1,3\right)T\left(N_2-3,0\right)p_{22}^3p_{11}^3. \end{aligned} \tag{4.9.6}$$

Thirdly, we considered the case the first triangle contains vertices only in $\mathcal{V}_1$ or $\mathcal{V}_2$, the second triangle has vertices both from $\mathcal{V}_1$ and $\mathcal{V}_2$. This is the 2nd row with 3rd and 4th column and 5th row with 3rd and 4th column.

For $\eta = 0$,

$$\begin{aligned} & E\left(f_2^2|\mathcal{V}_1,\mathcal{V}_1\mathcal{V}_2\right) + E\left(f_2^2|\mathcal{V}_2,\mathcal{V}_1\mathcal{V}_2\right) \\ =& T\left(N_1,3\right)T\left(N_2,0\right)T\left(N_1-3,2\right)T\left(N_2,1\right)p_{11}^3p_{11}p_{12}^2 \\ &+ T\left(N_1,3\right)T\left(N_2,0\right)T\left(N_1-3,1\right)T\left(N_2,2\right)p_{11}^3p_{11}p_{12}^2 \\ &+ T\left(N_1,0\right)T\left(N_2,3\right)T\left(N_1,2\right)T\left(N_2-3,1\right)p_{11}^3p_{11}p_{12}^2 \\ &+ T\left(N_1,0\right)T\left(N_2,3\right)T\left(N_1,1\right)T\left(N_2-3,2\right)p_{11}^3p_{11}p_{12}^2. \end{aligned}$$

For $\eta = 1$,

$$\begin{aligned} & E\left(f_2^2|\mathcal{V}_1,\mathcal{V}_1\mathcal{V}_2\right) + E\left(f_2^2|\mathcal{V}_2,\mathcal{V}_1\mathcal{V}_2\right) \\ =& T\left(N_1,3\right)T\left(3,1\right)T\left(N_1-3,2-1\right)T\left(N_2,1\right)p_{11}^3p_{11}p_{12}^2 \\ &+ T\left(N_1,3\right)T\left(3,1,\right)T\left(N_1-3,1-1\right)T\left(N_2,2\right)p_{11}^3p_{12}^2p_{22} \\ &+ T\left(N_2,3\right)T\left(3,1\right)T\left(N_1,2\right)T\left(N_2-3,1-1\right)p_{22}^3p_{11}p_{12}^2 \\ &+ T\left(N_2,3\right)T\left(3,1\right)T\left(N_1,\right)T\left(N_2-3,2-1\right)p_2^3p_{12}^2p_{22}. \end{aligned}$$

For $\eta = 2$,

$$\begin{aligned} & E\left(f_2^2|\mathcal{V}_1,\mathcal{V}_1\mathcal{V}_2\right) + E\left(f_2^2|\mathcal{V}_2,\mathcal{V}_1\mathcal{V}_2\right) \\ =& T\left(N_1,3\right)T\left(3,2,\right)T\left(N_1-3,2-2\right)T\left(N_2,1\right)p_{11}^3p_{12}^2 \\ &+ T\left(N_2,3\right)T\left(3,2\right)T\left(N_1,1\right)T\left(N_2-3,2-2\right)p_{22}^3p_{12}^2. \end{aligned}$$

We omit the proof for the case that first triangle has vertices both from $\mathcal{V}_1$ and $\mathcal{V}_2$, second triangle contains vertices only in $\mathcal{V}_1$ or $\mathcal{V}_2$, which is the 3rd and 4th rows with 2nd and 5th columns, since this is the opposite to the third case.

Lastly, we considered the case that both triangles contain vertices from block 1 and 2. This is the 3rd and 4th rows with 3rd and 4th columns in the table.

For $\eta = 0$,

$$\begin{aligned}E\left(f_2^2|\mathcal{V}_1\mathcal{V}_2\right) =& T\left(N_1,2\right)T\left(N_2,1\right)T\left(N_1-2,2\right)T\left(N_2-1,1\right)p_{11}p_{12}^2p_{11}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(N_1-2,1\right)T\left(N_2-1,2\right)p_{11}p_{12}^2p_{12}^2p_{22}\\ &+T\left(N_1,1\right)T\left(N_2,2\right)T\left(N_1-1,2\right)T\left(N_2-2,1\right)p_{22}p_{12}^2p_{11}p_{12}^2\\ &+T\left(N_1,1\right)T\left(N_2,2\right)T\left(N_1-1,1\right)T\left(N_2-2,2\right)p_{22}p_{12}^2p_{22}p_{12}^2.\end{aligned}$$

For $\eta = 1$,

$$\begin{aligned}&E\left(f_2^2|\mathcal{V}_1\mathcal{V}_2\right)\\ =&T\left(N_1,2\right)T\left(N_2,1\right)T\left(2,1\right)T\left(1,0\right)T\left(N_1-2,1\right)T\left(N_2-1,1\right)p_{11}p_{12}^2p_{11}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(2,0\right)T\left(1,1\right)T\left(N_1-2,2\right)T\left(N_2-1,0\right)p_{11}p_{12}^2p_{11}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(2,1\right)T\left(1,0\right)T\left(N_1-2,0\right)T\left(N_2-1,2\right)p_{11}p_{12}^2p_{22}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(2,0\right)T\left(1,1\right)T\left(N_1-2,1\right)T\left(N_2-1,1\right)p_{11}p_{12}^2p_{22}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(1,1\right)T\left(2,0\right)T\left(N_1-1,1\right)T\left(N_2-2,1\right)p_{22}p_{12}^2p_{11}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(1,0\right)T\left(2,1\right)T\left(N_1-1,2\right)T\left(N_2-2,0\right)p_{22}p_{12}^2p_{11}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(1,1\right)T\left(2,0\right)T\left(N_1-1,0\right)T\left(N_2-2,2\right)p_{22}p_{12}^2p_{22}p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(1,0\right)T\left(2,1\right)T\left(N_1-1,1\right)T\left(N_2-2,1\right)p_{22}p_{12}^2p_{22}p_{12}^2.\end{aligned}$$

For $\eta = 2$,

$$\begin{aligned}&E\left(f_2^2|\mathcal{V}_1\mathcal{V}_2\right)\\ =&T\left(N_1,2\right)T\left(N_2,1\right)T\left(2,2\right)T\left(1,0\right)T\left(N_2-1,1\right)p_{11}p_{12}^2p_{12}^2\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(2,1\right)T\left(1,1\right)T\left(N_1-2,1\right)p_{11}p_{12}^2p_1p_{12}\\ &+T\left(N_1,2\right)T\left(N_2,1\right)T\left(2,1\right)T\left(1,1\right)T\left(N_2-1,1\right)p_{11}p_{12}^2p_{22}p_{12}\\ &+T\left(N_1,1\right)T\left(N_2,2\right)T\left(1,1\right)T\left(2,1\right)T\left(N_1-1,1\right)p_{22}p_{12}^2p_{11}p_{12}\\ &+T\left(N_1,1\right)T\left(N_2,2\right)T\left(1,0\right)T\left(2,2\right)T\left(N_1-1,1\right)p_{22}p_{12}^2p_{12}^2\\ &+T\left(N_1,1\right)T\left(N_2,2\right)T\left(1,1\right)T\left(2,1\right)T\left(N_2-2,1\right)p_{22}p_{12}^2p_{22}p_{12}.\end{aligned}$$

For $\eta = 3$,

$$E\left(f_2^2|\mathcal{V}_1\mathcal{V}_2\right) = T\left(N_1,2\right)T\left(N_2,1\right)p_{11}p_{12}^2 + T\left(N_1,1\right)T\left(N_2,2\right)p_{22}p_{12}^2.$$

From these 5 steps, we can conclude that

$$\begin{aligned}
E\left(f_2^2\right) =& \sum_{\substack{a_1=0 \\ \sum_{r=1}^{2} a_r = k+1}}^{3}\sum_{a_2=0}^{3} T\left(N_1,a_1\right)T\left(N_2,a_2\right) \sum_{\eta=0}^{3}\sum_{\substack{\eta_1=0 \\ \sum_{r=1}^{2}\eta_r=\eta}}^{\eta}\sum_{\eta_2=0}^{\eta} T\left(a_1,\eta_1\right)T\left(a_2,\eta_2\right) \\
& \times \sum_{\substack{\alpha_1=0 \\ \sum_{r=1}^{2}\alpha_r=3-\eta}}^{3-\eta}\sum_{\alpha_2=0}^{3-\eta} T\left(N_1-a_1,\alpha_1-\eta_1\right)T\left(N_2-a_2,\alpha_2-\eta_2\right) \\
& \times p_{11}^{T(a_1,2)}p_{22}^{T(a_2,2)}p_{12}^{T(3,2)-T(a_1,2)-T(a_2,2)} \\
& \times p_{11}^{T(\alpha_1,2)-T(\eta_1,2)}p_{22}^{T(\alpha_2,2)-T(\eta_2,2)} \\
& \times p_{12}^{[T(3,2)-T(\alpha_1,2)-T(\alpha_2,2)]-[T(\eta,2)-T(\eta_1,2)-T(\eta_2,2)]} \\
=& f\left(a_r,3,2\right)\prod_{r=1}^{2} T\left(N_r,a_r\right)\sum_{\eta=0}^{3} f\left(\eta_r,\eta,2\right)\prod_{r=1}^{\zeta} T\left(a_r,\zeta_r\right) \\
& \times f\left(\alpha_r,3-\eta,2\right)\prod_{r=1}^{2} T\left(N_r-a_r,\alpha_r-\eta_r\right) \\
& \times \prod_{r=1}^{2} p_r^{T(a_r,2)+T(\alpha_r,2)-T(\eta_r,2)} \prod_{1\le r<s\le 2} p_{rs}^{\tau(a_r,a_s)+\tau(\alpha_r,\alpha_s)-\tau(\eta_r,\eta_s)}.
\end{aligned} \tag{4.9.7}$$

Furthermore, if we rewrite f_2 as f_k, the general second moment for f_k is shown as the following,

$$\begin{aligned}
E\left(f_k^2\right) =& f\left(a_r,k+1,2\right)\prod_{r=1}^{2} T\left(N_r,a_r\right)\sum_{\eta=0}^{k+1} f\left(\eta_r,\eta,2\right)\prod_{r=1}^{2} T\left(a_r,\eta_r\right) \\
& \times f\left(\alpha_r,k+1-\eta,2\right)\prod_{r=1}^{2} T\left(N_r-a_r,\alpha_r-\eta_r\right) \\
& \times \prod_{r=1}^{2} p_{rr}^{T(a_r,2)+T(\alpha_r,2)-T(\eta_r,2)} \prod_{1\le r<s\le 2} p_{ij}^{\tau(a_r,a_s)+\tau(\alpha_r,\alpha_s)-\tau(\eta_r,\eta_s)}.
\end{aligned}$$

Thus for the f_k in ζ-block model, by changing 2 to ζ we can conclude that

$$\begin{aligned}E\left(f_k^2\right) =& f\left(a_r, k+1, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r, a_r\right) \sum_{\eta=0}^{k+1} f\left(\eta_r, \eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(a_r, \eta_r\right) \\ & \times f\left(\alpha_r, k+1-\eta\right) \prod_{r=1}^{\zeta} T\left(N_r-a_r, \alpha_r-\eta_r\right) \\ & \times \prod_{r=1}^{\zeta} p_{rr}^{T(a_r,2)+T(\alpha_r,2)-T(\eta_r,2)} \prod_{1\le r<s\le\zeta} p_{ij}^{\tau(a_r,a_s)+\tau(\alpha_r,\alpha_s)-\tau(\eta_r,\eta_s)}.\end{aligned}$$

$e\left(f_1 f_2\right)$ for a 2-block model

For covariance, we take edge and triangle as an illustrative example, it is going to follow the same steps as the second moment for triangle.

Firstly, for all vertices that are in either $\mathcal{V}_1$ or $\mathcal{V}_2$.

$$\begin{aligned}E\left(f_1 f_2 | \mathcal{V}_1\right) &= T\left(N_1, 3\right) \sum_{\eta=0}^{2} T\left(3, \eta\right) T\left(N_1-3, 2-\eta\right) p_{11}^3 \cdot p_{11} \\ &= T\left(N_1, 3\right) T\left(3, 0\right) T\left(N_1-3, 2-0\right) p_{11}^4 \quad (\eta=0) \\ &\quad + T\left(N_1, 3\right) T\left(3, 1\right) T\left(N_1-3, 2-1\right) p_{11}^4 \quad (\eta=1) \\ &\quad + T\left(N_1, 3\right) T\left(3, 2\right) T\left(N_1-3, 2-2\right) p_{11}^3. \quad (\eta=2)\end{aligned}$$

Secondly, vertices for triangle are all in $\mathcal{V}_1$ and vertices for edges are all in $\mathcal{V}_2$ or vice versa. In this case, only case $\eta = 0$ exists.

$$\begin{aligned}E\left(f_1 f_2 | \mathcal{V}_1, \mathcal{V}_2\right) = & T\left(N_1, 3\right) T\left(N_2, 2\right) p_{11}^3 p_{22} \\ & + T\left(N_2, 3\right) T\left(N_1, 2\right) p_{22}^3 p_{11}.\end{aligned}$$

Thirdly, vertices for triangle are all in either $\mathcal{V}_1$ or $\mathcal{V}_2$, while vertices for edges are in both $\mathcal{V}_1$ and $\mathcal{V}_2$. We can have two possibilities for the intersection points which are $\eta = 0$ and $\eta = 1$.

For $\eta = 0$,

$$\begin{aligned}& E\left(f_1 f_2 | \mathcal{V}_1, \mathcal{V}_1 \mathcal{V}_2\right) + E\left(f_1 f_2 | \mathcal{V}_2, \mathcal{V}_1 \mathcal{V}_2\right) \\ =& T\left(N_1, 3\right) T\left(N_1-3, 1\right) T\left(N_2, 1\right) p_{11}^3 p_{12} \\ & + T\left(N_2, 3\right) T\left(N_1, 1\right) T\left(N_2-3, 1\right) p_{22}^3 p_{12}.\end{aligned}$$

For $\eta = 1$,

$$\begin{aligned} & E\left(f_1 f_2 | \mathcal{V}_1, \mathcal{V}_1 \mathcal{V}_2\right) + E\left(f_1 f_2 | \mathcal{V}_2, \mathcal{V}_1 \mathcal{V}_2\right) \\ =& T\left(N_1, 3\right) T\left(3, 1\right) T\left(N_2, 1\right) p_1^3 p_{12} \\ & + T\left(N_2, 3\right) T\left(3, 1\right) T\left(N_1, 1\right) p_2^3 p_{12}. \end{aligned}$$

Again, we omit the proof for the case that the triangle has vertices both from $\mathcal{V}_1$ and $\mathcal{V}_2$, and the edge contains vertices only in $\mathcal{V}_1$ or $\mathcal{V}_2$, which is the third case.

Lastly, we considered the case when both triangle and edge have vertices in both $\mathcal{V}_1$ and $\mathcal{V}_2$. In this situation, there are three possibilities for intersection points which are $\eta = 0, 1, 2$.

For $\eta = 0$,

$$\begin{aligned} & E\left(f_1 f_2 | \mathcal{V}_1 \mathcal{V}_2\right) \\ =& T\left(N_1, 2\right) T\left(N_2, 1\right) T\left(N_1 - 2, 1\right) T\left(N_2 - 1, 1\right) p_{11} p_{12}^2 p_{12} \\ & + T\left(N_1, 1\right) T\left(N_2, 2\right) T\left(N_1 - 1, 1\right) T\left(N_2 - 2, 1\right) p_{22} p_{12}^2 p_{12}. \end{aligned}$$

For $\eta = 1$,

$$\begin{aligned} & E\left(f_1 f_2 | \mathcal{V}_1 \mathcal{V}_2\right) \\ =& T\left(N_1, 2\right) T\left(N_2, 1\right) T\left(2, 1\right) T\left(N_1 - 2, 1\right) p_{11} p_{12}^2 p_{12} \\ & + T\left(N_1, 2\right) T\left(N_2, 1\right) T\left(1, 1\right) T\left(N_1 - 2, 1\right) p_{11} p_{12}^2 p_{12} \\ & + T\left(N_1, 1\right) T\left(N_2, 2\right) T\left(1, 1\right) T\left(N_2 - 2, 1\right) p_{22} p_{12}^2 p_{12} \\ & + T\left(N_1, 1\right) T\left(N_2, 2\right) T\left(2, 1\right) T\left(N_1 - 1, 1\right) p_{22} p_{12}^2 p_{12}. \end{aligned}$$

For $\eta = 2$,

$$\begin{aligned} & E\left(f_1 f_2 | \mathcal{V}_1 \mathcal{V}_2\right) \\ =& T\left(N_1, 2\right) T\left(N_2, 1\right) T\left(2, 1\right) T\left(1, 1\right) p_{11} p_{12}^2 \\ & + T\left(N_1, 1\right) T\left(N_2, 2\right) T\left(1, 1\right) T\left(2, 1\right) p_{22} p_{12}^2 \end{aligned}$$

From these five steps, we can conclude that

$$
\begin{aligned}
E\left(f_1 f_2\right) = & \sum_{\substack{a_1=0 \\ \sum_{r=1}^{2} a_r = k+1}}^{3} \sum_{a_2=0}^{3} T\left(N_1, a_1\right) T\left(N_2, a_2\right) \sum_{\eta=0}^{3} \sum_{\substack{\eta_1=0 \\ \sum_{r=1}^{2} \eta_r = \eta}}^{\eta} \sum_{\eta_2=0}^{\eta} T\left(a_1, \eta_1\right) T\left(a_2, \eta_2\right) \\
& \times \sum_{\substack{\alpha_1=0 \\ \sum_{r=1}^{2} \alpha_r = 2-\eta}}^{2-\eta} \sum_{\alpha_2=0}^{2-\eta} T\left(N_1 - a_1, \alpha_1 - \eta_1\right) T\left(N_2 - a_2, \alpha_2 - \eta_2\right) \\
& \times p_{11}^{T(a_1,2)} p_{22}^{T(a_2,2)} p_{12}^{T(3,2)-T(a_1,2)-T(a_2,2)} \\
& \times p_{11}^{T(\alpha_1,2)-T(\eta_1,2)} p_{22}^{T(\alpha_2,2)-T(\eta_2,2)} \\
& \times p_{12}^{[T(3,2)-T(\alpha_1,2)-T(\alpha_2,2)]-[T(\eta,2)-T(\eta_1,2)-T(\eta_2,2)]} \\
= & f\left(a_r, 3, 2\right) \prod_{r=1}^{2} T\left(N_r, a_r\right) \sum_{\eta=0}^{3} f\left(\eta_r, \eta, 2\right) \prod_{r=1}^{\zeta} T\left(a_r, \zeta_r\right) \\
& \times f\left(\alpha_r, 2-\eta, 2\right) \prod_{r=1}^{2} T\left(N_r - a_r, \alpha_r - \eta_r\right) \\
& \times \prod_{r=1}^{2} p_r^{T(a_r,2)+T(\alpha_r,2)-T(\eta_r,2)} \prod_{1 \le r < s < 2} p_{rs}^{\tau(a_r,a_s)+\tau(\alpha_r,\alpha_s)-\tau(\eta_r,\eta_s)}.
\end{aligned}
\tag{4.9.8}
$$

Similar to $E\left(f_k^2\right)$, if we change 2 blocks to ζ blocks and f_1, f_2 to f_k, f_{k+j} where $j \geq 0$, we can conclude the general formula for $E\left(f_k f_{k+j}\right)$ as

$$
\begin{aligned}
E\left(f_k f_{k+j}\right) = & f\left(a_r, k+j+1, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r, a_r\right) \sum_{\eta=0}^{k+1} f\left(\eta_r, \eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(a_r, \eta_r\right) \\
& \times f\left(\alpha_r, k+1-\eta\right) \prod_{r=1}^{\zeta} T\left(N_r - a_r, \alpha_r - \eta_r\right) \\
& \times \prod_{r=1}^{\zeta} p_{rr}^{T(a_r,2)+T(\alpha_r,2)-T(\eta_r,2)} \prod_{1 \le r < s \le \zeta} p_{ij}^{\tau(a_r,a_s)+\tau(\alpha_r,\alpha_s)-\tau(\eta_r,\eta_s)}.
\end{aligned}
$$

BINOMIAL IDENTITY

From the derivation of the each moment, if η is fixed, we can conclude the following binomial identity by setting $p = p_{ii} = p_{ij}$ for all i, j.

For the first moment, $E\left(f_k\right)$

$$T\left(\sum_{r=1}^{\zeta} N_r, k+1\right) = f\left(a_r, k+1, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r, a_r\right),$$

where $T\left(n,a\right) = \frac{n!}{a!(n-a)!}$ = binomial coefficient, $\tau\left(a,b\right) = T\left(a+b,2\right) - T\left(a,2\right) - T\left(b,2\right)$ and $f\left(a_r, k+1, \zeta\right) = \sum_{a_1=0}^{k+1} \sum_{a_2=0}^{k+1} \ldots \sum_{a_\zeta=0}^{k+1}$.

$\sum_{r=1}^{\zeta} a_r = k+1$

For the second moment $E\left(f_k^2\right)$

$$\begin{aligned} & T\left(\sum_{r=1}^{\zeta} N_r, k+1\right) T\left(k+1, \eta\right) T\left(\sum_{r=1}^{\zeta} N_r - \left(k+1\right), k+1-\eta\right) \\ =& f\left(a_r, k+1, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r, a_r\right) \sum_{\eta=0}^{k+1} f\left(\eta_r, \eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(a_r, \eta_r\right) \\ & \times f\left(\alpha_r, k+1-\eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r - a_r, \alpha_r - \eta_r\right). \end{aligned}$$

For the second moment $E\left(f_k f_{k+1}\right)$

$$\begin{aligned} & T\left(\sum_{r=1}^{\zeta} N_r, k+j+1\right) T\left(k+j+1, \eta\right) T\left(\sum_{r=1}^{\zeta} N_r - \left(k+1\right), k+1-\eta\right) \\ =& f\left(a_r, k+1, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r, a_r\right) \sum_{\eta=0}^{k+1} f\left(\eta_r, \eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(a_r, \eta_r\right) \\ & \times f\left(\alpha_r, k+1-\eta, \zeta\right) \prod_{r=1}^{\zeta} T\left(N_r - a_r, \alpha_r - \eta_r\right). \end{aligned}$$

5

TDA WITH SUBSAMPLING FOR POINT CLOUDS

5.1 INTRODUCTION

As mentioned earlier, one of the most important difficulties with computational Persistent Homology is that it is often not possible to use the full dataset when calculating Betti numbers. In this case, it is important to find out the relationship between the topological summary of the full sample and subsamples. In addition, the data points obtained by the specific methods should represent the topological information of the whole dataset as far as possible.

The aim of this chapter is to investigate numerically the relationship between the topological summaries computed under the full sample and subsamples via barcodes using different test statistics and sampling methods. The tests we consider are Kolmogorov–Smirnov (KS) test, Cramer-von Mises (CvM) test test and permutation test which are defined in Section 2.2. The selection methods for subsamples from the full sample which we consider in this chapter are: (i) completely random resampling without replacement (RC); and (ii) structured random resampling (explained later) without replacement (RS). RC is defined at the beginning of Section 5.2 and RS is explained in Section 5.3.

The outline of this chapter is as follows. In Section 5.2, we present simulation results for unit square which compare TDA for the full sample with TDA from subsamples selected according to RC. In Section 5.3, we analyze the human brains data considered by Bendich et al. (2016). Here, we compare structured subsampling, method RS, and purely random subsampling, method RC.

5.2 SIMULATED DATA

In this section, we consider the set $P = \{\mathbf{v}_i \in \mathbb{R}^d : i = 1, ..., N\}$ of data points, where each v_i is generated iid from F. We refer to P as the full sample. A subsample P_s of size $M < N$ selected from P according to subsampling method RC is defined as follows: $P_s = \{\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_M}\}$ where $i_1, \ldots, i_M$ are M distinct indices selected randomly with replacement from $\{1, \ldots, N\}$. As discussed in Section 2.5.4, (b, d) is the birth and death time for topological feature and $l = d - b$ is defined as the length of persistence and λ_1 is the landscape function for $\beta_k \geq 1$.

5.2.1 *Empirical Distribution*

As mentioned in Section 1, one of the most important challenges with Persistent Homology is that a dataset with a large number of data points may exceed the capability of the computational program at higher dimensional homology such as β_1, β_2, etc. Therefore, a small experiment that shows the relationship between time consumed and the size of the dataset is first given in the Table 5.2.1.

We choose the size of the full sample between $N = 50$ to 5000. Rips complex is considered as the filtration on $v_1, ..., v_N$ to ensure that the number of edges, $card(\mathcal{E})$, is tracked when maximum filtration value, $R_{\max}$ is changed between the maximum pairwise distance, R_{pair} and $\frac{R_{pair}}{2}$. Moreover, $\mathbf{v}_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$, the unit square is chosen as the example. Therefore, $R_{pair} = \sqrt{2}$. In this test, MATLAB package 'TDATools' is used for calculating β_1.

One of the most important factors for the sample size is the RAM of the computer. A better RAM directly implies a larger dataset. Unfortunately, my personal laptop has a much lower RAM than the university desktop. As a result, the sample size limitations for my own laptop is approximately 3000 data points for β_1. When the program broke down, MATLAB displays the error message: " Java exception occurred java.lang. OutOfMemoryError: Requested array size exceeds VM limit."

	$card(\mathscr{E})$	$\frac{R_{\max}}{2}$ (seconds)	$card(\mathscr{E})$	$R_{\max}$ (seconds)
50	794	2.81	1225	5.27
100	3251	4.30	4950	8.29
250	22454	11.13	31125	12.14
500	86996	16.48	124750	15.37
1000	356104	46.76	499500	68.61
2000	1448405	1028.81	1999000	574.25
3000	3269338	3757.49	4498500	error
4000	5863984	error	7998000	error

Table 5.2.1: Table of time in seconds needed for calculation of β_1 for $\mathbf{v}_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$ using MATALB package 'TDATools' where $card(\mathscr{E})$ is tracking the number of edges for different maximum filtration values.

In general, the sample size limitations of a computational program for a dataset is approximately 5000 data points for β_1, even though the full dataset may contain a far large number of data points. Therefore, we try to estimate Persistent Homology via the emprical cdf.

We define

$$G_M(x) = \frac{1}{\eta_M} \sum_{j=1}^{\eta_M} \mathbb{I}\{l_j \leq x\} \tag{5.2.1}$$

as the empirical cdf defined in (2.2.1) of the full sample, P, where $l_j = d_j - b_j$ is the length of the barcode, and η_M is the total number of bars. The theoretical cdf of l_j from the full sample is defined as

$$G_0 = E[G_M]. \tag{5.2.2}$$

We then randomly select points without replacement from the full sample P and construct empirical cdf for subsamples as

$$F_m(x) = \frac{1}{\eta_m} \sum_{j=1}^{\eta_m} \mathbb{I}\{l_{m,j} \leq x\} \tag{5.2.3}$$

where m is the m-th subsample, η_m is the total number of barcodes for the m-th subsample and $m = 1, ..., m_0$. Similarly, the theoretical cdf for subsamples is defined as

$$F_0 = E[F_m] \tag{5.2.4}$$

for each $m = 1, ..., m_0$.

However, it is worth noting that the l_j are not independent since each l_j is from the persistence diagram where the persistence diagram is a summary of the filtration of the complex. Moreover, l_j may not be identically distributed. However, the empirical cdf can still be defined and we try to exact some useful information from the ecdf of the full sample and subsamples.

Additionally, as mentioned earlier, in TDA, only long-lived topological features are considered as part of topological signals. Those which only appear for a short period are considered as topological noise.

In this section, full sample P are simulated data, therefore, when calculating G_M and F_m, all barcodes l_j and $l_{m,j}$ that are not zero have been included in the calculation. However, when the real dataset is analyzed, such as brain tree data which is used in Section 5.3, only 100 largest l_j are chosen to represent the entire persistence diagram. In another case, l_j and $l_{m,j}$ may be chosen at a same proportional rate, such as the top 25% of the barcodes. The examples for both 20 largest barcodes and top 25% are given later in this section.

We now consider the mean ecdf from m subsamples

$$\bar{F}_m(x) = \frac{1}{m_0} \sum_{m=1}^{m_0} F_m(x) \tag{5.2.5}$$

and the theoretical function

$$\bar{F}_0 = E_{\bar{F}_M}[\bar{F}_M] \tag{5.2.6}$$

which has the same definition as above. Then in the present setup, we can consider the KS statistic for the full sample and subsamples as follows,

$$d_m = \sup_{-\infty<x<\infty} |G_M - F_m| \tag{5.2.7}$$

where $m = 1, ..., m_0$.

It should be noted that it is very unlikely to find the underlying distributions for (5.2.2), (5.2.4) and (5.2.6).

Similarly, the standard two-sample KS test conditions defined in Section 2.2 are not satisfied here. The samples are not iid However, we still try to

construct KS statistic to access if there is anything we can learn from the result. Let

$$H_0 : \ F_0 = \bar{F}_0 \ \text{ VS } \ H_1 : \ F_0 \neq \bar{F}_0 \tag{5.2.8}$$

using the KS test statistic

$$d_{mm} = \sup_{-\infty<x<\infty} |F_m - \bar{F}_m| .$$

We also define a second hypothesis test, for each M

$$H_0 : \ G_0 = F_0 \ \text{ VS } \ H_1 : \ G_0 \neq F_0. \tag{5.2.9}$$

where the corresponding KS test statistic is (5.2.7).

As mentioned before, the numerical results for the empirical cdf and KS test given in this section are implemented using MATLAB package TDA-Tools. With the earlier results in Table 5.2.1, we choose the size of the full sample as $N = 1000$, i.e. $P = \left\{\mathbf{v}_i \in \mathbb{R}^d : i = 1, ..., 1000\right\}$. We use again the unit square, $\mathbf{v}_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$ as the example, and the Rips complex as the filtration to ensure that the maximum pairwise distance, $\sqrt{2}$ is included as the maximum filtration value, R_{max}. This means that all $\begin{pmatrix} 1000 \\ 2 \end{pmatrix}$ edges are included in the filtration process and each topological feature corresponding to β_1 is going to be captured. After computing G_M from the full sample, we then randomly select 200 points from P without replacement and computing F_m and d_m respectively. We also repeat subsampling for 1000 times i.e. $F_1, ..., F_{1000}$ and construct the corresponding statistics $d_1, ..., d_{1000}$.

Persistence diagrams which are defined in Section 2.5.4 for β_1 for the full sample and 1 subsample are presented for $\mathbf{v}_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$ in Figure 5.2.1. One thing we notice from Figure 5.2.1 is that in subsamples, there is relatively less topological noise than in the full sample. The major reason is that if there are fewer data points, there are going to form fewer numbers of edges, triangles, tetrahedron, etc. In TDA, the birth-and-death of an object indicates a β number, a β with a short lifetime indicates topological noise. Therefore, a smaller dataset implies fewer objects have short lifetimes which leads to less topological noise.

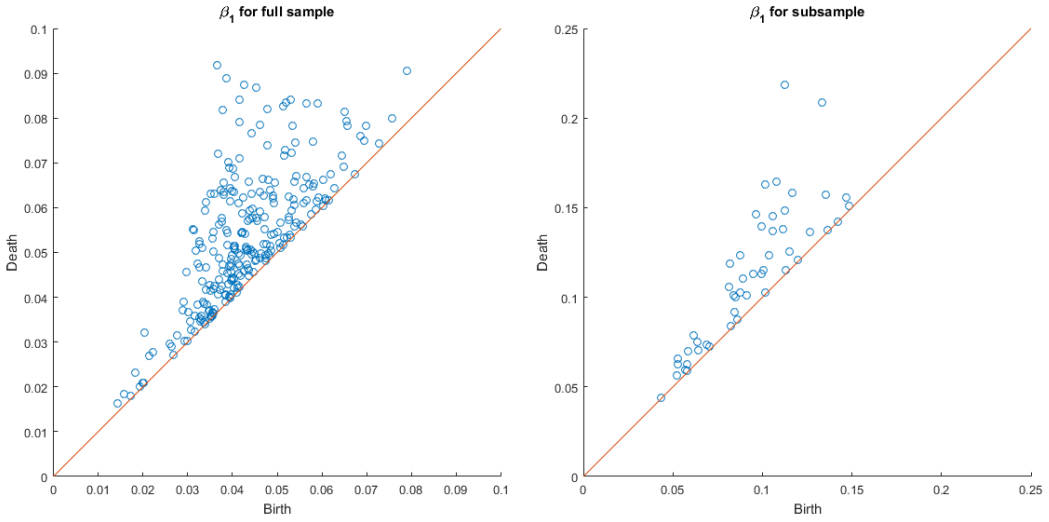

Figure 5.2.1: Persistence diagram for β_1 for full sample (left) and 1 subsample (right).

Moreover, in Figure 5.2.2, G_M, 10 randomly selected F_m and $\bar{F}_m$ are presented for $\mathbf{v}_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$.

The histogram and qqplot are presented in Figure 5.2.3. Both seem to suggest on the unit square, the KS statistic which is defined in 5.2.7 is normally distributed. This is surprising since in general the KS statistic d_m should not follow a normal distribution but extreme value distribution.

Therefore, we now define $\mu = E\left(d_m\right)$ and $s^2 = \frac{1}{m_0-1} Var\left(d_m\right)$ which is the sample mean and sample variance for d_m. Moreover, we also introduce a new hypotheses

$$H_0 : d_m \sim N\left(\mu, s^2\right) \tag{5.2.10}$$

For the first test $H_0 : F_0 = \bar{F}_0$ given in (5.2.8), we test for both 1% and 5% significant (sig.) level for the unit square under two-sample KS test. At 5% sig. level, 66 out of 1000 values are rejected by the null hypotheses, i.e. 66 F_m do not have the same distribution as $\bar{F}_m$. If we change the sig. level from 5% to 1%, the rejection rate decreases to 1.8%. The results from both sig. level suggest that most of the subsamples are from the same distribution.

For the second test $H_0 : G_0 = F_0$ given in (5.2.9), if one of the subsamples is randomly selected, then the $p-$value for two-sample KS test is 0.0093. Therefore, the null hypothesis is rejected under 1% sig. level, i.e. for this chosen subsample, two-sample KS test suggests that full sample and the subsample do not come from the same distribution.

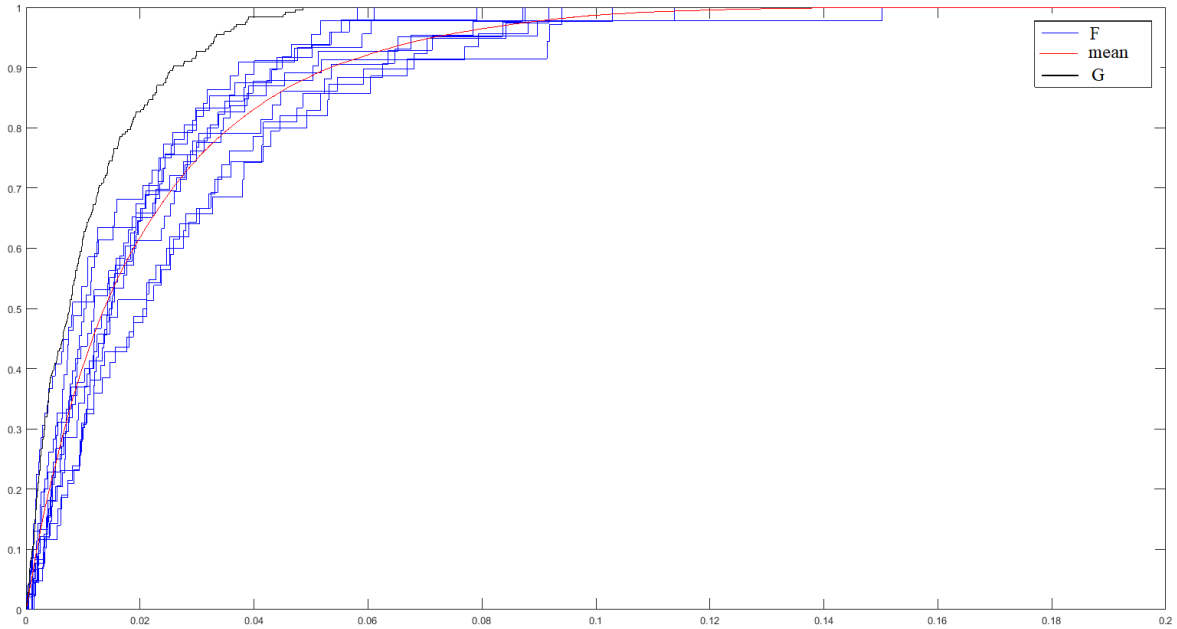

Figure 5.2.2: The empirical cdf of barcodes from full sample G_M (black) and the empirical cdf of the barcodes from 10 subsamples F_m (blue) and the mean function for ecdf of barcodes from 1000 subsamples $\bar{F}_m$ (red)

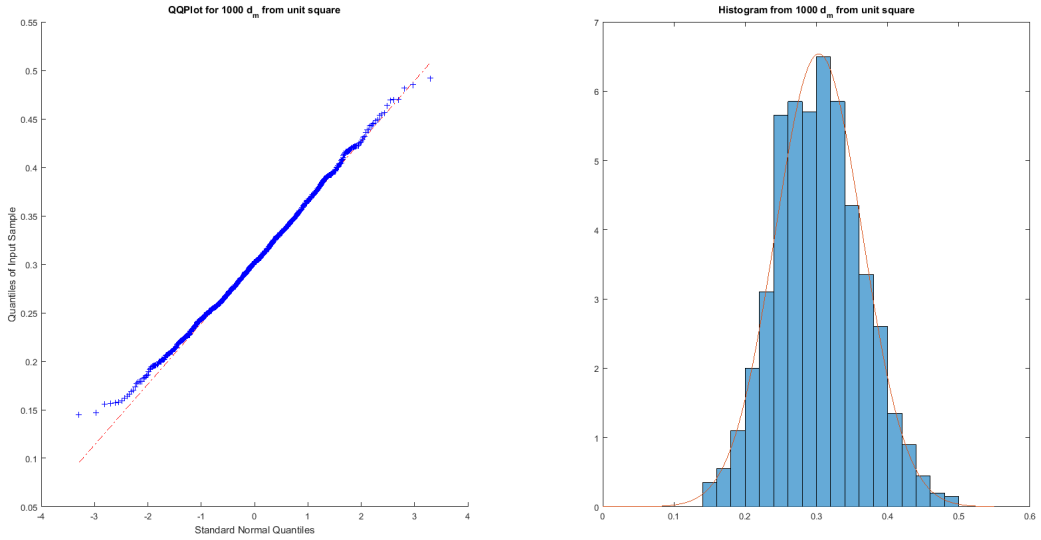

Figure 5.2.3: (left) qqplot and (right) histogram for 1000 d_m values for $v_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$. The $p-$values for both KS and CvM test are $p_{KS} = 0.7283$ and $p_{CvM} = 0.0929$.

Moreover, if the two-sample KS test is repeated for all 1000 subsamples, at 1% sig. level, 771 out of 1000 are rejected by the null hypotheses test. Instead, if we test the hypotheses $H_0 : G_0 = \bar{F}_0$ given in (5.2.10), the resulting $p-$value is 0. These two results suggest that in general, the full sample and subsamples do not have the same distribution for the length of barcodes where points are generated from $Uniform\left([0,1]^2\right)$.

Alternatively, we also use two-sample Cramer-von Mises (CvM) test which is defined in Section 2.2, where similar results are obtained for all three hypotheses tests for unit square. We accept the null hypotheses for (5.2.8) and reject the null hypotheses for (5.2.9).

The permutation test which is defined in Section 2.2 is another possible hypothesis test for (5.2.8) and (5.2.9). In this case, the KS statistic d_m defined in (5.2.7) is the test statistic T_0 defined in Section 2.2. By setting the number of permutations to be 1000, the resulting $p-$value is 0.0110 for (5.2.8) suggesting that we do not reject the null hypothesis at 1% sig. level.

The summary of the test statistic results for points generated from the unit square is displayed in table 5.2.2.

	$H_0: F_0 = \bar{F}_0$ (5.2.8)	$H_0: G_0 = F_0$ (5.2.9)
single 2−sample KS test	N/A	0.0093
1000 2−sample KS test	18	771
single 2−sample CvM test	N/A	0.0067
1000 2−sample CvM test	12	869
single permutation test	N/A	0.0110

Table 5.2.2: Table for the unit square with $p-$value and number of rejections out of 1000 for $H_0: G_0 = F_0$ and $H_0: F_0 = \bar{F}_0$ under RC where G_0, F_0 and $\bar{F}_0$ are defined as Equation 5.2.2, 5.2.4 and 5.2.6.

However, we have been unable to prove that d_m which is defined in (5.2.7) itself is normally distributed.

Since simulation result suggest that the d_m follows a normal distribution, we continue the investigation for this unusual situation for d_m. In the following calculation, we take d_m calculated from the simulated sample from the unit square as an example.

Let

$$x_m = \text{argmax} \left| G_M(x) - F_m(x) \right| \tag{5.2.11}$$

and

$$x_0 = \text{argamax} \left| G_M(x) - \bar{F}_m(x) \right|. \tag{5.2.12}$$

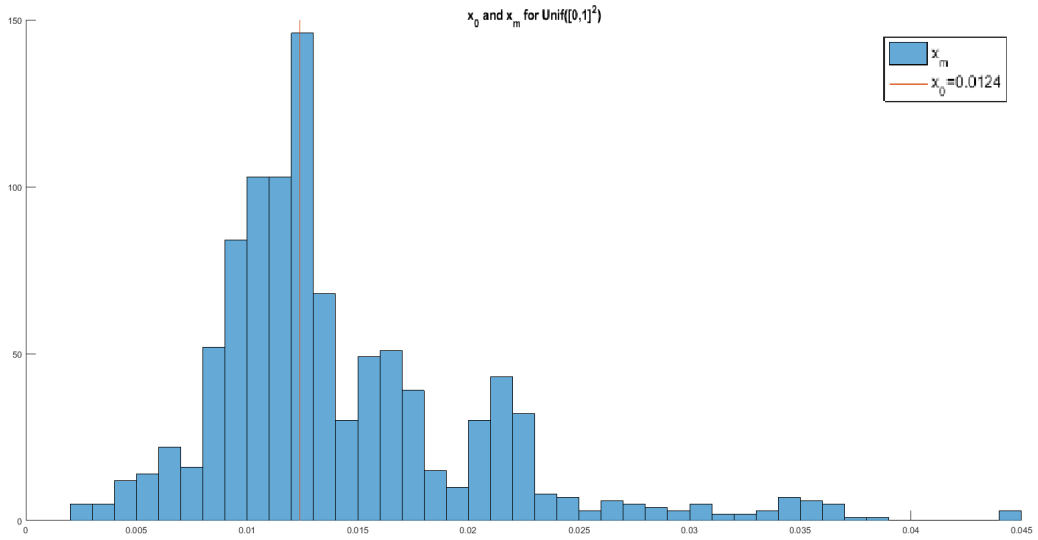

Figure 5.2.4: x_0 & x_m for $v_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$ where $x_0 = 0.0124$

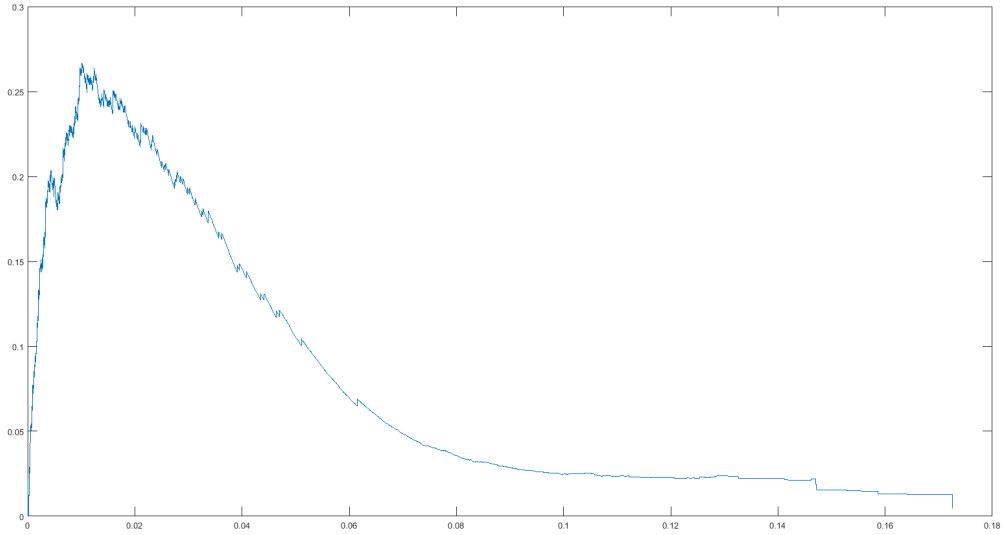

Figure 5.2.5: $G - \bar{F}_m$ for full samples $v_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$

It can been that in Figure 5.2.4 , x_m only assumes values in the range $(0, 0.045)$ comparing to the overall range. Moreover, x_0 is in the range where most x_m are i.e. $(0.008, 0.013)$. Moreover, nearly each value of $G_M(x_m) - F_m(x_m)$ is greater than 0 for points generated from the unit square, which implies that the d_m can be simplified as $d_m = \max(G_M - F_m)$ without the absolute sign. Therefore, we can rewrite d_m as follows

$$\begin{aligned} d_m &= \max |G_M(x) - F_m(x)| \\ &\approx \max (G_M(x) - F_m(x)) \\ &\approx G_M(x_m) - F_m(x_m) \quad (x_m = \operatorname{argmax} |G_M(x) - F_m(x)|) \\ &\approx [G_M(x_m) - \bar{F}_m(x_m)] - [F_m(x_m) - \bar{F}_m(x_m)]. \end{aligned} \tag{5.2.13}$$

Since $G_M(x_m) - \bar{F}_M(x_m)$ is unique for each x_m, the variation is only determined by the term $F_m(x_m) - \bar{F}_m(x_m)$. The simulation results are shown on the bottom row in Figure 5.2.5, where $F_m(x_m) - \bar{F}_M(x_m)$ is also normally distributed as d_m is normally distributed. Similarly, by replacing x_m with x_0, $F_m(x_0) - \bar{F}_m(x_0)$ presented on the top row in Figure 5.2.5 also appears to be normally distributed. Since the value of x_m ranges from 0 to 0.045, we then work out the value range in which $F_m - \bar{F}_m$ is normally distributed. By constructing order statistics $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(j)} \leq ...$ for all possible x values, the resulting range is found out to be $j = 9000$ to $j = 26000$, which is 44% of the total x values. Moreover, $x_0 = x_{16214}$ whereas $x_{(9000)} = 0.0059$

and $x_{(26000)} = 0.0266$, which again implies that x_0 falls within the range which normality appears to hold.

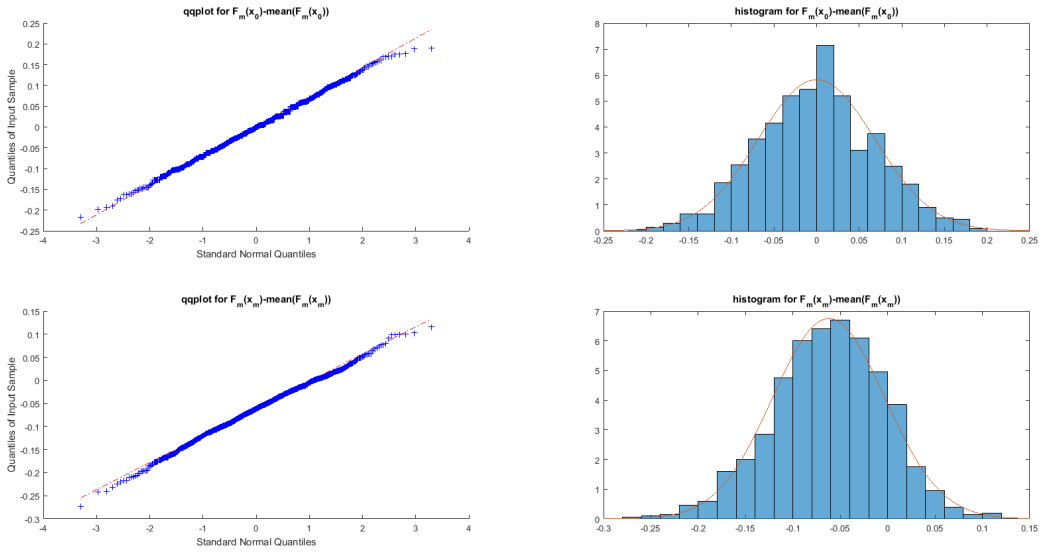

Figure 5.2.6: qqplot and histogram for $F_m(x_0) - \bar{F}_m(x_0)$ on top row whereas qqplot and histogram for $F_m(x_m) - \bar{F}_m(x_m)$ on bottom row

However, in Persistent Homology, only features that last for a long period are considered as the topological signal. Therefore, one possible modification of empirical cdf is that only the top n longest barcodes are chosen as the input data. Let $\{q_1, ..., q_n\} = \left\{l_{(1)}, ..., l_{(n)}\right\}$ where $l_{(1)} \geqslant l_{(2)} \geqslant ... \geqslant l_{(n)}$ i.e. the order statistics of lengths l_j. In this unit square example, the 20 longest barcodes are taken as the first example. In this case, (5.2.1) and (5.2.3) are modified as

$$\begin{aligned} G_M(x) &= \frac{1}{20} \sum_{j=1}^{20} \mathbb{I}\{q_j \leq x\} \\ F_m(x) &= \frac{1}{20} \sum_{j=1}^{20} \mathbb{I}\{q_{m,j} \leq x\} \end{aligned} \tag{5.2.14}$$

where q_j is the j-th longest barcode in the descending order for the full sample and $q_{m,j}$ is the j-th longest barcode in the descending order for the m-th subsample.

Another possibility is to retain the ratio of the barcodes between the full sample and subsamples, i.e. instead of always chosen the ξ longest barcodes, the $a\% \cdot \xi$ longest barcodes are considered. For example, in Table 5.2.2 the ratio is chosen to be 100%. However, the ratio is changed to the top 25%

of the longest barcodes as the input value. Then, (5.2.1) and (5.2.3) are modified as

$$\begin{aligned} G_M(x) &= \frac{1}{0.25\eta} \sum_{j=1}^{0.25\eta} \mathbb{I}\{q_j \leq x\} \\ F_m(x) &= \frac{1}{0.25\eta_m} \sum_{j=1}^{0.25\eta_m} \mathbb{I}\{q_{m,j} \leq x\} \end{aligned} \tag{5.2.15}$$

where q_j is the j-th longest barcode in the descending order for the full sample and $q_{m,j}$ is the $j-$th longest barcode in descending order for the m-th subsample.

In Figure 5.2.7, the top graph present the empirical cdf for G_M, 10 randomly selected F_m for 20 longest barcodes while the bottom graph display the empirical cdf for 25% longest barcodes for $\mathbf{v}_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$. Compare Figure 5.2.7 with Figure 5.2.2, there is not many difference between the bottom graph and Figure 5.2.2 due to the ratio of barcodes between the full sample and subsample stay unchanged. While between top graph of Figure 5.2.7 and Figure 5.2.2, as the size of the barcodes are fixed for the top graph, the empirical cdf for G_M is no longer dominating F_m.

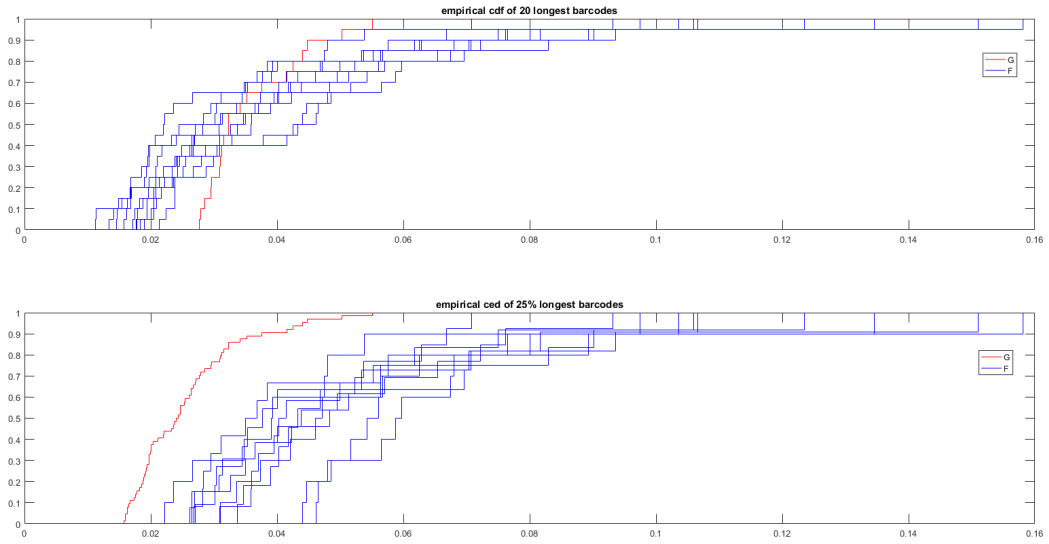

Figure 5.2.7: The empirical cdf of barcodes from full sample G_M (red) and the empirical cdf of the barcodes from 10 subsamples F_m (blue) for empirical cdf of 20 longest barcodes (top) and 25% longest barcodes (bottom).

The resulting summary of the test statistics results for $\mathbf{v}_i \overset{iid}{\sim} Uniform\left([0,1]^2\right)$ is displayed in table 5.2.3 for both 20 longest barcodes and top 25% of the barcodes.

	20 longest barcodes	25% top barcodes
1000 2−sample KS test	297	985
1000 2−sample CvM test	111	995
$p-$value for d_m	2.4260×10^{-51}	4.4274×10^{-7}

Table 5.2.3: Table for the unit square where both 20 longest barcodes and top 25% of the barcodes are included as input with null hypothesis $H_0: \ G_0 = F_0$ and $p-$values for d_m where G_0 and F_0 are defined as Equation 5.2.2 and 5.2.4.

From Table 5.2.3, the d_m is no longer normal distribution for neither 20 longest barcodes nor top 25% barcodes. This does not imply that the full sample and subsamples are from the same distribution. The qqplots for both d_m are shown in Figure 5.2.8.

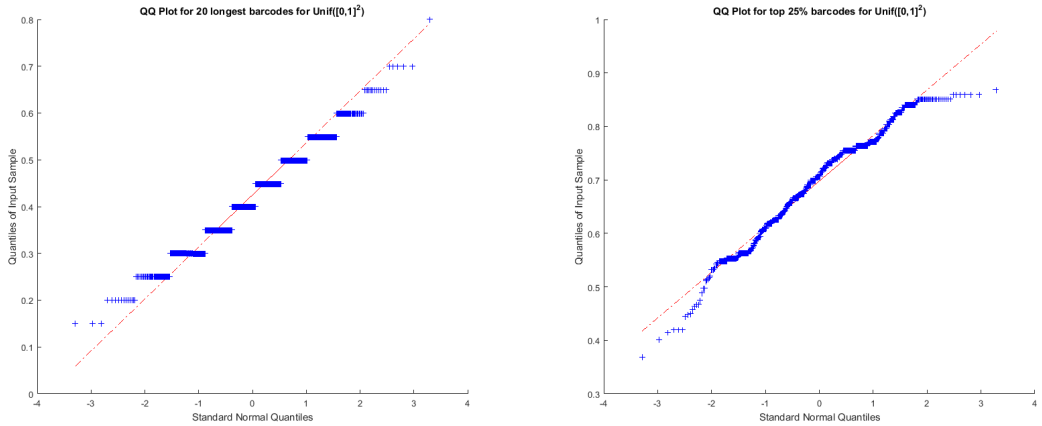

Figure 5.2.8: QQplots for unit square where 20 longest and top 25% of the barcodes are chosen as the input values

5.2.2 *Sum of Lengths*

In this section, another possible summary for Persistent Homology arises if we consider the barcode as a step function, then the integral of Betti number at radius r is

$$\int_0^r \beta(t)dt = \sum \left(d_j - b_j\right) \mathbb{I}\left\{d_j < r\right\} + \sum \left(r - b_j\right) \mathbb{I}\left\{\left(b_j < r < d_j\right)\right\}$$

where $\left(b_j, d_j\right)$ is the birth-and-death coordinate for the j-th topological feature.

Therefore

$$G_M = \sum_{j=1}^{\eta_M} l_j$$

is defined to be the sum of all line segments for full sample.

Then the total sum for the subsample is defined as

$$F_m = \sum_{j=1}^{\eta_m} l_{m,j}.$$

As a results, the test statistic is defined again as 5.2.7 $d_m = \sup_{-\infty<x<\infty} |G_M - F_m|$ for each m.

In this case, the test statistic $d_m = G_M - F_m$ is again normally distributed with $p = 0.6268$

5.3 BRAIN ARTERY TREE DATA

The dataset analysed in this study are tree of arteries in the brain of each of a number of human subjects which was introduced in Section 2.6.2, both male and female. The full data base consists of 98 datasets, and each dataset is called as BT_N where $N = 1, \ldots, 98$ with ages ranging from 18 to 72. Moreover, each dataset is presented as a MATLAB file that gives the (x, y, z) coordinates of each vertex and the connection information of the vertices. The data points in each dataset has an order of 10^5 and are spread among approximately 200 tree branches. The original dataset included more information, such as the branch thickness, people's handedness and the relationship between people, which are given as supplementary material. However, the additional information was not used by Bendich et al. (2016).

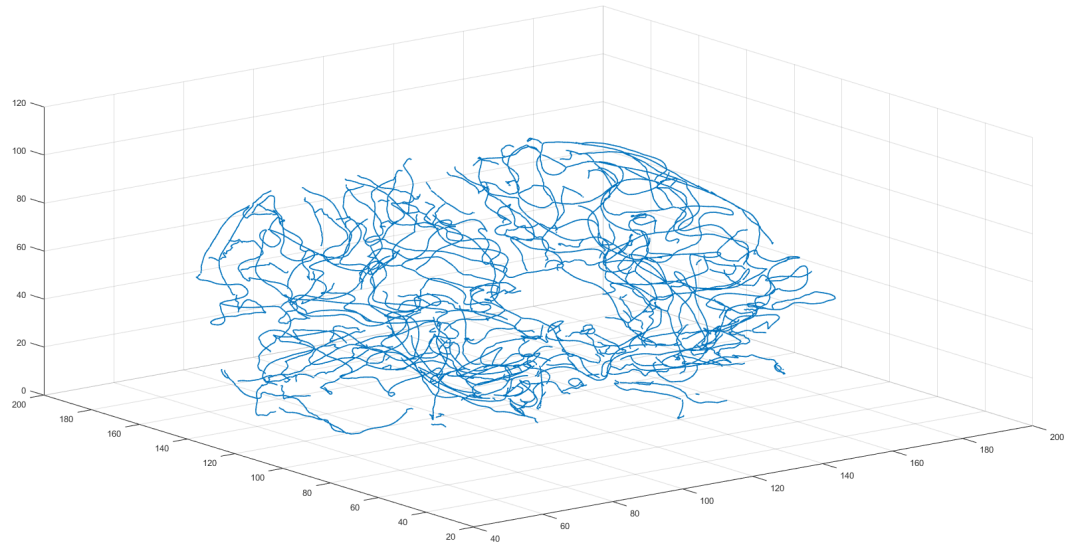

Figure 5.3.1: This is one of the brain tree datasets as a 3-dimensional plot.

Since it is not possible to compute the full tree dataset for β_1, which implies it is not possible to compute G_M defined in Section 5.2. Therefore, we can only analyses the data based on subsamples. However, the BT data is a tree sturcture data which suggest BT have a strong structure effect. Therefore, apart from the random selection without replacement (RC) , we introduce two other methods, which are maximin selection using Hausdorff distance and random selection with a fixed skeleton (RS). Note that in Section 5.2, we only considered completely random subsampling without replacement, referred to as method RC, while in this section, we consider both method RC and also structured random sampling, referred to as method RS. The latter involves random selection with a fixed skeleton. Additionally, a third selection method which is maximin selection is introduced in this section.

5.3.1 *Methods of Selection*

MAXIMIN SELECTION

Maximin selection which is introduced by choosing points using Hausdorff distance, which is defined as (2.7.1), between sets of points by selecting the first point randomly. Inductively, if B_{i-1} is the set of the first $i-1$ chosen points, then the i-th point $x \in A$ is selected to maximise the Hausdorff distance $d_H(A, B_{i-1})$ where A is the original given dataset.

The advantage of the maximin selection is that if Y' is the subset of Y, then $d_H(Y, Y')$ tends to 0 much faster than the random selection. From Matlab package 'JavaPlex', the resulting Hausdorff distance is 0.9736 at sample size of 3000.

However, there are some disadvantages of the maximin selection. Firstly, although the first point of maximin selection is chosen at random, the effect of such random selection decreases as sample size increases. Take $n = 50$ points as an example, if the $n = 50$ points are selected 50 times from BT_1, the Hausdorff distance lies within $(20, 22.5)$. Thus by selecting 10 to 50 data points from BT_1 and repeating this process for 50 times, the maximum difference for the Hausdorff distance for each sample size decreases from 9 to 2 by increasing the sample size from 10 to 50, as shown in Figure 5.3.2. Moreover, the time taken for choosing 3000 points is about 120 mins. Comparatively, the random selection method takes less than 1 min. As a result, the maximin selection is not a suitable sampling method for the data analysed in this report as it tends to be a non-random selection and the selection process is very time consuming.

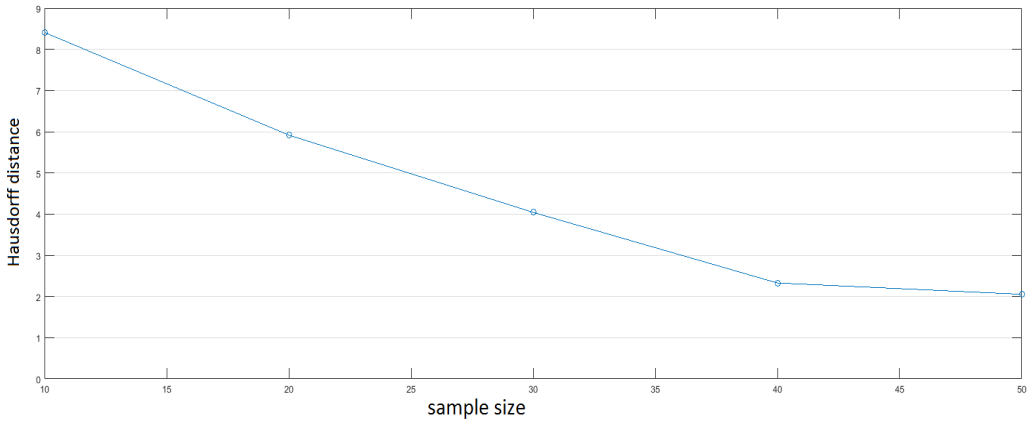

Figure 5.3.2: The effect of the maximin selection as sample size increase

RANDOM SELECTION

Two types of random selection methods are now considered: complete random sampling without replacement of any point (RC) and randomly selected points without replacement after including the basic skeleton of the tree (RS). For tree data, RS firstly selects the start and end points of each branch and then chooses the rest of the points randomly without replacement.

As mentioned before, (b_i, d_i) is the birth and death time for i-th topological feature, in this case, 1D hole and $l_i = d_i - b_i$ is defined as the length of the persistence. Then let $\{q_1, ..., q_n\} = \left\{l_{(1)}, ..., l_{(n)}\right\}$ where $l_{(1)} \geqslant l_{(2)} \geqslant ... \geqslant l_{(n)}$, which means that the length is sorted in descending order.

Take BT_{97} as an example, 3000 points are randomly selected using RC. In Figure 5.3.3, it can be seen that $q_j < 1$ from approximately $j = 100$. Additionally, $q_1 > 20$, so only the first 100 q_j are regraded as topological signals and written as $q = (q_1,\ q_2, ...,\ q_{100})$ where q_j is the j-th longest length of persistent in Dgm_1. Therefore q_j is also known as the j-th most persistent 1D hole.

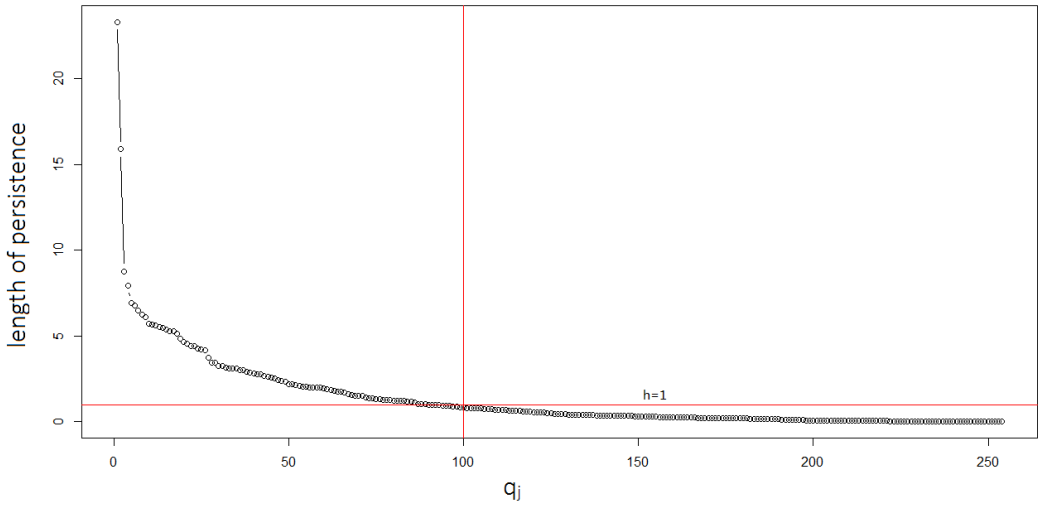

Figure 5.3.3: The sorted length of the persistence for BT_{97} by RS.

An example of the three methods RC, RS and maximin selection is shown in Figure 5.3.4. There is no significant difference in the line graph when sample size is 3000. On the other hand, 3000 points from one tree required about 5 mins to compute the 1 dimensional Persistent Homology, decreasing the sample points has been considered in the repeated process. In this case 1000 data points from the BT_{97} are chosen instead. However, the difference between the two lines is not significant in the line plot. As a result, instead of deciding which method to use at this point, both RS and RC are going to be used in the next section.

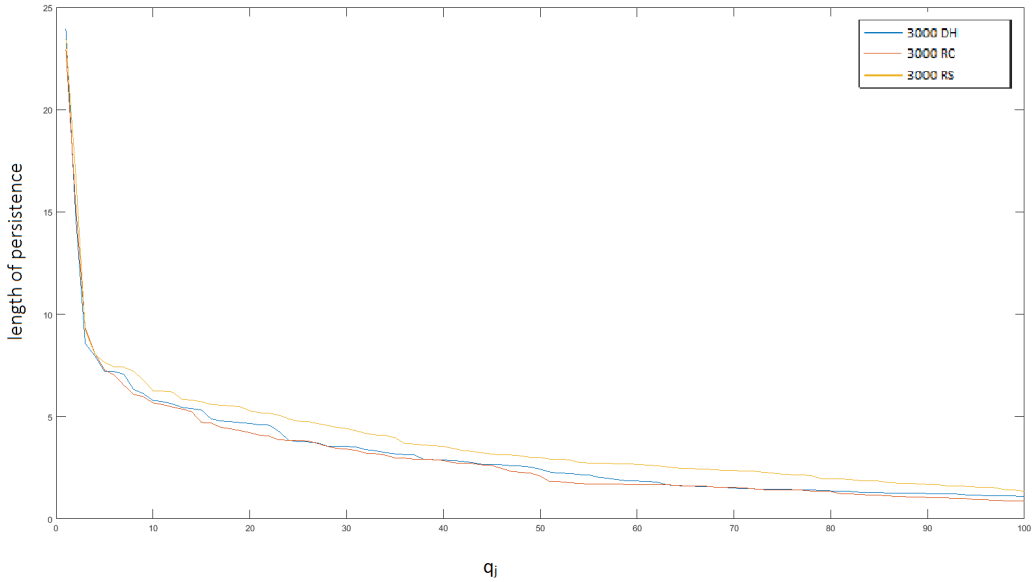

Figure 5.3.4: 3000 points seleted from RC, RS and maximin selection and the resulting q vector.

5.3.2 *Simulation Results for Brain Tree Data*

In this section, we use the BT data to illustrate the results from Section 5.2.1. One subsample with 6500 data points is generated using RS from the smallest dataset BT_{97} which has 64875 data points. We assume this subsample G_M is the replicated full sample for the BT_{97} data and conduct the rest of the simulation results. In the rest of this section, we call this G_M as the full sample.

5.3.2.1 *Summary of β_1 and β_2*

In this section, we present the summary of the β_1 and β_2 for BT_{97}. For full sample G_M, only β_1 is calculated from MATLAB package 'TDATools' while β_2 is generated from R package 'TDA' which can only perform the dataset with size $N \leq 1000$. Consequently, 1000 subsamples $F_{RC,1}, ..., F_{RC,1000}$ and $F_{RS,1}, ..., F_{RS,1000}$ are generated from G_M with $N = 1000$ using RC and RS respectively.

Firstly, we consider both persistence diagram and persistence landscape which are defined in Section 2.5.4. Noted that persistence landscape is also calculated by R package 'TDA'.

The persistence diagram and persistence landscape for β_1 for BT_{97} are given in Figure 5.3.5.

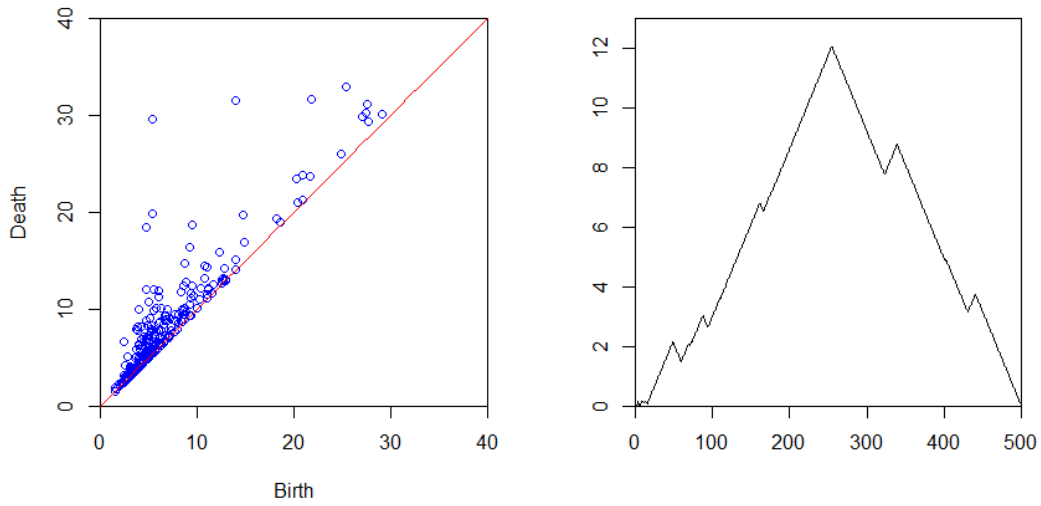

Figure 5.3.5: Persistence diagram for β_1 for G_M from BT_{97}.

In Figure 5.3.6, persistence diagrams are given for $F_{RC,1}$ (left) and $F_{RS,1}$ (right). In both graphs, the red triangles indicate β_1 and blue diamonds indicate β_2. Note that Figure 5.3.6 is plotted using R package 'TDA' as β_2 is included. In both graphs, they seem to suggest that the blue diamonds which represent β_2 are topological noise as they are very close to the diagonal line while some of the red triangles can be considered as topological signal as they are far away from the diagonal line.

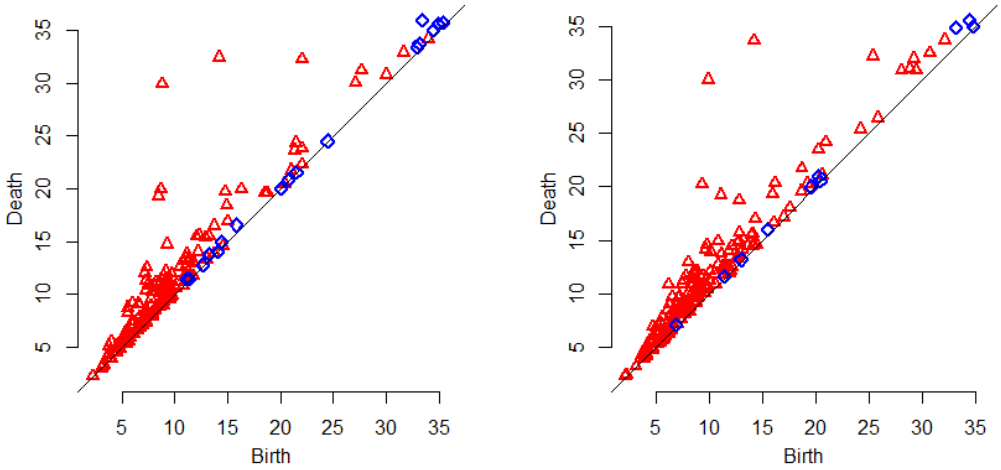

Figure 5.3.6: Persistence diagrams for β_1 and β_2 under RC (left) and RS (right).

Moreover, in Figure 5.3.7, persistence landscapes are given for RC (left) and RS (right). In both graphs, the black line indicates the λ_1 function for β_1 and the red line indicates the λ_1 function for β_2 where λ_1 is defined in Section 2.5.4. For β_2, λ_1 are very close to the x-axis, i.e. they point out again that the lifetime of the β_2 is very limited.

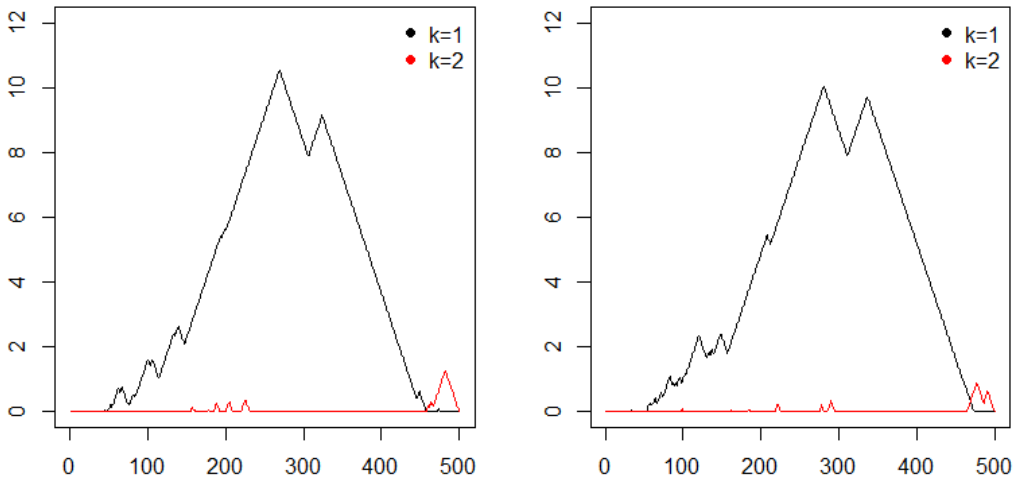

Figure 5.3.7: Persistence landscape for β_1 under RC (left) and RS (right).

Furthermore, Table 5.3.1 displays the number of β_1 for full sample, $F_{RC,1}$ and $F_{RS,1}$ which are 266, 152 and 140 in the second row. The mean and variance of 1000 subsamples of β_1 for RC and RS are given in the third and fourth row. While in the fourth row, β_2 are 17 and 10 for $F_{RC,1}$ and $F_{RS,1}$ respectively. The last two rows are mean and variance of 100 subsamples of β_2 for RC and RS.

β_k		full	RC	RS
β_1	number for 1 subsample	266	152	140
	mean of 1000 subsamples	N\A	149.13	144.25
	variance of 1000 subsamples	N\A	48.68	50.50
β_2	number for 1 subsample	N\A	17	10
	mean of 100 subsamples	N\A	14.15	13.71
	variance of 100 subsamples	N\A	9.93	8.19

Table 5.3.1: Table for BT_{97} for β_1 and β_2 for full sample and 1000 subsamples generated under both RC and RS.

In Table 5.3.1, RC has a larger mean value than RS. This may imply that RC generates more topological noise than RS. Whereas, RC has a smaller variance than RS.

As can be seen in Figure 5.3.6, Figure 5.3.7 and Table 5.3.1, $\beta_2 \leq 20$ for BT_{97} for both RC and RS. In addition, the lifetime for β_2 is very short that can be considered as topological noise. Therefore, in the rest of this chapter, we are not going to consider β_2 for BT data.

5.3.2.2 *Empirical Distribution*

Recall that the for each M, the null hypothesis is

$$H_0 : \ G_0 = F_0 \ \text{ VS } \ H_1 : \ G_0 \neq F_0$$

where the corresponding KS test statistic is

$$d_m = \sup_{-\infty<x<\infty} |G_M - F_m|$$

where $m = 1, ..., m_0$.

In Figure 5.3.8, G_M, 10 randomly selected F_m are presented for BT_{97} under RC while in Figure 5.3.9, G_M, 10 randomly selected F_m are presented for BT_{97} under RS.

It can be seen that in Figure 5.3.8, G_M first dominate the each of the subsamples then merge into the subsample while in Figure 5.3.9, there are no clear gap between G_M and F_m.

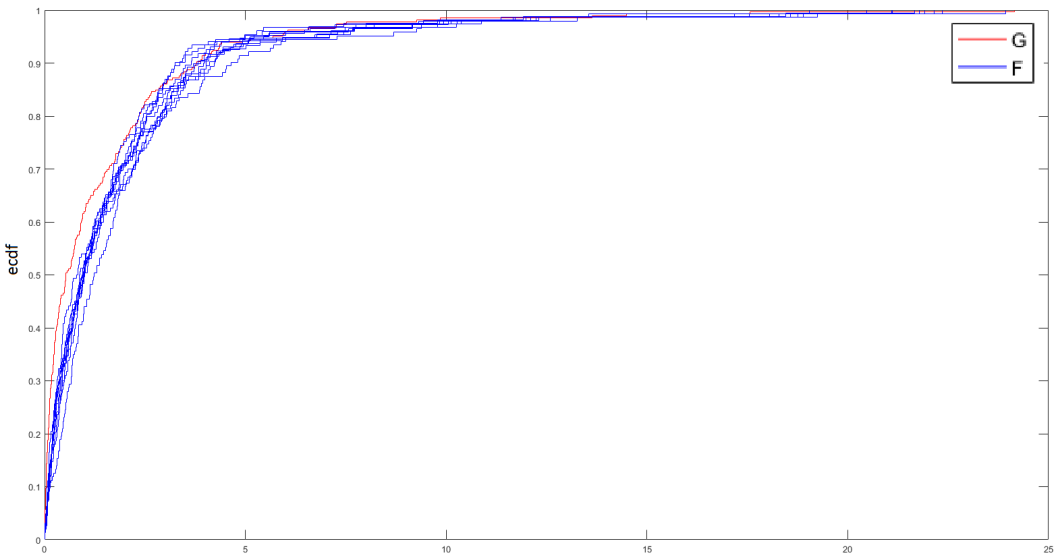

Figure 5.3.8: The empirical cdf of barcodes from full sample G_M (red) and the empirical cdf of the barcodes from 10 subsamples F_m (blue) for BT_{97} under RC method.

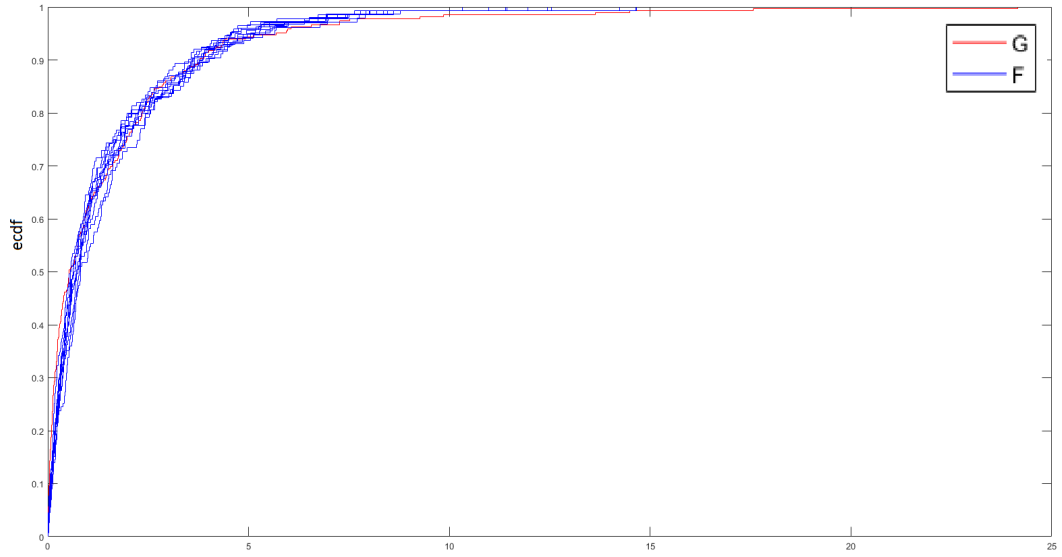

Figure 5.3.9: The empirical cdf of barcodes from full sample G_M (red) and the empirical cdf of the barcodes from 10 subsamples F_m (blue) for BT_{97} under RS methods.

Moreover, histograms and qqplots under RC (top row) and RS (bottom row) are presented in Figure 5.3.10. All seem to suggest that the F_m generated from G_M using RC and RS are normally distributed with heavy tail. However, the range of d_m from histogram suggest that d_m have a lower range for RS which is between 0.05 to 0.25 while the range for RC is 0.1 to 0.3. This seems to suggest that the empirical cdf for RS is closer to the full sample than RC which may imply that RS is a better simulation method than RC.

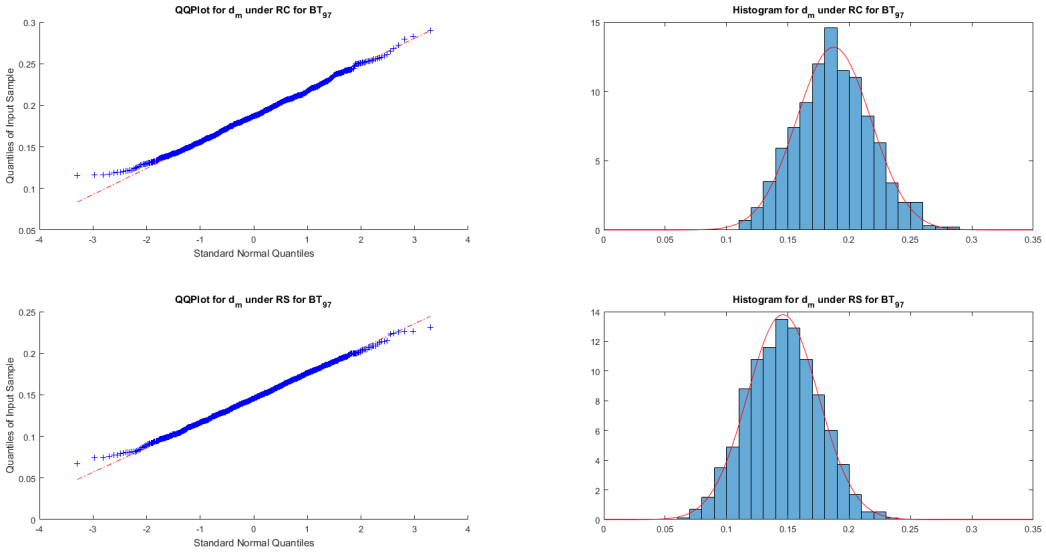

Figure 5.3.10: QQplot (top left) and histogram (top right) for 1000 d_m values for F_m under RC for BT_{97}. QQplot (bottom left) and histogram (bottom right) for 1000 d_m values for F_m under RS for BT_{97}. The $p-$values for RC for both KS and CvM test are $p_{KS} = 0.8550$ and $p_{CvM} = 0.8456$ while $p_{KS} = 0.9262$ and $p_{CvM} = 0.9512$ for RS.

We now try to replicate the other two methods in Section 5.2.1 for RC and RS which are 20 longest barcodes and 25% longest barcodes. However, this time we choose 100 longest barcodes instead of 20 while keeping 25% longest barcode unchanged. Therefore, (5.2.14) is modified as

$$\begin{aligned} G_M(x) &= \frac{1}{100}\sum_{j=1}^{100}\mathbb{I}\{q_j \le x\} \\ F_m(x) &= \frac{1}{100}\sum_{j=1}^{100}\mathbb{I}\{q_{m,j} \le x\} \end{aligned} \tag{5.3.1}$$

where q_j is the j-th longest barcode in the descending order for the full sample and $q_{m,j}$ is the j-th longest barcode in the descending order for the m-th subsample.

Summaries of the test statistics for F_m using both methods are displayed in Table 5.3.2. The top five rows are results for subsample generated under RC while the fifth to tenth rows are results under RS method. The three columns are full barcodes, i.e. all barcodes are considered, 100 longest barcodes and 25% top barcodes. The first and sixth row, third and ninth row indicate a single two-sample KS test or a single two-sample CvM test. The second and seventh row, fourth and ninth row indicate the number of sub-

samples out of 1000 that reject the null hypothesis $H_0 : \ G_0 = F_0$ which was define at (5.2.9) under two-sample KS test and two-sample CvM test respectively. The fifth and column of Table 5.3.2 is the p-values from one-sample KS test for

$$d_m = \sup_{-\infty<x<\infty} |G_M - F_m|$$

where null hypothesis $H_0 : d_m \sim N\left(\mu, s^2\right)$ as defined as (5.2.10).

As can be seen, two-sample KS tests and two-sample CvM tests suggest opposite conclusion for all three different cases for both RC and RS. Two-sample KS tests show that the subsamples and full sample are from same distribution while two-sample CvM tests displays an opposite result. However, d_m is normally distributed for both full barcodes and 100 longest barcodes for both RC and RS but d_m follows normal distribution only under RS for 25% top barcodes.

5.3.3 *Improved Results*

In this section we try to replicate and improve the result by Bendich et al. (2016) for sex effect. Apart from using mean-difference as test statistics, we also introduce the sum of length method to be the test statistics for the sex effect.

MEAN-DIFFERENCE

Recall that BT_N, $N = 1, \ldots, 98$, is the sample of brain trees. Let $\mathbf{m}_i$ be the i-th male person with vector $\mathbf{q}$ where $\mathbf{q}_i = (q_{1,i}, q_{2,i}, ..., q_{100,i})$ defined in Section 5.2.1, and similarly, define $\mathbf{f}_j$ to be the j-th female person with persistent vector $\mathbf{q}_j$. We use $T = \left\|\bar{\mathbf{m}} - \bar{\mathbf{f}}\right\|$ as the test statistics where

$$\bar{\mathbf{m}} = (\bar{q}_1, \bar{q}_2, ..., \bar{q}_{100})$$

and $\bar{q}_i = \frac{1}{100} \sum_{j=1}^{100} q_{j,i}$ is the mean vector for 100 longest persistent for male and similarly $\bar{\mathbf{f}}$ is the mean vector for female. The permutation test randomly splits the 98 $\mathbf{q}$ vectors into two groups of equal size and computes the difference of the mean values of the two groups as μ_{perm} and repeats this process for 1000 times. The empirical p-value of the permutation test is defined as

$$p = \frac{\mathbb{I}\left\{\mu_{perm} > T\right\}}{1000}.$$

		Full barcodes	100 longest barcodes (5.3.1)	25% top barcodes (5.2.15)
RC	single 2−sample KS test	1	1	1
	1000 2−sample KS test	0	0	0
	single 2−sample CvM test	1.1031×10^{-4}	0	2.6680×10^{-4}
	1000 2−sample CvM test	1000	1000	1000
	$p-$value for d_m	0.8550	0.073	4.2008×10^{-4}
RS	single 2−sample KS test	1	1	1
	1000 2−sample KS test	0	0	0
	single 2−sample CvM test	1.2463×10^{-4}	0	2.4447×10^{-4}
	1000 2−sample CvM test	1000	1000	1000
	$p-$value for d_m	0.9512	1	1

Table 5.3.2: Table for BT_{97} with $p-$value and number of rejections out of 1000 for $H_0 : \ G_0 = F_0$ under RC where G_0, F_0 are defined as Equation 5.2.2, 5.2.4.

The algorithm for repeating the mean-difference is given below as Algorithm 1.

ALGORITHM 1

1. Set $r = 1$
2. Compute the 98 q vectors where $q = (q_{1,}\, q_{2,}\, , ..., q_{100})$ is the 100 longest persistences for $\beta_1$as above
3. Let m_i be the i-th q vector for male and f_j be the j-th q vector for female
4. Calculate $\bar{m}$ and $\bar{f}$
5. Compute $T_{MD} = \left\|\bar{m} - \bar{f}\right\|$, and $p_r = \frac{\mathbb{I}\left\{\mu_{perm} > T\right\}}{1000}$
6. If $r < 100$, set $r = r + 1$, Goto step 2; else $r = 100$ stop.

Taking $\{BT_1\, , ...,\, BT_{20}\}$ as an example, resample 1000 times, and the T_{MD} statistics based on Algorithm 1, T_{MD} for both RC and RS are shown in Figure 5.3.11. It uses a red histogram to represent the data points sampled under RS method, and a blue histogram to represent the data points sampled under RC method. The vertical lines indicate the quantiles of the $T_{MD}-$values. This implies that the $T_{MD}-$values from RS is less variable than RC.

To conclude the results between RC and RS, all 98 datasets are considered, and instead of 1000 times of resampling, only 100 resamples were now taken. The resulting $T_{MD}-$values are shown in Figure 5.3.12. Compared to Figure 5.3.13, instead of towards to the right, the histogram of RS is more on the left hand side of the diagram. However, it can still be seen that the $T_{MD}-$values from RS selection is less variable than the one with RC.

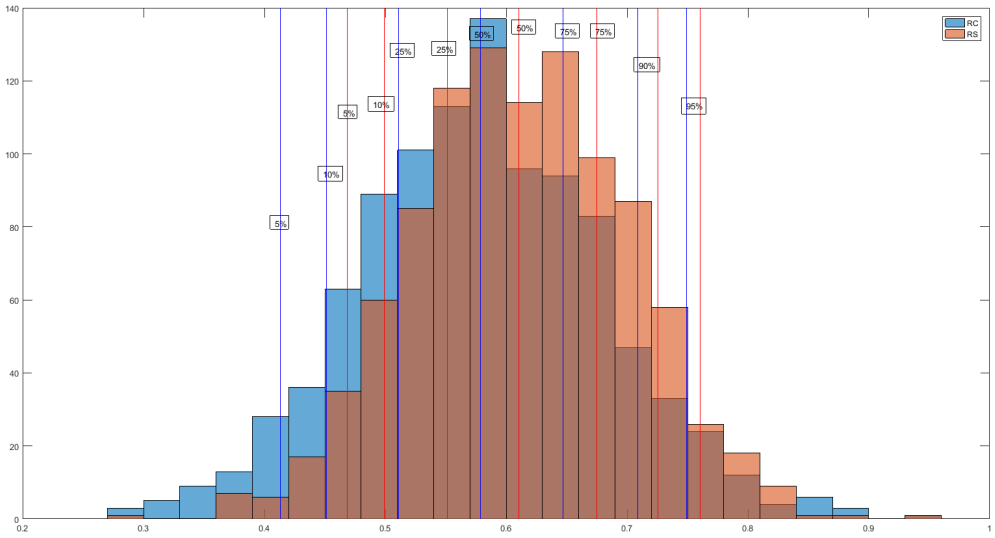

Figure 5.3.11: 1000 resampled $T_{MD}-$values for both RC and RS subsampling method. The histogram indicates that the RS is less variable than RC for the $\{BT_1, ..., BT_{20}\}$.

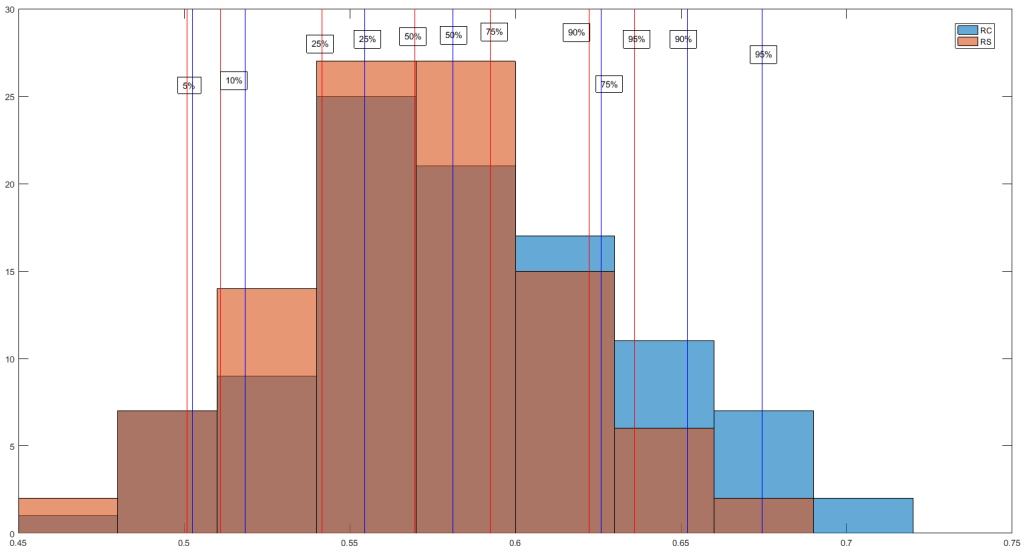

Figure 5.3.12: 100 resampled for $T_{MD}-$values for both RC and RS subsampling method. The histogram indicates that the RS is less variable than RC for the $\{BT_1, ..., BT_{98}\}$.

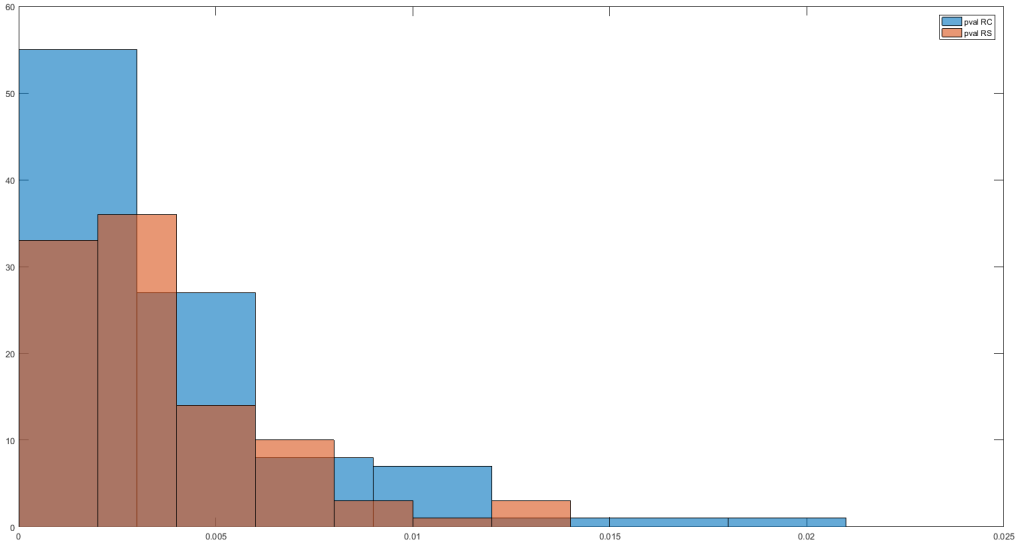

Figure 5.3.13: 100 resampled for p-values for both RC and RS subsampling method corresponding to the $T_{MD}-$values in Figure 5.3.12

The corresponding p-values for RC and RS for all 98 datasets are shown in Figure 5.3.13. The p-values are all less than 0.025 during this 100 resampling meaning the sex difference is statistically significant, on the basis of the mean difference method. This implies that we have shown a stronger significant difference result than those given by Bendich et al. (2016) which have a $p-$value as 0.031.

As the T-values show that the RS method is less variable than RC and includes the basic skeleton of the tree structure, it may be also considered that RS is a better method as it keeps the basic structure of the tree.

SUM OF LENGTHS

Instead of mean-difference statistics used by Bendich et al. (2016), we also considered the sum of lengths introduced in Section 5.2.2. We define L_{m_i} be the sum of all line segments on i-th male's barcode. Similarly, L_{f_j} is defined as the sum of all line segments on j-th female's barcode. The test statistic, $T_{SL}-$value, in this case is defined as $T_{SL} = \left\|\bar{L}_m - \bar{L}_f\right\|$ where $\bar{L}_m$ is the mean of the sum of male line segments and $\bar{L}_f$ is the mean of the sum of female line segments.

Resampling 98 datasets for 100 times using RS as the selection method and sum of lengths as the test statistic, the $p-$values shows a significant sex difference which also agrees with the results from previous method.

5.3.4 *Summary of Brain Tree Data*

In this section, we try to present the summary of β_1 for BT data under sex effect. Since Section 5.3.2 and Section 5.3.3 both suggest that RS is a better selected method than RC, we are going to present the results only for RS method.

Recall that BT_N, $N = 1, \ldots, 98$, is the sample of brain trees. As in original data, BT_1 is from female and BT_2 is from male. Therefore, these 2 datasets are used as examples in this section.

Let $\mathbf{m}_i$ be the i-th male person with vector $\boldsymbol{\beta}_1^i = \left(\beta_{1,1}^i, \beta_{1,2}^i, ..., \beta_{1,100}^i\right)$, and similarly, define $\mathbf{f}_j$ to be the j-th female person with vector $\mathbf{f}_j$. We use $T_{summary} = |\mu_{\mathbf{m}} - \mu_{\mathbf{f}}|$ as test statistics for permutation test where $\mu_{\mathbf{m}}$ is the mean value for vector $\bar{\mathbf{m}}$ where $\bar{\mathbf{m}}_i = \frac{1}{100} \sum_{j=1}^{100} \beta_{1,j}^i$ and similarly $\mu_{\mathbf{f}}$ is the mean value for female. The p-value of permutation test which defined in Section 2.2 is $p = 0.0040$ which suggest a significant sex difference for β_1 directly as well.

As can be seen, Table 5.3.3 displays the number of β_1 for female and male respectively.

	Female	Male
$\beta_{1,1}^1$	243	196
$E\left(\boldsymbol{\beta}_1^i\right)$	228.06	205.38
$Var\left(\boldsymbol{\beta}_1^i\right)$	81.99	102.32
μ	218.83	201.96

Table 5.3.3: Table for BT for $\beta_1$for female and male.

In addition, we consider both persistence diagram and persistence landscape which are defined in Section 2.5.4.

Persistence landscape for β_1 for the two sexes is given in Figure 5.3.14. The black line and the blue line indicate λ_1 for BT_1 and BT_2 respectively, whereas the red and green line indicate the mean landscape for female and male respectively. In Figure 5.3.14, it can be seen directly that β_1 has a longer lifetime for female than male.

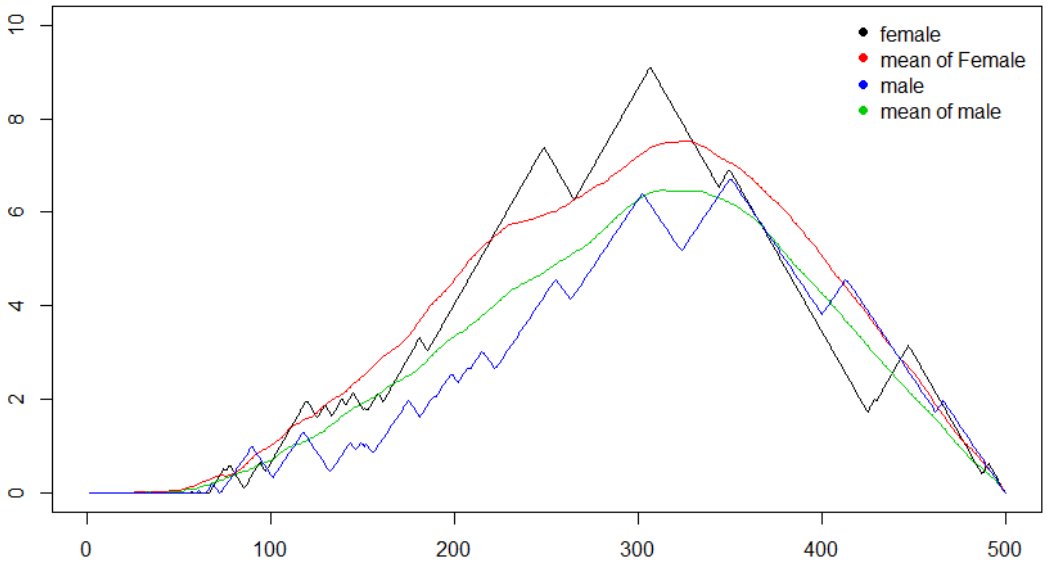

Figure 5.3.14: Persistence landscape for β_1 for BT data where black line and blue line indicate the persistence landscape for a female and a male, while red line and green line indicate the mean of the persistent landscape for female and male respectively.

Moreover, in Figure 5.3.15, persistence diagram are given for BT_1 (left) and BT_2 (right). Figure 5.3.15 may suggest that BT_2 has more topological data than BT_1 as there are more points away from the red line for BT_2 than BT_1.

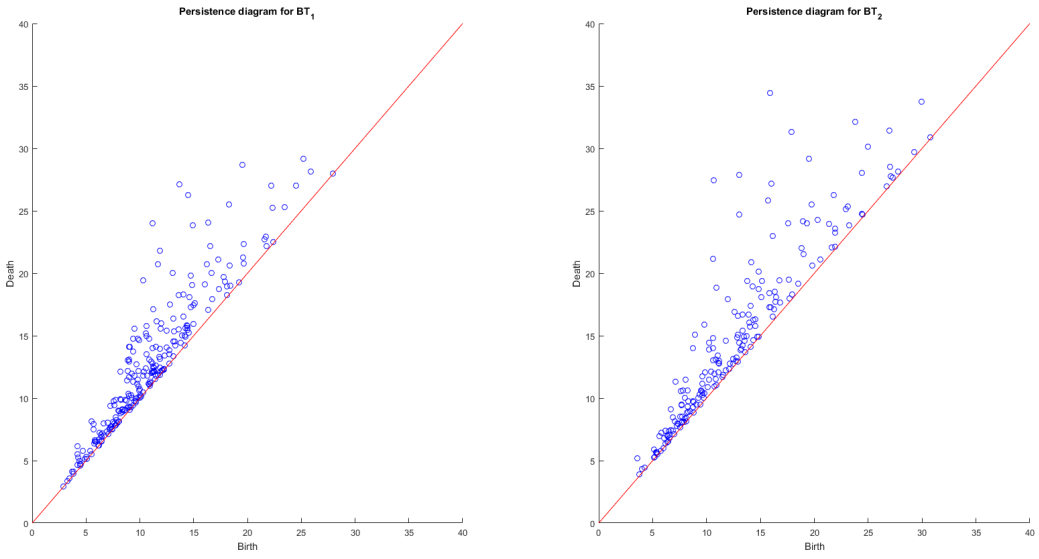

Figure 5.3.15: Persistence diagram for β_1 for BT_1(left) and BT_2 (right).

5.4 SUMMARY

When TDA is applied to point clouds in $\mathbb{R}^d$ it is often the case that it is not possible to use the full dataset due to problems with insufficient computational capacity and insufficient storage requirements. For this reason there is strong motivation for considering subsampling methods. In this chapter we have focused on two main methods of subsampling: completely random subsampling without replacement, referred to as method RC; and structured subsampling without replacement, referred to as method RS. We implemented RS by resampling a fixed skeleton of points combined with completely random resampling of other points. Our main finding in Section 5.3 is that when implemented appropriately, method RS has the potential to do much better than method RC.

6

CONCLUSION AND FUTURE WORK

In this chapter, we first summarize the main results of the thesis. Then we discuss possible directions for the future work.

6.1 SUMMARY OF THE THESIS

The new material in this thesis is contained in Chapter 3-5.

In Chapter 3, the key findings are Proposition 3.3.1 and Proposition 3.3.2. These results give the asymptotic spectral structure of the adjacency matrix under the asymptotic limit for the stochastic block model (SBM) with ζ blocks. Moreover, we have extended Proposition 3.3.1 from the adjacency matrix to the normalized graph Laplacian which is given as Proposition 3.4.1. The significance of these results is that they demonstrate that the strong form of the spectral gap theorem (SGT) will often fail to hold in the SBM. The implication of this finding is discussed further in Chapter 4.

In Chapter 4, we have first shown the difficulties of extending the SGT to SBM in Section 4.3. However, the full force of the SGT is not needed in the proof of central limit theorem (CLT) for Betti numbers in the SBM. Moreover, simulation evidence for the SGT which is included in Section 4.4, seem to suggest that λ_2, the second most dominant eigenvalue, converges to a value greater than the critical value 0.5 in the case of the second Betti number β_1 in the asymptotic regime for SBM. Therefore, by assuming that the SGT is true, we try to prove the rest of the CLT for Betti numbers in SBM. The lower and upper vanishing thresholds for Betti numbers for SBM are given as Theorem 4.5.4 and Theorem 4.6.2 respectively. Nevertheless, the lower vanishing threshold, Theorem 4.5.4 is proved under the assumption that SGT is true. Theorem 4.7.2, which is the main theorem in Chapter 4, is proved using Theorem 4.5.4 and Theorem 4.6.2. This implies that Theorem

4.7.2 is proved also under the assumption that SGT holds. However, the questions of whether the CLT holds for the SBM in broad generality, and if so how to prove this, remain open.

In Chapter 5, we have first shown the relationship between the topological summary of the full sample and subsamples in a unit square. If all barcodes are considered as the input values for the empirical cumulative distribution function (cdf), 2-sample Kolmogorov-Smirnov (KS) test and 2-sample Cramer-von Mises (CvM) tests both suggest that the full sample and subsamples are not from the same distribution. Moreover, d_m, the maximum distance between empirical cdf of the full sample and the empirical cdf of subsamples, appears empirically to be normally distributed. If only the 20 longest barcodes or only the top 25% of the barcodes are considered as topological signal, the results for the relationship between the topological summary of full sample and subsamples do not appear close. However, in these two cases, d_m no longer appears to follow a normal distribution. In Section 5.3, for the brain artery tree data, we have found out that the samples which include the basic skeleton performs better than the purely random sampling. We suspect that, in general, structured resampling of this kind has the potential to do much better than purely random resampling.

6.2 DISCUSSION AND FUTURE WORK

We now briefly discuss several directions for future research.

The asymptotic spectral structure of the adjacency matrix for the ζ-block model has been derived in Proposition 3.3.1 and Proposition 3.3.2. We have partially extended the results for the normalized graph Laplacian in Section 3.4, in that we have proved Proposition 3.4.1, which is an analogue of Proposition 3.3.1 . However, due to the introduction of the complex dependencies in the normalized graph Laplacian, we have not yet completed the proof of the analogue of Proposition 3.3.2 for the normalized graph Laplacian. This is an immediate goal of future research and we believe that the proof can be completed without fundamental difficulty though the calculations involved are quite substantial.

Preliminary study has shown that by assuming SGT is true, we are able to prove the CLT for Betti numbers in SBM. However, from the results in Chapter 3, we know that the strong form of the SGT used by Kahle and Meckes (2013, 2015) does not hold in general with the SBM. A key question

of interest is whether the failure to extend the CLT to the general SBM is only due to the breakdown of the method of proof; or is it because the CLT in fact fails to hold in generality in the SBM. This remains an open question. It would be very interesting to know if an answer can be found in future work.

We have found empirically that d_m, the maximum distance between the empirical cdf of the full sample and subsamples, is approximately normally distributed in our numerical examples. This is a rather surprising result since in general d_m should follow the distribution of a maximum-type statistic. It would be interesting to investigate this finding in future research and to see whether or not the approximately normality holds more broadly.

One more direction for future work will now be discussed. We have mentioned that the most commonly used simplicial complex in TDA programming is the Rips complex. However, De Silva and Carlsson (2004) have suggested another complex which is called the Witness complex. It would be interesting to figure out the difference and similarity between the Witness complex and the Rips complex. Moreover, a related question of interest is whether the Witness complex has the potential to be cheaper in computing time compared to the Rips complex, while still retaining important topological information. If so then it may provide a useful alternative to subsampling.

BIBLIOGRAPHY

Adler, R. (2014a). Topos, and why you should care about it. *IMS Bulletin*, 43(2).

Adler, R. (2014b). Topos: Applied topologists do it with persistence. *IMS Bulletin*, 43(6).

Adler, R. (2014c). Topos: Pinsky was wrong, euler was right. *IMS Bulletin*, 43(8).

Adler, R. (2015). Topos: Lenot make the same mistake twice. *IMS Bulletin*, 44(2).

Adler, R. J., Agami, S., and Pranav, P. (2017). Modeling and replicating statistical topology and evidence for cmb nonhomogeneity. *Proceedings of the National Academy of Sciences*, 114(45):11878–11883.

Anderson, T. W. (1962). On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, pages 1148–1159.

Ballmann, W. and Świątkowski, J. (1997). On l2-cohomology and property (t) for automorphism groups of polyhedral cell complexes. *Geometric and Functional Analysis*, 7(4):615–645.

Barbour, A. D., Karoński, M., and Ruciński, A. (1989). A central limit theorem for decomposable random variables with applications to random graphs. *Journal of Combinatorial Theory, Series B*, 47(2):125–145.

Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). Persistent homology analysis of brain artery trees. *The annals of applied statistics*, 10(1):198.

Bernstein, S. (1924). On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49.

Biscio, C. A. and Møller, J. (2019). The accumulated persistence function, a new useful functional summary statistic for topological data analysis,

with a view to brain artery trees and spatial point process applications. *Journal of Computational and Graphical Statistics*, pages 1–21.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Carlsson, G., Ishkhanov, T., De Silva, V., and Zomorodian, A. (2008). On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1):1–12.

Carstens, C. J. and Horadam, K. J. (2013). Persistent homology of collaboration networks. *Mathematical problems in engineering*, 2013.

Chernoff, H. et al. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. (2010). Lipschitz functions have l p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

De Silva, V. and Carlsson, G. E. (2004). Topological estimation using witness complexes. *SPBG*, 4:157–166.

De Silva, V. and Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358.

Dryden, I. and Mardia, K. (2016). *Statistical Shape Analysis: With Applications in R*. Wiley Series in Probability and Statistics. Wiley.

Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Soc.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen*, 6:290–297.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Erdős, P. and Rényi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1-2):261–267.

Fasy, B. T., Kim, J., Lecci, F., Maria, C., Millman, D. L., included GUDHI is authored by Clement Maria, V. R. T., by Dmitriy Morozov, D., by Ulrich Bauer, P., Kerber, M., and Reininghaus., J. (2019). *TDA: Statistical Tools for Topological Data Analysis*. R package version 1.6.9.

Feige, U. and Ofek, E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275.

Fisher, R. A. (1936). Design of experiments. *Br Med J*, 1(3923):554–554.

Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75.

Ghrist, R. and Muhammad, A. (2005). Coverage and hole-detection in sensor networks via homology. In *Proceedings of the 4th international symposium on Information processing in sensor networks*, page 34. IEEE Press.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Hoffman, C., Kahle, M., and Paquette, E. (2019). Spectral gaps of random graphs and applications. *International Mathematics Research Notices*.

Kahle, M. (2009). Topology of random clique complexes. *Discrete Mathematics*, 309(6):1658–1671.

Kahle, M. (2014). Sharp vanishing thresholds for cohomology of random flag complexes. *Annals of Mathematics*, pages 1085–1107.

Kahle, M. and Meckes, E. (2013). Limit theorems for betti numbers of random simplicial complexes. *Homology, Homotopy and Applications*, 15(1):343–374.

Kahle, M. and Meckes, E. (2015). Erratum: Limit theorems for betti numbers of random simplicial complexes. *arXiv preprint arXiv:1501.03759*.

Knuth, D. E. (1997). *The art of computer programming*, volume 3. Pearson Education.

Kolaczyk, E. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer New York.

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

Kovacev-Nikolic, V., Bubenik, P., Nikolić, D., and Heo, G. (2016). Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology*, 15(1):19–38.

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17.

Richardson, E. and Werman, M. (2014). Efficient classification using the euler characteristic. *Pattern Recognition Letters*, 49:99–106.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281.

Sutherland, W. A. (2009). *Introduction to metric and topological spaces*. Oxford University Press.

Tausz, A., Vejdemo-Johansson, M., and Adams, H. (2014). JavaPlex: A research software package for persistent (co)homology. In Hong, H. and Yap, C., editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136. Software available at `http://appliedtopology.github.io/javaplex/`.

van de Weygaert, R., Platen, E., Vegter, G., Eldering, B., and Kruithof, N. (2010). Alpha shape topology of the cosmic web. In *2010 International Symposium on Voronoi Diagrams in Science and Engineering*, pages 224–234. IEEE.

van de Weygaert, R., Pranav, P., Jones, B. J., Bos, E., Vegter, G., Edelsbrunner, H., Teillaud, M., Hellwing, W. A., Park, C., and Hidding, J. (2011). Probing dark energy with alpha shapes and betti numbers. *arXiv preprint arXiv:1110.5528*.

Van de Weygaert, R., Vegter, G., Edelsbrunner, H., Jones, B. J., Pranav, P., Park, C., Hellwing, W. A., Eldering, B., Kruithof, N., and Bos, E. (2011). Alpha, betti and the megaparsec universe: on the topology of the cosmic

web. In *Transactions on computational science XIV*, pages 60–101. Springer-Verlag.

Von Mises, R. (1928). Statistik und wahrheit. *Julius Springer*, 20.

Williams, D. (1991). *Probability with Martingales*. Cambridge mathematical textbooks. Cambridge University Press.