Normalized Laplacian for Random Graphs: Spectral Properties and Applications

Yip, Colin

Universiteit
Leiden
The Netherlands

Normalized Laplacian for Random Graphs: Spectral Properties and Applications

Colin Shing Yum Yip

Thesis Advisors: Dr. R.S. Hazra
Dr. O. Kanavetas

Defended on June 30, 2025

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

Graphs are powerful mathematical objects for modelling the interaction and evolution of complex physical systems and networks. While these objects are comprised of a set of vertices with a corresponding set of connecting edges, it is more mathematically convenient to represent them in matrix form, for example with an adjacency matrix with binary entries indicating the presence of an edge connecting a given pair of vertices. These structures can be further encoded with assumptions of randomness, where the probability of connection between any two vertices in the graph is a random variable, giving rise to random graphs, which themselves can be further delineated between homogeneous and inhomogeneous random graphs. Homogeneous random graphs have a uniform probability of connection across all edges, and are classically known as Erdös-Rényi Random Graphs. In contrast, inhomogeneous random graphs have edge specific probabilities of connection, which can be dependent on the two vertices a given edge connects. This work focuses on the latter delineation of random graphs, specifically looking at the asymptotic behaviour of the distribution of eigenvalues, which we refer to as the empirical spectral distribution, of the normalized Laplacian matrix of these graphs. The normalized Laplacian matrix can intuitively be viewed as a matrix describing how well-connected a given node is relative to the rest of the graph, with its eigenvalues then being akin to the stationary distribution of a random walk on the graph.

This feature of the empirical spectral distribution of the normalized Laplacian matrix makes it a valuable mathematical object to use in a variety of statistical and machine learning applications. One of those applications is in centrality measures that describe how well-connected a vertex or set of vertices is to the rest of the graph, such as PageRank, which was originally designed to measure the relative importance of a webpage as a function of how connected that webpage is to others [34]. The empirical spectral distribution is also valuable in graph-based learning problems, such as in graphical neural networks, which rely on the vector of eigenvalues as a key input in learning the underlying structure or node features of a graph [30]. While there are many other uses for the spectral components of the normalized Laplacian matrix, this work will focus on the two aforementioned applications.

This thesis proves the convergence of the empirical spectral distribution of the normalized Laplacian matrix of an inhomogeneous random graph to a deterministic limiting measure, which was previously defined by [6], as the number of the vertices in the graph tends to infinity. In the process of completing this proof, several inequalities governing the behaviour of the expected connectedness of the vertices of the graph are also proven. These inequalities can then be extended to approximation of PageRank, as in a real world setting, the true PageRank vector is unknown, and so must instead be estimated. We evaluate an approximation based upon the connectedness of a given vertex relative to the total number of edges in a graph and prove that this methodology is in fact asymptotically close to the true PageRank vector. Finally, we provide a novel mathematically motivated eigenvalue correction strategy for graphical neural networks that aims to improve model performance by diversifying the range of potential values produced. When tested on well-connected real world datasets, this method yields significantly better model accuracy against existing correction strategies and models that do not perform any correction.

While these results are generalizable for a number of commonly used inhomogeneous network models, such as the Chung-Lu model, we operate under the assumption of a well-connected graph where an edge path can be drawn between any two vertices, which we refer to as "dense". As such, these results are not applicable for "sparse" graphs which potentially have isolated communities. Proving these asymptotic results for more widely generalizable classes of inhomogeneous graphs, such as kernel-based random graphs, remains an open question.

Contents

Chapter 1
Introduction

This thesis analyzes the spectral behavior of the normalized Laplacian matrix of both homogeneous and inhomogeneous random graphs. Beginning with a theoretical perspective, the methods and results derived from the first two chapters will then be applied to a centrality measure, PageRank, and then to graphical neural networks in the final two chapters.

1.1 Graphs and Summary Matrices

1.1.1 Graphs as Mathematical Objects

Real world systems are oftentimes comprised of individual elements interacting with each other, with examples of these systems including gas particles in a vacuum, users in a social network [1], and global trade between countries [21]. Graph structures, also referred to as networks, are useful tools in the mathematical representation of these systems, and are widely used in statistical physics, biology, and computer science. In this section, we will explore some fundamentals of the analysis of graphs.

A graph is composed of two sets of objects: vertices and edges. Vertices are the individual elements within a graph, also known as nodes, and edges denote the connection between these vertices. Graphs are highly flexible structures, and so there can exist graphs with self-loops, where a vertex is connected to itself. These representations can also encode directions, leading to the distinction between directed and undirected graphs. Whereas directed graphs have edges with an orientation from one node to another, undirected graphs have no such information encoded in their edge set. In this work, we will focus on the latter category of graphs. We visualize a simple example of a graph below in Figure 1.

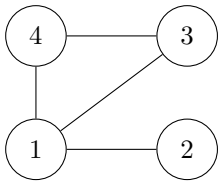


Figure 1: An undirected graph with 4 vertices.

This can be represented mathematically as $G \coloneqq (V, E)$, where G is the graph, V is the set of vertices, and E is the set of edges. The graph in Figure 1 would have $V = \{1, 2, 3, 4\}$ and $E = \{(1, 2), (1, 3), (1, 4), (3, 4)\}$. An adjacency matrix summarizes the connection between the vertices in a graph, with an entry in the matrix being binary to represent the existence of an edge,

$$\mathbf{A}(i, j) \coloneqq \begin{cases} 1, & (i, j) \in E \\ 0, & (i, j) \notin E. \end{cases}$$

Note that for undirected graphs, this produces a symmetric matrix, where $\mathbf{A}(i, j) = \mathbf{A}(j, i)$, and a self-loop connecting $\mathbf{A}(i, i)$ to itself would yield a 1 on the diagonal. The adjacency matrix of the graph in Figure 1 would be,

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

To represent how well-connected an individual vertex is, we can calculate the degree of that vertex, which is the number of other vertices that are connected to it. Finding the degree of all nodes in a graph

with N vertices produces a vector of length N, which we denote as $\mathbf{d}$, where each entry is calculated as follows, assuming there are no self-loops,

$$\mathbf{d}_i = \sum_{j \neq i}^{N} \mathbf{A}(i,j),$$

and so the degree matrix is then defined as $\mathbf{D} = \text{diag}(\mathbf{d}_1, \ldots, \mathbf{d}_N)$, with 0 in all off-diagonal entries. Figure 1 has a degree vector, $\mathbf{d} = (3,1,2,2)$. The Laplacian matrix is defined as $\tilde{\mathbf{L}} \coloneqq \mathbf{D} - \mathbf{A}$, and can be interpreted as the matrix form of the discrete Laplacian operator [31]. While we do not discuss the Laplacian matrix in this work, we refer readers to [25] for more discussion on the behavior of these matrices. This thesis will instead focus on the normalized Laplacian matrix, which is defined,

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \tag{1.1}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{D}^{-1/2}(i,i) = \frac{1}{\sqrt{\mathbf{D}(i,i)}}$ if $\mathbf{d}_i$ is non-zero, and 0 otherwise. Figure 1 produces a normalized Laplacian matrix as follows,

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

For a graph with no self-loops, $\text{diag}(\mathbf{L}) = 1$, and off-diagonal entries are the proportionate connectivity of an individual vertex relative to the rest of the graph.

1.1.2 Spectral Analysis of Graphs

The spectral quantities of a graph G are the eigenvalues and eigenvectors of one of the graph matrices, such as its adjacency, Laplacian, or normalized Laplacian matrix. These quantities prove to be easier objects to work with in describing the behavior of graphs, particularly as the number of vertices grows to infinity. We denote the ordered eigenvalues of a matrix of size N,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N,$$

with corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$. A useful result relying on these quantities is the Perron-Frobenius theorem, which states that if G is a connected graph where all nodes are connected to one another, then there is a clear leading eigenvalue of magnitude greater than the trailing eigenvalue, that is, $\lambda_1 > \lambda_2$ and $\lambda_1 \geq -\lambda_N$, and this leading eigenvalue has a strictly positive eigenvector [44].

Another useful result from spectral analysis is spectral decomposition, which allows a Hermitian matrix, also referred to as a self-adjoint matrix, to be expressed as,

$$\mathbf{A} = U\Lambda U^T,$$

where U is a unitary matrix with the eigenvectors as the columns, and $\Lambda \coloneqq \text{diag}(\lambda_1, \ldots, \lambda_N)$. In order for a matrix to be Hermitian, it must satisfy that it is equivalent to its conjugate transpose,

$$\mathbf{A} = \mathbf{A}^T,$$

where $\mathbf{A}^T$ is the conjugate transpose. As graph matrices have real entries as byproducts of the adjacency matrix, this is then equivalent to requiring a symmetric matrix. $\mathbf{A}$ is symmetric, and as the identity and degree matrices are both diagonal, both the Laplacian and normalized Laplacian matrices are also symmetric, thus allowing for spectral decomposition of $\tilde{\mathbf{L}}$ and $\mathbf{L}$. As its entries are normalized with respect to the adjacency matrix, the normalized Laplacian and its spectra are often used to communicate the probability of connection between vertices, such as for random walks or centrality measures [2], the latter of which will be discussed later.

1.2 Random Graphs

1.2.1 The Erdös-Rényi Random Graph Model

Graphs with fixed edges are useful mathematical tools, but real world physical systems follow probabilistic behavior. For example, whether an illness spreads between individuals is a probabilistic event, and similarly

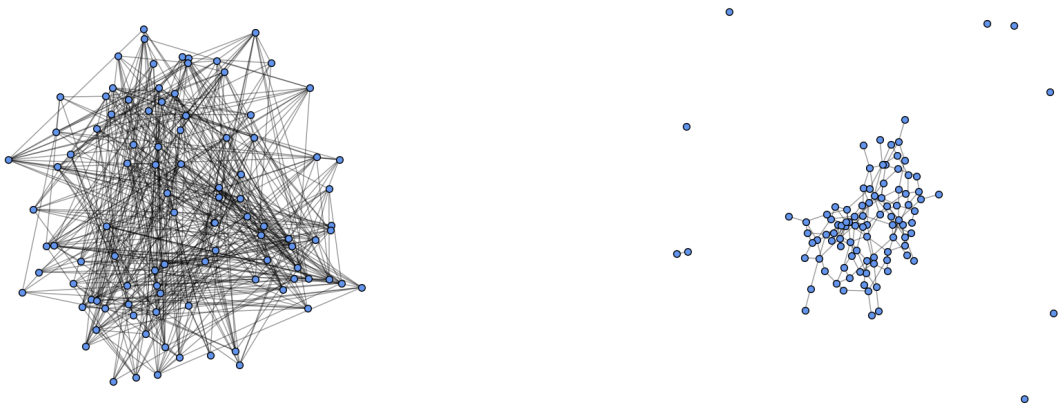
the likelihood of two arbitrary users on a social media network to connect is a random event when viewed at scale. The study of these random graphs thus extends the usefulness of graphical models in representing complex settings.

The Erdös-Rényi random graph (ERRG) as described in [17] is the simplest random graph model. In this thesis we will address the ERRG model of the form $G_N(N, p_N)$, where the produced graph is parameterized by N vertices with a connection probability between any two of these vertices given as a Bernoulli random variable with probability p_N, Ber(p_N). Moving forward, we will include the subscript N in the notation of G and p to note that the behavior of both the graph and its connection probability are contingent on the value of N. This relationship between N and p_N in controlling the overall connectivity of G_N provides two overarching regimes for ERRG models:

1. *Dense*: A graph where $Np_N \to \infty$ as $p_N \to 0$ or p_N is held as a constant is considered dense, as the average degree of the graph grows proportionately with N, the size of the graph [45].

2. *Sparse*: Conversely, if a graph has that $p_N \to 0$ but $Np_N \to \lambda \in (0, \infty)$, the graph is within the sparse regime, as the average degree does not grow proportionately with N, thus leading to isolated individual or neighborhoods of nodes [45].

Additionally, we have that in order for an ERRG to be well-connected, it must satisfy that $Np_N > \log(N)$. Figure 2 illustrates this difference using simulated graphs. This is an assumption that we will rely on throughout the thesis.



(a) Dense regime ERRG with $p_N = 0.10, Np_N = 10$.

(b) Sparse regime ERRG with $p_N = 0.03, Np_N = 3$.

Figure 2: Visualizations of a dense and sparse ERRG, both simulated with $N = 100, log(N) \approx 4.61$. The dense regime has one large connected component while the sparse regime has several isolated vertices.

The ERRG model can be further extended by removing the assumption of a homogeneous or uniform p_N across all edges, allowing the model to represent more real world systems that are often highly non-uniform. An inhomogeneous ERRG instead has a probability dependent on the two vertices it connects, defined as $p_{ij} = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)$, where $f : [0,1]^2 \to [0, \infty)$ is a bounded, Riemann integral function, that is, continuous almost everywhere, and ε_N is a scaling constant dependent on N ensuring that p_{ij} is bounded as a probability [12]. As such, the behavior of the graph becomes dependent on that of f. This family of graphs includes well known models such as Stochastic Block models [27], and Chung-Lu models [11].

1.2.2 Random Matrix Theory

A relevant field of study in mathematics is that of random matrix theory, which seeks to analyze the behavior of matrices with entries each being a random variable. It is evident that graph matrices, such as the adjacency matrix, of random graphs must themselves be random. As such, we will use tools from random matrices to aid in the analysis of random graphs, specifically in the spectral analysis of these objects. These include results and methods for studying the distribution of eigenvalues of a matrix and the behavior of the largest eigenvalue. The empirical spectral distribution (ESD) will be used heavily to describe the observed eigenvalue behavior of a random graph, and is the primary object of analysis for

existing results from random matrix theory that seek to determine the limiting behavior of the ESD as the graph size grows to ∞, $N \to \infty$. Defining this formally, let $\delta_x(\cdot)$ being the Dirac delta function with a mass of 1 at a value x, and λ_i being the i-th eigenvalue of a matrix $\mathbf{M}_N$ of dimension $N \times N$, we define its ESD as follows,

$$\mathrm{ESD}\left(\mathbf{M}_N\right)(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i(\mathbf{M}_N)}(\cdot). \tag{1.2}$$

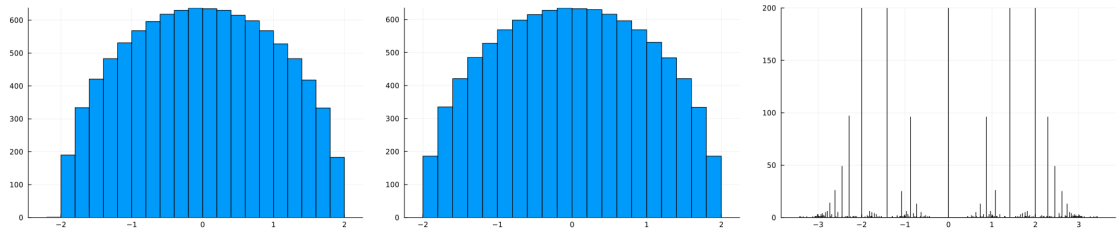
Of course, if $\mathbf{M}_N$ is a random matrix, then $\mathrm{ESD}\left(\mathbf{M}_N\right)(\cdot)$ is a random measure. Additionally, let d be a metric on a space $\mathcal{M}_1(\mathbb{R})$. Convergence of $\mathrm{ESD}\left(\mathbf{M}_N\right)$ to a limiting measure, μ, weakly in probability can be defined then as follows, for $\varepsilon > 0$,

$$\mathbb{P}\left(d(\mathrm{ESD}\left(\mathbf{M}_N\right), \mu) \geq \varepsilon\right) \to 0, \quad N \to \infty. \tag{1.3}$$

This convergence to a limiting measure will be stated as $\lim_{N\to\infty} \mathrm{ESD}\left(\mathbf{M}_N\right) = \mu$, weakly in probability. Figure 3 illustrates how the ESD of the adjacency and normalized Laplacian matrices behave with respect to the dense and sparse regimes. In the dense regime, the ESD of a homogeneous $\mathbf{A}_N$, as outlined in [43], and an appropriately centered homogeneous $\mathbf{L}_N$, as proven in [29], converge weakly almost surely to the semicircle law, μ_{sc}, the density of which is:

$$\mu_{sc}(\mathrm{d}\, x) = \frac{1}{2\pi}\sqrt{4 - x^2}\mathbb{1}_{|x|\leq 2}\,\mathrm{d}\, x. \tag{1.4}$$

While the semicircle law is a universal limiting measure in the homogeneous regime for ERRG models, it is itself a special case of a more general limiting measure in the inhomogeneous regime dependent on the behavior of f in $p_{ij} \coloneqq \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)$, as μ_{sc} can be recovered by fixing $f\left(\frac{i}{N}, \frac{j}{N}\right) = 1$ uniformly across all edges [12]. This then motivates a more general definition of a limiting measure encompassing both homogeneous and inhomogeneous ERRG models, which as discussed in [6] and [12], is a deterministic, compactly supported, and symmetric measure, μ_f. The convergence of both $\mathrm{ESD}\left(\mathbf{A}_N\right)$ and a centered $\mathrm{ESD}(\tilde{\mathbf{L}}_N)$ to a deterministic limit, μ_f, weakly in probability, was proven in [7]. In this thesis, we will consider an extension of this result for $\mathbf{L}_N$.



(a) $\mathrm{ESD}\left(\mathbf{A}_N\right)$, $p \approx 0.06$ (Dense). (b) $\mathrm{ESD}\left(\mathbf{L}_N\right)$, $p \approx 0.06$ (Dense). (c) $\mathrm{ESD}\left(\mathbf{A}_N\right)$, $p = 5e^{-5}$ (Sparse).

Figure 3: Comparison of the empirical spectral distributions of $\mathbf{A}_N$ in both dense and sparse regimes and $\mathbf{L}_N$ in the dense regime, $N = 10,000$.

1.3 PageRank and Random Graphs

1.3.1 Centrality Measures and the Normalized Laplacian

A key statistic in network analysis for computer science and statistics is how well-connected an individual vertex is relative to the remainder of the graph beyond first order connections, that is, how central a vertex is to a community of other vertices, up to and including the remainder of the graph. Understanding the centrality measure of a graph is critical in reputation systems, machine learning, and graph partitioning [1]. This gives rise to the category of centrality measures which are solutions to the eigenvalue problem,

$$\mathbf{x}\mathbf{M} = \lambda\mathbf{x},$$

for some input row vector $\mathbf{x}$ and a matrix $\mathbf{M}$ [1]. In the context of random graphs, $\mathbf{M}$ is typically taken to be the adjacency matrix or a matrix derived from it, such as the normalized Laplacian matrix.

One such use case for these centrality measures is the in the analysis of internet traffic between web pages, which motivates the measure PageRank [34], which represents the probability of a walk from a starting vertex to a connected vertex with probability α, and the chance to restart this walk from a randomly selected node in the graph with probability $1-\alpha$ [1]. The Perron-Frobenius theorem surrounding

the magnitude of the leading eigenvalue in a connected graph is another such centrality measure, but in this work we will focus on the analysis of PageRank. A PageRank vector $\boldsymbol{\pi}$, can thus be interpreted as the stationary distribution of a random walk under these conditions, and furthermore is the solution to,

$$\boldsymbol{\pi} = \alpha\boldsymbol{\pi}\mathbf{P}_N + (1-\alpha)\mathbf{v},$$

where $\mathbf{P}_N$ is the unsymmetrized normalized Laplacian, $\mathbf{P}_N \coloneqq \mathbf{D}_N^{-1}\mathbf{A}_N$ and $\mathbf{v}$ is a vector of uniformly distributed probabilities corresponding to all N vertices in G_N. Note that for $\alpha = 1$, $\boldsymbol{\pi}$ is the stationary measure of the Markov chain defined by $\mathbf{P}_N$. This definition also yields a Personalized PageRank statistic, as $\mathbf{v}$ can be restricted to a subset of V_N, which allows for measures of centrality relative to a specific subgraph. In this context, $\mathbf{v}$ is then referred to as the personalization distribution [1].

It is evident then, that PageRank relies on the information surrounding vertex connectivity encoded in the normalized Laplacian to inform its random walk process. As such, understanding the behavior of $\mathbf{L}_N$ for the underlying graph is critical in understanding the subsequent behavior of PageRank and its approximations. As it is not possible to have an explicit form for $\boldsymbol{\pi}$ in real-world settings, approximations have been proposed, such as using the degree vector normalized by the volume of the graph, $\text{vol}(G_N) = \sum_{i=1}^{N} \mathbf{d}_i$, in its place [2],

$$\overline{\boldsymbol{\pi}} = \frac{\alpha\mathbf{d}}{\text{vol}(G_N)} + (1-\alpha)\mathbf{v}. \tag{1.5}$$

[2] proves that substituting the PageRank vector with the simple random walk, $\mathbf{d}/\text{ vol}(G_N)$, provides an approximation where the distance in total variation between the two vectors is of the order of magnitude $o(1)$ as $N \to \infty$. This was demonstrated in their paper to hold for both the Chung-Lu and Stochastic Block Model settings by showing that these families of graphs adhere to certain conditions that allow for this closeness in distance to hold. As such, the behavior of PageRank in inhomogeneous regimes has been studied empirically in a number of studies [38], and under some specific regimes [2], but analysis of PageRank in a generalized inhomogeneous setting has not been done. In this work, we will focus on this more generalized setting with respect to the conditions and volume based approximation laid out in [2].

1.4 Neural Networks and the Normalized Laplacian

1.4.1 Graph Neural Networks

Graph neural networks are a class of convolutional neural networks that approach learning problems on graph-like structures, such as citation networks [48]. These datasets are thus composed of vertex and edge sets that can be represented in standard graph form, $G = (V, E)$ with V being the set of vertices and E being the set of edges. In this context, a graph of size N can be represented by its binary adjacency matrix, $\mathbf{A} \in \mathbb{R}^{N\times N}$, and its matrix of node features, $\mathbf{X} \in \mathbb{R}^{N\times d}$, for d being the length of the node features. As such, we can define the normalized Laplacian matrix of this graph-like structure as, $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. We can thus define a neural network accepting a graph as an input as a graph neural network, $f(\mathbf{X}, \mathbf{A})$. This work will focus specifically on spectral graph neural networks, which map node features, $\mathbf{X}$, to the Fourier domain to perform filtering or attenuation before transforming the output back to the spatial domain.

To begin, we provide a mathematical definition for graph neural networks (GNNs) and how messages are passed between the layers of these structures. Intuitively, GNNs produce outputs at the individual node level by aggregating information from neighboring nodes, with the number of layers in a neural network dictating how many neighbors away from the vertex to select from (i.e. how many "hops" away from which to start collecting information). Let $\mathbf{x}_n$ denote the output vector of the n-th node in G, $1 \le n \le N$ and ne$[n]$ be the neighborhood around node n. We also define the behavior of the other nodes in ne$[n]$ as $\mathbf{x}_{\text{ne}[n]}$ being the corresponding node features and $\mathbf{e}_{\text{ne}[n]}$ being the edge states surrounding n. We introduce a differentiable, permutation invariant, aggregating function over ne$[n]$, $\bigoplus$, such as a sum or mean, and differentiable functions ϕ and ψ over corresponding appropriate domains. Thus, the message passing function at each layer, $\mathbf{x}_n$, can be defined as follows, with the 0-th layer typically being the original input vector $\mathbf{x}$ [30]:

$$\mathbf{x}_n^{(k)} = \phi^{(k)}\left(x_n^{(k-1)}, \bigoplus \psi^{(k)}\left(\mathbf{x}_n^{(k-1)}, \mathbf{x}_{\text{ne}[n]}^{(k-1)}, \mathbf{e}_{\text{ne}[n]}\right)\right), \tag{1.6}$$

where k is the specific layer number, $1 \le k \le K$ for K being the total number of layers in a network. Given their respective purposes, ϕ is often referred to as the update function, as in an update of $\mathbf{x}_n$ between layer $k-1$ and k, and ψ is referred to as the message function, passing information from the other nodes in ne$[n]$ to the node of interest, n [22]. In this way, by summing across all K layers, a final output can

be created corresponding to all N nodes. We note here that this final summation is also potentially a learnable process as we will discuss later.

1.4.2 Graph Convolutional Networks and Signal Filters

We provide a more specific definition for graph convolutional networks (GCNs) that rely on convolving input signals with filters to produce the relevant output vector, effectively using convolutions to perform both the update and message functions. A convolution is defined for two integrable functions, $f, g : \mathbb{R} \to \mathbb{R}$, as follows:

$$(f \star g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau. \tag{1.7}$$

This is the product of the two functions, including a reflection about an axis and a shift by a factor of τ. However, as GCNs, and neural networks more broadly, take matrices as inputs, we instead use the discrete definition of the above to represent summations across the grid of these matrices [22]. Thus, for the matrices $f, g : \mathbb{Z} \to \mathbb{R}$, we instead sum across all $\mathbb{Z}$ to produce a single dimensional convolution:

$$(f \star g)(t) = \sum_{m \in \mathbb{Z}} f(m)g(t-m). \tag{1.8}$$

Extending this definition to two dimensions, as is done in practice for matrices in GCNs, we arrive at the following definition for convolution of a matrix by a filter of equal or smaller dimensions, where the domain of the matrix is $[H] \times [W]$, and the filter is of the size $H^{\dim(f)} \times W^{\dim(f)}$. For notational consistency, we replace the function g with the filter ω [22]. Thus, the convolution of f by the filter ω can be expressed as follows over the domain of ω:

$$(f \star \omega)(x, y) = \sum_{v=1}^{H^{\dim(f)}} \sum_{w=1}^{W^{\dim(f)}} \omega(v, w)f(x-v, y-w). \tag{1.9}$$

Having now established a mathematical definition for a convolution, we can integrate this into our earlier definition of GNNs from (1.6). Let $\phi^{(k)}$ be a general activation function, such as $\text{ReLU}(\cdot) = \max(0, \cdot)$, and $\bigoplus \psi^{(k)}(\cdot)$ be $\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}(\cdot)^{(k-1)}\mathbf{W}^{(k-1)}$, for $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ being the adjacency matrix of the input graph with self-connections imposed, and $\tilde{\mathbf{D}}$ being the degree matrix of $\tilde{\mathbf{A}}$. Additionally, $\mathbf{W}^{(k-1)}$ is a learnable weight matrix for a given k-th layer, where its entries are trainable parameters that are adjusted during the GCN's training process such that a loss function is minimized. We also note that the filter function, ω, is also learnable for specific implementations of GCNs, in that they also rely on a learnable parameter or vector of parameters which control how aggregation across or within layers is handled. These are also adjusted during the GCN training process, and we will look at some of these implementations in more detail later on. Thus, by using summary information given by the graph's normalized uncentered Laplacian matrix, we can encode the information required for the behavior of a neighborhood around a node n, $\mathbf{x}_{\text{ne}[n]}^{(k-1)}$ and $\mathbf{e}_{\text{ne}[n]}$. Thus, a GCN's propagation function for each layer can be expressed as:

$$\mathbf{H}^{(k)} = \sigma\left(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(l-1)}\mathbf{W}^{(l-1)}\right), \tag{1.10}$$

where $\mathbf{H}^{(0)} = X$, the initial matrix of node feature vectors [30]. The propagation function uses a form of the normalized Laplacian as the normalization process is helpful in computation given that it is a real symmetric positive definite matrix, thus having N many orthonormal eigenvectors and real non-negative eigenvalues [13]. As such, this presents a useful starting point for spectral functions which operate on the spectral components of $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = U\Lambda U^T$, where $U = [u_0, \ldots, u_{N-1}] \in \mathbb{R}^{N \times N}$ is the matrix of eigenvectors and $\Lambda = \text{diag}(\lambda_0, \ldots, \lambda_{N-1})$ is a vector of the corresponding ordered eigenvalues. Transformation to a domain other than that of the original graph is necessary for computing filters, as transformations such as translations and reflections, as outlined in (1.7), are not meaningful in the space of a graph. Thus, signals $\mathbf{x} \in \mathbb{R}^N$ are transformed to the Fourier domain $\hat{\mathbf{x}} = U^T\mathbf{x}$ where the transformations required for filtering can be done, after which the signals are transformed back to the graph domain, via $\mathbf{x} = U\hat{\mathbf{x}}$ [42]. Let $h_\theta(\Lambda)$ be a function of the eigenvalues of $\mathbf{L}$ parameterized by a vector of possibly learnable weights, θ, of size K, and $\phi^{(k)}$ be defined as in (1.6). The full process of applying a filter to a graph signal $\mathbf{x}$, can be expressed as the following convolution,

$$h_\theta(\Lambda) \star \mathbf{x} = Uh_\theta(\Lambda)U^T\mathbf{x} = U\sum_{k=0}^{K} \theta_k\phi^{(k)}(\Lambda)U^T\mathbf{x}. \tag{1.11}$$

However, the direct calculation of this expression is expensive, with the multiplication of U by the filter h and the input signal $\mathbf{x}$ being $\mathcal{O}(N^2)$ in terms of computation time. Thus, there are many alternative approximations of this computation, such as Chebyshev polynomials, that significantly reduce the computational load of applying these filters [23]. Specifically, they seek to approximate the application of the filter on the vector of eigenvalues. We will evaluate the performance of some of these filters in a later part of this work. As stated by Theorem 4.1 of [46], the performance of a GCN is fundamentally limited by the number of unique eigenvalues in Λ, as filter functions will produce the same output for equivalent eigenvalues. Thus, in settings where the spectra of the graph has high multiplicities, as is seen in real world datasets [33], strategies that reduce these multiplicities can improve GCN performance. Existing methods to perform these reductions are limited, with a uniform correction strategy done across the interval $[0, 2]$ as proposed by [33] showing notable performance improvements when using polynomial filters on dense networks. However, this method assumes convergence of the spectra of the normalized Laplacian to the semicircular law [29, 10], which holds theoretically only for the homogeneous setting [7]. Thus, in this work we will provide an empirical degree driven correction methodology that is more broadly applicable to both homogeneous and inhomogeneous regimes. The code for the computational results and figures seen in this work can be found at the following GitHub repository: https://github.com/colinsyyip/Normalized-Laplacian-ESD.

Chapter 2
The Normalized Laplacian in the Erdös-Rényi Regime

In this section, we will outline and prove the convergence of the empirical spectral distribution (ESD) of the normalized Laplacian matrix to the semicircular law in a homogeneous setting, where the probability of connection between any two nodes in a graph is governed by the same probability.

2.1 Notation

Let $G_N = (V, E)$ be an Erdös-Rényi Random Graph (ERRG) with $V = [N] = \{1, 2, \ldots, N\}$, and E being the set of edges, which are given by any two vertices being connected with a probability p_N, which is independent of other vertices. We represent its adjacency matrix, $\mathbf{A}_N$ as an $N \times N$ matrix with its entries, $\{\mathbf{A}_N(i,j); 1 \leq i < j \leq N, N \geq 2\}$, being independent random variables such that $\mathbb{P}(\mathbf{A}_N(i,j) = 1) = p_N = 1 - \mathbb{P}(\mathbf{A}_N(i,j) = 0)$ for $p \in (0,1)$. This is a symmetric matrix, as having node i connected to node j also implies, so we have that $\mathbf{A}_N(i,j) = \mathbf{A}_N(j,i)$, and we additionally impose that $\mathbf{A}_N(i,i) = 0$. Setting the diagonal of $\mathbf{A}_N$ to be 0 imposes no self-loops on G_N. The degree matrix of $\mathbf{A}_N$ is defined as $\mathbf{D}_N \coloneqq \operatorname{diag}(\mathbf{d}_1, \ldots, \mathbf{d}_N)$, where each $\mathbf{d}_i = \sum_{j \neq i}^{N} \mathbf{A}_N(i,j)$ is the degree of vector i. Let $\mathbf{I}_N$ be the identity matrix of size $N \times N$, and from here we can define the normalized Laplacian matrix as follows:

$$\mathbf{L}_N = \mathbf{I}_N - \mathbf{D}_N^{-1/2} \mathbf{A}_N \mathbf{D}_N^{-1/2}. \tag{2.1}$$

Additionally, let the centered and scaled normalized Laplacian of $\mathbf{A}_N$ be defined,

$$\mathbf{B}_N \coloneqq \sqrt{\frac{Np_N}{1-p_N}} (\mathbf{I}_N - \mathbf{L}_N), \tag{2.2}$$

and assume that $\sup\{p_N; N \geq 2\} < 1$. We define the spectral measure of $\mathbf{B}_N$ as follows,

$$\operatorname{ESD}(\mathbf{B}_N) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i(\mathbf{B}_N)}. \tag{2.3}$$

Finally, we define an approximation between two sequences, a_n and b_n, $a_n \approx b_n$, as follows for some constants c_1 and c_2,

$$c_1 \leq \liminf \frac{a_n}{b_n} \leq \limsup \frac{a_n}{b_n} \leq c_2, \tag{2.4}$$

which we also extend to inequalities, $a \lesssim b$,

$$\limsup \frac{a_n}{b_n} \leq c_2. \tag{2.5}$$

We are now in a position to state our first theorem regarding the spectral behavior of the normalized Laplacian matrix of G_N.

2.2 Convergence of the ESD of the Normalized Laplacian

We will first state the overarching theorem regarding the convergent behavior of $\mathbf{L}_N$ for an ERRG and then outline our strategy for proving this statement, including the intuition of the supporting lemmas that must be proven as well.

Theorem 2.2.1. *[29] If G_N is well-connected, that is, if $Np/\log(N) \to \infty$ as $N \to \infty$, then we have that:*

$$\lim_{N\to\infty} \mathrm{ESD}\left(\boldsymbol{B}_N\right) = \mu_{sc}, \tag{2.6}$$

weakly in probability, where μ_{sc} is the semicircle law as defined in (1.4).

To aid in the proof, we establish some tools. Let the bounded Lipschitz distance between two probability measures, μ and ν on $\mathbb{R}$ be:

$$d_{BL}(\mu,\nu) = \sup\{|\int f d\mu - \int f d\nu| : \|f\|_\infty + \|f\|_L \leq 1\}, \tag{2.7}$$

where $\|f\|_\infty$ is the infinity norm and $\|f\|_L = \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|}$. Additionally, we introduce the Hoffman-Wielandt inequality.

Lemma 2.2.1. *[3] For $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ being Hermitian matrices of equal dimension, the cubed bounded Lipschitz distance between their spectral measures can be bounded by the normalized trace of their squared differences,*

$$d_{BL}^3(\mathrm{ESD}\left(\boldsymbol{M}_1\right), \mathrm{ESD}\left(\boldsymbol{M}_2\right)) \leq \frac{1}{N}\,\mathrm{Tr}\left(\left(\boldsymbol{M}_1 - \boldsymbol{M}_2\right)^2\right). \tag{2.8}$$

We first define how the deviations of the entries of G_N behave in the ERRG regime and extend this to the behavior of the degrees of the graph, showing that it is possible to approximate the behavior of a node's degree with its expected value. This is critical, as $\mathbf{A}_N$ and $\mathbf{D}_N$ are dependent, given that the latter is calculated directly from the adjacency matrix. Thus, "decoupling" $\mathbf{D}_N$ from the entry-level behavior of $\mathbf{A}_N$ by using a fixed degree breaks this dependency. We then use that when scaled by an appropriate normalizing constant, $\mathbf{A}_N$ converges to μ_{sc} [15]. Tying these two convergences together completes the proof for (2.6).

2.3 Bounded Deviations of Homogeneous Graphs

We evaluate the node level behavior of a given $\mathbf{A}_N(i,j)$ in G_N by generalizing to any Binomial random variable parameterized by N and p, $\mathrm{Bin}(N,p)$.

Lemma 2.3.1. *Let $\epsilon_1,\ldots,\epsilon_N$ be i.i.d random variables with $\mathbb{P}(\epsilon_i = 1) = 1 - \mathbb{P}(\epsilon_i = 0) = p \in (0,1)$. Then:*

$$P\left(\left|\frac{\sum_{i=1}^N \epsilon_i}{Np} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{(Np)\delta^2}{4}\right) \tag{2.9}$$

for all $N \geq 1$, and for all $\delta \in (0,1)$.

In order to prove this statement, we will use the relative error Chernoff bound [14]. For $X = \sum_{i=1}^N X_i$ being a sum of independent random variables, X_i, the upper tail bound is:

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}[X]) \leq \exp\left(-\frac{\delta^2\mathbb{E}[X]}{2+\delta}\right), \tag{2.10}$$

and the lower tail bound is:

$$\mathbb{P}(X \leq (1-\delta)\mathbb{E}[X]) \leq \exp\left(-\frac{\delta^2\mathbb{E}[X]}{2}\right). \tag{2.11}$$

Proof of Lemma 2.3.1:

Let $X = \sum_{i=1}^N \epsilon_i$. As each ϵ_i is an independent Bernoulli random variable, X is a binomial random variable and thus, $\mathbb{E}[X] = Np$. We now assess $\mathbb{P}\left(\left|\frac{X}{Np} - 1\right| \geq \delta\right)$, knowing that the tail behavior of the binomial distribution is symmetric. To begin, we can decompose the absolute value into the corresponding sides of the deviation from Np by a factor of $1 \pm \delta$,

$$\mathbb{P}\left(\left|\frac{X}{Np} - 1\right| \geq \delta\right) \leq \mathbb{P}\left(X \geq (1+\delta)Np\right) + \mathbb{P}\left(X \leq (1-\delta)Np\right).$$

We can thus substitute the first individual tail probability into the upper tail bound, (2.10) and do the same for the latter tail probability, using the lower tail bound, (2.11). Additionally, δ is an error

term, so we take this to be small, with the 2 in the denominator of the upper tail bound dominating the contribution. As such, the earlier decomposition yields the following final symmetric bound,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{X}{Np}-1\right|\geq\delta\right)&\leq\mathbb{P}\left(X\geq(1+\delta)Np\right)+\mathbb{P}\left(X\leq(1-\delta)Np\right)\\&\leq\exp\left(-\frac{\delta^2Np}{2+\delta}\right)+\exp\left(-\frac{\delta^2Np}{2}\right)\\&\leq 2\exp\left(-\frac{\delta^2Np}{2}\right).\end{aligned}$$

This is precisely the inequality laid out in (2.9), so we have completed the proof. ■

Lemma 2.3.1 states that the deviation of a binomial variable from its expected value is a function of N and p, and can be quantified uniformly on both tails by the above bound. The deviation from the mean is thus finite and vanishing as $Np\to\infty$. We can now use this to explicitly bound the degrees of G_N, which are also binomial random variables as the sum of individual $\text{Ber}(p_N)$ entries.

2.4 Degree Bounds for Homogeneous Graphs

There are three degree properties which we seek to bound in this section to ensure that G_N has relatively well-behaved nodes, where the degree of the i-th node is $\mathbf{d}_i\coloneqq\sum_{j\neq i}^N\mathbf{A}_N(i,j)$. The first property is the maximum degree deviation of the graph,

$$a_N\coloneqq\max_{1\leq i\leq N}\{|\mathbf{d}_i-(N-1)p_N|\}. \tag{2.12}$$

Ensuring that this deviation has vanishing probability relative to an error term δ guarantees that the graph remains relatively uniform as $N\to\infty$. The second property concerns the minimum degree of G_N,

$$b_N\coloneqq\min_{1\leq i\leq N}\{\mathbf{d}_i\}, \tag{2.13}$$

which converges with the expected degree as $N\to\infty$, again to ensure that there is uniform behavior, similar to the convergence of a_N. Finally, we look to prove the convergence of the total number of connections in the graph to N^2p_N, which is to say that the graph is well-connected with the connection probability between any two nodes being p_N. Having provided some intuition, we formally state our next lemma.

Lemma 2.4.1. *If $Np_N/\log(N)\to\infty$ as $N\to\infty$, then:*

$$\frac{a_N}{Np_N}\to 0,\quad \frac{b_N}{Np_N}\to 1,\quad \frac{1}{N^2p_N}\sum_{1\leq i\neq j\leq N}\boldsymbol{A}_N(i,j)\to 1,\quad \textit{almost surely.} \tag{2.14}$$

Proving this statement will also rely upon Lemma 2.3.1 and the Borel-Cantelli lemma [4], which we restate. Let $\{E_i\}_{1\leq i}$ denote a set of events and the series of the probabilities of these events occurring is finite,

$$\sum_{i=1}^{\infty}\mathbb{P}\left(E_i\right)<\infty.$$

The probability of infinitely many occurrences of E_i is thus 0,

$$\mathbb{P}\left(\limsup_{i\to\infty}E_i\right)=\mathbb{P}\left(\bigcap_{i=1}^{\infty}\bigcup_{j=i}^{\infty}E_j\right)=0.$$

Intuitively, for a sequence of events, if the summed probability of all events occurring is finite, then the probability that an infinitely long sequence of these events occur is 0. The lim sup can be interpreted as the set of events where infinitely many of the sequences occur.

Proof of Lemma 2.4.1:

Looking first at $\frac{a_N}{Np_N}\to 0$, we apply the result bounding deviations in Binomial variables from (2.9), where the expected value of the degree is $(N-1)p_N$, in conjunction with a union bound to account for taking a maximum over N terms, absorbing the leading 2 from (2.9) into N,

$$\mathbb{P}\left(a_N \geq \delta(N-1)p_N\right) = \mathbb{P}\left(\max_{1\leq i\leq N}\left|\frac{\sum_{j=1}^{N}\mathbf{A}_N(i,j)}{(N-1)p_N} - 1\right| \geq \delta\right) \lesssim N\exp\left(-\frac{\delta^2(N-1)p_N}{4}\right).$$

Moving the leading N into the exponential as $\log(N)$ allows us to then use that $\frac{Np_N}{\log N} \to \infty$. We can then introduce an error ϵ representing the gap in the ratio between Np_N ad $\log(N)$, $Np_N > \epsilon\log(N)$. We choose $\epsilon = \frac{8}{\delta^2}$, and substitute this into the above inequality,

$$N\exp\left(-\frac{\delta^2(N-1)p_N}{4}\right) \approx \exp\left(\log(N) - \frac{\delta^2 Np_N}{4}\right) \leq \exp\left(-\frac{8}{\delta^2}\log(N)\frac{\delta^2}{4}\right) = N^{-2}.$$

As such, $\mathbb{P}\left(a_N \geq \delta(N-1)p_N\right)$ is finite and vanishing at a rate of N^{-2}, interpretable as stating that the event that the deviation exceeds δ is finite. This allows us to apply the Borel-Cantelli lemma to imply that $\frac{a_N}{(N-1)p_N} \to 0$ a.s., thus proving the first convergence in (2.14).

We now look at the second convergence, $\frac{b_N}{Np_N} \to 1$, beginning by rearranging the statement of convergence:

$$\left|\frac{b_N}{(N-1)p_N} - 1\right| = \left|\frac{\min_{1\leq i\leq n}\{\mathbf{d}_i\}}{(N-1)p_N} - 1\right| = \left|\frac{\min_{1\leq i\leq N}\{\mathbf{d}_i - (N-1)p_N\}}{(N-1)p_N}\right|.$$

Note that this is now of the form of the definition of a_N, and so we can bound this from above by the result we have just proven for $\frac{a_N}{(N-1)p_N} \to 0$,

$$\left|\frac{\min_{1\leq i\leq N}\{\mathbf{d}_i - (N-1)p_N\}}{(N-1)p_N}\right| \leq \left|\frac{\max_{1\leq i\leq N}\{\mathbf{d}_i - (N-1)p_N\}}{(N-1)p_N}\right| = \frac{a_N}{Np_N} \overset{a.s.}{\to} 0,$$

which then implies $\frac{b_N}{(N-1)p_N} \overset{a.s.}{\to} 1$, thus proving the second convergence.

Looking finally at the third convergence, $\frac{1}{N^2p_N}\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j) \to 1$, we can apply the symmetry of $\mathbf{A}_N$, and instead show that:

$$\frac{2}{N(N-1)p_N}\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j) \to 1,$$

as we can group the upper and lower triangles of the matrix and omit the diagonal entries. Applying (2.9) from above, we have that:

$$\mathbb{P}\left(\left|\frac{2}{N(N-1)p_N}\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j) - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{N(N-1)p_N\delta^2}{4}\right).$$

Applying the same approach as for the first convergence, we now set the gap term $\epsilon = \frac{8}{N\delta^2}$ in $Np_N > \epsilon\log(N)$ to account for the total degree across N^2 entries as opposed to the maxima, which then yields the same bound of N^{-2} for the right-hand side of the above expression,

$$2\exp\left(-\frac{N(N-1)p_N\delta^2}{4}\right) \lesssim \exp\left(-Np_N\left(\frac{N\delta^2}{4} - \frac{\log(2)}{Np_N}\right)\right) \leq \exp\left(-\frac{8}{N\delta^2}\log(N)\left(\frac{N\delta^2}{4}\right)\right) = N^{-2}.$$

We have thus shown that $\sum_{N=1}^{\infty}\mathbb{P}\left(\left|\frac{2}{N(N-1)p_N}\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j) - 1\right| \geq \delta\right)$ is finite, and so we again apply the Borel-Cantelli lemma to show convergence to 1, a.s. As such, we have proven all three convergences regarding the behavior of degree deviations, the minimum degree, and the total degree of G_N as laid out in (2.14). ■

Having shown that in the setting laid out at the beginning of this section, that the degree behavior of $\mathbf{A}_N$ will be well-behaved, we will restate Theorem 3 from [15] and its Corollary before completing the proof. The result states that the spectral measure of an ERRG converges to μ_{sc} given that it is centered by the expectation of its entries and scaled by a function of the variance of its entries.

Lemma 2.4.2. *Let $\boldsymbol{\omega}_N$ be an $N \times N$ matrix where $\omega_N(i,j) \coloneqq (\boldsymbol{A}_N(i,j) - \mu_N)/\sigma_N$ for all i,j,N where $\mu_N = \mathbb{E}[\boldsymbol{A}_N(i,j)]$ and $\sigma_N^2 = \mathrm{Var}(\boldsymbol{A}_N(i,j))$. Assume that*

$$\max_{1\leq i<j\leq N}\mathbb{E}\left[(\omega_N(i,j))^2\,\mathbb{1}\left(|\omega_N(i,j)| \geq \varepsilon\sqrt{N}\right)\right] \to 0$$

as $N \to \infty$ for some $\epsilon > 0$. We additionally define its spectral distribution,

$$\mathrm{ESD}\left(\frac{\boldsymbol{A}_N - \mu_N \boldsymbol{J}_N}{\sqrt{N}\sigma_N}\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i(\cdot)}, \tag{2.15}$$

for $\lambda_i(\cdot)$ being the eigenvalues of $\frac{\boldsymbol{A}_N - \mu_N \boldsymbol{J}_N}{\sqrt{N}\sigma_N}$ and $\boldsymbol{J}_N$ being an $N \times N$ matrix of all 1s. Under the same setting as before, we assert that $\mathrm{ESD}\left(\frac{\boldsymbol{A}_N - \mu_N \boldsymbol{J}_N}{\sqrt{N}\sigma_N}\right)$ converges to the semicircular law,

$$\lim_{N\to\infty} \mathrm{ESD}\left(\frac{\boldsymbol{A}_N - \mu_N \boldsymbol{J}_N}{\sqrt{N}\sigma_N}\right) = \mu_{sc}, \tag{2.16}$$

weakly, almost surely.

The assumption in this statement can be interpreted as stating that the variance of the deviations that exceed a given error, dependent on $\sqrt{N}$, vanishes, implying that excessively large deviations will vanish as $N \to \infty$ for a sufficiently large N. Note that this is similar to the Lindeberg condition for the Lindeberg-Feller central limit theorem [4]. Having these well-behaved deviations ensures that we can show convergent behavior between $\mathrm{ESD}\left(\frac{\mathbf{A}_N - \mu_N \mathbf{J}_N}{\sqrt{N}\sigma_N}\right)$ and μ_{sc} without mass escaping. Note that for a homogeneous ERRG, $\mu_N = p_N$ and $\sigma_N = \sqrt{p_N(1-p_N)}$, with σ_N arising from the properties of the Bernoulli distribution. This thus leads to the following corollary [15].

Corollary 2.4.1. *For $\alpha_N \coloneqq (Np_N(1-p_N))^{1/2} \to \infty$ as $N \to \infty$, the empirical spectral distribution of $\boldsymbol{A}_N/\alpha_N$, $\mathrm{ESD}\left(\boldsymbol{A}_N/\alpha_N\right)$, converges weakly to the semicircular law μ_{sc}, almost surely,*

$$\lim_{N\to\infty} \mathrm{ESD}\left(\boldsymbol{A}_N/\alpha_N\right) = \mu_{sc}. \tag{2.17}$$

Having stated these results concerning convergence under proper scaling, we can now complete the proof of the overarching theorem.

2.5 Proof of Theorem 2.2.1

Having established all the various supporting results, we can now directly prove (2.6). To do so, we will use the triangle inequality and the following set-up to prove that $\lim_{N\to\infty} \mathrm{ESD}\left(\mathbf{B}_N\right) = \mu_{sc}$ weakly in probability,

$$d_{BL}(\mathrm{ESD}\left(\mathbf{B}_N\right), \mu_{sc}) \le d_{BL}\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) + d_{BL}\left(\mathrm{ESD}\left(\hat{\mathbf{B}}_N\right), \mu_{sc}\right), \tag{2.18}$$

where $\hat{\mathbf{B}}_N = \sqrt{\frac{Np_N}{1-p_N}}(\mathbf{I} - \hat{\mathbf{L}}_N)$ and $\hat{\mathbf{L}}_N = \mathbf{I}_N - (Np_N)^{-1}\mathbf{A}_N$ being the "decoupled" matrices for $\mathbf{B}_N$ and $\mathbf{L}_N$, respectively. We then seek to show that both components on the right-hand side of the inequality converge to 0, thus yielding that $\mathrm{ESD}\left(\mathbf{B}_N\right)$ converges to μ_{sc}. To begin, we set up Lemma 2.2.1 for the first term, $d_{BL}\left(\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right)\right.$:

$$d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) \le \frac{1}{N}\mathrm{Tr}\left((\mathbf{B}_N - \hat{\mathbf{B}}_N)^2\right). \tag{2.19}$$

We assess this by first expanding and simplifying the definitions of $\mathbf{B}_N$ and $\hat{\mathbf{B}}_N$,

$$\frac{1}{N}\mathrm{Tr}\left((\mathbf{B}_N - \hat{\mathbf{B}}_N)^2\right) = \frac{1}{N}\mathrm{Tr}\left(\frac{Np_N}{1-p_N}(\hat{\mathbf{L}}_N - \mathbf{L}_N)^2\right) = \frac{Np_N}{N(1-p_N)}\sum_{i,j=1}^{N}((\hat{\mathbf{L}}_N - \mathbf{L}_N)(i,j))^2,$$

the final equality arising from the element-wise expansion of the trace. We use the (i,j) notation to denote the entry-wise value of a given matrix. Looking at the squared difference between $\hat{\mathbf{L}}_N$ and $\mathbf{L}_N$ within the summation, we plug in the definitions of the non-decoupled and decoupled variations of the normalized Laplacian matrix,

$$\begin{aligned}((\hat{\mathbf{L}}_N - \mathbf{L}_N)(i,j))^2 &= ((1-(Np)^{-1}\mathbf{A}_N) - (1-\mathbf{D}_N^{-1/2}\mathbf{A}_N\mathbf{D}_N^{-1/2}))^2(i,j) \\ &= \left(\frac{1}{\sqrt{\mathbf{d}_i}}\mathbf{A}_N(i,j)\frac{1}{\sqrt{\mathbf{d}_j}} - \frac{1}{Np_N}\mathbf{A}_N(i,j)\right)^2 \\ &= \mathbf{A}_N^2(i,j)\left(\frac{1}{\sqrt{\mathbf{d}_i}\sqrt{\mathbf{d}_j}} - \frac{1}{Np_N}\right)^2.\end{aligned}$$

Of course, $\mathbf{A}_N$ is a binary matrix, taking either 1 or 0 as the entry values. As such, in the event that the (i,j)-th entry is 0, this term does not contribute to the resulting summation, so we can ignore those cases. However, if the entry takes the value 1, we can reduce $\mathbf{A}_N^2(i,j) = \mathbf{A}_N(i,j)$. Additionally, we can use the inequality $|\sqrt{x}-\sqrt{y}|^2 \leq \frac{|x-y|^2}{\sqrt{y}}$ to simplify the degree terms in the brackets, letting $Np_N = \sqrt{y}$ and $\sqrt{\mathbf{d}_i}\sqrt{\mathbf{d}_j} = \sqrt{x}$. This yields the following:

$$\mathbf{A}_N^2(i,j)\left(\frac{1}{\sqrt{\mathbf{d}_i}\sqrt{\mathbf{d}_j}} - \frac{1}{Np_N}\right)^2 = \mathbf{A}_N(i,j)\frac{(Np_N - \sqrt{\mathbf{d}_i}\sqrt{\mathbf{d}_j})^2}{N^2p_N^2\mathbf{d}_i\mathbf{d}_j} \leq \mathbf{A}(i,j)\frac{|\mathbf{d}_i\mathbf{d}_j - N^2p_N^2|^2}{N^4p_N^4\mathbf{d}_i\mathbf{d}_j}.$$

Observe that we can provide a bound on any given degree using (2.4.1) as follows:

$$\mathbf{d}_i \leq |\mathbf{d}_i + \mathbb{E}[\mathbf{d}_i]| - \mathbb{E}[\mathbf{d}_i] \leq a_N + Np_N + \delta.$$

This bound states that any individual degree deviations will be near the maximum given deviation as denoted by a_N, with a reasonably close error term δ. We can now insert this bound into the numerator of our earlier statement,

$$\mathbf{A}(i,j)\frac{|\mathbf{d}_i\mathbf{d}_j - N^2p_N^2|^2}{N^4p_N^4\mathbf{d}_i\mathbf{d}_j} \lesssim \mathbf{A}(i,j)\frac{|(Np_N+a_N)(Np_N+a_N) - N^2p^2|^2}{N^4p_N^4\mathbf{d}_i\mathbf{d}_j} = \mathbf{A}(i,j)\frac{|2a_Np_NN + a_N^2|^2}{N^4p_N^4\mathbf{d}_i\mathbf{d}_j}.$$

This decoupling from the individual node degrees, $\mathbf{d}_i$, thus allows us to bound the element-wise difference between $\mathbf{L}_N$ and $\hat{\mathbf{L}}_N$ by applying the bounds derived in Lemma 2.3.1. We now do the same for the degree terms in the denominator using b_N, observing that any $\mathbf{d}_i \geq b_N = \min_{1\leq i\leq N}\{\mathbf{d}_i\}$,

$$\mathbf{A}(i,j)\frac{|2a_Np_NN + a_N^2|^2}{N^4p_N^4\mathbf{d}_i\mathbf{d}_j} \leq \mathbf{A}(i,j)\frac{|2a_Np_NN + a_N^2|^2}{N^4p_N^4b_N^2} \leq \mathbf{A}(i,j)\frac{4a_N^2p_N^2N^2 + 2a_N^4}{N^4p_N^4b_N^2}.$$

The final inequality arises from the identity $(x+y)^2 \leq 2x^2 + 2y^2$. Having now re-expressed this difference without any reference to an individual $\mathbf{d}_i$, we can evaluate the original trace of the difference between $\mathbf{B}_N$ and $\hat{\mathbf{B}}_N$,

$$\frac{Np_N}{N(1-p_N)}\sum_{i,j=1}^{N}((\hat{\mathbf{L}}_N - \mathbf{L}_N)(i,j))^2 \leq \frac{Np_N}{N(1-p_N)}\frac{4a_N^2p_N^2N^2 + 2a_N^4}{N^4p_N^4b_N^2}\sum_{i,j=1}^{N}\mathbf{A}(i,j).$$

We can now apply Lemma 2.4.1, using the result that $b_N \approx Np$, given that $\frac{b_N}{Np_N} \to 1$ as $N \to \infty$. Additionally, note that $\mathbf{A}_N$ is composed of approximately N^2 entries, each of which is a Bernoulli random variable with expectation p_N. We can thus rearrange to find a normalizing constant that removes this term from the statement, leaving us with a purely deterministic bound.

$$\begin{aligned}\frac{Np_N}{N(1-p_N)}\frac{4a_N^2p_N^2N^2 + 2a_N^4}{N^4p_N^4b_N^2}\sum_{i,j=1}^{N}\mathbf{A}(i,j) &= \frac{Np_N}{N(1-p_N)}\frac{4a_N^2p_N^2N^2 + 2a_N^4}{N^4p_N^4(Np_N)^2}\sum_{i,j=1}^{N}\mathbf{A}(i,j) \\ &= \frac{Np_N}{N(1-p_N)}\frac{4a_N^2p_N^2N^2 + 2a_N^4}{N^4p_N^5}\frac{\sum_{i,j=1}^{N}\mathbf{A}(i,j)}{N^2p_N} \\ &\approx \frac{4}{1-p_N}\left[\frac{a_N^2}{N^2p_N^2} + \frac{a_N^4}{2N^4p_N^4}\right].\end{aligned}$$

The first convergence from Lemma 2.4.1, $\frac{a_N}{Np_N} \to 0$ as $Np_N \to \infty$, can now be applied to the fractions in the brackets, thus showing that this final quantity goes to 0, almost surely. We have thus shown that the normalized difference in the trace on the right-hand side of the initial Hoffman-Wielandt inequality from

(2.19) goes to 0, which then yields that the squared bounded Lipschitz distance of the spectral measures of $\mathbf{B}_N$ and $\hat{\mathbf{B}}_N$ also goes to 0,

$$d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) \leq \frac{1}{N}\,\mathrm{Tr}\left((\mathbf{B}_N - \hat{\mathbf{B}}_N)^2\right) \overset{a.s.}{\to} 0.$$

By extension this states that $d_{BL}\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) \to 0$, which can be interpreted as stating that using the decoupled $\hat{\mathbf{B}}_N$ in place of $\mathbf{B}_N$ is admissible. We now look to prove that the second distance measure in (2.18) also goes to 0, which then completes the proof. To do so, we can inspect the definition of $\hat{\mathbf{B}}_N$ to verify it is of the form of (2.17),

$$\hat{\mathbf{B}}_N = \sqrt{\frac{Np_N}{1-p_N}}(\mathbf{I}_N - \hat{\mathbf{L}}_N) = \sqrt{\frac{Np_N}{1-p_N}}\left(\frac{1}{Np_N}\mathbf{A}_N\right) = \frac{\mathbf{A}_N}{\sqrt{Np_N(1-p_N)}}.$$

As per the statement of the Corollary, we see that $\mathbf{A}_N$ is normalized by $\alpha_N \coloneqq \sqrt{Np_N(1-p_N)}$, and as such we can directly apply the result to state that,

$$\lim_{N\to\infty} \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right) = \lim_{N\to\infty} \mathrm{ESD}\left(\mathbf{A}_N/\alpha_N\right) = \mu_{sc}, \quad \text{weakly in probability.}$$

Thus, the BL distance between the two spectral measures also goes to 0, as is required to demonstrate that the right-hand side of (2.18) goes to 0. This then completes the proof in the homogeneous ERRG setting. ■

We have now established a baseline result for the convergent behavior of the ESD of a homogeneous random graph, that is, a graph with a connection probability p_N uniform across all pairs of nodes. It is useful to frame this as one case of a more generalized inhomogeneous class of random graphs, where the connection probability between nodes is a function of the nodes themselves, and so a uniform connection probability p_N is simply a constant function. This will be discussed in the next chapter.

Chapter 3
The Normalized Laplacian in the Inhomogeneous Regime

In this section, we will first provide some additional notation for the inhomogeneous setting before stating our main result for the convergence of the ESD of the normalized Laplacian to a deterministic measure and its proof.

3.1 Set-Up

We continue to use the graph in prior section defined as $G_N = (V, E)$ of size N. However, we re-define the connection probability as follows. Let $f : [0,1]^2 \to [0,\infty)$, where f is bounded and Riemann integrable. The probability of connection between two vertices is,

$$p_{ij} = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right),$$

where $\varepsilon_N \in \mathbb{R}^+$ is a tuning parameter, and because f is bounded and $\varepsilon_N \to 0, \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right) \leq 1$. Thus, any entry in the matrix $\mathbf{A}_N$ is a Bernoulli random variable:

$$\mathbf{A}_N(i,j) = \text{Ber}\left(\varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\right).$$

Additionally, we assume that f can be decomposed multiplicatively into two composite relatively simple polynomial functions, such that $f(x,y) = r(x)r(y)$, $r(x) \in \mathbb{R}$. Finally, we introduce a general measure, μ_f which is non-random and compactly supported on $\mathbb{R}$ [7]. We will provide a more robust definition of this measure later on in the section. The remainder of the definitions from the prior chapter, such as for the descriptive matrices of G_N, remain the same. Additionally, we note that many of the following calculations will follow the same steps for a general f.

3.2 Convergence of ESD $(\mathbf{B})$ to μ_f

We first state Theorem 2 before briefly discussing the differences in approaching the proof in this regime as opposed to the homogeneous regime in the previous section. Additionally, we will calculate an explicit form for the normalized expected value of a randomly drawn vertex, which will then be used to find an appropriate scaling constant, which in Theorem 2.2.1 was taken to be $\sqrt{\frac{Np_N}{1-p_N}}$.

Theorem 3.2.1. *Let the centered and scaled normalized Laplacian of $\boldsymbol{A}_N$ be defined as $\boldsymbol{B}_N \coloneqq \gamma_N(\boldsymbol{I}_N - \boldsymbol{L}_N)$ for an appropriate scaling constant, $\gamma_N \coloneqq \sqrt{\frac{N\varepsilon_N}{1-\varepsilon_N}} \in \mathbb{R}$. If $N\varepsilon_N \gg \log(N), N \to \infty, \varepsilon_N \to 0$, such that there exists a non-random and compactly supported measure, μ_f, on $\mathbb{R}$,*

$$\lim_{N\to\infty} \text{ESD}\left(\boldsymbol{B}_N\right) = \mu_f, \tag{3.1}$$

weakly in probability.

We will see that in order to prove this convergence of ESD $(\mathbf{B}_N)$ to μ_f, that the proof will broadly follow the same structure as in the homogeneous setting by first bounding element-wise variance, bounding the degree behavior of the graph to ensure a well-behaved object, and finally using these to show that the spectra of the scaled normalized Laplacian converges to μ_f. However, in this final step, there will be multiple perturbations that need to be applied prior to a statement of final convergence, like was done in the final proof of Theorem 2.2.1. To start, we evaluate an explicit form for the normalized expectation of a randomly selected vertex of G_N.

Lemma 3.2.1. *For u_N being uniformly drawn from vertices in $\mathbf{A}_N$,*

$$\lim_{N\to\infty} \frac{\mathbb{E}[\boldsymbol{d}_{u_N}]}{N\varepsilon_N} = \left(\int_0^1 r(x)\,dx\right)^2 = \int_{[0,1]^2} f(x,y)dxdy. \tag{3.2}$$

The proof of this statement is relies on the properties of the Bernoulli random variable.

Proof of Lemma 3.2.1: We start by defining $\mathbb{E}[\mathbf{d}_{u_N}]$ for a uniformly drawn u_N, with the probability of drawing a specific node i being denoted $\mathbb{P}(u=i)=\frac{1}{N}$,

$$\frac{1}{N\varepsilon_N}\mathbb{E}[\mathbf{d}_{u_N}] = \frac{1}{N\varepsilon_N}\sum_{i=1}^N \mathbb{E}[\mathbf{d}_i]\,\mathbb{P}(u=i) = \frac{1}{N^2\varepsilon_N}\sum_{i=1}^N \sum_{1\le j\neq i\le N} \mathbb{E}[\mathbf{A}_N(i,j)].$$

Each entry of $\mathbf{A}_N$ is a Bernoulli random variable, so its expectation is the probability of an edge at that index, which is $p_{ij}=\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)$. We can additionally use the multiplicative structure of f to decompose it to the product of r,

$$\begin{aligned}\frac{1}{N^2\varepsilon_N}\sum_{i=1}^N \sum_{1\le j\neq i\le N} \mathbb{E}[\mathbf{A}_N(i,j)] &= \frac{1}{N^2}\sum_{i=1}^N \sum_{1\le j\neq i\le N} f\left(\frac{i}{N},\frac{j}{N}\right)\\ &= \frac{1}{N^2}\sum_{i=1}^N \sum_{1\le j\neq i\le N} r\left(\frac{i}{N}\right) r\left(\frac{j}{N}\right)\\ &\to \int_{[0,1]^2} f(x,y)dxdy.\end{aligned}$$

With the last convergence from f being Riemann integrable and as such, r also being Riemann integrable. We can thus split the denominator of the leading fraction to yield two Riemann sums of r, thus providing the integrals in (3.2) as $N\to\infty$. This thus provides a convenient definition of the expectation of a normalized drawn entry. ■

Note that the above argument can be made for any general continuous function f. We can also impose that $r(x)$ close to $r(y)$ by a factor of $o(1)$. Let $m \coloneqq \sup_{x\in[0,1]} r(x)$ and $m_1 \coloneqq \int_0^1 r(x)dx \le m$. Additionally, let $m_2 \coloneqq \int_0^1 r^2(x)dx$ be the second moment of $r(x)$. We can express the above statement then as,

$$\begin{aligned}\mathbb{E}[\mathbf{d}_{u_N}] = N\varepsilon_N\left(\int_0^1 r(x)\right)^2 &= N\varepsilon_N\left(\frac{\sum_{i=1}^N r(i/N)}{N}\frac{\sum_{j=1}^N r(j/N)}{N}\right)\\ &\le \frac{\varepsilon_N}{N}\left(\sum_{i=1}^N m\sum_{j=1}^N m\right)\\ &= N\varepsilon_N m^2(1+o(1)).\end{aligned} \tag{3.3}$$

We are now able to provide an explicit form for γ_N. As a scaling constant, we know that,

$$\gamma_N \coloneqq \frac{\mathbb{E}[\mathbf{d}_i]}{\sqrt{\mathrm{Var}[\mathbf{d}_i]}}.$$

As established above, $\mathbb{E}[\mathbf{d}_i] = \varepsilon_N r(\frac{i}{N})\sum_{1\le j\neq i\le N} r(\frac{j}{N})$. We thus look to evaluate $\mathrm{Var}[\mathbf{d}_i]$, using that the individual $\mathbf{A}_N(i,j)$ entries are independent Bernoulli random variables, $\mathrm{Ber}(p_{ij})$, with variance $p_{ij}(1-p_{ij})$

$$\mathrm{Var}[\mathbf{d}_i] = \sum_{1\le j\neq i\le N}\mathrm{Var}[\mathbf{A}_N(i,j)] = \sum_{1\le j\neq i\le N}\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\left[1-\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\right].$$

We can then expand this expression and use that r and m_2 are both bounded to provide a final form for the variance as $N\to\infty$,

$$\begin{aligned}
&\sum_{1\le j\neq i\le N} \varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\left[1-\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\right]\\
=&\varepsilon_N r\left(\frac{i}{N}\right)\left[\sum_{1\le j\neq i\le N} r\left(\frac{j}{N}\right)-\varepsilon_N r\left(\frac{i}{N}\right)\sum_{1\le j\neq i\le N} r\left(\frac{j}{N}\right)^2\right]\\
\approx&N\varepsilon_N r(\frac{i}{N})[m_1-\varepsilon_N r(\frac{i}{N})m_2]\approx N\varepsilon_N r\left(\frac{i}{N}\right)m_1.
\end{aligned}$$

Substituting this final expression into $\gamma_N \coloneqq \frac{\mathbb{E}[\mathbf{d}_i]}{\sqrt{\mathrm{Var}[\mathbf{d}_i]}}$, we arrive at the scaling constant for $\mathbf{B}_N$:

$$\frac{\mathbb{E}[\mathbf{d}_i]}{\sqrt{\mathrm{Var}[\mathbf{d}_i]}}=\frac{N\varepsilon_N r(i/N)\sum_{i\le j\neq i\le N} r(j/N)}{\sqrt{N\varepsilon_N r(i/N)m_1}}=\frac{\sqrt{N\varepsilon_N r(i/N)}\sum_{i\le j\neq i\le N} r(j/N)}{\sqrt{m_1}}\approx\sqrt{N\varepsilon_N}.$$

Thus, we now have an appropriate scaling constant that ensures that the mass of ESD $(\mathbf{B}_N)$ accumulates to the correct shape. Incorrect scaling would result in overly diffuse or potentially skewed asymptotic distributions of this ESD. Having now established these fundamental properties of an inhomogeneous G_N, we can begin assessing the behavior of its degree matrix.

3.3 Degree Bounds for Inhomogeneous Graphs

As in Lemma 2.4.1, we seek to bound the maximal degree deviation, the minimum degree, and the total degree of G_N to enable the decoupling process in later parts of the proof. However, as connection behavior in the graph is no longer a consistent p_N throughout, the calculations become slightly more involved. To begin, we restate a_N and b_N for the inhomogeneous setting,

$$a_N \coloneqq \max_{1\le i\le N}\{|\mathbf{d}_i-\mathbb{E}[\mathbf{d}_i]|\},\quad b_N \coloneqq \min_{1\le i\le N}\{\mathbf{d}_i\}, \tag{3.4}$$

where $\mathbb{E}[\mathbf{d}_i]=N\varepsilon_N r(i/N)\sum_{i\le j\neq i\le N} r(j/N)\lesssim N\varepsilon_N m m_1$.

Lemma 3.3.1. *If $N\varepsilon_N/\log(N)\to\infty$ as $N\to\infty$, then:*

$$\frac{a_N}{N\varepsilon_N}\to 0,\ \frac{b_N}{N\varepsilon_N}\to 1,\ \frac{1}{N^2 m m_1\varepsilon_N}\sum_{1\le i\neq j\le N}\boldsymbol{A}_N(i,j)\to 1, \tag{3.5}$$

almost surely.

We will use a slightly different form of the Chernoff bound from Theorem 2.21 of [45] for this proof compared to the bounds used for the homogeneous setting, as we no longer assume symmetric tail behavior in the inhomogeneous case. These tail bounds are defined as follows for an error term $t\in\mathbb{R}$ and $X=\sum_{i=1}^N X_i$ being the sum of independent Bernoulli random variables, starting first with the upper tail,

$$P(X\ge\mathbb{E}[X]+t)\le\exp\left(-\frac{t^2}{2(\mathbb{E}[X]+t/3)}\right), \tag{3.6}$$

and the lower tail,

$$P(X\le\mathbb{E}[X]-t)\le\exp\left(-\frac{t^2}{2\mathbb{E}[X]}\right). \tag{3.7}$$

Proof of Lemma 3.3.1:

Let t in the above Chernoff bound be defined as $t=\delta_N\varepsilon_N N$. Starting with $\frac{a_N}{N\varepsilon_N}\overset{a.s.}{\to}0$, it is useful to state the convergence we are assessing, $\left|\frac{a_N}{N\varepsilon_N}\right|\to 0$, explicitly as a probability in line with the upper tail Chernoff bound:

$$\mathbb{P}\left(\frac{\max_{1\le i\le N}\{\mathbf{d}_i-\mathbb{E}[\mathbf{d}_i]\}}{N\varepsilon_N}\ge\delta_N\right)=\mathbb{P}\left(\max_{1\le i\le N}\{\mathbf{d}_i\}\ge\mathbb{E}[\mathbf{d}_i]+N\varepsilon_N\delta_N\right).$$

This is now in the form of the left-hand side of (3.6), which we can now apply to bound the upper tail behavior for an arbitrary $\mathbf{d}_i$, additionally using the upper bound on the expectation of a given degree,

$$\mathbb{P}\left(\mathbf{d}_i \geq \mathbb{E}[\mathbf{d}_i] + N\varepsilon_N\delta_N\right) \leq \exp\left(-\frac{\delta_N^2\varepsilon_N^2N^2}{2(Nmm_1\varepsilon_N + \frac{\delta_N\varepsilon_N}{3})}\right) \leq \exp\left(-\frac{\delta_N^2\varepsilon_NN}{2(mm_1+\delta_N/3)}\right).$$

We now let the relative gap between $N\varepsilon_N$ and $\log(N)$ to be $N\varepsilon_N > (2mm_1+2\delta_N/3)\frac{3\log N}{\delta_N^2}$, and use that to again bound above,

$$\exp\left(-\frac{\delta_N^2\varepsilon_NN}{2(mm_1+\delta_N/3)}\right) \leq \exp\left(-\frac{\delta_N^2}{2(mm_1+\delta_N/3)}\frac{3(2mm_1+2\delta_N/3)\log N}{\delta_N^2}\right) = \exp(-3\log N).$$

We can now extend this to $\max_{1\leq i\leq n}\{\mathbf{d}_i\}$ by applying a union bound, yielding the final behavior of the upper tail probability of this convergence,

$$\mathbb{P}\left(\max_{1\leq i\leq N}\{\mathbf{d}_i\} \geq \mathbb{E}[\mathbf{d}_i] + N\varepsilon_N\delta_N\right) \leq N\cdot\mathbb{P}\left(\mathbf{d}_i \geq \mathbb{E}[\mathbf{d}_i] + N\varepsilon_N\delta_N\right) \leq N\cdot N^{-3} = N^{-2}.$$

Thus, the upper tail probability is finitely bounded, and so now we look to prove the finiteness of the lower tail using the same process. To begin, we first rearrange the convergence to fit the lower tail Chernoff bound, noting that we are now assessing the undershoot of an arbitrary degree relative to $\mathbb{E}[\mathbf{d}_i]$:

$$\mathbb{P}\left(\frac{\max_{1\leq i\leq N}\{\mathbf{d}_i - \mathbb{E}[\mathbf{d}_i]\}}{N\varepsilon_N} \leq -\delta_N\right) = \mathbb{P}\left(\max_{1\leq i\leq N}\{\mathbf{d}_i\} \leq \mathbb{E}[\mathbf{d}_i] - N\varepsilon_N\delta_N\right).$$

Again, we use the gap between $N\varepsilon_N$ and $\log(N)$, letting $N\varepsilon_N > 3\log(N)(2mm_1/\delta^2)$. Using this inequality in (3.7) as well as a union bound as above to handle the max, we have then,

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq i\leq N}\{\mathbf{d}_i\} \leq \mathbb{E}[\mathbf{d}_i] - N\varepsilon_N\delta_N\right) \leq N\exp\left(-\frac{\delta_N^2N^2\varepsilon_N^2}{2Nmm_1\varepsilon_N}\right) &\leq N\exp\left(-\frac{\delta_N^2}{2mm_1}\frac{3\log N(2mm_1)}{\delta_N^2}\right)\\ &= N\exp(-3\log(N))\\ &= N^{-2}.\end{aligned}$$

Having now shown the finiteness of both tails, with the upper tail converging at a rate $\frac{6(mm_1+\delta_N/3)}{\delta_N^2}$ and the lower tail converging at a rate $\frac{6mm_1}{\delta_N^2}$, we can now make a state that the probability of $\frac{a_N}{N\varepsilon_N}$ exceeding an error δ is finite,

$$\mathbb{P}\left(\frac{\max_{1\leq i\leq N}|\mathbf{d}_i - \mathbb{E}[\mathbf{d}_i]|}{N\varepsilon_N} \geq \delta_N\right) \leq N\mathbb{P}(|\mathbf{d}_i - \mathbb{E}[\mathbf{d}_i]| \geq \delta_N N\varepsilon_N) \leq N^{-2}.$$

This then allows us to apply the Borel-Cantelli lemma, which then yields the final convergence that $\frac{a_N}{N\varepsilon_N} \overset{a.s.}{\to} 0$. The proof of $\frac{b_N}{N\varepsilon_N} \overset{a.s.}{\to} 1$ follows similarly, with the same application of the Chernoff bounds for each tail individually. Starting with the upper tail, we do a similar rearrangement in order to apply (3.6),

$$\mathbb{P}\left(\frac{b_N}{N\varepsilon_N} - 1 \geq \delta_N\right) = \mathbb{P}\left(\min_{1\leq i\leq N}\{\mathbf{d}_i\} \geq N\varepsilon_N\delta_N + N\varepsilon_N\right) \leq \sum_{i=1}^{N}\mathbb{P}\left(\mathbf{d}_i \geq N\varepsilon_N\delta_N + N\varepsilon_N\right).$$

Using the appropriate tail bound and having that $N\varepsilon_N > \log(N)\frac{6(mm_1+\delta_N/3)}{\delta_N^2}$, we arrive at a statement of finiteness for the upper tail,

$$\begin{aligned}\sum_{i=1}^{N}\mathbb{P}\left(\mathbf{d}_i \geq N\varepsilon_N\delta_N + N\varepsilon_N\right) &\leq \sum_{i=1}^{N}\exp\left(-\frac{N^2\varepsilon_N^2\delta_N^2}{2(Nmm_1\varepsilon_N + \delta_N\varepsilon_N N/3)}\right)\\ &\leq \sum_{i=1}^{N}\exp\left(-\frac{N\varepsilon_N\delta_N^2}{2(mm_1+\delta_N/3)}\frac{6\log N(mm_1+\delta_N/3)}{\delta_N^2}\right)\\ &= N\exp(-3\log(N))\\ &= N^{-2}.\end{aligned}$$

We repeat the process again for the lower tail using a gap between $N\varepsilon_N$ and $\log(N)$ of $3\left(\frac{2mm_1}{\delta_N^2}\right)$. Here we shorten the calculation for brevity, as it is the same as the prior calculations, again using (3.7),

$$\mathbb{P}\left(\frac{b_N}{N\varepsilon_N} - 1 \leq \delta_N\right) \leq \sum_{i=1}^{N}\mathbb{P}\left(\mathbf{d}_i \leq N\varepsilon_N - N\varepsilon_N\delta_N\right) \leq N\exp\left(-\frac{N\varepsilon_N\delta_N^2}{2mm_1}\right) \leq N^{-2}.$$

Thus, both tails are finite and converge to a value of 1, allowing us to apply the Borel-Cantelli lemma to arrive at our final statement for this convergence,

$$\mathbb{P}\left(\left|\frac{b_N}{N\varepsilon_N} - 1\right| \geq \delta_N\right) \leq N^{-2}, \text{ which implies } \frac{b_N}{N\varepsilon_N} \overset{a.s.}{\to} 1.$$

Finally, we prove the convergence of the total degree of $\mathbf{A}_N$,

$$\frac{\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j)}{N^2mm_1\varepsilon_N} \overset{a.s.}{\to} 1,$$

where we look first at the upper tail. We restate the event that we are looking to assess the probability of, rearranging as before to apply (3.6),

$$\mathbb{P}\left(\frac{\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j)}{\varepsilon_N mm_1N^2} - 1 \geq \delta_N\right) = \mathbb{P}\left(\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j) \geq \varepsilon_N mm_1N^2 + \delta_N\varepsilon_N mm_1N^2\right).$$

Applying the bound and having $N\varepsilon_N > 2\log(N)\frac{2(1+\delta_N/3)}{\delta_N^2Nm^2}$, we arrive at a finite upper bound,

$$\begin{aligned}\mathbb{P}\left(\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j) \geq \varepsilon_N mm_1N^2 + \delta_N\varepsilon_N mm_1N^2\right) &\leq \exp\left(-\frac{\delta_N^2\varepsilon_N^2(mm_1)^2N^4}{2\left(N^2mm_1\varepsilon_N + \delta_N\varepsilon_N mm_1N^2/3\right)}\right)\\ &\leq \exp\left(-\frac{\delta_N^2mm_1N}{2(1+\delta_N/3)}\frac{4\log(N)(1+\delta_N/3)}{\delta_N^2Nmm_1}\right)\\ &= N^{-2}.\end{aligned}$$

We do the same for the lower tail after rearrangement, applying (3.7) and having that $\varepsilon_N N > \frac{4\log(N)}{\delta_N^2Nmm_1}$, which again yields a finite bound,

$$\begin{aligned}\mathbb{P}\left(\sum_{i\neq j}\mathbf{A}_N(i,j) \leq \varepsilon_N mm_1N^2 - \delta_N\varepsilon_N mm_1N^2\right) &\leq \exp\left(-\frac{\delta_N^2\varepsilon_N^2(mm_1)^2N^4}{2\varepsilon_N mm_1N^2}\right)\\ &\leq \exp\left(-\frac{4\log(N)(\delta_N^2Nmm_1)}{2(\delta_N^2Nmm_1)}\right)\\ &= N^{-2}.\end{aligned}$$

Having now shown that both lower tails are finitely bounded, we can bound the entire probability from above:

$$\mathbb{P}\left(\left|\frac{\sum_{i\neq j}\mathbf{A}_N(i,j)}{\varepsilon_N mm_1N^2} - 1\right| \geq \delta\right) \leq N^{-2},$$

thus satisfying the requirements for the Borel-Cantelli lemma, which in turn completes the proof for the convergence of the total degree of $\mathbf{A}_N$:

$$\frac{\sum_{i\neq j}\mathbf{A}_N(i,j)}{\varepsilon_N m m_1 N^2} \xrightarrow{a.s.} 1.$$

As such, we have used the Chernoff bounds as defined in [45] and the Borel-Cantelli lemma to show that the three convergences outlined in the statement of this lemma indeed hold, almost surely. ■

We note that the degree properties have varying rates of convergence across their corresponding tails. These rates are a function of N, the size of the graph, but also of m, which is the behavior of r, and by extension, f, which implies that the well-behaved nature of the degrees of G_N are highly dependent on the underlying connection function. Having now equipped ourselves with useful tools with which to decouple the empirical degree matrix using a fixed degree matrix, we move to proving that perturbations of the summary matrices do not change the validity of the overall convergence to μ_f. We again note that this computation goes through for any general bounded and continuous f.

3.4 Approximation by Perturbation of $\mathbf{A}_N$

The structure of this section of the proof will be to show that through a series of convergences, slight perturbations of the matrix $\mathbf{A}_N$ do not change the convergent spectral measure. We will arrive at a final perturbed matrix such that we can then show convergence of the moments of the perturbed matrix to the moments of μ_f by a method of moments approach. We first recall from 3.2.1 that $\mathbf{B}_N \coloneqq \sqrt{N\varepsilon_N}(\mathbf{I}_N - \mathbf{L}_N)$, and so its decoupled counterpart for $\hat{\mathbf{L}}_N \coloneqq (N\varepsilon_N)^{-1}\mathbf{A}_N$ is,

$$\hat{\mathbf{B}}_N \coloneqq \sqrt{N\varepsilon_N}(\mathbf{I}_N - \hat{\mathbf{L}}_N). \tag{3.8}$$

The overarching triangle inequality, similar to that used in the homogeneous case, that we bound is as follows:

$$d_{BL}(\mathrm{ESD}\left(\mathbf{B}_N\right), \mu_f) \leq d_{BL}\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) + d_{BL}\left(\mathrm{ESD}\left(\hat{\mathbf{B}}_N\right), \mu_f\right). \tag{3.9}$$

To begin, we show that the spectra of the decoupled $\hat{\mathbf{B}}_N$ converges to the spectra of the original $\mathbf{B}_N$.

Lemma 3.4.1. *The spectra of the decoupled $\hat{\boldsymbol{B}}_N$ converges weakly in probability to the spectra of $\boldsymbol{B}_N$,*

$$d_{BL}\left(\mathrm{ESD}\left(\boldsymbol{B}_N\right), \mathrm{ESD}\left(\hat{\boldsymbol{B}}_N\right)\right) \to 0. \tag{3.10}$$

We will again use Lemma 2.2.1 to aid in bounding the distance between these two measures from above by showing that the difference in the normalized trace between these two matrices goes to 0.

Proof of Lemma 3.4.1:

To begin, we restate Lemma 2.2.1,

$$d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) \leq \frac{1}{N}\mathrm{Tr}\left((\mathbf{B}_N - \hat{\mathbf{B}}_N)^2\right). \tag{3.11}$$

As before, we will seek to show the right-hand side of the expression can be bounded such that it goes to 0. Rearranging, we again arrive at the trace of the squared difference between the empirical and uncoupled normalized Laplacian matrices,

$$\mathrm{tr}\left((\mathbf{B}_N - \hat{\mathbf{B}}_N)^2\right) = \mathrm{tr}\left(\left(\sqrt{N\varepsilon_N}\left(\hat{\mathbf{L}}_N - \mathbf{L}_N\right)\right)^2\right) = \frac{N\varepsilon_N}{N}\sum_{i,j=1}^{N}\left((\hat{\mathbf{L}}_N - \mathbf{L}_N)(i,j)\right)^2.$$

This again puts in a position to evaluate the difference between the normalized Laplacian and its uncoupled counterpart, using the inequalities established in Lemma 3.3.1 to bound this distance as a function of $N\varepsilon_N$. To begin, we re-express the two element-wise forms of the matrices in terms of the adjacency matrix, $\mathbf{A}_N$,

$$\left((\hat{\mathbf{L}}_N - \mathbf{L}_N)(i,j)\right)^2 = \left(\mathbf{A}_N\left(\frac{1}{\sqrt{\mathbf{d}_i}}\frac{1}{\sqrt{\mathbf{d}_j}} - (N\varepsilon_N)^{-1}\right)\right)^2(i,j).$$

However, as the statement in the latter inner bracket is in isolation of a specific set of entry indices (i,j), we can pull $\mathbf{A}_N$ out of the squared brackets, applying that the entries of $\mathbf{A}_N$ are binary to have that $\mathbf{A}_N^2 = \mathbf{A}_N$,

$$\left(\mathbf{A}_N(i,j)\left(\frac{1}{\sqrt{\mathbf{d}_i}}\frac{1}{\sqrt{\mathbf{d}_j}} - (N\varepsilon_N)^{-1}\right)\right)^2 = \mathbf{A}_N(i,j)\frac{\left(N\varepsilon_N - \sqrt{\mathbf{d}_i\mathbf{d}_j}\right)^2}{N^2\varepsilon_N^2\mathbf{d}_i\mathbf{d}_j}.$$

We can then bound the final quantity from above by applying the identity $|\sqrt{x}-\sqrt{y}|^2 \leq \frac{|x-y|^2}{\sqrt{y}}$, for $x = \frac{(N\varepsilon_N)^2}{N^2\varepsilon_N^2\mathbf{d}_i\mathbf{d}_j}$ and $y = \frac{(\sqrt{\mathbf{d}_i\mathbf{d}_j})^2}{N^2\varepsilon_N^2\mathbf{d}_i\mathbf{d}_j}$. Additionally, we use the absolute value to swap the order of the degree terms and $N^2\varepsilon_N^2$. However, working with $\mathbf{d}_i$ and $\mathbf{d}_j$ is still difficult, as they are still dependent on the behavior of $\mathbf{A}_N$, so we now seek to use the convergence inequalities from Lemma 3.3.1 to bound them with quantities independent of the adjacency. For notational convenience, we focus on the degree fraction in the next two calculations, holding $\mathbf{A}_N(i,j)$ to be fixed for now,

$$\frac{\left(N\varepsilon_N - \sqrt{\mathbf{d}_i\mathbf{d}_j}\right)^2}{N^2\varepsilon_N^2\mathbf{d}_i\mathbf{d}_j} \leq \frac{|\mathbf{d}_i\mathbf{d}_j - N^2\varepsilon_N^2|^2}{N^4\varepsilon_N^4\mathbf{d}_i\mathbf{d}_j} \leq \frac{|(N\varepsilon_N + a_N)(N\varepsilon_N + a_N) - N^2\varepsilon_N^2|^2}{N^4\varepsilon_N^4\mathbf{d}_i\mathbf{d}_j}.$$

The final inequality in the numerator above is an application of the first inequality from Lemma 3.3.1, which states that any degree deviation from the expected degree, $N\varepsilon_N$, is at most a_N. We then apply the second inequality in Lemma 3.3.1 to decouple the denominator, which states the minimum degree scales like $N\varepsilon_N$ as $N \to \infty$,

$$\frac{|(N\varepsilon_N + a_N)(N\varepsilon_N + a_N) - N^2\varepsilon_N^2|^2}{N^4\varepsilon_N^4\mathbf{d}_i\mathbf{d}_j} \leq \frac{|N^2\varepsilon_N^2 + 2N\varepsilon_N a_N + a_N^2 - N^2\varepsilon_N^2|^2}{N^4\varepsilon_N^4 b_N^2} \approx \frac{4N^2\varepsilon_N^2 a_N^2 + 2a_N^4}{N^6\varepsilon_N^6}.$$

We have thus arrived at a bounding constant independent of any specific matrix index (i,j), which we can now plug back into the normalized trace stated above:

$$\frac{N\varepsilon_N}{N}\sum_{i,j=1}^{N}\left((\hat{\mathbf{L}}_N - \mathbf{L}_N)(i,j)\right)^2 \leq \frac{N\varepsilon_N}{N}\frac{4N^2\varepsilon_N^2 a_N^2 + 2a_N^4}{N^6\varepsilon_N^6}\sum_{i,j=1}^{N}\mathbf{A}_N(i,j).$$

However, as there are no self-loops in G_N, the diagonal terms of $\mathbf{A}_N(i,j)$ where $i=j$ do not contribute the final summation, so we can more specifically index the sum as $\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j)$.

$$\frac{N\varepsilon_N}{N}\frac{4N^2\varepsilon_N^2 a_N^2 + 2a_N^4}{N^6\varepsilon_N^6}\sum_{i,j=1}^{N}\mathbf{A}_N(i,j) = \frac{4N^2\varepsilon_N^2 a_N^2 + 2a_N^4}{N^6\varepsilon_N^5}\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j).$$

This then puts us in a position to apply the final convergence concerning the total degree of the graph from Lemma 3.3.1, $\frac{\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j)}{N^2 m m_1 \varepsilon_N} \xrightarrow{a.s.} 1$,

$$\frac{4N^2\varepsilon_N^2 a_N^2 + 2a_N^4}{N^6\varepsilon_N^5}\sum_{1\leq i\neq j\leq N}\mathbf{A}_N(i,j) \approx \frac{4N^2\varepsilon_N^2 a_N^2 + 2a_N^4}{N^6\varepsilon_N^5}N^2 m m_1\varepsilon_N = \frac{m m_1(4N^2\varepsilon_N^2 + 2a_N^2)}{N^2\varepsilon_N^2}\frac{a_N^2}{N^2\varepsilon_N^2}.$$

However, as established in the first convergence of Lemma 3.3.1, $\frac{a_N}{N\varepsilon_N} \to 0$, and thus this entire quantity goes to 0, almost surely, $N \to \infty$. Therefore, we have proven that,

$$d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{B}_N\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) \leq \mathrm{tr}\left((\mathbf{B}_N - \hat{\mathbf{B}}_N)^2\right) \leq \frac{m m_1(4N^2\varepsilon_N^2 + 2a_N^2)}{N^2\varepsilon_N^2}\frac{a_N^2}{N^2\varepsilon_N^2} \xrightarrow{a.s.} 0,$$

which satisfies the initial lemma, thus stating that the use of the decoupled $\hat{\mathbf{B}}_N$ in place of $\mathbf{B}_N$ does not change the spectral behavior, as $N \to \infty$. ■

Having now decoupled the normalized Laplacian from the degree matrix, we now show that the spectra of $\hat{\mathbf{B}}_N$ is unaffected by another perturbation, namely by instead using the spectral measure of the normalized and centered adjacency matrix,

$$\mathbf{A}_N^0 \coloneqq \frac{\mathbf{A}_N - \mathbb{E}[\mathbf{A}_N]}{\sqrt{N\varepsilon_N(1-\varepsilon_N)}}. \tag{3.12}$$

Lemma 3.4.2. *The distance between the spectra of* $\boldsymbol{A}_N^0 \coloneqq \frac{\boldsymbol{A}_N - \mathbb{E}[\boldsymbol{A}_N]}{\sqrt{N\varepsilon_N}}$ *and* $\hat{\boldsymbol{B}}_N \coloneqq \sqrt{\frac{N\varepsilon_N}{1-\varepsilon}}(\boldsymbol{I}_N - \hat{\boldsymbol{L}}_N)$ *converges to 0 weakly in probability,*

$$\lim_{N\to\infty} d_{BL}\left(\mathrm{ESD}\left(\boldsymbol{A}_N^0\right), \mathrm{ESD}\left(\hat{\boldsymbol{B}}_N\right)\right) = 0. \tag{3.13}$$

The structure of this proof relies on setting up Lemma 2.2.1 to bound the difference in the normalized trace of the two matrices and then using the definitions of the two matrices to arrive at a result which goes to 0 as $N\varepsilon_N \to \infty$.

Proof of Lemma 3.4.2:

To begin, we set up Lemma 2.2.1 for $\mathbf{A}_N^0$ and $\hat{\mathbf{B}}_N$,

$$d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{A}_N^0\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) \leq \mathrm{tr}\left((\mathbf{A}_N^0 - \hat{\mathbf{B}}_N)^2\right).$$

Note that $\hat{\mathbf{B}}_N$ can be expressed as the following, for easier use later on:

$$\hat{\mathbf{B}}_N \coloneqq \sqrt{N\varepsilon_N}\left(\mathbf{I}_N - \left(\mathbf{I}_N - \frac{\mathbf{A}_N}{N\varepsilon_N}\right)\right) = \frac{\mathbf{A}_N}{\sqrt{N\varepsilon_N}}.$$

We now look to simplify the quantity in the normalized trace in order to arrive at a quantity more conducive to being worked as a summation. Using that $\varepsilon_N \to 0$,

$$\mathbf{A}_N^0 - \hat{\mathbf{B}}_N = \frac{\mathbf{A}_N - \mathbb{E}[\mathbf{A}_N]}{\sqrt{N\varepsilon_N(1-\varepsilon_N)}} - \frac{\mathbf{A}_N}{\sqrt{N\varepsilon_N}} \approx -\frac{\mathbb{E}[\mathbf{A}_N]}{\sqrt{N\varepsilon_N}}.$$

Substituting this condensed definition back into the original statement from the right-hand side of the initial Hoffman-Wielandt inequality and expanding the definition of the normalized trace, we arrive at the following statement:

$$\mathrm{tr}\left((\mathbf{A}_N^0 - \hat{\mathbf{B}}_N)^2\right) = \frac{1}{N}\sum_{i,j=1}^{N}\left[\frac{\mathbb{E}[\mathbf{A}_N(i,j)]}{\sqrt{N\varepsilon_N}}\right]^2 = \frac{\sum_{i=1}^N\sum_{j\neq i}^N \mathbb{E}[\mathbf{A}_N(i,j)]^2}{N^2\varepsilon_N}.$$

The summation is split into two, with the outer summation accounting for all i-th row in the matrix, and the inner summation accounting for all j-th column terms, omitting the scenario where $i = j$, as those entries are 0, and thus non-contributing. Substituting in that $\mathbb{E}[\mathbf{A}_N(i,j)] = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)$ by virtue of being a Bernoulli random variable, we can then use the approximation given in (3.3),

$$\frac{\sum_{i=1}^N\sum_{j\neq i}^N \mathbb{E}[\mathbf{A}_N(i,j)]^2}{N^2\varepsilon_N} = \frac{\sum_{i=1}^N\sum_{j\neq i}^N \varepsilon_N^2 f\left(\frac{i}{N}, \frac{j}{N}\right)^2}{N^2\varepsilon_N} \leq \frac{\varepsilon_N^2(1+o(1))m^4}{N^2\varepsilon_N} \approx \varepsilon_N m^4(1+o(1)).$$

Of course, as $N\varepsilon_N \to \infty$, as defined in our setting, and $N \to \infty$, then we can take that $\varepsilon_N \to 0$ while m being bounded, which then allows us to take this final quantity to go to 0,

$$d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{A}_N^0\right), \mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right) \leq \mathrm{tr}\left((\mathbf{A}_N^0 - \hat{\mathbf{B}}_N)^2\right) = \varepsilon_N m^4(1+o(1)) \to 0. \quad \blacksquare$$

From this, we have shown that yet another small change to the underlying matrix does not fundamentally change the resulting behavior of the spectral measure. We continue with this approach by showing that a Gaussianized $\mathbf{A}_N^0$, which we will denote as $\mathbf{A}_N^g$, is equivalent in spectral behavior to the above defined $\mathbf{A}_N^0$. For $X_{ij} \overset{iid}{\sim} N(0,1)$ being a class of standard normal random variables, $\mathbf{A}_N^g$ is defined as,

$$\mathbf{A}_N^g(i,j) = \begin{cases} \sqrt{\frac{\varepsilon_N f(\frac{i}{N},\frac{j}{N})(1-\varepsilon_N f(\frac{i}{N},\frac{j}{N}))}{N(1-\varepsilon_N)}}X_{ij}, & i < j \\ 0, & i = j \end{cases} \tag{3.14}$$

Gaussianization allows us to characterize the behavior of the variation of the expected values of individual entries of $\mathbf{A}_N$, critically showing that assuming a matrix with finite variance does not change the underlying spectral properties. This then gives way to the last perturbation prior to defining the final convergent measure, μ_f. This is an application of the general result from [8], in using a matrix with standard normal entries to establish a point from which the variance statement of $\mathbf{A}_N^g$ can be further simplified. First, in order to validate the use of $\mathbf{A}_N^g$, we will use the Stieltjes transform of a measure as our analytical tool. This transform uniquely characterizes a measure [4], μ, and for $z \in \mathbb{C}^+, \Im(z) > 0$,

$$S_\mu(z) \coloneqq \int_{\mathbb{R}} \frac{1}{x-z}\mu(dz). \tag{3.15}$$

Lemma 3.4.3. *Let $h : \mathbb{R} \to \mathbb{R}$ be a function which is thrice differentiable:*

$$\max_{0\le j\le 3} \sup_{x\in\mathbb{R}} |h^{(j)}(x)| < \infty$$

Thus for $S_{\boldsymbol{A}_N^0}(z)$ and $S_{\boldsymbol{A}_N^g}$ being the Stieltjes transforms of $\boldsymbol{A}_N^0$ and $\boldsymbol{A}_N^g$, respectively,

$$\lim_{N\to\infty} |\mathbb{E}[h(\Re(S_{\boldsymbol{A}_N^0}(z))) - h(\Re(S_{\boldsymbol{A}_N^g}(z)))]| = 0, \tag{3.16}$$

$$\lim_{N\to\infty} |\mathbb{E}[h(\Im(S_{\boldsymbol{A}_N^0}(z))) - h(\Im(S_{\boldsymbol{A}_N^g}(z)))]| = 0. \tag{3.17}$$

The central claim of the lemma is that the distance in expectation between the transforms of the measures of the two matrices goes to 0 as a function of h. It is also important to note that the normalized trace of the resolvent of a matrix $\mathbf{M}$, $\mathrm{tr}((\mathbf{M}-z\mathbf{I})^{-1})$, converges in probability to the Stieltjes transform of its spectral measure [35], and thus,

$$S_{\mathrm{ESD}(\mathbf{M})}(z) = \frac{1}{N}\,\mathrm{Tr}((\mathbf{M}-z\mathbf{I})^{-1}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i - z}.$$

As such, the proof will seek to show that this quantity converges between the two matrices, in both the real and imaginary planes, via Stieltjes transform. The motivation behind the use of Stieltjes transform as the analytical tool of choice is that first, the Stieltjes transform of a measure is analytic on $\mathbb{C}^+$, and that secondly, by showing that two the transform of the two measures converge in the limit $\forall z \in \mathbb{C}^+$, we can conclude weak convergence of the two measures [43]. Thus, the Stieltjes transform of a measure a powerful tool, as it is a more convenient object to work than the measure itself.

Proof of Lemma 3.4.3:

We introduce a function, $\phi : \mathbb{R}^n \to \mathbb{C}, n = \frac{N(N-1)}{2}$, as shorthand for the resolvent of the centered and Gaussianized matrices, respectively,

$$\phi^0(\mathbf{x}) = \mathrm{tr}(\mathbf{A}_N^0(\mathbf{x}) - z\mathbf{I})^{-1}$$
$$\phi^g(\mathbf{x}) = \mathrm{tr}(\mathbf{A}_N^g(\mathbf{x}) - z\mathbf{I})^{-1}.$$

These can be intuitively be considered to be the distance of a complex variable, z, from the spectra of $\mathbf{A}_N^0$ and $\mathbf{A}_N^g$, with $\mathbf{x} \coloneqq \{x_{ij}\}_{1\le i\le j\le N}$ referring to a matrix such that $\mathbf{A}_N(x_{ij})$ yields the (i,j)-th element of $\mathbf{A}_N$. Thus, we look to evaluate the behavior of these deviations by evaluating its derivatives, which is a similar approach to the proof approach of using the Lindeberg condition in proving the Central Limit Theorem. This is the same approach as in [8], where the result involves taking the derivative of ϕ. We must first define the upper bound behavior of the first three derivatives of ϕ for a general ϕ, irrespective of being ϕ^0 or ϕ^g. We start with the first derivative, $\frac{\partial}{\partial x_{ij}}\phi$, reminding that the derivative of a matrix is its element-wise derivative,

$$\frac{\partial\phi}{\partial x_{ij}} = \frac{\partial}{\partial x_{ij}}\mathrm{tr}\left((\mathbf{A}_N(x) - z\mathbf{I})^{-1}\right) = \mathrm{tr}\left(\frac{\partial}{\partial x_{ij}}(\mathbf{A}_N(x) - z\mathbf{I})^{-1}\right) = -\mathrm{tr}\left(\frac{\partial\mathbf{A}_N(x)}{\partial x_{ij}}(\mathbf{A}_N(x) - z\mathbf{I})^{-2}\right).$$

The final derivation is done by applying chain rule to the derivative within the trace. Let $K = (\mathbf{A}_N - z\mathbf{I})^{-1}$, for notational simplicity. The first derivative of ϕ is thus,

$$\frac{\partial\phi(x)}{\partial x_{ij}} = -\mathrm{tr}\left(\frac{\partial\mathbf{A}_N(x)}{\partial x_{ij}}K^2\right).$$

Note that the eigenvalues of $(\mathbf{A}_N - z\mathbf{I})^{-1}$ are upper-bounded by $|\Im(z)|^{-1}$, for $\mathbf{A}_N$ being binary, and the derivative of $\mathbf{A}_N$ has eigenvalues bounded $N^{-1/2}$ given that $\sqrt{N}$ dominates the denominator of the scaling constant of both $\mathbf{A}_N^g$ and $\mathbf{A}_N^0$ for $N\varepsilon_N \to \infty$. These are then applied through the following identity for two matrices $\mathbf{B}, \mathbf{C}$ of dimension N, where $\max\{\|\mathbf{BC}\|, \|\mathbf{CB}\|\} \le \max_{1\le i\le N}|\lambda_i|\|\mathbf{C}\|$, λ_i being the eigenvalues of $\mathbf{B}$. In conjunction with the Cauchy-Schwarz inequality for matrices, this then yields,

$$\frac{1}{N}\left\|\mathrm{Tr}\left(\frac{\partial\mathbf{A}_N(x)}{\partial x_{ij}}(\mathbf{A}_N(x) - z\mathbf{I})^{-2}\right)\right\|_\infty \le \frac{1}{N}\frac{1}{\sqrt{N}}\frac{2}{|\Im(z)|^{-2}} = \frac{2}{N^{3/2}|\Im(z)|^{-2}},$$

This same approach can be taken for the higher order derivatives, first calculating the derivative and then bounding from above using its infinity norm. We thus move to calculate the second derivative as follows,

$$\frac{\partial^2\phi(x)}{\partial x_{ij}^2} = \frac{\partial^2}{\partial x_{ij}^2}\text{tr}(K^1) = -\frac{1}{N}\frac{\partial}{\partial x_{ij}}\text{Tr}\left(\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right) = -\frac{2}{N}\text{Tr}\left(\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\left[-2K^2\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\right]\right),$$

using a similar expansion via chain rule as done in the first derivative. We can now bound this from above using its infinity norm and the Cauchy-Schwarz inequality:

$$\left\|\frac{\partial^2\phi(x)}{\partial x_{ij}^2}\right\|_\infty = \left\|\frac{2}{N}\text{Tr}\left(\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right)\right\|_\infty \leq \frac{2}{N}\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\right\|_\infty\left\|K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right\|_\infty.$$

As established, we can bound $\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\right\|_\infty \leq N^{-1/2}$ and $\|K\|_\infty \leq |\Im(z)|^{-1}$, and by applying these individual bounds and the properties of the infinity norm of matrices, we get that,

$$\frac{2}{N}\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\right\|_\infty\left\|K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right\|_\infty \leq \frac{4}{N}\frac{1}{\sqrt{N}}\frac{1}{\sqrt{N}}\frac{1}{|\Im(z)|^3} = \frac{4}{N^2|\Im(z)|^3},$$

again with a constant 2 contributed by the symmetry of the matrix. The upper bound of the final derivative can then be calculated using the same approach. We start by plugging in the previous result for the second derivative and then applying the product rule,

$$\frac{\partial^3\phi(x)}{\partial x_{ij}^3} = \frac{2}{N}\text{Tr}\left(\frac{\partial}{\partial x_{ij}}\left[\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\right]\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2 + \frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial}{\partial x_{ij}}\left[\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right]\right),$$

after which via applying chain rule to the various components in the brackets we have,

$$\frac{2}{N}\text{Tr}\left(-3\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right) = -\frac{6}{N}\text{Tr}\left(\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right).$$

We then finally assess its infinity norm and apply the previously mentioned bounds over the individual components, first splitting the larger expression into individual component norms,

$$\begin{aligned}\left\|\frac{\partial^3\phi(x)}{\partial x_{ij}^3}\right\|_\infty &= \left\|\frac{6}{N}\text{Tr}\left(\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right)\right\|_\infty \\ &\leq \frac{6}{N}\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\right\|_\infty\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\right\|_\infty\|K\|_\infty\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right\|_\infty,\end{aligned}$$

and then bounding the individual components accordingly,

$$\begin{aligned}\frac{6}{N}\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\right\|_\infty\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}\right\|_\infty\|K\|_\infty\left\|\frac{\partial \mathbf{A}_N(x)}{\partial x_{ij}}K^2\right\|_\infty &\leq \frac{6}{N}\frac{1}{\sqrt{N}}\frac{1}{\sqrt{N}}|\Im(z)|^{-1}|\Im(z)|^{-3}\frac{1}{\sqrt{N}}2 \\ &= \frac{12}{N^{5/2}}|\Im(z)|^{-4}.\end{aligned}$$

Having now bounded all three derivatives, we provide some summary values across these derivatives, which in themselves act as bounds on the behavior across higher order derivatives,

$$\lambda_1(\phi) = \left\| \frac{\partial^2}{\partial x_{ij}^2}\phi \right\|_\infty \leq \frac{2}{N^{3/2}|\Im(z)|^{-2}},$$
$$\lambda_2(\phi) = \sup\left\{ \left\| \frac{\partial}{\partial x_{ij}}\phi \right\|_\infty^2, \left\| \frac{\partial^2}{\partial x_{ij}^2}\phi \right\|_\infty \right\} \leq \frac{4\max\{\Im(z)^{-4}, \Im(z)^{-3}\}}{N^2},$$
$$\lambda_3(\phi) = \sup\left\{ \left\| \frac{\partial}{\partial x_{ij}}\phi \right\|_\infty^3, \left\| \frac{\partial^2}{\partial x_{ij}^2}\phi \right\|_\infty^2, \left\| \frac{\partial^3}{\partial x_{ij}^3}\phi \right\|_\infty \right\} \leq \frac{12\max\{\Im(z)^{-6}, \Im(z)^{-5}\}}{N^{5/2}}.$$

To aid in bounding the distance in expectation between the two matrices, we introduce two more terms for notational clarity that represent the real components of the normalized trace of the two respective resolvents,

$$U = \Re\left(\frac{1}{N}\operatorname{Tr}(\mathbf{A}_N^0 - z\mathbf{I})^{-1}\right) = \Re(S_{\mathbf{A}_N^0}(z))$$
$$V = \Re\left(\frac{1}{N}\operatorname{Tr}(\mathbf{A}_N^g - z\mathbf{I})^{-1}\right) = \Re(S_{\mathbf{A}_N^g}(z)).$$

We are now in a position to apply Theorem 1.1 from [8], which provides a bound for the maximum difference between these two quantities, with $\{\lambda_r(\phi)\}_{1\leq r\leq 3}$ being the maximum influence of a single point of difference on ϕ. This is thus conditional on λ_2 and λ_3 being sufficiently small, implying that the higher order fluctuations are well-behaved, similar to the requirement for Lindeberg's argument for proof of the Central Limit Theorem. For $C_1(h) = \|h'\|_\infty + \|h''\|_\infty$, $C_2(h) = \frac{1}{6}\|g'\|_\infty + \frac{1}{2}\|g''\|_\infty + \frac{1}{6}\|g'''\|_\infty$, and $L > 0$, the bounding value is thus,

$$\begin{aligned}|\mathbb{E}[h(U) - h(V)]| \leq & C_1(h)\lambda_2(\phi)\sum_{i\neq j}^N \left[\mathbb{E}[\mathbf{A}_N^0(i,j)^2\mathbb{1}_{\{|\mathbf{A}_N^0(i,j)|>L\}}] + \mathbb{E}[\mathbf{A}_N^g(i,j)^2\mathbb{1}_{\{|\mathbf{A}_N^g(i,j)|>L\}}]\right] \\ & + C_2(h)\lambda_3(\phi)\sum_{i\neq j}^N \left[\mathbb{E}[\mathbf{A}_N^0(i,j)^3\mathbb{1}_{\{|\mathbf{A}_N^0(i,j)|\leq L\}}] + \mathbb{E}[\mathbf{A}_N^g(i,j)^3\mathbb{1}_{\{|\mathbf{A}_N^g(i,j)|\leq L\}}]\right].\end{aligned}$$

We thus look to bound the right-hand side of this statement to 0. To do this, we approach $\mathbf{A}_N^0$ and $\mathbf{A}_N^g$ independently. We start by seeking to bound $\mathbb{E}[\mathbf{A}_N^0(i,j)^2\mathbb{1}_{\{|\mathbf{A}_N^0(i,j)|>L\}}]$, noting that the expectation of an indicator function is interpretable as a probability via the Dirac delta function definition of a probability density. Thus, the latter half of the term becomes a centered Bernoulli random variable, and as such via the definition of $\mathbf{A}_N^0$ can take at most a value of 2. So for $L > 2$, this quantity goes to 0,

$$\sum_{i\neq j}^N \mathbb{E}[\mathbf{A}_N^0(i,j)^2\mathbb{1}_{\{|\mathbf{A}_N^0(i,j)|>L\}}] = 0.$$

We now move to the second term, $\mathbb{E}[\mathbf{A}_N^g(i,j)^2\mathbb{1}_{\{|\mathbf{A}_N^g(i,j)|>L\}}]$, which we can approach using the Cauchy-Schwarz inequality and Markov's inequality. We first use Cauchy-Schwarz to decompose the statement into an expectation and a probability,

$$\sum_{i\neq j}^N \left[\mathbb{E}[\mathbf{A}_N^g(i.j)^2\mathbb{1}_{\{|\mathbf{A}_N^g(i,j)|>L\}}]\right] \leq \sum_{i\neq j}^N \sqrt{\mathbb{E}[\mathbf{A}_N^g(i,j)^4]\mathbb{P}(|\mathbf{A}_N^g(i,j)| > L)},$$

from which we evaluate the expectation by using the definition of $\mathbf{A}_N^g$ in (3.14),

$$\mathbb{E}[\mathbf{A}_N^g(i,j)^4] = \frac{\mathbb{E}\left[\left(\varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\left(1 - \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\right)X_{ij}\right)^2\right]}{N^2(1-\varepsilon_N)^2}.$$

Taking $\varepsilon_N \to 0$ for $N\varepsilon_N \to \infty$ as $N \to \infty$ and because f is bounded, we can then simplify this expression further, as $(1 - \varepsilon_N) \to 1$ and similarly $(1 - \varepsilon_N f(\cdot)) \to 1$,

$$\frac{\mathbb{E}\left[\left(\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\left(1-\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\right)\right)^2\right]}{N^2(1-\varepsilon_N)^2} = \frac{\mathbb{E}\left[\left(\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\right)^2\right]}{N^2} = O(N^{-2}).$$

This then yields that the expectation of $\mathbf{A}_N^g$ goes to 0 at a rate of $O(N^{-2})$. Now, to evaluate the probability, we use Markov's inequality,

$$\mathbb{P}(|\mathbf{A}_N^g(i,j)| > L) = \mathbb{P}\left(\left|\sqrt{\frac{\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\left(1-\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)\right)}{N(1-\varepsilon_N)}}X_{ij}\right| > L\right) \leq \frac{\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)}{NL^2}.$$

Thus, we again use that f is bounded and take $N \to \infty$ to have that this term is $O(N^{-1}L^{-2})$. Combining these two quantities, we have that the Gaussianized term in the first component of the summation is bounded,

$$\sqrt{\mathbb{E}[\mathbf{A}_N^g(i,j)^4]\mathbb{P}(|\mathbf{A}_N^g(i,j)| > L)} \leq \sqrt{O(N^{-2})O(N^{-1}L^{-2})} = O(N^{-3/2}L^{-1}).$$

In order to evaluate the higher moments in the latter half of the inequality, we make use of the fact that for a random variable X, $\mathbb{E}\left[|X|^3\mathbb{1}_{\{|X|\leq L\}}\right] \leq L\mathbb{E}\left[X^2\right]$, which gives us that,

$$\sum_{i\neq j}^{N}\left[\mathbb{E}[\mathbf{A}_N^0(i,j)^3\mathbb{1}_{\{|\mathbf{A}_N^0(i,j)|\leq L\}}] + \mathbb{E}[\mathbf{A}_N^g(i,j)^3\mathbb{1}_{\{|\mathbf{A}_N^g(i,j)|\leq L\}}]\right] \leq L\sum_{i\neq j}^{N}\left[\mathbb{E}[\mathbf{A}_N^0(i,j)^2] + \mathbb{E}[\mathbf{A}_N^g(i,j)^2]\right].$$

Observe however, that both $\mathrm{Var}[\mathbf{A}_N^0(i,j)]$ and $\mathrm{Var}[\mathbf{A}_N^0(i,j)]$ are less than $\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right)$, by definition. We can thus upper bound this as follows,

$$L\sum_{i\neq j}^{N}\left[\mathbb{E}[\mathbf{A}_N^0(i,j)^2] + \mathbb{E}[\mathbf{A}_N^g(i,j)^2]\right] \leq 2L\sum_{i\neq j}^{N}\varepsilon_N f\left(\frac{i}{N},\frac{j}{N}\right) = O(LN).$$

Thus, for a sufficient L, we can then see that the right-hand side of the initial inequality indeed goes to 0,

$$\begin{aligned}
&C_1(h)\lambda_2(\phi)\sum_{i\neq j}^{N}\left[\mathbb{E}[\mathbf{A}_N^0(i,j)^2\mathbb{1}_{\{|\mathbf{A}_N^0(i,j)|>L\}}] + \mathbb{E}[\mathbf{A}_N^g(i,j)^2\mathbb{1}_{\{|\mathbf{A}_N^g(i,j)|>L\}}]\right]\\
&+ C_2(h)\lambda_3(\phi)\sum_{i\neq j}^{N}\left[\mathbb{E}[\mathbf{A}_N^0(i,j)^3\mathbb{1}_{\{|\mathbf{A}_N^0(i,j)|\leq L\}}] + \mathbb{E}[\mathbf{A}_N^g(i,j)^3\mathbb{1}_{\{|\mathbf{A}_N^g(i,j)|\leq L\}}]\right]\\
=&C_1(h)O(N^{-2})O(N^{-3/2}L^{-1}) + C_2(h)O(N^{-5/2})O(LN)\\
=&C_1(h)O(N^{-7/2}) + C_2(h)O(N^{-3/2}L).
\end{aligned}$$

As h and its three derivates are bounded by definition, C_1 and C_2 are by extension also bounded. We can also select an L such that $N^{-3/2}L \to 0$. Thus, this entire quantity goes to 0 at their respective rates dependent on N, yielding then that the difference in expectation between the two test functions parameterized by U and V are is indeed vanishing,

$$|\mathbb{E}[h(U) - h(V)]| \to 0, N \to \infty.$$

This thus satisfies the first statement set out at the beginning of the proof of the lemma for the real plane. To complete the proof, we can see that for the statement for the imaginary part,

$$\lim_{N\to\infty}|\mathbb{E}[h(\Im(S_{\mathbf{A}_N^0}(z))) - h(\Im(S_{\mathbf{A}_N^g}(z)))]| = 0,$$

that the same methods can be applied directly, as the deviations in the imaginary follow the same behavior as in the real. Thus, by extending the above steps to this statement, we can verify the above.

Having shown that the real and imaginary components of $S_{\mathbf{A}_n^H}(z)$ and $S_{\mathbf{A}_N^0}(z)$ are convergent, and given that the Stieltje's transform of these matrices uniquely characterize their spectral measures, we

can conclude that given these two individual convergences, the spectra of these two matrices do indeed converge, as $N \to \infty$. ■

We have thus shown that the spectra of the Gaussianized adjacency matrix, $\mathbf{A}_N^g$, is a valid approximation for the spectra of the centered and scaled adjacency matrix $\mathbf{A}_N^0$. Continuing along this track, we now look to show that the behavior of the spectra is dominated by its leading order variance, still operating using a Gaussianized $\mathbf{A}_N$.

Lemma 3.4.4. *Let $\overline{\boldsymbol{A}}_N$ be an $N \times N$ matrix defined as:*

$$\overline{\boldsymbol{A}}_N(i,j) = \sqrt{\frac{f\left(\frac{i}{N}, \frac{j}{N}\right)}{N}} X_{i\wedge j, i\vee j}, 1 \leq i, j, \leq N \tag{3.18}$$

The spectral measure of this matrix and that of the previously defined Gaussianized matrix $\boldsymbol{A}_N^g$ converge to one another,

$$\lim_{N\to\infty} d_{BL}\left(\mathrm{ESD}\left(\boldsymbol{A}_N^g\right), \mathrm{ESD}\left(\overline{\boldsymbol{A}}_N\right)\right) = 0, \quad \textit{in probability.} \tag{3.19}$$

The proof of this lemma follows the structure of setting up Lemma 2.2.1, which we will then use to bound the distance in expectation between these two measures by expanding the normalized trace. As both matrices are scaled by standard normal variables X_{ij}, they each have a mean of 0 for all entries. We thus focus on the behavior of their respective variances.

Proof of Lemma 3.4.4:

To begin, we apply Lemma 2.2.1 to show the closeness of these measures. The inequality is set up as follows given that the differences between the two matrices are independent, thus allowing us to move the expectation within the summation,

$$\mathbb{E}\left[d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{A}_N^g\right), \mathrm{ESD}\left(\overline{\mathbf{A}}_N\right)\right)\right] \leq \mathbb{E}\left[\mathrm{tr}\left((\mathbf{A}_N^g - \overline{\mathbf{A}}_N)^2\right)\right] = \frac{1}{N}\sum_{i=1}^N \sum_{j\neq i}^N \mathbb{E}\left[(\mathbf{A}_N^g - \overline{\mathbf{A}}_N)^2(i,j)\right].$$

We consider only the entries where $i \neq j$, as the diagonal of both matrices are defined to be 0, so they do not contribute to the resulting final values. We thus inspect the difference, $\mathbf{A}_N^g - \overline{\mathbf{A}}_N$,

$$\begin{aligned}(\mathbf{A}_N^g - \overline{\mathbf{A}}_N)(i,j) &= \left(\sqrt{\frac{\varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\left(1 - \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\right)}{N(1-\varepsilon_N)}} X_{ij} - \sqrt{\frac{f\left(\frac{i}{N}, \frac{j}{N}\right)}{N}} X_{ij}\right)^2 \\ &= \frac{X_{ij}^2 f\left(\frac{i}{N}, \frac{j}{N}\right)}{N}\left(\sqrt{\frac{\varepsilon_N\left(1 - \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\right)}{1-\varepsilon_N}} - 1\right)^2.\end{aligned}$$

Plugging this quantity back into the initial expectation, $\mathbb{E}[X_{ij}^2] = 1$ as X_{ij} is a standard normal variable, and so we can remove the expectation as the remaining terms are dependent on the indexes i and j,

$$\frac{1}{N}\sum_{i=1}^N \sum_{j\neq i}^N \mathbb{E}\left[(\mathbf{A}_N^g - \overline{\mathbf{A}}_N)^2\right] = \frac{1}{N^2}\sum_{i=1}^N \sum_{j\neq i}^N f\left(\frac{i}{N}, \frac{j}{N}\right)\left(\sqrt{\frac{\varepsilon_N\left(1 - \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\right)}{1-\varepsilon_N}} - 1\right)^2.$$

Taking $\varepsilon_N \to 0$ and noting that for x being bounded as a probability that $(\sqrt{1-x} - 1)^2 \leq |x|$, we can simplify this statement further,

$$\begin{aligned}&\frac{1}{N^2}\sum_{i=1}^N \sum_{j\neq i}^N f\left(\frac{i}{N}, \frac{j}{N}\right)\left(\sqrt{\frac{\varepsilon_N\left(1 - \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\right)}{1-\varepsilon_N}} - 1\right)^2 \\ \approx &\frac{1}{N^2}\sum_{i=1}^N \sum_{j\neq i}^N f\left(\frac{i}{N}, \frac{j}{N}\right)\left(\sqrt{\varepsilon_N\left(1 - \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)\right)} - 1\right)^2 \\ \leq &\frac{\varepsilon_N}{N^2}\sum_{i=1}^N \sum_{j\neq i}^N f\left(\frac{i}{N}, \frac{j}{N}\right)\left(\left|f\left(\frac{i}{N}, \frac{j}{N}\right)\right|\right) = O(\varepsilon_N).\end{aligned}$$

Given that $f\left(\frac{i}{N}, \frac{j}{N}\right)$ is bounded, the final quantity then goes to 0 as $\varepsilon_N \to 0$ while $N \to \infty$. Finally, by Lemma 2.2.1, we have shown that:

$$\mathbb{E}\left[d_{BL}^3\left(\mathrm{ESD}\left(\mathbf{A}_N^g\right), \mathrm{ESD}\left(\overline{\mathbf{A}}_N\right)\right)\right] \leq \mathbb{E}\left[\mathrm{tr}(\mathbf{A}_N^g - \overline{\mathbf{A}}_N)^2)\right] \to 0,$$

and by Markov's inequality, for a $\delta > 0$,

$$\mathbb{P}\left(d_{BL}\left(\mathrm{ESD}\left(\mathbf{A}_N^g\right), \mathrm{ESD}\left(\overline{\mathbf{A}}_N\right)\right) > \delta\right) \leq \frac{\mathbb{E}[d_{BL}\left(\mathrm{ESD}\left(\mathbf{A}_N^g\right), \mathrm{ESD}\left(\overline{\mathbf{A}}_N\right)\right)]}{\delta^2} \to 0. \quad \blacksquare$$

Thus, by focusing only on the leading order variance in place of $\mathbf{A}_N^g$, we have shown that there is no impact on the resulting spectra. $\overline{\mathbf{A}}_N$ is now the final matrix with which we will show convergence to μ_f. Prior to showing this convergence, we will describe the characteristics of μ_f in more detail and then make the traditional combinatorial arguments for convergence from random matrix theory.

3.5 Characterization of μ_f

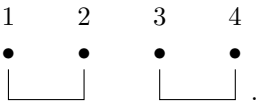
In this section, we will establish some combinatorial results that will be used to both better define μ_f and later to show the convergence of $\overline{\mathbf{A}}_N$ to this measure. We first introduce some additional notational elements. Wick's formula will be used to assist in counting the contributions of the elements of a matrix to its spectra [47]. Let $Y_1, \ldots, Y_p$ be independent standard Gaussian random variables, and let $x_1, \ldots, x_n \in \{Y_1, \ldots, Y_p\}$ be a subset of those random variables. Wick's formula is given as:

$$\mathbb{E}\left[x_1 \ldots x_n\right] = \sum_{\pi \in P_2(N)} \mathbb{E}_\pi\left[x_1, \ldots, x_n\right] \coloneqq \sum_{\pi \in P_2(n)} \prod_{(r,s) \in \pi} \mathrm{Cov}(x_r, x_s) = \sum_{\pi \in P_2(n)} \prod_{(r,s) \in \pi} \mathbb{E}\left[x_r, x_s\right], \qquad (3.20)$$

for π being a pairing of $[n] = \{1, 2, \ldots, n\}$. A pairing is defined as a disjoint pair, so groups of size 2, of $[n]$, with the set of all pairings, $P_2(n)$ being formally defined as,

$$P_2(n) := \{\pi : \pi \text{ pairings of } [n]\}.$$

Note that for an odd value of n, $P_2(n) = 0$, as it is not possible to create a set of disjoint pairs without having an atom left over, and thus the set cannot be characterized only by pairings. We also introduce the cyclic permutation operator, γ, which is the shift by 1 modulo $2k$ for $k \in \mathbb{N}, k \geq 1$, such that the $(2k+1)$-th element becomes the first element in the set [3]. This cyclic permutation can be applied to π, which can be interpreted as a walk along the graph created by the individual combinations of $\pi \in P_2(n)$. To illustrate, let $N = 4$, so we are looking to work on the pairings between the elements of $[n] = \{1, 2, 3, 4\}$. For example, take $\pi_1 \in P_2(4)$,



This yields the pairs (1 2)(3 4), which we can then notate as follows, with each vertical alignment in the right-hand side matrix given by the two corresponding nodes in a pairing. In other words, the first row is one pair element, and the second row is the other connected element,

$$\pi_1 = (1\ 2)\ (3\ 4) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}.$$

We can now see how applying a cyclic permutation changes this matrix. A shift by 1 mod $2k$ can be applied by first creating the γ permutation matrix, as can be seen below, which takes and shifts the bottom register of a matrix of indexes by 1,

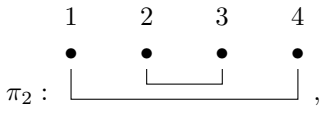
$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} \overset{1 \bmod 4}{\Longrightarrow} \gamma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}.$$

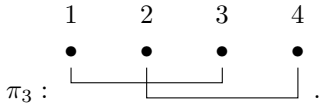
This re-indexed permutation can then be multiplied by the initial pairing, π_1, with the multiplication being applied by following the index pairs from π_1 to the index pairs in γ. For example, we find the new

connection pair of 1 by taking $1 \to 2$, as defined in π_1, and then taking $2 \to 3$ from γ, thus yielding that $1 \to 3$ in $\gamma\pi_1$. This yields a new matrix as can be seen below,

$$\gamma\pi_1 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix}. \tag{3.21}$$

This produces one connected pair (1 3), and two atoms which are self-connected, (2) and (4). Thus, $\pi_1 \in P_2(4)$ has a contribution of 3 elements: a pair and 2 singletons. We count the number of these elements to assess the contribution of a given $\pi \in P_2(2k)$ towards the spectra of matrix for its $2k$-th moment. Additionally, we define a subset of these pairs, namely, the set of non-crossing pairs, $NC_2(2k)$, on $P_2(2k)$, which we will illustrate by assessing all the combinations of $P_2(4)$. Having defined π_1 above, we illustrate π_2 and π_3,





Note that π_1 and π_2 do not exhibit crossing pairs as the pairings between indexes can be drawn without having their connectors overlap, whereas there is no way to draw π_3 without intersecting the connecting lines. Following our earlier steps for applying γ across the index sets of these pairs, we can find the contributions of each π,

$$\gamma\pi_2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} = (1)(2\ 4)(3),$$

$$\gamma\pi_3 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix} = (4\ 1\ 2\ 3).$$

The $NC_2(4)$ pairings each produce 3 components after permutation by γ, whereas the crossing pair produces only 1. This then motivates our first result in this section, which establishes an upper bound of the number of contributions from crossing pairs possible for a given $\pi \in P_2(2k)$. Prior to stating our next lemma, we formally define $NC_2(2k)$ as follows:

$$NC_2(2k) := \{\pi \in P_2(2k) | \pi \text{ is non-crossing}\}. \tag{3.22}$$

Lemma 3.5.1. *The number of components contributed by a modulo 1 shift, γ, of a pairing $\pi \in NC_2(2k)$, is $\#\gamma\pi = k+1$, for $\#$ being the counting of components in each $\gamma\pi$. If $\pi \in P_2(2k)$ is crossing, then $\#\gamma\pi < k+1$.*

Proof of Lemma 3.5.1:

Let $(i, i+1)$ be a pair in π that corresponds to an entry in $\gamma\pi$. Walking through the permutation, we see the following steps, as illustrated earlier for $P_2(4)$,

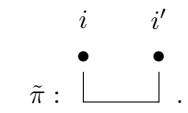
$$i+1 \xrightarrow{\pi} i \xrightarrow{\gamma} i+1,$$

and similarly,

$$i \xrightarrow{\pi} i+1 \xrightarrow{\gamma} i+2.$$

This provides at least 2 components in $\gamma\pi$: a singleton, $(i+1)$; and a pairing, $(\dots, i, i+2, \dots)$. We now seek to prove $\#\ \gamma\pi = k+1$ if π is non-crossing, NC_2. We remove the pair $(i, i+1)$ to produce another pair set, $\tilde{\pi}$. As such, the component $(i+1)$ drops out and the point i in $(\dots, i, i+2, \dots)$ must be removed. This reduces the total number of indexes, $2k$, by 2, and subsequently reduces $\#\ \gamma\pi$ by 1 as $(i+1)$ no longer contributes. As seen in the example for $P_2(4)$, these individual components that drop out only occur for pairs that do not cross, as crossing pairs do not produce atoms. However, this is not

to say that $\pi \in NC_2(2k)$ contribute only 1, as they will contribute as many atoms as possible, but will result in an irreducible crossing component which will produce one large final component. We continue this removal of pairs until $\tilde{\pi}$ has only 2 points remaining:



Applying the permutation given by γ to this final set,

$$\gamma\tilde{\pi} = \begin{pmatrix} i & i' \\ i' & i \end{pmatrix}\begin{pmatrix} i & i' \\ i' & i \end{pmatrix} = \begin{pmatrix} i & i' \\ i & i' \end{pmatrix} = (i)(i'),$$

we have that $\#\ \gamma\tilde{\pi} = 2$. Of course, there can be at most $k-1$ removals of 2 points per removal from the starting count of $2k$ points such that $\tilde{\pi}$ remains non-empty, and each reduction of pairs leads to a removal of 1 component from $\gamma\pi$. As such, counting these removals and adding the final 2 components yields our result,

$$\#\gamma\pi = (k-1)\cdot 1 + 2 = k+1.$$

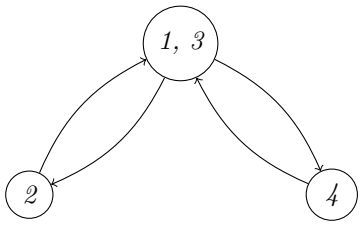
Looking now at $\pi \in P_2(2k)$ where $\pi \notin NC_2(2k)$, we similarly execute this removal process until no longer possible. The key difference is when a crossing pair is in $\tilde{\pi}$, in which case each π can only produce atoms until it is irreducible with 2 or more indexes remaining, thus yielding,

$$\#\gamma\pi \leq k < k+1,$$

which shows the maximal contribution of the case where there are crossing pairs is k. ■

We now introduce a graphical form for this cyclic permutation, upon which counting of edges will be done.

Definition 3.5.1. *Let $G_{\gamma\pi} = (V_{\gamma\pi}, E_{\gamma\pi})$, which is the graph for which $\gamma\pi$ walks. To illustrate, we use the case of $P_2(4)$. We take π_1, where $\gamma\pi_1 = (1\ 3)(2)(4)$, as discussed in (3.21), to produce the following $G_{\gamma\pi_1}$:*



Note that the cyclic walk above follows the ordering of the indexes: $1 = 3 \to 2$, $2 \to 1 = 3$, $1 = 3 \to 4$, $4 \to 1 = 3$. "Gluing" these directed edges together produces the undirected edges in the corresponding graph. This directed structure is actually a rooted planar tree, and walking up the corresponding sides of each branch starting and ending at the root is the count of contribution that we are seeking from each π. Finally, we define a function τ_π, which maps $\{1, 2, \ldots, 2k\}$ to $\{1, 2, \ldots, k+1\}$, satisfying $i \in V_{\tau_\pi(i)}, 1 \leq i \leq 2k$, where $V_{\tau_\pi(i)}$ is the Kreweras complement of π [6].

Now that we have defined the objects we will be counting and established the difference between crossing and non-crossing pairs from a combinatorial lens, we can define μ_f, which does not have an explicit form, but has characterizable even moments [6].

Lemma 3.5.2. *Let there be a function $f : [0,1]^2 \to \mathbb{R}$, a mapping τ_π, and $G_{\gamma\pi} = (V_{\gamma\pi}, E_{\gamma\pi})$ be the graph upon which the permutation walks. The odd moments of μ_f are zero, and the even moments of μ_f are defined as follows,*

$$m_{2k} \coloneqq \int_{\mathbb{R}} x^{2k}\mu_f(dx) = \sum_{\pi\in NC_2(2k)} \int_{[0,1]^{k+1}} \prod_{(u,v)\in E_{\gamma\pi}} f(x_{\tau_\pi(u)}, x_{\tau_\pi(v)}), \tag{3.23}$$

and furthermore, m_{2k} satisfies Carleman's condition,

$$\limsup_{k\to\infty} m_k^{1/(2k)} < \infty. \tag{3.24}$$

Proof of Lemma 3.5.2:

We first note that $\sum_{\pi\in NC_2(2k)}$ and $\prod_{(u,v)\in E_{\gamma\pi}}$ are equivalent to enumerating all components from all $\pi \in NC_2(2k)$ via Lemma 3.5.1. Thus, if $f(\cdot) = 1$, this quantity could be bounded above by $\#NC_2(2k)$. However, for a more general f, we use the integral over $k+1$ unit squares, $\int_{(0,1)^{k+1}}$ for each contribution. Thus, assuming only f being bounded, we bound using the infinity norm of $f, M \coloneqq \|f\|_\infty$, which is counted at most k times, yielding f^k. However, m_k is taken to the $\frac{1}{2k}$-th power, so,

$$m_k^{1/(2k)} \leq \|f\|_\infty (\#NC_2(2k))^{1/(2k)} = M(\#NC_2(2k))^{1/(2k)}.$$

We note that $\#NC_2(2k)$ is a factorial quantity characterized by the Catalan numbers, $C_k = \frac{1}{k+1}\binom{2k}{k}$ [12]. Expanding the binomial coefficient of the Catalan numbers and applying Stirling's approximation [19], $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$, we arrive at a more workable approximation of this quantity,

$$\frac{1}{k+1}\binom{2k}{k} = \frac{1}{k+1}\left(\frac{2k!}{(k!)^2}\right) \approx \frac{1}{k+1}\frac{\sqrt{4\pi k}2^{2k}}{2\pi k} \approx \frac{4^k}{k^{3/2}\sqrt{\pi}},$$

with the final approximation as $k \to \infty$ absorbing the leading fraction into the latter. Plugging this back into the inequality above in place of $\#NC_2(2n)$, we arrive at our result, given that $M < \infty$ as f is bounded,

$$m_k^{1/(2k)} \leq M(\#NC_2(2k))^{1/(2k)} \approx M\left(\frac{4^k}{k^{3/2}\sqrt{\pi}}\right)^{1/(2k)} = \frac{2M}{k^{3/4k}\pi^{1/4k}} \to 2M,$$

again as $k \to \infty$. ■

We note that given f is bounded, we have then that μ_f is also compactly supported, implying that all of its moments exist and are finite, thus making it a mathematically useful measure to show convergence to for a general class of matrices. Having now proved the finiteness of the even moments of μ_f, we can now show that the moments of the spectral measure $\mathrm{ESD}\left(\overline{\mathbf{A}}_N\right)$ converge to and are characterized uniquely by that of μ_f using a combinatorial approach.

Lemma 3.5.3. *The spectral measure of the leading order variance matrix* $\mathrm{ESD}\left(\overline{\boldsymbol{A}}_N\right)$ *converges to the non-zero probability measure* μ_f*,*

$$\lim_{N\to\infty} \mathrm{ESD}\left(\overline{\boldsymbol{A}}_N\right) = \mu_f, \tag{3.25}$$

weakly in probability. Furthermore, μ_f *characterizes this convergent measure uniquely.*

The proof of this claim will rely on expressing the spectral measure of $\overline{\mathbf{A}}_N$ as a combinatorial quantity for all even moments, which we then show is equivalent to the definition of μ_f from (3.23).

Proof of Lemma 3.5.3:

We use Wick's theorem to characterize the even moments of the expectation of the trace of $\overline{\mathbf{A}}_N$, $\mathbb{E}[\mathrm{tr}(\overline{\mathbf{A}}_N)^{2k}]$. To help illustrate, we use the second moment, $k = 1$, to demonstrate how this counting occurs,

$$\mathbb{E}\left[\mathrm{tr}\left((\overline{\mathbf{A}}_N)^2\right)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i_1,i_2=1}^{N}\overline{\mathbf{A}}_N(i_1,i_2)\overline{\mathbf{A}}_N(i_2,i_1)\right] = \frac{1}{N}\sum_{i_1,i_2=1}^{N}\mathbb{E}\left[\overline{\mathbf{A}}_N(i_1,i_2)^2\right].$$

The final equality arises as the indexes of $\overline{\mathbf{A}}_N$ are undirected, so a connection from i_2 to i_1 is equivalent to a connection from i_1 to i_2 in terms of the resulting variance. We can thus collapse these indexes together into one term. The variance of $\overline{\mathbf{A}}_N(i,j)$ is $\frac{f\left(\frac{i}{N},\frac{j}{N}\right)}{N}$, as defined in (3.18), which we can use to reintroduce our dependence on f into the equation,

$$\frac{1}{N}\sum_{i_1,i_2=1}^{N}\mathbb{E}\left[\overline{\mathbf{A}}_N(i_1,i_2)^2\right] = \frac{1}{N}\sum_{i_1,i_2=1}^{N}\frac{f\left(\frac{i}{N},\frac{j}{N}\right)}{N} = \frac{1}{N^2}\sum_{i_1,i_2=1}^{N}f\left(\frac{i_1}{N},\frac{i_2}{N}\right).$$

Given f is Riemann integrable, we can thus state that as $N \to \infty$, the summation can be expressed as a double integral over the two respective indexes of f, for $x = \frac{i_1}{N}$, $y = \frac{i_2}{N}$,

$$\frac{1}{N^2}\sum_{i=1}^{N} f\left(\frac{i_1}{N},\frac{i_2}{N}\right) = \frac{1}{N}\sum_{i_1=1}^{N}\frac{1}{N}\sum_{i_2=1}^{N} f\left(\frac{i_1}{N},\frac{i_2}{N}\right) \approx \int_0^1\int_0^1 f(x,y)\,dx\,dy.$$

We now extend this to even moments beyond the second moment where $k=1$, by generalizing for all $2k$ moments using 3.20,

$$\mathbb{E}\left[\operatorname{tr}\left((\overline{\mathbf{A}}_N)^{2k}\right)\right] = \frac{1}{N^{k+1}}\sum_{i_1,\ldots,i_{2k}=1}^{N}\sum_{\pi\in P_2(2k)}\prod_{(r,s)\in\pi}\mathbb{E}\left[\overline{\mathbf{A}}_N(i_r,i_s)\overline{\mathbf{A}}_N(i_s,i_r)\right],$$

where the second summation over $\pi\in P_2(2k)$ is not visible in the case of $k=1$ as there is only one possible pair from i_1 to i_2. However, as established prior, we count only the components generated by applying the cyclic permutation γ to a given π. As such, for a given index element i_r, it is only counted if the element contributes in $\gamma\pi(r)$, which is then scaled by $\sqrt{f\left(\frac{i_r}{N},\frac{i_s}{N}\right)}$ in order to yield the correct total contribution of f,

$$\begin{aligned}&\frac{1}{N^{k+1}}\sum_{i_1,\ldots,i_{2k}=1}^{N}\sum_{\pi\in P_2(2k)}\prod_{(r,s)\in\pi}\mathbb{E}\left[\overline{\mathbf{A}}_N(i_r,i_s)\overline{\mathbf{A}}_N(i_s,i_r)\right]\\ =&\frac{1}{N^{k+1}}\sum_{i_1,\ldots,i_{2k}=1}^{N}\sum_{\pi\in P_2(2k)}\prod_{(r,s)\in\pi}\delta_{\left(i_r=\gamma\pi(r)\right)}\sqrt{f\left(\frac{i_r}{N},\frac{i_s}{N}\right)}.\end{aligned}$$

δ is the Dirac delta function representing the contributions of a specific combination of nodes within a given π, and γ is constant over π. We can in fact upper bound this statement by specifying that we count only the $\pi\in NC_2(2k)$. Recalling from Lemma 3.5.1 that $\pi\in NC_2(2k)$ produces $k+1$ components and $\pi\notin NC_2(2k)$ produces at most k components, we first re-express the prior expression in slightly more compact notation, with the enumeration of the sum across individual indexes in a π summarized as $N^{\#\gamma\pi}$ and the effect of $\sqrt{f\left(\frac{i_r}{N},\frac{i_s}{N}\right)}$ summarized in a constant c,

$$\frac{1}{N^{k+1}}\sum_{\pi\notin NC_2(2k)}\sum_{i_1,\ldots,i_{2k}=1}^{N}\prod_{(r,s)\in\pi}\left[\delta_{\left(i_r=\gamma\pi(r)\right)}\sqrt{f\left(\frac{i_r}{N},\frac{i_s}{N}\right)}\right] \approx \frac{1}{N^{k+1}}\sum_{\pi\notin NC_2(2k)} cN^{\#\gamma\pi}.$$

The contribution of $\pi\notin NC_2(2k)$ is thus bounded as follows given that $\#\gamma\pi\leq k$,

$$\frac{1}{N^{k+1}}\sum_{\pi\notin NC_2(2k)} cN^{\#\gamma\pi} \leq \frac{1}{N^{k+1}}\sum_{\pi\notin NC_2(2k)} cN^{k} \to 0,$$

as $N\to\infty$, and the leading normalizing constant outpaces the quantity in the summation. Ergo, pairs which include crossing elements do not contribute in the limit to the total number of components counted, leading to the narrowing of contributing pairs to just those from $\pi\in NC_2(2k)$. We will again use $2k=4$ to illustrate how this contribution arises specifically for non-crossing pairs. Specifically, we will focus on π_1, with $\gamma\pi_1=(1\ 3)\ (2)\ (4)$. For simplicity, we use $i_1,\ldots,i_4$ to correspond to each point,

$$\mathbb{E}\left[\operatorname{tr}\left((\overline{\mathbf{A}}_N)^4\right)\right] = \frac{1}{N^3}\sum_{\pi\in NC_2(4)}\sum_{i_1,\ldots,i_4=1}\sqrt{f\left(\frac{i_1}{N},\frac{i_2}{N}\right)f\left(\frac{i_2}{N},\frac{i_1}{N}\right)f\left(\frac{i_3}{N},\frac{i_4}{N}\right)f\left(\frac{i_4}{N},\frac{i_3}{N}\right)}.$$

As $(1,3)$ indicates $i_1=i_3$ in the connection of the individual elements, all pairs of the form $(i,i+1)$ are re-indexed accordingly. We then simplify the above statement by collapsing the indexes if they share an edge on a walk on $G_{\gamma\pi}$,

$$\begin{aligned}
&\frac{1}{N^3}\sum_{\pi\in NC_2(4)}\sum_{i_1,i_2=i_3,i_4=1}\sqrt{f\left(\frac{i_1}{N},\frac{i_2}{N}\right)f\left(\frac{i_2}{N},\frac{i_1}{N}\right)f\left(\frac{i_3}{N},\frac{i_4}{N}\right)f\left(\frac{i_4}{N},\frac{i_3}{N}\right)}\\
\approx&\frac{1}{N^3}\sum_{\pi\in NC_2(4)}\sum_{i_1,i_2,i_4=1}f\left(\frac{i_1}{N},\frac{i_2}{N}\right)f\left(\frac{i_1}{N},\frac{i_4}{N}\right)\\
\approx&\sum_{\pi\in NC_2(4)}\iiint f(x,y)f(x,z)dxdydz,
\end{aligned}$$

with the final substitution being $x=\frac{i_1}{N}, y=\frac{i_2}{N}, z=\frac{i_4}{N}$. As before, this result extends for all values of $2k$, leading to the following generalized result,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{tr}\left((\overline{\mathbf{A}}_N)^{2k}\right)\right]&=\frac{1}{N^{k+1}}\sum_{\pi\in NC_2(2k)}\sum_{i_1,\ldots,i_{2k}=1}^{N}\prod_{(r,s)\in\pi}\left[\delta_{\left(i_r=i_{\gamma\pi(r)}\right)}\sqrt{f\left(\frac{i_r}{N},\frac{i_s}{N}\right)}\right]\\
&=\sum_{\pi\in NC_2(2k)}\int_{[0,1]^{k+1}}\prod_{(r,s)\in\pi}f\left(\frac{x_r}{N},\frac{x_s}{N}\right)dx_1\ldots dx_{k+1}.
\end{aligned}$$

Again, in the homogeneous setting where $f(x,y)=1$, this would provide the Catalan numbers, which characterize the moments of the semicircle distribution [12]. In the inhomogeneous setting, we identify that $\lim_{N\to\infty}\mathbb{E}\left[\mathrm{tr}(\overline{\mathbf{A}}_N)^{2k}\right]$ is in fact equivalent to m_k from (3.23), and so,

$$\lim_{N\to\infty}\mathbb{E}\left[\mathrm{tr}(\overline{\mathbf{A}}_N)^{2k}\right]\to m_{2k}.$$

Having then defined in Lemma 3.5.2 that m_k is the even moments of μ_f, we have thus shown that in the limit, the spectral measure of $\overline{\mathbf{A}}_N$ is indeed characterized by μ_f. In order to prove the unique nature of this limiting characterization, we will rely on Carleman's condition, which assesses if a measure on $\mathbb{R}$ with finite moments, a, satisfies,

$$\sum_{k=1}^{\infty}a_{2k}^{-1/(2k)}=+\infty.$$

If a measure meets the above condition, that any other measure that has the same moments must also then be the same measure [16]. As highlighted in the proof of Lemma 3.5.2, $m_k^{1/(2k)}$ is bounded above by MC_k, for C_k being the Catalan numbers, which do indeed satisfy Carleman's condition. Thus, the limiting measure of $\mathbb{E}\left[\mathrm{tr}(\overline{\mathbf{A}}_N)^{2k}\right]$ must be unique, then implying that μ_f is indeed the unique limiting measure for this quantity. As such, by showing this convergence in the limit, we have shown that,

$$\lim_{N\to\infty}\mathrm{ESD}\left(\overline{\mathbf{A}}_N\right)=\mu_f,$$

weakly in probability. ■

3.6 Proof of Theorem 3.2.1

Having now gathered all the requisite supporting results, we can now work backwards to prove the initial theorem (3.1), using several perturbations of the initial matrix $\mathbf{B}_N$. We have shown from (3.19) that $\lim_{N\to\infty}d_{BL}\left(\mathrm{ESD}\left(\mathbf{A}_N^g\right),\mathrm{ESD}\left(\overline{\mathbf{A}}_N\right)\right)=0$, in probability. We can extend this assertion using Lemma 3.5.3 to state that:

$$\lim_{N\to\infty}\mathrm{ESD}\left(\mathbf{A}_N^g\right)=\mu_f,$$

weakly in probability. Similarly, as we have proven in Lemma 3.4.3,

$$\lim_{N\to\infty}\left|\mathbb{E}\left[h(\Re(S_{\mathbf{A}_N^g}(z)))-h(\Re(S_{\mathbf{A}_N^0}(z)))\right]\right|=0.$$

And so, by $S_{\mathbf{A}_N^g}(z)$ being the resolvent of $\mathbf{A}_N^g$,

$$\lim_{N\to\infty}\mathbb{E}\left[h(\Re(S_{\mathbf{A}_N^g}(z)))\right]=h\left(\int_{\mathbb{R}}\frac{1}{x-z}\mu_f(dx)\right),$$

with the above equality in expectation concluding that this is then equivalent to,

$$\lim_{N\to\infty}\mathbb{E}\left[h(\Re(S_{\mathbf{A}_N^0}(z)))\right]=h\left(\int_{\mathbb{R}}\frac{1}{x-z}\mu_f(dx)\right),$$

in probability. The right-hand side is now a deterministic quantity, and because h is thrice differentiable by definition, we have that:

$$\lim_{N\to\infty}\Re(S_{\mathbf{A}_N^0}(z))=\int_{\mathbb{R}}\frac{1}{x-z}\mu_f(dx),$$

in probability. These equivalences similarly hold for the imaginary part, as we also demonstrated synonymous behaviour between both the real and imaginary components of the transform of these measures,

$$\lim_{n\to\infty}\left|\mathbb{E}\left[h(\Im(S_{\mathbf{A}_N^g}(z)))-h(\Im(S_{\mathbf{A}_N^0}(z)))\right]\right|=0.$$

Using the above logic that was applied to the real plane, the following statements are then also equal:

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}\left[h(\Im(S_{\mathbf{A}_N^g}(z)))\right]=h\left(\int_{\mathbb{R}}\frac{1}{x-z}\mu_f(dx)\right)\\ \iff&\lim_{n\to\infty}\mathbb{E}\left[h(\Im(S_{\mathbf{A}_N^0}(z)))\right]=h\left(\int_{\mathbb{R}}\frac{1}{x-z}\mu_f(dx)\right)\\ \iff&\lim_{n\to\infty}\Im(S_{\mathbf{A}_N^0}(z))=\int_{\mathbb{R}}\frac{1}{x-z}\mu_f(dx).\end{aligned}$$

And so, having shown convergence to the Stieltjes transform of μ_f in both the real and imaginary parts, we can thus state:

$$\lim_{N\to\infty}\mathrm{ESD}\left(\mathbf{A}_N^0\right)=\mu_f$$

weakly in probability. We have also shown that $d_{BL}\left(\mathrm{ESD}\left(\mathbf{A}_N^0\right),\mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right)\to 0$ in (3.13), which then yields,

$$d_{BL}\left(\mathrm{ESD}\left(\hat{\mathbf{B}}_N\right),\mu_f\right)\to 0,$$

weakly in probability, thus completing the initial triangle inequality we initially set up in (3.9),

$$d_{BL}(\mathrm{ESD}\left(\mathbf{B}_N\right),\mu_f)\leq d_{BL}\left(\mathrm{ESD}\left(\mathbf{B}_N\right),\mathrm{ESD}\left(\hat{\mathbf{B}}_N\right)\right)+d_{BL}\left(\mathrm{ESD}\left(\hat{\mathbf{B}}_N\right),\mu_f\right)\to 0,$$

weakly in probability. This then concludes the proof that,

$$\lim_{N\to\infty}\mathrm{ESD}\left(\mathbf{B}_N\right)=\mu_f,$$

weakly in probability. ■

We have now shown that with appropriate scaling and centering, the spectral measure of the normalized

Laplacian in a general inhomogeneous case indeed converges weakly to the measure defined in [6] and [12], μ_f. The next sections will focus on applying this result and the methods used in its proof first in the context of a centrality measure, PageRank, and secondly on graphical neural networks. Both of these applications rely on the normalized Laplacian to describe the connectedness of nodes in G_N.

Chapter 4
PageRank Approximation in the Inhomogeneous Regime

In this section we apply the methods and results derived in the prior chapter to extend the use of PageRank approximation by average graph volume to inhomogeneous random graphs.

4.1 Notation

PageRank and the broader family of centrality measures are useful descriptive tools for conveying the connectedness of a graph or a subset of vertices within a graph. However, these measures cannot be expressed in an explicit form empirically, and as such, approximations must be used. This section will focus specifically on the approximation of PageRank using the degrees and volume of a graph in a generalized inhomogeneous setting. We continue to carry forward the above setting where $p_{ij} = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right) = \varepsilon_N r\left(\frac{i}{N}\right) r\left(\frac{j}{N}\right)$, with f, and by extension $r : [0,1] \to [0,\infty)$, being Riemann integrable and bounded. Finally, we will denote the PageRank vector as $\boldsymbol{\pi}_N$, and a preference vector $\mathbf{v}$. $\mathbf{v}$ is also a probability vector, playing a role such that it balances against $\mathbf{P}$ as follows, for a damping factor $\alpha \in (0,1)$,

$$\tilde{\mathbf{P}} = \alpha\mathbf{P} + (1-\alpha)\mathbf{v}\mathbf{1}^T, \tag{4.1}$$

where $\mathbf{1}$ is a vector of ones. $\tilde{\mathbf{P}}$ is then the stochastic transition matrix of a modified Markov chain, which over multiple iterations of the calculation of PageRank would yield a dominating eigenvector of $\tilde{\mathbf{P}}$, which is equivalently its stationary distribution [24]. We denote this vector as $\boldsymbol{\pi}_N$ as follows,

$$\boldsymbol{\pi}_N = \tilde{\mathbf{P}}\boldsymbol{\pi}_N,$$

or, equivalently when $\alpha < 1$,

$$\boldsymbol{\pi}_N = (1-\alpha)[\mathbf{I} - \alpha\mathbf{P}]^{-1}\mathbf{v}. \tag{4.2}$$

The approximation of $\boldsymbol{\pi}_N$ using the volume of a graph, as defined earlier in (1.5), is given as:

$$\overline{\boldsymbol{\pi}}_N = \frac{\alpha\mathbf{d}}{\mathrm{vol}(G_N)} + (1-\alpha)\mathbf{v}.$$

In order to assess the validity of this approximation, we use total variation distance, which for two discrete probability distributions $\mathbf{p} = \{p_i\}_{i\geq 0}$ and $\mathbf{q} = \{q_i\}_{i\geq 0}$ is defined to be:

$$d_{TV}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\sum_i |p_i - q_i|. \tag{4.3}$$

Finally, an event, E, will hold with high probability (w.h.p), if there exists a C such that $\mathbb{P}[E] \geq 1 - O(c^{-m})$ for an $m > 0, \forall c > C$.

4.2 PageRank Approximation in General Inhomogeneous Graphs

Prior to stating the main result of this chapter, we first require two conditions regarding the behavior of the graph, such that the degrees of the graph are reasonably well-behaved while the underlying structure still exhibits some level of centrality across its nodes. Formally then, let G_N satisfy the following conditions:

1. The maximum and minimum degrees, $\mathbf{d}_{\max}$ and $\mathbf{d}_{\min}$, are both finite and somewhat close. That is, $\mathbf{d}_{\max} \leq K\mathbf{d}_{\min}$ for a constant $K > 0$ not dependent on N.

2. The remaining eigenvalues beyond the leading eigenvalue of $\mathbf{Q}_N \coloneqq \mathbf{I}_N - \mathbf{L}_N = \mathbf{D}_N^{-1/2}\mathbf{A}_N\mathbf{D}_N^{-1/2}$ are vanishingly small,

$$\max\{|\lambda_2(\mathbf{Q}_N)|, |\lambda_N(\mathbf{Q}_N)|\} \leq \frac{2}{\sqrt{N\varepsilon_N}\delta},$$

with high probability, for δ being the infinimum of $r(x), \delta \coloneqq \inf_{x\in[0,1]} r(x)$. We denote $\lambda_i(\mathbf{Q}_N)$ as λ_i, unless otherwise stated. This is equivalent to stating that the spectral gap, $1 - \max\{|\lambda_2|, |\lambda_N|\}$, is bounded away from 0 [2].

Theorem 4.2.1. *Let $\|\mathbf{v}\|_2 = O(1/\sqrt{N})$. Under the assumptions $p_{ij} = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)$ and $N\varepsilon_N \gg (\log(N))^\xi$, $\xi > 6$, and additionally that the graph G_N satisfies conditions (1) and (2), the PageRank measure of this graph, $\boldsymbol{\pi}_N$, can be approximated in total variation by $\overline{\boldsymbol{\pi}}_N$, as defined in* (1.5)*,*

$$d_{TV}(\boldsymbol{\pi}_N, \overline{\boldsymbol{\pi}}_N) \leq \frac{K}{\sqrt{N\varepsilon_N}} = o(1), w.h.p, \tag{4.4}$$

as $N \to \infty$.

[2] proved that under the above conditions that (4.4) holds for Chung-Lu and Stochastic Block inhomogeneous graphs. We will approach this problem for a more general class of inhomogeneous graphs and additionally validate that convergence rate is dependent on the average degree of the graph via $\sqrt{N\varepsilon_N}$.

In order to prove this theorem, we must first validate that the general inhomogeneous random graphs in our setting meet the two conditions, with the final proof following as a result. As such, we begin by individually assessing these two claims, starting with the behavior of the minimum and maximum degrees. We additionally impose the assumption that $N\varepsilon_N \gg (\log(N))^\xi$, for a $\xi \geq 1$, thus ensuring a minimal degree of connectivity in the graph. As a reminder, we have already proved that the maximum degree deviations from the expected degree disappear asymptotically in Lemma 3.3.1, but we will restate it here for convenience:

$$\max_{1\leq i\leq N} \left| \frac{\mathbf{d}_i - \mathbb{E}[\mathbf{d}_i]}{N\varepsilon_N} \right| \to 0,$$

as $N \to \infty$, almost surely. Similarly, we have proved that the minimum degree is convergent to the expected degree,

$$\min_{1\leq i\leq N} \left(\frac{\mathbf{d}_i}{N\varepsilon_N} \right) \to 1,$$

$N \to \infty$. We have thus already addressed the well-behaved nature of the degrees of G_N with respect to the average degree asymptotically. In the following two sections we show that G_N satisfies (1) and (2), first focusing on (1).

4.3 Relative Degree Bounds on an Inhomogeneous Graph

As mentioned in the prior section, we require that a graph meet the sufficient condition that $\mathbf{d}_{\max} \leq K\mathbf{d}_{\min}$ for the result to hold. This leads to our first lemma concerning PageRank.

Lemma 4.3.1. *If for a graph G_N where the connection probability is given $p_{ij} = \varepsilon_N r\left(\frac{i}{N}\right) r\left(\frac{j}{N}\right)$ and $N\varepsilon_N \gg (\log(N))^\xi$ for $\xi \geq 1$, then,*

$$\frac{\boldsymbol{d}_{\max}}{\boldsymbol{d}_{\min}} \leq K, \tag{4.5}$$

for a $K > 1$ and not dependent on N.

The proof seeks to show that the probability $\frac{\mathbf{d}_{\max}}{\mathbf{d}_{\min}} > K$ goes to 0 as $N \to \infty$. This probability can be decomposed into the intersect of two events, the event that $\mathbf{d}_{\max} \geq K\mathbf{d}_{\min}$ and the behaviour of $\mathbf{d}_{\min}$ relative to $\tilde{\delta}N\varepsilon_N$ for some $\tilde{\delta} > 0$.

Proof of Lemma 4.3.1:

To start, we illustrate the decomposition of the probability of the event in interest, namely that the ratio between the maximum and minimum degrees exceeds an arbitrary K,

$$\mathbb{P}(\mathbf{d}_{\max} \geq K\mathbf{d}_{\min}) = \mathbb{P}(\mathbf{d}_{\max} \geq K\mathbf{d}_{\min}, \mathbf{d}_{\min} \geq \tilde{\delta}N\varepsilon_N) + \mathbb{P}(\mathbf{d}_{\max} \geq K\mathbf{d}_{\min}, \mathbf{d}_{\min} < \tilde{\delta}N\varepsilon_N). \tag{4.6}$$

We assess the probability of the two intersected events individually. First, looking at $\mathbb{P}(\mathbf{d}_{\max} \geq K\mathbf{d}_{\min}, \mathbf{d}_{\min} \geq \tilde{\delta}N\varepsilon_N)$, we note that the intersect in the latter half of the probability is the more restrictive condition on $\mathbf{d}_{\min}$ as K is arbitrary. We can thus use this to more accurately bound the probability of this intersection, with $\mathbf{d}_{\min}$ being replaced by the more restrictive $\tilde{\delta}N\varepsilon_N$ in the first inequality,

$$\mathbb{P}(\mathbf{d}_{\max} \geq K\mathbf{d}_{\min}, \mathbf{d}_{\min} \geq \tilde{\delta}N\varepsilon_N) \leq \mathbb{P}(\mathbf{d}_{\max} \geq K\tilde{\delta}N\varepsilon_N) \leq \sum_{i=1}^{N} \mathbb{P}(\mathbf{d}_i \geq K\tilde{\delta}N\varepsilon_N).$$

Having now replaced $\mathbf{d}_{\max}$ with a union bound over all $\mathbf{d}_i$, we can rearrange the inequality in the probability and apply that $\mathbb{E}[\mathbf{d}_i] = \varepsilon_N r(i/N) \sum_{j\neq i}^{N} r(j/N)$ which yields then $N\varepsilon_N\delta^2 \leq \mathbb{E}[\mathbf{d}_i] \leq N\varepsilon_N m^2$, for $\delta \coloneqq \inf_{x\in[0,1]} r(x)$ and $m \coloneqq \sup_{x\in[0,1]} r(x)$. As such, we must pick a $\tilde{\delta} < \delta$ where $K > \frac{\delta}{\tilde{\delta}} > 1$ and that $K > \frac{m^2}{\delta^2} > 1$, so,

$$\sum_{i=1}^{N} \mathbb{P}(\mathbf{d}_i \geq K\tilde{\delta}N\varepsilon_N) = \sum_{i=1}^{N} \mathbb{P}(\mathbf{d}_i - \mathbb{E}[\mathbf{d}_i] \geq K\tilde{\delta}N\varepsilon_N - \mathbb{E}[\mathbf{d}_i]) \leq \sum_{i=1}^{N} \mathbb{P}(\mathbf{d}_i \geq \mathbb{E}[\mathbf{d}_i] + N\varepsilon_N(K\tilde{\delta} - m^2))$$

We see that we are now in a position to apply an upper tail Chernoff bound to this probability, as described in [45],

$$\sum_{i=1}^{N} \mathbb{P}(\mathbf{d}_i \geq \mathbb{E}[\mathbf{d}_i] + N\varepsilon_N(K\tilde{\delta} - \delta)) \leq \sum_{i=1}^{N} \exp\left(-\frac{(K\tilde{\delta} - m^2)^2 N^2 \varepsilon_N^2}{2(N\varepsilon_N + (K\tilde{\delta} - m^2)N\varepsilon_N/3)}\right).$$

In order to simplify the denominator, we can bound from above by using $4(K\tilde{\delta}-m^2)N\varepsilon_N$, as $K\tilde{\delta}-m^2 > 1$. Note that $N\varepsilon_N \gg (\log(N))^{\xi}$, so we specify that $N\varepsilon_N > \frac{12\log^{\xi}(N)}{K\tilde{\delta}-m^2}$ and simplify accordingly,

$$\begin{aligned}
\sum_{i=1}^{N} \exp\left(-\frac{(K\tilde{\delta} - m^2)^2 N^2 \varepsilon_N^2}{2(N\varepsilon_N + (K\tilde{\delta} - m^2)N\varepsilon_N/3)}\right) &\leq \sum_{i=1}^{N} \exp\left(-\frac{(K\tilde{\delta} - m^2)^2 N^2 \varepsilon_N^2}{4(K\tilde{\delta} - m^2)N\varepsilon_N}\right) \\
&\leq N \exp\left(-\frac{(K\tilde{\delta} - m^2)}{4} \frac{12\log^{\xi}(N)}{(K\tilde{\delta} - m^2)}\right) \\
&= N^{-2} \to 0.
\end{aligned}$$

We have thus shown that the probability that the maximum degree is greater than the minimum degree scaled by any constant is vanishing. We now assess the second probability, $\mathbb{P}(\mathbf{d}_{\max} \geq K\mathbf{d}_{\min}, \mathbf{d}_{\min} < \tilde{\delta}N\varepsilon_N)$, and rearrange to apply a lower tail Chernoff bound, as in [45]. Focusing here on the latter half of the second probability, namely that the minimum degree is less than our slightly scaled expected degree where $\delta > \tilde{\delta}$,

$$\mathbb{P}(\mathbf{d}_{\min} < \tilde{\delta}N\varepsilon_N) = \mathbb{P}(\mathbf{d}_{\min} \leq \mathbb{E}[\mathbf{d}_i] - (\mathbb{E}[\mathbf{d}_i] - \tilde{\delta}N\varepsilon_N)) \leq \mathbb{P}(\mathbf{d}_{\min} \leq \mathbb{E}[\mathbf{d}_i] - N\varepsilon_N(\delta - \tilde{\delta})).$$

Note that $\delta - \tilde{\delta} > 0$, and so the term $N\varepsilon_N(\delta - \tilde{\delta})$ is a valid positive error term for the Chernoff bound. Applying the lower tail bound:

$$\mathbb{P}(\mathbf{d}_{\min} \leq \mathbb{E}[\mathbf{d}_i] - N\varepsilon_N(\delta - \tilde{\delta})) \leq \sum_{i=1}^{N} \mathbb{P}(\mathbf{d}_i \leq \mathbb{E}[\mathbf{d}_i] - N\varepsilon_N(\delta - \tilde{\delta})) \leq N \exp\left(-\frac{(\delta - \tilde{\delta})^2 N^2 \varepsilon^2}{2N\varepsilon_N}\right).$$

We bound $N\varepsilon_N$ from below by $N\varepsilon_N > \frac{6\log^{\xi}(N)}{(\delta-\tilde{\delta})^2}$ to achieve our result,

$$N \exp\left(-\frac{(\delta - \tilde{\delta})^2 N^2 \varepsilon^2}{2N\varepsilon_N}\right) = N \exp\left(-\frac{(\delta - \tilde{\delta})^2 N\varepsilon_N}{2}\right) \leq N \exp\left(-\frac{(\delta - \tilde{\delta})^2}{2} \frac{6\log^{\xi}(N)}{(\delta - \tilde{\delta})^2}\right) \to 0.$$

Finally, via (4.6) we have our result that $\mathbb{P}\left(\frac{\mathbf{d}_{\max}}{\mathbf{d}_{\min}} < K\right) \to 1$. ■

This proves that the degrees of our inhomogeneous graph are well-behaved and somewhat close, thus

satisfying the first condition underpinning Theorem 4.2.1. We now look to prove the second condition regarding the behavior of the spectral gap and that the eigenvalues beyond the leading one are vanishing.

4.4 Characterizing the Spectral Gap

Similar to our previous results which interact with the degree behavior of G_N, we look to work with the expected degree instead of the actual degree matrix in assessing spectral behavior. To that end, we introduce a slightly reworked form of the normalized and centered adjacency matrix. Let $\mathbf{w} \coloneqq (\mathbb{E}[\mathbf{d}_1], \ldots, \mathbb{E}[\mathbf{d}_N])$ be a vector of expected degrees and $\mathbf{W}_N \coloneqq \text{diag}(\mathbf{w})$ be a diagonal matrix of expected degrees as in [10]. Additionally, let the following be a row vector:

$$\boldsymbol{\chi}_N \coloneqq \left(\sqrt{\frac{r(1/N)}{\sum_{i=1}^N r(i/N)}}, \ldots, \sqrt{\frac{r(N/N)}{\sum_{i=1}^N r(i/N)}} \right).$$

We use these to define the normalized and centered adjacency matrix, $\mathbf{C}_N \coloneqq \mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2} - \boldsymbol{\chi}_N^T\boldsymbol{\chi}_N$ [10]. Observe that this can also be rewritten as follows in order to better see this centering and scaling in action,

$$\begin{aligned}\mathbf{C}_N = \mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2} - \boldsymbol{\chi}_N^T\boldsymbol{\chi}_N &= \frac{\mathbf{A}_N(i,j) - \varepsilon_N r(i/N)r(j/N)}{\sqrt{\varepsilon_N r(i/N)\sum_{i'\neq i} r(i'/N)}\sqrt{\varepsilon_N r(j/N)\sum_{j'\neq j} r(j'/N)}} \\ &= \mathbf{W}_N^{-1/2}(\mathbf{A}_N - \mathbb{E}[\mathbf{A}_N])\mathbf{W}_N^{-1/2}.\end{aligned}$$

The last equality is via the definition of $\mathbb{E}[\mathbf{A}_N(i,j)] = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right) = \varepsilon_N r\left(\frac{i}{N}\right) r\left(\frac{j}{N}\right)$. In the following lemma that assesses the behaviour of the leading eigenvalue, we will use this readjusted form as it is more conducive to element-wise operations [20].

Lemma 4.4.1. *Let G_N be as described above and let $N\varepsilon_N \gg (\log(N))^\xi$, $\xi > 6$. Then we have,*

$$\|\boldsymbol{C}_N\|_2 \leq (1 + o(1))\frac{2}{\sqrt{N\varepsilon_N}\delta}, \tag{4.7}$$

for δ being the infinimum of $r(x)$.

In order to bound the leading eigenvalue of $\mathbf{C}_N$, we first state the quantities that we will be working with. For an integer k, via properties of the trace:

$$\mathbb{E}\left[\lambda_1(\mathbf{C}_N)^{2k}\right] \leq \sum_{i=1}^N \mathbb{E}\left[\lambda_i(\mathbf{C}_N)^{2k}\right] = \mathbb{E}\left[\text{Tr}(\mathbf{C}^{2k})\right]. \tag{4.8}$$

To bound $\mathbb{E}[\text{Tr}(\mathbf{C}^{2k})]$ then, we will again turn to combinatorial arguments to define and quantify the spectral behavior of a random matrix. This is a similar argument to that made in (3.25), albeit adjusted to bound just the leading eigenvalue. To restate, this argument has two components. First, determining the contribution of a step in a valid walk of length $2k$ across the complete graph defined in size by N, such that each edge is walked with multiplicity of at least 2 to be considered valid, and second, the enumeration of these steps and walks. Additionally, k can now potentially be dependent on N.

Proof of Lemma 4.4.1:

We start by determining the contribution per step in a valid walk. This quantity is defined by the expectation of an entry in $\mathbf{C}_N$ in its m-th moment, for a general integer m. To start, we define the distribution of an entry $\mathbf{C}_N(i,j)$, where $\mathbf{w}_i$ is the i-th element in $\mathbf{w}$,

$$\mathbf{C}_N(i,j) = \begin{cases} \frac{1}{\sqrt{\mathbf{w}_i}}(1 - \varepsilon_N r(i/N)r(j/N))\frac{1}{\sqrt{\mathbf{w}_j}}, & \text{with probability } p_{ij} = \varepsilon_N r(i/N)r(j/N), \\ \frac{1}{\sqrt{\mathbf{w}_j}}(-\varepsilon_N r(i/N)r(j/N))\frac{1}{\sqrt{\mathbf{w}_j}}, & \text{with probability } q_{ij} = 1 - p_{ij}. \end{cases} \tag{4.9}$$

We can thus calculate the expectation of $\mathbf{C}_N$ for a moment m by multiplying the outcomes by their corresponding supports, which in turn can then be bounded as the numerator is of the form $|p(1-p)^m + (1-p)(-p)^m|$, which is less than or equal to p itself. Thus, for $m \geq 2$,

$$|\mathbb{E}[\mathbf{C}_N^m(i,j)]| = \left|\frac{p\,(q)^m + (q)(-p)^m}{(\mathbf{w}_i\mathbf{w}_j)^{m/2}}\right| \leq \frac{\varepsilon_N r(i/N)r(j/N)}{(\mathbf{w}_i\mathbf{w}_j)^{m/2}}.$$

From here, we substitute in the definition of $\mathbf{w}_i$.

$$\frac{\varepsilon_N N r(i/N) r(j/N)}{(\mathbf{w}_i \mathbf{w}_j)^{m/2}} = \frac{\varepsilon_N r(i/N) r(j/N)}{\left(\varepsilon_N r(i/N) \sum_{i' \neq i} r(i'/N) \varepsilon_N r(j/N) \sum_{j' \neq j} r(j'/N)\right)^{m/2}},$$

and apply that $r(x)$ is also Riemann integrable as $f(x, y)$ is Riemann integrable, so we can group the integrals of $r(x)$,

$$\frac{\varepsilon_N r(i/N) r(j/N)}{\left(\varepsilon_N r(i/N) \sum_{i' \neq i} r(i'/N) \varepsilon_N r(j/N) \sum_{j' \neq j} r(j'/N)\right)^{m/2}} \approx \frac{\varepsilon_N}{\left(N \varepsilon_N \int_0^1 r(x) dx\right)^m \left(r\left(\frac{i}{N}\right) r\left(\frac{j}{N}\right)\right)^{\frac{m}{2}-1}},$$

as $N \to \infty$. As δ is the infinimum of $r(x)$, we can then upper bound this using δ, which then provides an index independent bound on $\mathbb{E}[\mathbf{C}_N^m(i, j)]$,

$$\frac{\varepsilon_N}{\left(N \varepsilon_N \int_0^1 r(x) dx\right)^m \left(r\left(\frac{i}{N}\right) r\left(\frac{j}{N}\right)\right)^{\frac{m}{2}-1}} \leq \frac{\varepsilon_N}{\left(N \varepsilon_N \int_0^1 r(x) dx\right)^m \delta^{m-2}}.$$

As such, for any given moment m, the expected contribution of an edge can be bounded,

$$|\mathbb{E}[\mathbf{C}_N^m(i, j)]| \leq \varepsilon_N \left[\left(N \varepsilon_N \int_0^1 r(x) dx\right)^m \delta^{m-2}\right]^{-1}. \tag{4.10}$$

It is important to reiterate that in using a combinatorial approach to quantify these moments that only edges that are walked at least twice contribute, as this is equivalent to counting only the connected nodes in the graph of a cyclic walk, such as in the example given in Definition 3.5.1. This graph can be represented as a rooted planar tree, for which a walk is invalid if each edge is not walked once on the walk away from the root and again back towards the root, leading to only edges walked twice being considered, and only "good" walks comprised of these "good" edges being counted in totality. In order to have a walk only of edges walked with a multiplicity of at least 2, let l be the number of unique edges in a valid walk, $e_1, \ldots, e_l$ be the edges walked, and $W_{l,k}$ be the set of all good walks. The maximal contribution of a component across these walks can be expressed by directly using (4.10) where the indexes (i, j) are given from the two sides of an edge e_h, $e_h(1)$ and $e_h(2)$,

$$\prod_{h=1}^{l} \left|\mathbb{E}\left[\mathbf{C}_N^{2k}(e_h(1), e_h(2))\right]\right| \leq \prod_{h=1}^{l} \frac{\varepsilon_N}{\left(N \varepsilon_N \int_0^1 r(x) dx\right)^{2k} \delta^{2k-2}},$$

after which the product can be re-expressed as a power over all terms given that it no longer has a relevant index. Thus, the overall contribution of a good walk is bounded by,

$$\prod_{h=1}^{l} \frac{\varepsilon_N}{\left(N \varepsilon_N \int_0^1 r(x) dx\right)^{2k} \delta^{2k-2}} = \varepsilon_N^l \delta^{2l} \prod_{h=1}^{l} \frac{1}{\left(N \varepsilon_N \int_0^1 r(x) dx \delta\right)^{2k}} = \frac{\varepsilon_N^l}{\left(N \varepsilon_N \int_0^1 r(x) dx\right)^{2k} \delta^{2k-2l}},$$

as an edge can be walked at most $2k$ times for a walk of $2k$ length. Given that this counting characterizes $\mathrm{Tr}(\mathbf{C}_N^{2k})$, we can then bound the expected trace of $\mathbf{C}_N$ in any even $2k$-th moment as the above quantity scaled by the total number of good walks, $|W_{l,k}|$,

$$\mathbb{E}\left[\mathrm{Tr}(\mathbf{C}_N^{2k})\right] \leq \sum_{l=0}^{k} |W_{l,k}| \frac{\varepsilon_N^l}{\left(N \varepsilon_N \int_0^1 r(x) dx\right)^{2k} \delta^{2k-2l}}. \tag{4.11}$$

Of course, we must quantify this enumeration, $|W_{l,k}|$, and derive an upper bound for it. To do so, we will rely on the following result from Füredi and Komlós [20], which states for all $l < N$,

$$|W_{l,k}| \leq N(N-1)\ldots(N-l)\binom{2k}{2l}\binom{2l}{l}\frac{1}{l+1}(l+1)^{4(k-l)} \leq N^{l+1} 4^l \binom{2k}{2l}(l+1)^{4(k-l)},$$

with the last inequality having shown to be a cleaner form of the initial result which can be used with no impact on the final result [10]. We now use this quantity to directly bound the inequality from (4.11) from above,

$$\mathbb{E}\left[\mathrm{Tr}(\mathbf{C}^{2k})\right] \leq \sum_{l=0}^{k} |W_{l,k}| \frac{\varepsilon_N^l}{\left(N\varepsilon_N \int_0^1 r(x)dx\right)^{2k} \delta^{2k-2l}} \leq \sum_{l=0}^{k} \frac{\varepsilon_N^l N^{l+1} 4^l \binom{2k}{2l} (l+1)^{4(k-l)}}{\left(N\varepsilon_N \int_0^1 r(x)dx\right)^{2k} \delta^{2k-2l}}. \tag{4.12}$$

For convenience of notation, we define $S(N,k,l)$ to be the quantity in the summation for a given edge count l,

$$S(N,k,l) \coloneqq \frac{\varepsilon_N^l N^{l+1} 4^l \binom{2k}{2l} (l+1)^{4(k-l)}}{\left(N\varepsilon_N \int_0^1 r(x)dx\right)^{2k} \delta^{2k-2l}}.$$

We now look to prove that this sum is dominated by its leading term, the k-th summand given as $S(N,k,k)$. In order to do so, we will see that for any $l \leq k-1$, that $S(N,k,l)/S(N,k,k)$ is vanishing, thus implying that $S(N,k,k)$ dominates the entire sum. Setting up this ratio and performing some simplification,

$$\frac{S(N,k,l)}{S(N,k,k)} = \frac{\varepsilon_N^l N^{l+1} 4^l \binom{2k}{2l} (l+1)^{4(k-l)}}{\left(N\varepsilon_N \int_0^1 r(x)dx\right)^{2k} \delta^{2k-2l}} \frac{\left(N\varepsilon_N \int_0^1 r(x)dx\right)^{2k} \delta^{2k-2k}}{\varepsilon_N^k N^{k+1} 4^k \binom{2k}{2k} (k+1)^{4(k-k)}} = \frac{\varepsilon_N^l N^{l+1} 4^l \binom{2k}{2l} (l+1)^{4(k-l)}}{\delta^{2k-2l} \varepsilon_N^k N^{k+1} 4^k}.$$

From here, we rearrange the expression to express as many terms as bases for the power $l-k$, and we use the identity $\binom{2k}{2l} \leq 2k^{2(k-l)}$ to bound with a simpler form. Additionally, we use that $l+1 \leq k$ to group the latter two terms, given our initial restriction on l,

$$\frac{\varepsilon_N^l N^{l+1} 4^l \binom{2k}{2l} (l+1)^{4(k-l)}}{\delta^{2k-2l} \varepsilon_N^k N^{k+1} 4^k} \leq \left(4\varepsilon_N N\delta^2\right)^{l-k} 2k^{2(k-l)} (l+1)^{4(k-l)} \leq 2\left(\frac{4\varepsilon_N N\delta^2}{k^6}\right)^{l-k}.$$

We can thus express a bound on $S(N,k,l)$ in terms of a factor of $S(N,k,k)$ by rearranging the initial ratio,

$$S(N,k,l) \leq 2\left(\frac{4\varepsilon_N N\delta^2}{k^6}\right)^{l-k} S(N,k,k).$$

We choose $k = g(N)\log(N)$, where $g(N)$ is an arbitrarily slowly growing function of N such that $\log(N) \gg g(N)$. Using then that $N\varepsilon_N \gg (\log(N))^{\xi}, \xi > 6$, we have that,

$$\left(\frac{4\varepsilon_N N\delta^2}{k^6}\right)^{l-k} \leq \left(\frac{4\delta^2 \log^{\xi}(N)}{g^6(N)\log^6(N)}\right)^{l-k} \leq \left(\frac{4\delta^2 \log^{\xi-6}(N)}{g^6(N)}\right)^{l-k}.$$

As $l-k \leq -1$, we can then state that $S(N,k,l)$ is a vanishingly small portion of $S(N,k,k)$. This can thus be absorbed within the sum in (4.11) as $(1+o(1))$ multiplied by the leading term,

$$\mathbb{E}[\mathrm{Tr}(\mathbf{C}^{2k})] \leq \sum_{l=0}^{k} S(N,k,l) \leq (1+o(1))S(N,k,k).$$

Substituting back in the initial definition of $S(N,k,k)$, we arrive at a bound expressed in terms of our graph properties,

$$(1+o(1))S(N,k,k) = (1+o(1))\frac{N(4\varepsilon_N N)^k}{\left(N\varepsilon_N \int_0^1 r(x)dx\right)^{2k}} = (1+o(1))N\left(\frac{2}{\sqrt{N\varepsilon_N}\int_0^1 r(x)dx}\right)^{2k}.$$

We can also use that $\int_0^1 r(x)dx \geq \delta$, by definition. Thus, this expression can be simplified again, as follows,

$$\mathbb{E}\left[\mathrm{Tr}(\mathbf{C}_N^{2k})\right] \leq (1+o(1))N\left(\frac{2}{\sqrt{N\varepsilon_N}\int_0^1 r(x)dx}\right)^{2k} \leq (1+o(1))N\left(\frac{2}{\sqrt{N\varepsilon_N}\delta}\right)^{2k}. \tag{4.13}$$

Having now determined the convergent value of the expectation of the trace of $\mathbf{C}$ for an even moment

$2k$, we now apply Markov's inequality to prove convergence for the leading eigenvalue, $\lambda_1(\mathbf{C}_N^{2k})$, as an extension of (4.8). For an error term Δ dependent on N and using the bound found in (4.13):

$$\mathbb{P}\left(\lambda_1(\mathbf{C}_N^{2k}) \geq (1+\Delta)^{2k}\left(\frac{2}{\sqrt{N\varepsilon_N}\delta}\right)^{2k}\right) \leq \frac{\mathbb{E}[\lambda_1(\mathbf{C}_N^{2k})]}{(1+\Delta)^{2k}\left(\frac{2}{\sqrt{N\varepsilon_N}\delta}\right)^{2k}} \leq \frac{(1+o(1))N\left(\frac{2}{\sqrt{N\varepsilon_N}\delta}\right)^{2k}}{(1+\Delta)^{2k}\left(\frac{2}{\sqrt{N\varepsilon_N}\delta}\right)^{2k}} = \frac{(1+o(1))N}{(1+\Delta)^{2k}}.$$

As k is of the order of $\log(N)$ and Δ is vanishing with N, this final quantity can be reduced as follows to show that this probability is $o(1)$,

$$\frac{(1+o(1))N}{(1+\Delta)^{2k}} \approx \frac{N}{(1+\Delta)^{2k}} = \exp\left(\log(N) - 2k(1+\Delta)\right) \leq \exp(-\log(N)) = N^{-1}.$$

We have thus shown that the probability the leading eigenvalue of the $2k$-th moment of $\mathbf{C}_N$ exceeds the bound defined in (4.13) is vanishing as a function of N, thus completing the proof that $|\lambda_1(\mathbf{C}_N)| \leq (1+o(1))\frac{2}{\sqrt{N\varepsilon_N}\delta}$, almost surely. ■

Having now shown the behaviour of the leading eigenvalue, we can examine the behaviour of the remaining eigenvalues to determine the overall behaviour of the spectra and quantify the spectral gap.

Lemma 4.4.2. *Let there be a graph satisfying $N\varepsilon_N \gg (\log(N))^{\xi}, \xi \geq 6$. For $\mathbf{P}$ being a Markov matrix,*

$$\max(\lambda_2(\mathbf{P}), -\lambda_N(\mathbf{P})) \leq \frac{2}{\sqrt{N\varepsilon_N}\delta}, \tag{4.14}$$

with high probability.

The approach to proving this statement is similar to the prior proof in bounding the degree of normalized adjacency matrices. We will first show that replacing $\mathbf{D}_N$ with its expectation, $\mathbf{W}_N$ is valid, and then bound the subsequent replaced statement using Lemma 4.3.1.

Proof of Lemma 4.4.2:

To start, we would like to show that the deviation between the empirical degree and the expected degree is vanishing with respect to their L^2-norm where for a matrix $\mathbf{X} \in \mathbb{R}^{N\times N}$, $\|\mathbf{X}\|_2 = \sqrt{\max_{1\leq i\leq N}\lambda_i(\mathbf{X}^T\mathbf{X})}$,

$$\|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2}\|_2 = o(1), \text{ w.h.p,}$$

recalling that $\mathbf{Q}_N \coloneqq \mathbf{D}_N^{-1/2}\mathbf{A}_N\mathbf{D}_N^{-1/2}$. To do so, observe that the following rearrangement of the above can use (3.5) to yield, almost surely:

$$\|\mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2} - \mathbf{I}_N\|_2 = \max_{1\leq i\leq N}\left|\sqrt{\frac{\mathbf{D}_i}{\mathbf{w}_i}} - 1\right| \leq \max_{1\leq i\leq N}\left|\sqrt{\frac{\mathbf{D}_i}{\varepsilon_N r(i/N)\sum_{j\neq i} r(j/N)}} - 1\right| = o(1). \tag{4.15}$$

Additionally, given that $\mathbf{P}$ is the unsymmetrized version of $\mathbf{Q}$, and both are stochastic matrices with a maximum eigenvalue of 1, it is clear that,

$$\|\mathbf{Q}_N\|_2 = \max_{1\leq i\leq N}|\lambda_i(\mathbf{Q}_N)| = \max_{1\leq i\leq N}|\lambda_i(\mathbf{P}_N)| = 1. \tag{4.16}$$

With these two equivalences established, we can now approach the convergence in L^2 of $\mathbf{Q}_N$ and $\mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2}$ directly, first by manipulating to create another instance of $\mathbf{Q}_N$ in the difference,

$$\begin{aligned}\|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2}\|_2 &= \|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{D}_N^{-1/2}\mathbf{A}_N\mathbf{D}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2}\|_2 \\ &= \|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N\mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2}\|_2.\end{aligned}$$

We additionally introduce $-\mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N + \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N$ into the equation above,

$$\|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N\mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2}\|_2 = \|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N + \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N\mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2}\|_2.$$

We can now factor out some common terms and group this one large norm into separate norms by applying the triangle inequality of norms followed by the submultiplicativity of matrix norms.

$$\begin{aligned}&\|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N + \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N\mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2}\|_2 \\ =&\|(\mathbf{I}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2})\mathbf{Q}_N\|_2 + \|\mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\mathbf{Q}_N(\mathbf{I} - \mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2})\|_2 \\ \leq&\|(\mathbf{I}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2})\|_2\|\mathbf{Q}_N\|_2 + \|\mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\|_2\|\mathbf{Q}_N\|_2\|\mathbf{I}_N - \mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2}\|_2.\end{aligned} \tag{4.17}$$

Observe that we can now apply (4.15) and (4.16) in combination with the triangle identity for norms, $\|\mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\|_2 \leq \|\mathbf{I}_N\|_2 + \|\mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2} - \mathbf{I}_N\|_2 = 1 + o(1)$ a.s., again by the above observation and the straightforward result that $\|\mathbf{I}_N\|_2 = 1$. Putting these results together, we get that, almost surely,

$$\|(\mathbf{I}_N - \mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2})\|_2\|\mathbf{Q}_N\|_2 + \|\mathbf{W}_N^{-1/2}\mathbf{D}_N^{1/2}\|_2\|\mathbf{Q}_N\|_2\|\mathbf{I}_N - \mathbf{D}_N^{1/2}\mathbf{W}_N^{-1/2}\|_2 = o(1).$$

We can now apply the identity for the spectra of square Hermitian matrices where for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, $|\lambda_i(\mathbf{A}+\mathbf{B}) - \lambda_i(\mathbf{A})| \leq \|\mathbf{B}\|_2$, where i are the indexes of the ordered eigenvalues. This provides a bound for the eigenvalue-wise difference between the two matrices,

$$|\lambda_i(\mathbf{Q}_N) - \lambda_i(\mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2})| \leq \|\mathbf{Q}_N - \mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2}\|_2 = o(1).$$

As such, we have proven that a substitution of $\mathbf{D}_N$ by $\mathbf{W}_N$ makes a negligible difference in L^2 and that by extension the spectra of the two matrices are also close. We now seek to show that normalizing the adjacency matrix by $\mathbf{W}_N$ yields vanishingly small eigenvalues beyond the largest one. To do so, we first observe that $\boldsymbol{\chi}_N^T\boldsymbol{\chi}_N$ yields one a rank-one matrix, and as such $\lambda_i(\boldsymbol{\chi}_N^T\boldsymbol{\chi}_N) = 0$ for $i \geq 2$. Thus, for $i \geq 2$:

$$|\lambda_i(\mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2})| = |\lambda_i(\mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2}) - \lambda_i(\boldsymbol{\chi}_N^T\boldsymbol{\chi}_N)|.$$

Again, applying the identity for square Hermitian matrices and subsequently using the result from (4.7), we can bound all remaining eigenvalues,

$$|\lambda_i(\mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2}) - \lambda_i(\boldsymbol{\chi}_N^T\boldsymbol{\chi}_N)| \leq \|\mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2} - \boldsymbol{\chi}_N^T\boldsymbol{\chi}_N\|_2 \leq \frac{2}{\sqrt{N\varepsilon_N}\delta}.$$

Having now established that both the spectra of $\mathbf{Q}_N$ and the spectra of $\mathbf{W}_N^{-1/2}\mathbf{A}_N\mathbf{W}_N^{-1/2}$ are close, and additionally that the spectra of the latter for $\lambda_i, i \geq 2$ are at most $\frac{2}{\sqrt{N\varepsilon_N}\delta}$, we can combine these results to state that the eigenvalues of $\mathbf{Q}_N$ beyond the leading eigenvalue must also be of the order:

$$\max_{i\geq 2}|\lambda_i(\mathbf{Q}_N)| \leq \frac{2}{\sqrt{N\varepsilon_N}\delta}, \text{ w.h.p.}$$

Finally, as $\max(\lambda_2(\mathbf{P}_N), -\lambda_N(\mathbf{P}_N)) = \max_{i\geq 2}|\lambda_i(\mathbf{Q}_N)|$ given that the two are the symmetrized and unsymmetrized counterparts of the same matrix, we have that, w.h.p.,

$$\max(\lambda_2(\mathbf{P}_N), -\lambda_N(\mathbf{P}_N)) \leq \frac{2}{\sqrt{N\varepsilon_N}\delta}. \quad \blacksquare$$

Remark 4.4.1. *The bound derived in Lemma 4.4.2 is both the upper and lower bound for the behaviour of the spectral gap. We can see this through a straightforward manipulation of* $\mathrm{ESD}(\mathbf{P}_N)$*,*

$$\mathrm{ESD}(\mathbf{P}_N) = \mathrm{ESD}\left(\mathbf{D}_N^{-1}\mathbf{A}_N\right) \approx \mathrm{ESD}\left(\mathbf{D}_N^{-1/2}\mathbf{A}_N\mathbf{D}_N^{-1/2}\right) \approx \frac{1}{\sqrt{N\varepsilon_N}}\mathrm{ESD}\left(\frac{\mathbf{A}_N - \mathbb{E}[\mathbf{A}_N]}{\sqrt{N\varepsilon_N}}\right). \tag{4.18}$$

$\mathrm{ESD}\left(\frac{\mathbf{A}_N - \mathbb{E}[\mathbf{A}_N]}{\sqrt{N\varepsilon_N}}\right)$ *recovers the semicircle law which is bounded strictly to* $[-2, 2]$ *[43], thus yielding that the bounds of* $\mathrm{ESD}(\mathbf{P}_N)$ *are approximately* $\left(-\frac{2}{\sqrt{N\varepsilon_N}}, \frac{2}{\sqrt{N\varepsilon_N}}\right)$*, as in Lemma 4.4.2. In fact, due to the strictness of the bound from the semicircle law, this in turn is an optimal bound for the spectral gap.*

4.5 Proof of Theorem 4.2.1

We have now shown that in our setting for inhomogeneous random graphs with connection probabilities $p_{ij} = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)$ that we meet the two criteria required in Theorem 4.2.1. Namely, that:

1. $\mathbf{d}_{\max} \leq K\mathbf{d}_{\min}, K > 0$, as shown in the proof of Lemma 4.3.1.
2. $\max\{|\lambda_2|, |\lambda_N|\} \leq \frac{2}{\sqrt{N\varepsilon_N\delta}}$, as shown in the proof of Lemma 4.4.2.

Thus, we can now complete the proof of the theorem.

Proof of Theorem 4.2.1:

We first note that α as defined in (4.1) is a dampening parameter, and when set to 0 or 1 the theorem can be seen to hold with minimal manipulation. This is illustrated below.

$\underline{\alpha = 0:}$

We restate the definition of $\tilde{\mathbf{P}}_N$ prior to substitution of $\alpha = 0$,

$$\tilde{\mathbf{P}}_N = \alpha\mathbf{P}_N + (1-\alpha)\mathbf{v}\mathbf{1}^T = \mathbf{v}\mathbf{1}^T.$$

Using this definition of $\tilde{\mathbf{P}}_N$ in that of $\boldsymbol{\pi}_N$ then gives,

$$\boldsymbol{\pi}_N = \tilde{\mathbf{P}}_N\boldsymbol{\pi}_N = \mathbf{v}\mathbf{1}^T\boldsymbol{\pi}_N = \mathbf{v}.$$

The final equality arises as $\mathbf{v}\mathbf{1}^T = 1$ given that $\mathbf{v}$ is a probability vector. $\overline{\boldsymbol{\pi}}_N$, the approximation of $\boldsymbol{\pi}_N$, is thus, for $\alpha = 0$:

$$\overline{\boldsymbol{\pi}}_N = \frac{\alpha\mathbf{d}}{\text{vol}(G_N)} + (1-\alpha)\mathbf{v} = \mathbf{v}.$$

Clearly then, as both $\boldsymbol{\pi}_N$ and $\overline{\boldsymbol{\pi}}_N$ are equal to $\mathbf{v}$, the theorem holds.

$\underline{\alpha = 1:}$

Plugging this into $\tilde{\mathbf{P}}_N$, we simply have:

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

which when placed into the definition of $\boldsymbol{\pi}_N$, simply yields $\boldsymbol{\pi}_N = \mathbf{P}_N\boldsymbol{\pi}_N$. Again, we can recalculate $\overline{\boldsymbol{\pi}}_N$ to provide $\overline{\boldsymbol{\pi}}_N = \frac{\mathbf{d}}{\text{vol}(G_N)}$. In the case of a connected graph, such as when $N\varepsilon_N \gg \log(N)$ as in this setting, this is equivalent to $\mathbf{P}_N \coloneqq \mathbf{A}_N\mathbf{D}_N^{-1}$, and thus the theorem holds.

$\underline{0 < \alpha < 1:}$

We now look to assess the more complicated scenario where $0 < \alpha < 1$. To begin, we decompose $\mathbf{Q}_N = \mathbf{D}_N^{-1/2}\mathbf{A}_N\mathbf{D}_N^{-1/2}$ into its spectral components for $1 = \lambda_1 \geq \cdots \geq \lambda_N$ being the eigenvalues corresponding to the eigenvectors $\{\mathbf{u}_i\}_{i=1}^N$, $\mathbf{u}_i \in \mathbb{R}^N, \|\mathbf{u}_i\|_2 = 1$,

$$\mathbf{Q}_N = \lambda_1\mathbf{u}_1\mathbf{u}_1^T + \sum_{i=2}^N \lambda_i\mathbf{u}_i\mathbf{u}_i^T = \mathbf{u}_1\mathbf{u}_1^T + \sum_{i=2}^N \lambda_i\mathbf{u}_i\mathbf{u}_i^T.$$

We then substitute this decomposition into the resolvent definition of $\boldsymbol{\pi}_N$:

$$\boldsymbol{\pi}_N = (1-\alpha)\mathbf{D}_N^{1/2}\left[\mathbf{I}_N - \alpha\mathbf{Q}_N\right]^{-1}\mathbf{D}_N^{-1/2}\mathbf{v} = (1-\alpha)\mathbf{D}_N^{1/2}\left[\frac{\mathbf{u}_1\mathbf{u}_1^T}{1-\alpha} + \sum_{i=1}^N \frac{\mathbf{u}_i\mathbf{u}_i^T}{1-\alpha\lambda_i}\right]\mathbf{D}_N^{-1/2}\mathbf{v}.$$

Expanding the quantity in the brackets by multiplying in $(\mathbf{I}_N - \alpha)$ at each entry i for each corresponding orthogonal eigenvectors $\mathbf{u}_i\mathbf{u}_i^T$ yields a separated leading eigenvector quantity and a remaining spectral quantity,

$$(1-\alpha)\mathbf{D}_N^{1/2}\left[\frac{\mathbf{u}_1\mathbf{u}_1^T}{1-\alpha} + \sum_{i=1}^N \frac{\mathbf{u}_i\mathbf{u}_i^T}{1-\alpha\lambda_i}\right]\mathbf{D}_N^{-1/2}\mathbf{v} = \mathbf{D}_N^{1/2}\mathbf{u}_1\mathbf{u}_1^T\mathbf{D}_N^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}_N^{1/2}\sum_{i=2}^N(1-\alpha\lambda_i)^{-1}\mathbf{u}_i\mathbf{u}_i^T\mathbf{D}_N^{-1/2}\mathbf{v}.$$

Having now rearranged the initial form of $\boldsymbol{\pi}_N$, we introduce the error between $\boldsymbol{\pi}_N$ and $\overline{\boldsymbol{\pi}}_N$,

$$\Delta = \boldsymbol{\pi}_N - \overline{\boldsymbol{\pi}}_N.$$

Performing similar manipulations to $\overline{\boldsymbol{\pi}}$ as we have done for $\boldsymbol{\pi}$ provides a form that is easier to work with when evaluating the error in terms of total variation. Using that $\mathbf{v}$ is a probability vector, we can thus expand the approximation,

$$\overline{\boldsymbol{\pi}}_N = \alpha\frac{\mathbf{d}}{\text{vol}(G_N)} + (1-\alpha)\mathbf{v} = \alpha\frac{\mathbf{D}_N\mathbf{1}\mathbf{1}^T\mathbf{v}}{\text{vol}(G_N)} + (1-\alpha)\mathbf{v} = \alpha\frac{\mathbf{D}_N\mathbf{1}\mathbf{1}^T\mathbf{v}}{\text{vol}(G_N)} + (1-\alpha)\mathbf{D}_N^{1/2}\mathbf{D}_N^{-1/2}\mathbf{v}.$$

$\overline{\boldsymbol{\pi}}_N$ can thus be expressed in terms of $\mathbf{u}_i$ with an additional factor:

$$\begin{aligned}\overline{\boldsymbol{\pi}}_N =& \alpha\frac{\mathbf{D}_N\mathbf{1}\mathbf{1}^T\mathbf{v}}{\text{vol}(G_N)} + (1-\alpha)\mathbf{D}_N^{1/2}\mathbf{D}_N^{-1/2}\mathbf{v}\\ =& \alpha\mathbf{D}_N^{1/2}\frac{\mathbf{D}_N^{1/2}\mathbf{1}}{\sqrt{\text{vol}(G_N)}}\frac{\mathbf{1}^T\mathbf{D}_N^{1/2}}{\sqrt{\text{vol}(G_N)}}\mathbf{D}_N^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}_N^{1/2}\mathbf{D}_N^{-1/2}\mathbf{v}.\end{aligned}$$

We note that the Perron-Frobenius theorem allows us to express the leading eigenvector as,

$$\mathbf{u}_1 = \frac{\mathbf{D}^{1/2}\mathbf{1}}{\sqrt{\mathbf{1}^T\mathbf{D}_N\mathbf{1}}}\frac{\mathbf{D}_N^{1/2}\mathbf{1}}{\sqrt{\text{vol}(G_N)}},$$

which then can be used in the prior expression to express the first quantity in terms of the eigenvectors of the leading eigenvalue,

$$\begin{aligned}\overline{\boldsymbol{\pi}}_N =& \alpha\mathbf{D}_N^{1/2}\frac{\mathbf{D}_N^{1/2}\mathbf{1}}{\sqrt{\text{vol}(G_N)}}\frac{\mathbf{1}^T\mathbf{D}_N^{1/2}}{\sqrt{\text{vol}(G_N)}}\mathbf{D}_N^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}_N^{1/2}\mathbf{D}_N^{-1/2}\mathbf{v}\\ =& \alpha\mathbf{D}_N^{1/2}\mathbf{u}_1\mathbf{u}_1^T\mathbf{D}_N^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}_N^{1/2}\mathbf{D}_N^{-1/2}\mathbf{v}.\end{aligned}$$

Having now rearranged the original PageRank measure and its approximation in terms of their eigenvector decompositions while accounting for the necessary degree normalization, we can evaluate the error Δ, directly,

$$\begin{aligned}\Delta = \boldsymbol{\pi}_N - \overline{\boldsymbol{\pi}}_N =& \mathbf{D}_N^{1/2}\mathbf{u}_1\mathbf{u}_1^T\mathbf{D}_N^{-1/2}\mathbf{v} + (1-\alpha)\mathbf{D}_N^{1/2}\sum_{i=2}^{N}(1-\alpha\lambda_i)^{-1}\mathbf{u}_i\mathbf{u}_i^T\mathbf{D}_N^{-1/2}\mathbf{v}\\ &- \alpha\mathbf{D}_N^{1/2}\mathbf{u}_1\mathbf{u}_1^T\mathbf{D}_N^{-1/2}\mathbf{v} - (1-\alpha)\mathbf{D}_N^{1/2}\mathbf{D}_N^{-1/2}\mathbf{v}\\ =& (1-\alpha)\mathbf{D}_N^{1/2}\left(\left(\sum_{i=2}^{N}\frac{\mathbf{u}_i\mathbf{u}_i^T}{1-\alpha\lambda_i}\right) - \mathbf{I}_N + \mathbf{u}_1\mathbf{u}_1^T\right)\mathbf{D}_N^{-1/2}\mathbf{v}.\end{aligned}$$

As $\mathbf{u}_i$ form an orthonormal basis, $\sum_i^N \mathbf{u}_i\mathbf{u}_i^T = \mathbf{I}_N$, and so removing the leading eigenvector readjusts the starting index the summation across the remaining eigenvectors. Thus, we can rewrite purely in terms of the trailing eigenvectors, as follows :

$$\begin{aligned}\Delta =& (1-\alpha)\mathbf{D}_N^{1/2}\left(\sum_{i=2}^{N}\left(\frac{\mathbf{u}_i\mathbf{u}_i^T}{1-\alpha\lambda_i}\right) - \mathbf{I}_N + \mathbf{u}_1\mathbf{u}_1^T\right)\mathbf{D}_N^{-1/2}\mathbf{v}\\ =& (1-\alpha)\mathbf{D}_N^{1/2}\left(\left(\sum_{i=2}^{N}\frac{\mathbf{u}_i\mathbf{u}_i^T}{1-\alpha\lambda_i}\right) - \left(\sum_{i=2}^{N}\mathbf{u}_i\mathbf{u}_i^T\right)\right)\mathbf{D}_N^{-1/2}\mathbf{v}\\ =& (1-\alpha)\mathbf{D}_N^{1/2}\left(\sum_{i=2}^{N}\mathbf{u}_i\mathbf{u}_i^T\left(\frac{\alpha\lambda_i}{1-\alpha\lambda_i}\right)\right)\mathbf{D}_N^{-1/2}\mathbf{v}.\end{aligned}$$

We look to bound the error across all remaining eigenvectors, and to do so, we will use the L^1 norm. Thus, applying the Cauchy-Schwarz inequality for the L^1 norm, where for a vector $\mathbf{x} \in \mathbb{R}^N$, $\|\mathbf{x}\|_1 \leq \sqrt{N}\|\mathbf{x}\|_2$, and additionally normalizing by $1-\alpha$, we have,

$$\frac{\|\Delta\|_1}{1-\alpha} \leq \sqrt{N}\frac{\|\Delta\|_2}{1-\alpha}.$$

Expanding $\|\Delta\|_2$ via its simplified form above and using the submultiplicativity of matrix norms, we have,

$$\frac{\|\Delta\|_1}{1-\alpha} \leq \|\mathbf{D}_N^{1/2}\|_2 \left\| \sum_{i=2}^{N} \mathbf{u}_i \mathbf{u}_i^T \left(\frac{\alpha\lambda_i}{1-\alpha\lambda_i} \right) \right\|_2 \|\mathbf{D}_N^{-1/2}\|_2 \|\mathbf{v}\|_2 = \sqrt{N}\sqrt{\frac{\mathbf{d}_{\max}}{\mathbf{d}_{\min}}} \max_{i>1} \left| \frac{\alpha\lambda_i}{1-\alpha\lambda_i} \right| \|\mathbf{v}\|_2.$$

The last equality arises from taking the L^2-norm over the diagonalized degree vector, which yields either $\mathbf{d}_{\max}$ or $\mathbf{d}_{\min}$. The summation can now be bounded, as the L^2-norm of a symmetric matrix is bounded above by its maximum eigenvalue. We now apply the second property regarding the spectral gap, $\max\{|\lambda_2|, |\lambda_N|\} \leq \frac{2}{\sqrt{N\varepsilon_N}\delta}$ to state that any $\lambda_i = o(1)$ w.h.p. $\forall i > 1$. Additionally, as the spectral gap is bounded away from 0, as established in the proof of Lemma 4.4.2, there is a constant $C, |1-\alpha\lambda_i| > C$, and as assumed earlier, $\|\mathbf{v}\|_2 = O(1/\sqrt{N})$. This yields the following upper bound,

$$\sqrt{N}\sqrt{\frac{\mathbf{d}_{\max}}{\mathbf{d}_{\min}}} \max_{i>1} \left| \frac{\alpha\lambda_i}{1-\alpha\lambda_i} \right| \|\mathbf{v}\|_2 \leq C\sqrt{\frac{\mathbf{d}_{\max}}{\mathbf{d}_{\min}}} \max\{|\lambda_2|, |\lambda_N|\},$$

Having now established that the ratio between the maximum and minimum degree is bounded by a constant K w.h.p via the first property, and that $\max\{|\lambda_2|, |\lambda_N|\} \leq \frac{2}{\sqrt{N\varepsilon_N}\delta}$ w.h.p via the second, the final inequality can be expressed as,

$$\|\Delta\|_1 \leq C\sqrt{\frac{\mathbf{d}_{\max}}{\mathbf{d}_{\min}}} \max\{|\lambda_2|, |\lambda_N|\} \leq CK\frac{2}{\sqrt{N\varepsilon_N}\delta} = \frac{K}{\sqrt{N\varepsilon_N}},$$

w.h.p, absorbing the constants into K. And as such the initial theorem holds given the definition of d_{TV} as an L^1 distance measure between two probability vectors,

$$d_{TV}(\boldsymbol{\pi}_N, \overline{\boldsymbol{\pi}}_N) = \frac{K}{\sqrt{N\varepsilon_N}} = o(1), \quad \text{w.h.p, as } N \to \infty. \quad \blacksquare$$

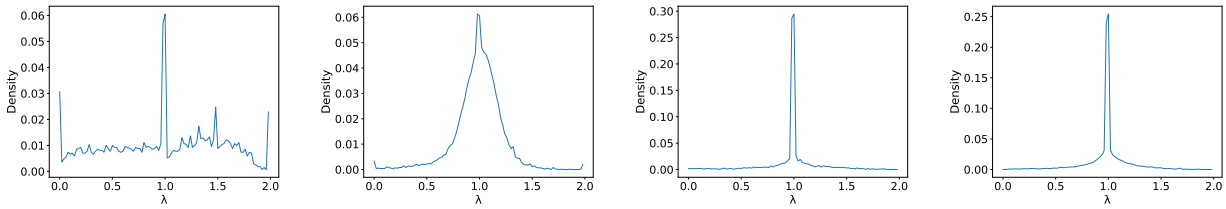
This result thus shows that the approximation method using the volume of G_N in place of directly calculating the PageRank measure as proposed by [2] can be extended beyond both the Chung-Lu and Stochastic Block settings in their initial work to all generalized inhomogeneous graphs, contingent on the graph being well-connected such that $N\varepsilon_N \gg \log^6(N)$. Additionally, we have quantified that the error, as measured by d_{TV}, is not only $o(1)$, but vanishes at a rate governed by $\sqrt{N\varepsilon_N}$, again supporting that we require a well-connected graph for this approximation to hold. Finally, we note that this final bound in distance between the two vectors is tight, as δ can be selected to have no dependence on $r(i/N)$. We now move to our final result concerning graphical neural networks.

Chapter 5
Degree Based Eigenvalue Correction

This section will discuss the filter functions that have been selected for testing, including definitions of how these filters produce a final prediction vector for classification. We then discuss our implementation of these filters, the eigenvalue correction strategies assessed, and how the datasets were masked for training. Finally, we will define how the performance of these filters was assessed.

5.1 Selected Polynomial Spectral Filters

GNNs are powerful predictive tools when applied to real-world networks, but are fundamentally dependent on the spectral behavior of the graph upon which they operate. This is evident when inspecting the formal definition of the application of a spectral filter function, h_θ, as in (1.11). However, polynomial methods to approximate h_θ have outputs limited by the diversity of eigenvalues being passed through them as these functions will produce the same output for identical eigenvalues, reducing the potential diversity of information provided by the filter. Thus, applying spectral filters in GNNs to data yields suboptimal results in scenarios where there are eigenvalues with multiplicities greater than 1 [33]. This issue of high multiplicity is particularly prevalent in real-world datasets where there are typically a bulk of eigenvalues seen near 1 for $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, as can be seen in Figure 4, thus reducing the expressive power of the final resulting network.



(a) Cora (b) Photo (c) Chameleon (d) Squirrel

Figure 4: Eigenvalue distributions of the normalized Laplacian matrix, $\mathbf{L}$, of real-world graph datasets. Note that the distributions have concentration of mass near 1, indicating a large number of unconnected nodes.

There is a wide range of learnable polynomial filter functions for spectral GNNs that approximate the behavior of signal filters, such as low-pass, high-pass, or band filters. We have selected three learnable filters with which to assess the performance of various eigenvalue correction strategies, as learnable filters are more responsive to the complexities of inhomogeneous real-world data [33]. As a reminder, these filters are used as an approximation of the full computation $Uh(\Lambda)U^T$ where $h : [0, 2] \to [0, 1]$ is a filter function, and so we can express polynomial filters that adhere to this form as follows, for $\mathbf{Z}$ being the final fully convolved signal, $\mathbf{X} \in \mathbb{R}^{N\times d}$ being the input signal, $\mathbf{W}$ being a weight matrix, and K being the number of layers in the GNN:

$$\mathbf{Z} = h(Uh(\Lambda)U^T)\mathbf{X}\mathbf{W} = U\sum_{k=0}^{K} h_k(\Lambda)U^T\mathbf{X}\mathbf{W}. \tag{5.1}$$

The selected filters for comparison are in line with those tested in [33], with additional depth on their methodology from their corresponding papers [9, 26, 46].

5.1.0.1 Generalized PageRank

PageRank measures offer means to quantify the connectivity of individual nodes in a graph, with traditional PageRank measures using heuristically chosen weights for each length of random walk [32]. We note that we will switch from the notation used in the prior chapter for PageRank to remain in line with the literature surrounding the use of PageRank in the context of learning problems [33]. Let s be a starting seed node, such that $s \in V$, and that there exists an arbitrary node $v \in V$, such that $s \neq v$. The PageRank score for v is then the summed landing probabilities across all random walks of length k that begin on s and end on v,

$$pr(\gamma, x^{(0)}) = \sum_{k=0}^{\infty} \gamma_k x_v^{(k)}, \gamma = \{\gamma_k\}_{k \geq 0}, \tag{5.2}$$

where $\gamma \in \mathbb{R}^K$ is the vector of non-negative weights for each random walk length and $x^{(0)}$ is the initial distribution, $x : V \to [0, 1]$, where $\sum_{v \in V} x_v = 1$ so x_v is the probability of a vertex v. In PageRank measures such as Personalized PageRank, γ is selected as a fixed value, such as $\gamma_k = (1-\alpha)\alpha^k, \alpha \in (0, 1)$[28]. In contrast, application of Generalized PageRank (GPR) in polynomial graph filters sets γ to be a vector of learnable parameters for the model, and so in learning the optimal GPR weights to best represent the underlying graph structure an optimal polynomial graph filter is also learned [9]. Integrating this with the full filtering process described in (5.1), $h_k(\cdot)$ can be replaced by $pr(\cdot)$ in (5.2):

$$\mathbf{Z} = U pr(\gamma, \Lambda) U^T \mathbf{X}\mathbf{W} = U \sum_{k=0}^{K} \text{diag}\left(\gamma_k \Lambda^k\right) U^T \mathbf{X}\mathbf{W}. \tag{5.3}$$

Note that in passing Λ to $pr(\cdot)$, pr operates element-wise across the vector, where $pr(\gamma, \Lambda)$ yields $pr(\gamma, \lambda_i) = \sum_{k=0}^{K} \gamma_k \lambda_i^k$ for each λ_i. Additionally, applying GPR to the GNN architecture helps to mitigate the impacts of over-smoothing, a phenomenon which occurs when higher order elements of the polynomial dominate in contribution to the final values of $\mathbf{Z}$. In having order-specific coefficients via γ, higher order summands can be down-weighted, and so the weights learned by GPR should decrease as the order grows [9]. In practice, this is also dependent on the optimization algorithm selected, as the learning rate of the components of γ must also have a suitable learning rate and corresponding learning rate decay, lest higher order contributors be penalized too heavily. The implementation of GPR-GNN into the forward propagation function in PyTorch can be seen in the pseudocode below [36]. The $\odot$ operator is the element-wise product between the output vector and $U^T\mathbf{X}$, otherwise known as the Hadamard product.

Algorithm 1 GPR_Prop: Propagation Layer for GPR-GNN Message Passing

Initializing Values: Polynomial maximum order K, transition probability α
Initialize:
$\gamma_k \leftarrow \alpha(1-\alpha)^k, \quad 0 \leq k < K$
$\gamma_K \leftarrow (1-\alpha)^K$
function FORWARD($\mathbf{X}$, Λ, U)
    Data: Node features $\mathbf{X}$, eigenvalues of $\mathbf{L}$ Λ, eigenvectors of $\mathbf{L}$ U
    `out` $\leftarrow \gamma_0 \cdot (\Lambda)^0$
    for $k = 1$ to K **do**
        `out` $\leftarrow$ `out` $+ \gamma_k \cdot (\Lambda)^k$
    end for
    return $U \cdot$ (`out` $\odot (U^T\mathbf{X})$)
end function

Centrality measures such as PageRank are thus useful in approximating filter functions on GNNs, as they are directly motivated by the connectedness of the underlying graph. Integrating a learnable vector of weights, γ_k, as opposed to a fixed coefficient, for each step in a walk in GPR further allows these graphical methods to approximate complex filter functions.

5.1.0.2 BernNet

Polynomial filters are in themselves polynomial functions, and so methods to approximate complicated polynomial functions are also applicable in the context of signal filters in GNNs. One such approximation can be done using Bernstein polynomials, which aim to re-express a polynomial function as a linear combination of Bernstein basis polynomials [18]. A basis polynomial of the order k takes the following form for a maximum order K and $t \in [0, 1]$:

$$b_k^K(t) \coloneqq \binom{K}{k}(1-t)^{K-k}t^k. \tag{5.4}$$

Thus, to approximate an arbitrary continuous function f, the linear combination can be constructed as follows, with $\{\theta_k\}_{1\leq k\leq K} \in \mathbb{R}^K$ being the coefficients for their corresponding order bases,

$$f(t) \approx p_K(t) \coloneqq \sum_{k=0}^{K} \theta_k \cdot b_k^K(t) = \sum_{k=0}^{K} \theta_k \cdot \binom{K}{k}(1-t)^{K-k}t^k. \tag{5.5}$$

For $\theta_k = f\left(\frac{k}{K}\right)$, it can be shown that $p_K(t) \to f(t), K \to \infty$ [26]. As the bases themselves are explicitly defined as a function of the order and the function input, the problem then focuses on defining θ_k in the event that f is also unknown, which is typically the case. We first reintroduce the function, $h : [0,2] \to [0,1]$, that we are seeking to approximate from (5.1). Let t in the above definition be $\frac{\lambda}{2}$, for an eigenvalue λ, thus restricting the input of f to be applicable for the Bernstein polynomial approximation. This then yields the following form for $p_K(t)$,

$$p_K\left(\frac{\lambda}{2}\right) = \sum_{k=0}^{K} \theta_k \binom{K}{k}\left(1-\frac{\lambda}{2}\right)^{K-k}\left(\frac{\lambda}{2}\right)^k = \sum_{k=0}^{K} \theta_k \binom{K}{k}\frac{1}{2^K}(2-\lambda)^{K-k}\lambda^k. \tag{5.6}$$

In order for this approximation to achieve $p_K\left(\frac{\lambda}{2}\right) \to h(\lambda), K \to \infty$, it must be that $\theta_k = h\left(\frac{2k}{K}\right)$, in the event that h is an explicitly defined function. This then yields the following filtering process to produce the final output signal $\mathbf{Z}$ as in (5.1) where $\lambda_1, \ldots, \lambda_N$ are the individual eigenvalue entries of Λ,

$$\mathbf{Z} = U \operatorname{diag}\left[p_K\left(\frac{\lambda_1}{2}\right), \ldots, p_K\left(\frac{\lambda_N}{2}\right)\right] U^T\mathbf{X}\mathbf{W}. \tag{5.7}$$

In the case of approximating a linear high-pass filter of the form $\frac{\lambda}{2}$ or a linear low-pass filter of the form $1-\frac{\lambda}{2}$, this can be calculated directly to be $\frac{k}{K}$ and $1-\frac{k}{K}$, respectively. However, for more complex filters or filters for which the final form is unknown, the parameter vector θ_k can be approximated using gradient descent methods [26]. Within this optimization process, there are additional constraints placed on the behavior of the filter, namely that the filter itself must be non-negative. To motivate this constraint, we view the task of optimizing p_K to be a graph optimization problem, which we define as the following loss function,

$$\min_{\mathbf{Z}} f(\mathbf{Z}) = (1-\alpha)\mathbf{Z}^T\eta(\mathbf{L})\mathbf{Z} + \alpha\|\mathbf{Z}-\mathbf{X}\|_2^2, \tag{5.8}$$

where $\alpha \in [0,1)$ and η is a filtering function operating on the spectra of the normalized Laplacian, $\mathbf{L}$. As in standard optimization problems, we thus seek the minima of this function by finding the value of $\mathbf{Z}$ at which its derivative is zero, $\frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}} = 0$ [49]. When evaluating the optimization problems, there are thus requirements on the loss function such that a solution can be found, such as convexity. In this case, for $f(\mathbf{Z})$ to be convex, $\eta(\mathbf{L})$ must also be positive semidefinite to avoid saddle points and $f(\mathbf{Z})$ going to $-\infty$. This is clear, as in the event that $\mathbf{Z}^T\eta(\mathbf{L})\mathbf{Z} < \mathbf{0}$ for some signal $\mathbf{Z}$, the loss has the potential to be non-convex or unbounded from below, which potentially then yields a nonsensical solution. Therefore, we must restrict the problem to conditions in which $\eta(\mathbf{L})$ is positive semidefinite, which we can express as follows for a general polynomial filter $g(\lambda) = \sum_{k=0}^{K}\gamma_k\lambda^k$. We note that all eigenvalues of $\mathbf{L}$ will lie in the interval $[0,2]$ [43], and so we can restrict the behavior of g to just the interval $[0,2]$,

$$0 \leq g(\lambda) \leq 1, \forall\lambda \in [0,2]. \tag{5.9}$$

The upper bound on this filter arises from taking the solution to the derivative of (5.8). The first derivative is as follows,

$$f'(\mathbf{Z}) = 2(1-\alpha)\eta(\mathbf{L})\mathbf{Z} + 2\alpha(\mathbf{Z}-\mathbf{X}), \tag{5.10}$$

after which we can then set the equation to $\mathbf{0}$ to find the optima.

$$\begin{aligned}
\mathbf{0} &= 2(1-\alpha)\eta(\mathbf{L})\mathbf{Z} + 2\alpha(\mathbf{Z}-\mathbf{X}) \\
\alpha\mathbf{X} &= \alpha\mathbf{Z} + \eta(\mathbf{L})\mathbf{Z} - \alpha\eta(\mathbf{L})\mathbf{Z} \\
\mathbf{Z} &= \alpha\left[\alpha + \eta(\mathbf{L})(1-\alpha)\right]^{-1}\mathbf{X},
\end{aligned}$$

for which we know that this is the minima as $\eta(\mathbf{L})$ is positive definite, thus yielding a convex loss function. Applying this element-wise to each λ_i in Λ, for Λ being the diagonalized eigenvalues of $\mathbf{L}$, we have that for $\lambda_i \in [0,2]$, $\alpha\left[\alpha+\lambda_i(1-\alpha)\right]^{-1} \leq 1$, specifically in the scenario where $\lambda_i = 0$. This then yields the upper bound given in (5.9), as for $g(\lambda)$ being an approximation of the function h, it should abide by the same bounds. The lower bound arises from h being defined as strictly non-negative. In order to extend this non-negativity to the use of the Bernstein polynomials in approximating h, we address first the upper bound of 1 by selecting an appropriate γ_k such that $\sum_{k=0}^{K}\gamma_k\lambda^k \leq 1$ [26]. To ensure that $g(\lambda) \geq 0$, we instead rely on the fact that any polynomial that is non-negative over an interval can be represented using a non-negative linear combination of the Bernstein bases [39]. Thus, in selecting a non-negative vector θ_k, the positive semi-definite form for $\eta(\mathbf{L})$ can be retained, subsequently ensuring a convex loss function.

Pseudocode of the implementation using PyTorch for this methodology is as follows [33]. Note that the constraint of non-negativity on the coefficient of the bases is done via the ReLU function, which is defined as $\mathrm{ReLU}(x) = \max(x, 0)$, thus guaranteeing non-negativity.

Algorithm 2 Bern_Prop: Propagation Layer for Bernstein Polynomial Message Passing

Initializing Values: Polynomial maximum order K
Initialize:
Initialize θ
function FORWARD($\mathbf{X}$, Λ, U)
    $\theta \leftarrow \mathrm{ReLU}(\theta)$
    $\texttt{out} \leftarrow \binom{K}{0} \cdot \frac{1}{2^K} \cdot \theta_0 \cdot (1+\Lambda)^K$
    for $i = 0$ to $K-1$ **do**
        $\tilde{\Lambda} \leftarrow (1+\Lambda)^{K-i-1} \cdot (1-\Lambda)^{i+1}$
        $\texttt{out} \leftarrow \texttt{out} + \binom{K}{i+1} \cdot \frac{1}{2^K} \cdot \theta_{i+1} \cdot \tilde{\Lambda}$
    end for
    return $U \cdot (\texttt{out} \odot (U^T\mathbf{X}))$
end function

BernNet thus utilizes the Bernstein polynomial bases to produce a linear combination of basis functions with which to approximate the true signal function h. This method is unique, as the flexibility of the Bernstein bases to represent any non-negative filter function h over the relevant interval $[0,2]$ ensures a convex loss function in optimizing the learning parameters, θ_k.

5.1.0.3 JacobiConv

We begin the motivation for use of Jacobi basis polynomials by assessing the Hessian of the squared loss function, $R = \frac{1}{2}\|\mathbf{Z}-\mathbf{Y}\|_F^2$, where $\mathbf{Y}$ is the true values of the output signal and α_k is a vector of layer weight parameters. Note we are taking the Frobenius norm, which for a matrix $\mathbf{A}$ is $\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A}\mathbf{A}^T)}$. The Hessian matrix, $\mathbf{H}$, is of size $K \times K$, corresponding to the number of layers or bases in the filter, and each (k_1, k_2) element can be expressed as the following with respect to the parameter vector α_k,

$$\mathbf{H}_{k_1,k_2} = \frac{\partial}{\partial\alpha_{k_1}\partial\alpha_{k_2}}\frac{1}{2}\|\mathbf{Z}-\mathbf{Y}\|_F^2 = (\mathbf{X}\mathbf{W})^T g_{k_1}(\mathbf{L})g_{k_2}(\mathbf{L})\mathbf{X}\mathbf{W} = \tilde{\mathbf{W}}\mathbf{X}^T g_{k_1}(\mathbf{L})g_{k_2}(\mathbf{L})\mathbf{X}, \tag{5.11}$$

with the final equality derived by holding $\mathbf{W}$ constant thus allowing us to pull out a constant $\tilde{\mathbf{W}} = \mathbf{W}^T\mathbf{W}$. We then apply this to the spectra of $\mathbf{L}$ as a sum over $g_{k_1}(\lambda_i)g_{k_2}(\lambda_i)$, scaled by the signal density of $\mathbf{X}$ for that frequency, λ_i, which we denote as $f(\lambda_i)$ [46]. This can then be approximated as a Riemann integral over the interval $[0,2]$ for all possible eigenvalues of $\mathbf{L}$,

$$\mathbf{H}_{k_1,k_2} = \mathbf{X}^T g_{k_1}(\mathbf{L})g_{k_2}(\mathbf{L})\mathbf{X} = \sum_{i=1}^{N} g_{k_1}(\lambda_i)g_{k_2}(\lambda_i)f(\lambda_i) \approx \int_{\lambda=0}^{2} g_{k_1}(\lambda)g_{k_2}(\lambda)f(\lambda)d\lambda. \tag{5.12}$$

We seek to minimize the condition number of $\mathbf{H}$, $\kappa(\mathbf{H}) = \frac{|\lambda_{\max}(\mathbf{H})|}{|\lambda_{\min}(\mathbf{H})|}$, as it is the change in loss for a change in inputs, and we are seeking a solution in a convex problem. Thus, orthogonal basis functions, g_k, in $\mathbf{H}$, yield the best setting for fast and stable solution finding [46]. We now introduce the Jacobi bases,

$$P_k^{a,b}(z) = \begin{cases} 1, & k = 0 \\ \frac{a-b}{2} + \frac{a+b+2}{2}z, & k = 1 \\ (\theta_k z + \theta'_k)P_{k-1}^{a,b}(z) - \theta''_k P_{k-2}^{a,b}(z), & k \geq 2. \end{cases} \tag{5.13}$$

The individual coefficients, dependent on k as well as two additional hyperparameters, a and b, are defined as,

$$\theta_k = \frac{(2k+a+b)(2k+a+b-1)}{2k(k+a+b)}, \quad \theta'_k = \frac{(2k+a+b-1)(a^2-b^2)}{2k(k+a+b)(2k+a+b-2)}, \\ \theta''_k = \frac{(k+a-1)(k+b-1)(2k+a+b)}{k(k+a+b)(2k+a+b-2)}. \tag{5.14}$$

Furthermore, JacobiConv mitigates issues with linear GNNs and their limited ability to express all possible results in multidimensional problems where $d > 1$ by fitting a different vector of α_k for each individual output channel, $1 \leq l \leq d$. For the l-th column of a matrix $\mathbf{A}$ being denoted $\mathbf{A}_{:l}$, this then yields the following final filtering process for a specific output channel l,

$$\mathbf{Z}_{:l} = \sum_{k=0}^{K} \alpha_{kl} P_k^{a,b}(\mathbf{L})\mathbf{X}\mathbf{W}_{:l}. \tag{5.15}$$

Let $\mathbf{P}$ be a $k \times N$ matrix of the polynomial coefficient values for each individual node, with each column representing the coefficients for the corresponding order k. See below for a pseudocode implementation of the basis calculations, where the function is passed to the forward propagation functionality from PyTorch modules, as in the other polynomial bases, with α being initialized according to Polynomial Coefficient Decomposition as described in [46].

Algorithm 3 JacobiConv: Recursive Jacobi Polynomial Basis

function JacobiConv($k, \mathbf{P}, \Lambda, \alpha, a, b$)
  if $k == 0$ **then return 1**
  else if $k == 1$ **then**
    `coef1` $\leftarrow \alpha_0 \left(\frac{a-b}{2}\right)$
    `coef2` $\leftarrow \alpha_0 \left(\frac{a+b+2}{2}\right) \cdot \Lambda$
    return `coef1` + `coef2`
  else
    $\theta_k \leftarrow \frac{(2k+a+b)(2k+a+b-1)}{2k(k+a+b)}$
    $\theta'_k \leftarrow \frac{(2k+a+b-1)(a^2-b^2)}{2k(k+a+b)(2k+a+b-2)}$
    $\theta''_k \leftarrow \frac{(k+a-1)(k+b-1)(2k+a+b)}{k(k+a+b)(2k+a+b-2)}$
    `coef1` $\leftarrow \alpha_{k-1}(\theta_k \cdot \Lambda + \theta'_k)\mathbf{P}_{:k-1}$
    `coef2` $\leftarrow \alpha_{k-1}\theta''_k\mathbf{P}_{:k-2}$
    return `coef1` − `coef2`
  end if
end function

Extending upon BernNet, JacobiConv thus uses the orthogonality of the Jacobi basis functions to drive a faster and more stable solution finding algorithm. Additionally, the algorithm accounts for multichannel problems by incorporating channel-specific coefficients across all layers, enabling more flexible estimation of filter functions in categorization problems with more than two categories.

5.2 Eigenvalue Correction

In order to address the issue of high eigenvalue multiplicity in spectral filtering of GNNs, correction strategies must satisfy two criteria. They should first reduce the multiplicity of the spectra in order to improve filter performance, but they must also retain the general shape of the initial empirical eigenvalue distribution in order to propagate the correct information regarding the frequency of each signal. [33] proposes a uniform eigenvalue correction methodology, combining the empirical eigenvalue distribution with a set of equally spaced eigenvalues, $\mathbf{v}$. The i-th entry in $\mathbf{v}$ is defined,

$$\mathbf{v}_i = \frac{2i}{N-1}, i \in \{0, 1, 2, \ldots, N-1\}. \tag{5.16}$$

This monotonic series is then joined with the ordered set of eigenvalues with a trade-off parameter $\beta \in [0, 1)$, to produce a new set of ordered eigenvalues, $\boldsymbol{\mu}$:

$$\boldsymbol{\mu}_i = \beta\lambda_i + (1-\beta)\mathbf{v}_i. \tag{5.17}$$

In the case that $\beta = 0$, the new eigenvalues are simply the uniformly spaced series $\mathbf{v}$, whereas if $\beta = 1$, the original eigenvalues are returned. As [33] states in Theorem 2, given that both Λ and $\mathbf{v}$ are monotonically increasing, $\boldsymbol{\mu}$ must also be monotonic with $\mu_i \neq \mu_j, i \neq j$. Additionally, β can then be tuned to ensure that the correct amount of information is retained from the original spectra, thus satisfying the two criteria set out above. A key feature of this correction strategy is that it operates uniformly across $[0, 2]$, even in cases where regions of high multiplicity in the spectra are concentrated in regions smaller than that interval. As such, the number of effected eigenvalues may be excessive, thus losing too much of the initial information from the empirical spectra. This then motivates a more precise correction strategy in the scenarios where we see a well-defined bulk of eigenvalues. Formally, this occurs in graphs where $Np \geq \log(N)$, as a reasonably well-connected graph is required for the eigenvalues of the normalized Laplacian to begin exhibiting concentrated behavior. In regimes where $Np < \log(N)$, the eigenvalue distribution exhibits atom-like behavior, with a particularly large atom at 0 [5].

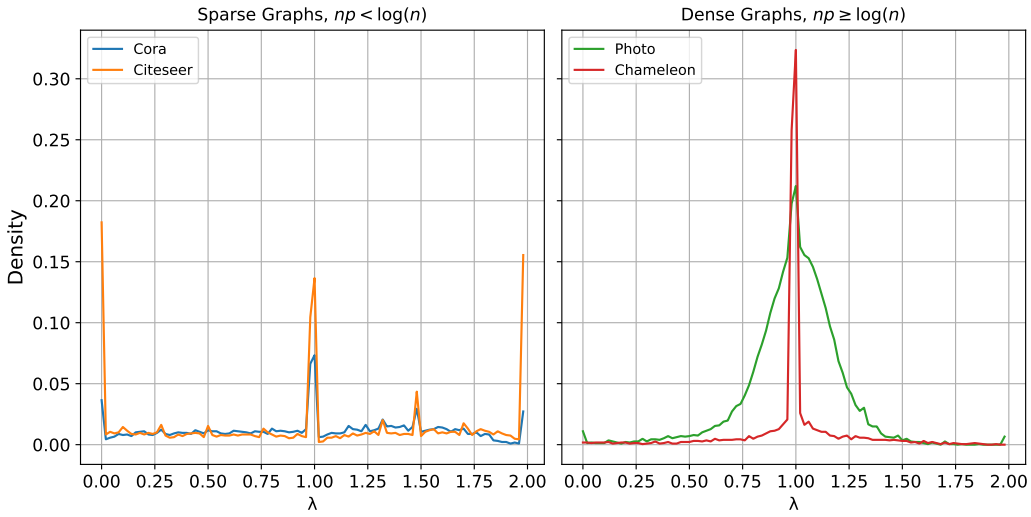


Figure 5: Eigenvalue distributions of $\mathbf{L}$ in sparse and dense regimes, with sparse graphs having $Np < \log(N)$, whereas dense graphs have $Np \geq \log(N)$. Sparse graphs exhibit atom-like behaviour in their spectra, with a relatively reduced mass at 1. Dense graphs begin to exhibit an accumulation of mass near 1, as can notably be seen in the Photo dataset (green).

This motivates our proposed correction strategy, which retains the uniform spacing to ensure that the multiplicities are reduced to 1 in the affected interval, but instead relies on the empirical degree of the graph to determine the interval upon which this spacing is imposed. As is proven in [29], the spectra of the normalized Laplacian converges to the semicircular law in the homogeneous setting, that is, for a fixed value of p across all N nodes. It is critical to note that the semicircular law is strictly bounded on the interval $[-2, 2]$. While the true connection probability of a real-world graph is most likely not represented by an individual fixed p, and is more akin to the inhomogeneous setting where a connection probability is defined as a scaled function $\varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)$, the empirical estimation of both ε_N and f is complicated and beyond the scope of this work. We look to bound the behavior of the maximum eigenvalue of $\mathbf{L}_N$ to form the interval on which a more specific correction should be done. We start with fixing a consistent p across all nodes such that Theorem 2 from [29] can be applied:

$$\max_i \left| \sqrt{\frac{Np}{1-p}} \left((1 - \lambda_i(\mathbf{L}_N) \right) \right| \approx 2$$

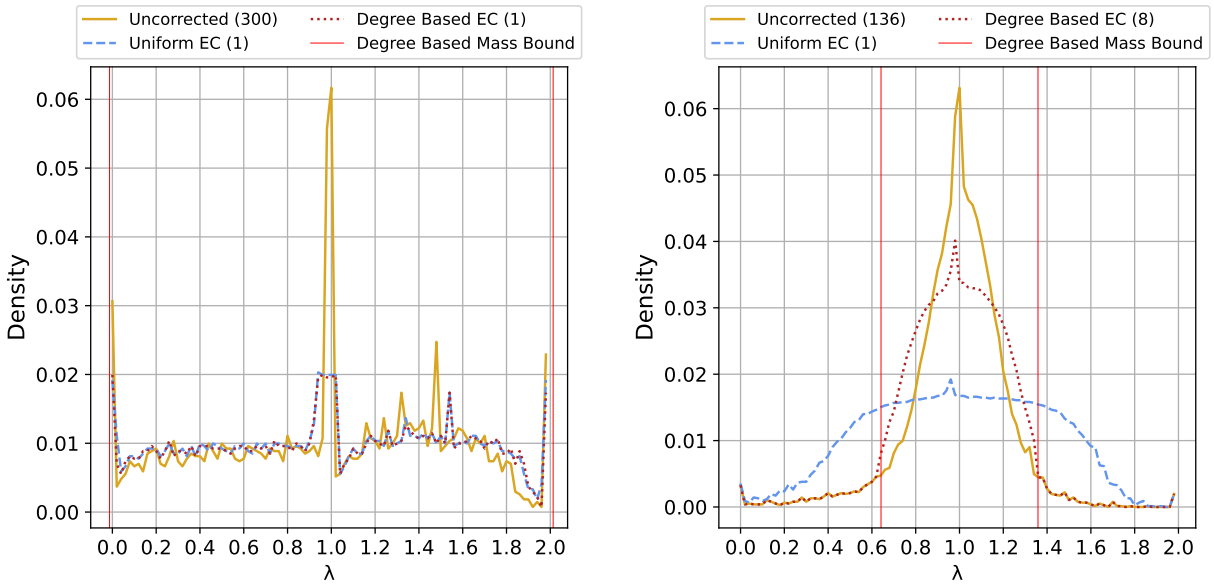
$$\max_i |1 - \lambda_i(\mathbf{L}_N)| \approx 2\sqrt{\frac{1-p}{Np}},$$

with high probability. We can then take $p \downarrow 0$, thus simplifying the right-hand side to $\frac{2}{\sqrt{Np}}$, and yield our final interval,

$$\max_i |\lambda_i(\mathbf{L}_N)| \approx 1 \pm \frac{2}{\sqrt{Np}}, \tag{5.18}$$

with high probability. Applying the original uniform eigenvalue correction strategy within this interval in dense graph settings thus sufficiently reduces the multiplicity of the eigenvalue spectrum while retaining

more of the initial empirical eigenvalue distribution structure than can be done by correcting the entire $[0, 2]$ interval. It is key to note that in sparse, $Np < \log(N)$ regimes, that this method will under-perform compared to the original uniform strategy, as the bounds of correction will become larger than $[0, 2]$, thus losing more of the original eigenvalue structure. The impact on the eigenvalue distribution of $\mathbf{L}_N$ can be seen in (6) for both sparse and dense regimes.



(a) Cora (Sparse)

(b) Photo (Dense)

Figure 6: Comparison of three eigenvalue correction methodologies: (1) no correction, (2) uniform correction ($\beta = 0.5$), and (3) degree based correction ($\beta = 0.5$). The highest eigenvalue multiplicity after each correction is applied is in parentheses in each corresponding legend. Note that in the dense regime, that the degree based bounds offer a more precise region of correction, while in the sparse regime, this yields an interval of correction exceeding that of the uniform method.

As such, let our new corrected eigenvalue vector be $\mathbf{u}$, where:

$$\mathbf{u}_i = \frac{2i}{N-1}, i \in \{a, a+1, \ldots, b-1, b\}, \tag{5.19}$$

for a being the index of the eigenvalue immediately greater than $1 - \frac{2}{\sqrt{Np}}$ and b being the index of the eigenvalue immediately less than $1 + \frac{2}{\sqrt{Np}}$. $\mathbf{u}$ is thus of size at most N, allowing for the same combination via the trade-off parameter β to be done as in (5.17), producing a corrected eigenvalue vector $\boldsymbol{\nu}$,

$$\boldsymbol{\nu}_i = \beta\lambda_i + (1-\beta)\mathbf{u}_i. \tag{5.20}$$

The following pseudocode is an implementation of this degree-based eigenvalue correction (DEC) methodology, where $\#\Lambda_{\text{bounded}}$ is the number of elements in the masked eigenvalue vector, Λ_{bounded} and `linspace` is a function producing a vector of $\#\Lambda_{\text{bounded}}$ equally spaced values between $1 \pm$ `bound_value`.

Algorithm 4 DEC: Degree-Based Eigenvalue Correction

function DEC(N, p, Λ, β)
  `bound_value` $\leftarrow \frac{2}{\sqrt{Np}}$
  `eigenvalue_mask` $\leftarrow (\Lambda \geq (1 -$ `bound_value`$)) \bigwedge \Lambda\,(\leq (1 +$ `bound_value`$))$
  $\Lambda_{\text{bounded}} \leftarrow \Lambda$[`eigenvalue_mask`]
  $\mathbf{u} \leftarrow$ `linspace`$(1 -$ `bound_value`$, 1 +$ `bound_value`$, \#\Lambda_{\text{bounded}})$
  $\boldsymbol{\nu}_{\text{bounded}} \leftarrow \beta \times \Lambda_{\text{bounded}} + (1-\beta) \times \mathbf{u}$
  $\boldsymbol{\nu} \leftarrow \Lambda$
  $\boldsymbol{\nu}$[`eigenvalue_mask`] $\leftarrow \boldsymbol{\nu}_{\text{bounded}}$
  return $\boldsymbol{\nu}$
end function

5.3 Real World Graph Data

Real world benchmark datasets were used to compare the performance of the three eigenvalue correction methodologies across the aforementioned polynomial filter types. These include three citation datasets, Cora, CiteSeer, and PubMed [48], two purchase datasets, Computers and Photo [41], two Wikipedia connection graphs, Chameleon and Squirrel [40], a co-occurence graph for actors, Actor, and webpage graphs Texas and Cornell [37]. These graphs have individual node labels, and as such, the model is evaluated based on its ability to classify the nodes correctly.

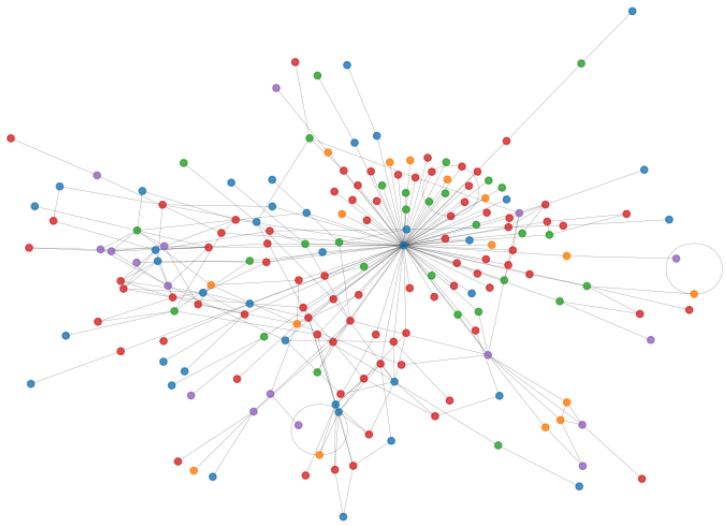


Figure 7: A visualization of the Cornell graph dataset ($N = 183$). Nodes representing web pages are colored according to their respective classes. Edges represent hyperlinks between the individual web pages [37].

Additionally, these datasets differ greatly in size and the characteristics of their connection structure, thus giving a wide array of settings on which to evaluate how flexible a model is in adjusting to a variety of graphs. Some of these characteristics are highlighted in Table 1.

Datasets	Cora	CiteSeer	PubMed	**Computers**	**Photo**	Texas	Cornell	**Squirrel**	**Chameleon**	Actor
N	2708	3327	19717	13752	7650	183	183	5201	2277	7600
$\|\|E\|\|$	5278	4552	44324	245861	119081	279	277	198353	31371	26659
$\Lambda_{distinct}$	2204	1885	7596	13351	7477	106	115	3275	1122	6420
$Np/\log(N)$	0.493	0.336	0.459	3.752	3.482	0.602	0.548	8.917	3.567	0.783

Table 1: Summary statistics of the datasets used for eigenvalue correction performance evaluation. N is the number of nodes, $\|\|E\|\|$ the number of edges, $\Lambda_{distinct}$ the number of unique eigenvalues, and $Np/\log(N)$ determines if the graph falls in the sparse regime (< 1), or the dense regime (> 1).

5.4 Model Training and Evaluation

In line with other work that uses these benchmark datasets to evaluate GNNs, such as [33] and [46], we mask all graphs with the same split of 60% for training, 20% for validation, and 20% for testing. The loss function used for optimization is the cross entropy loss, which is used specifically for classification problems in PyTorch [36]. Cross entropy loss is defined as follows for C being the number of classes, w_c being class weights, and x and y being inputs and targets respectively:

$$\ell(x, y) = \frac{1}{N} \sum_{n=1}^{N} \left[-\sum_{c=1}^{C} w_c \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} y_{n,c} \right]. \tag{5.21}$$

Use of cross entropy loss in place of binary classification error during training provides a more nuanced

error term, incorporating how confident the model is in its softmax classification by measuring the log-loss between the predicted class probabilities as opposed to a binary outcome. This in turn avoids overfitting by directly addressing the underlying distances as opposed to final labels. However, for ease of interpretation, the final score for each model against its testing mask is expressed using the mean classification accuracy over 10 fitting runs.

In order to have a valid comparison of the performance of the three polynomial filters, we fix the filter specific hyperparameters across all eigenvalue correction strategies for a given dataset. Thus, the only hyperparameter tuned is β from (5.17), which controls the balance between the uncorrected eigenvalues and the redistributed eigenvalues, as in (5.17) and (5.20). This will be tested on intervals in $\beta \in [0, 1)$ spaced by $0.05, \beta \in \{0, 0.05, 0.10, \ldots, 0.95\}$, with the value of β that maximizes the classifying accuracy being selected. Values of β for the uniform EC are held constant from Appendix D of [33].

All experiments are conducted on a machine with an AMD Radeon RX 7800 XT 16GB GPU, an AMD Ryzen 5 5600X-6 Core Processor 3.70GHz CPU, and 64GB of memory.[1]

5.5 Hyperparameter Tuning and Selection

As outlined in (5.20), the hyperparameter of interest in our degree-based eigenvalue correction strategy for polynomial filters in GNNs is β, which operates as a trade-off parameter between the uncorrected empirical eigenvalue distribution, $\boldsymbol{\lambda}$, and a set of size N containing an equidistantly spaced interval of eigenvalues of length at most N, $\mathbf{u}$. For $\beta = 0$, $\boldsymbol{\nu}$ is then simply $\mathbf{u}$, whereas in the setting where $\beta = 1$, we recover the original empirical eigenvalue distribution, that is, the scenario with no eigenvalue correction done at all. As such, the range of interest for β can be restricted to $\beta \in [0, 1)$.

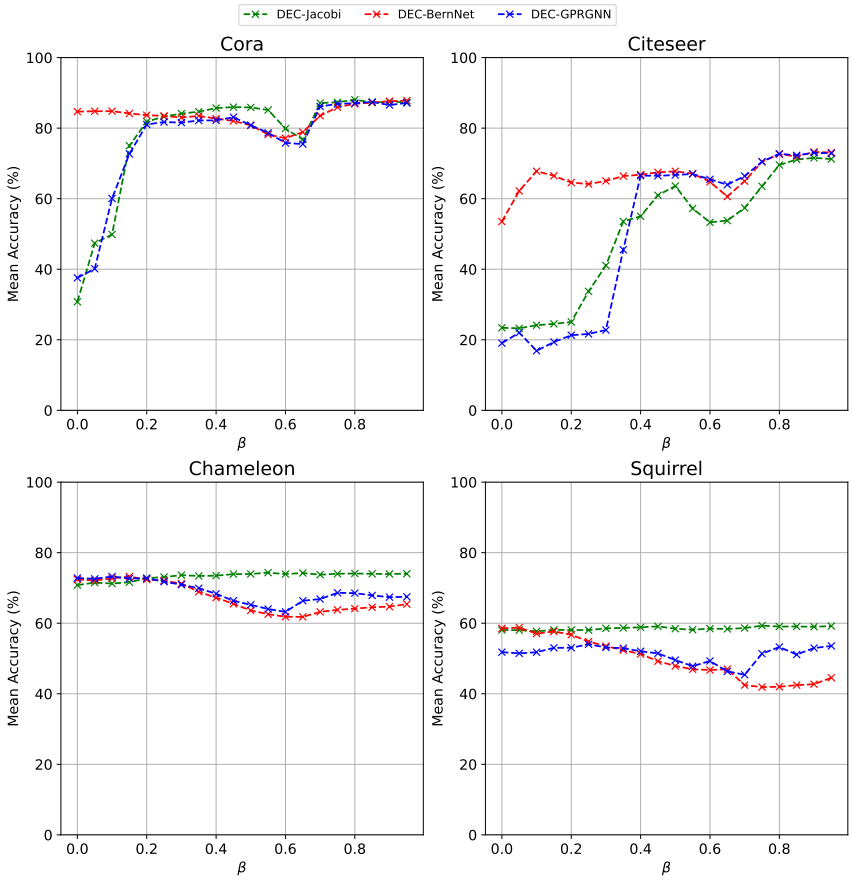


Figure 8: Model sensitivity to β for $\beta \in [0, 1)$. Note that Cora and Citeseer are in the sparse regime, whereas Chameleon and Squirrel are in the dense regime, so model accuracy is relatively stable for dense graphs, whereas in sparse graphs, tuning β is critical to improving model performance.

Figure 8 shows the effect of β in both dense and sparse regimes. Looking first at the sparse datasets, Cora and Citeseer, we see significant increases in performance when applying degree based eigenvalue

[1] ZLUDA (https://github.com/vosen/ZLUDA) was used to wrap CUDA in order for PyTorch and Torch Geometric to run on an AMD GPU.

correction (DEC) to Jacobi and GPR at value of β greater than 0.3. In contrast, BernNet sees relatively stable performance even at lower values of β, although model performance does increase slightly in line with increasing β across all three filters. Looking back to Figure 6, it is evident that in sparse settings, DEC does not yield a very informative interval on which to perform correction, and given the small value of Np, we retrieve a potentially overly wide interval to correct. Given that in a sparse setting, we see primarily atom-like behavior in the empirical spectral distribution, slight corrections of lower multiplicity eigenvalues has reduced benefits for model outcomes. However, as is discussed in [33], multiplicity reduction near 1 is still valuable in improving model performance, and so we see that there is a localized improvement near $\beta = 0.5$ for both Cora and Citeseer. This highly localized behavior in the eigenvalues drives the increased performance for higher values of β, which more closely emulate the original distribution, indicating that retaining the original shape is more beneficial in sparse regimes than correcting eigenvalues with multiplicities greater than 1. Additionally, we see improved performance of BernNet at lower values of β as it is inherently less sensitive to overfitting, which is beneficial when the spectra has been over flattened, as in the case of applying DEC in sparse regime graphs. This is due to BernNet's use of the Bernstein basis polynomials, which are themselves smooth functions, and thus less likely to overfit to the small sudden changes in the spectral distribution, as seen in Figure 6. In contrast, GPRGNN uses node-level adjustments given its reliance on PageRank, which leads to significant overfitting, and JacobiConv's use of orthogonal polynomials to increase the efficiency of learning works against its performance on a relatively uninformative spectra in this scenario. As such, sparse graphs require careful tuning of β for DEC to perform well.

Chameleon and Squirrel are both dense, as can be seen in Table 1 by their corresponding ratios for $Np/\log(N)$. Figure 8 shows that in this setting, modulating the value of β produces a more stable result for model performance. This is driven by the accumulation of mass near, but not precisely at 1, as can be seen in Figure 6, which provides a strong baseline of performance, even in the fully DEC driven ($\beta = 0$) setting. In contrast, for the uniform method proposed by [33] there is significant drop-off in performance for values of β near the bounds 0 and 1. As such, it is evident that DEC's restriction of the interval of correction to an interval smaller than $[0, 2]$, yields more stable results in dense settings, given that the method specifically focuses on regions of high multiplicity without distorting lower density regions in the eigenvalue distribution. A full list of selected β values across all filters and datasets can be seen in Table 2.

Filter	Actor	Chameleon	Citeseer	Computers	Cora	Cornell	Photo	PubMed	Squirrel	Texas
BernNet	0.10	0.15	0.90	0.10	0.95	0.10	0.75	0.95	0.05	0.15
GPRGNN	0.20	0.10	0.90	0.75	0.85	0.35	0.75	0.95	0.25	0.90
JacobiConv	0.65	0.55	0.90	0.35	0.80	0.90	0.95	0.90	0.75	0.50

Table 2: Selected values of hyperparameter β for the proposed DEC method for each corresponding filter and dataset. Note that $\beta \in [0, 1)$, and values were selected by taking the value of β that corresponded to the highest training accuracy.

5.6 Evaluated Model Performance

Using the values of β listed in Table 2 for each combination of filter, eigenvalue correction strategy, and dataset yields the mean classification accuracy values given in Table 3. The corresponding confidence intervals for these accuracies can be found in Appendix A.1.

Dataset	**BernNet**			**GPRGNN**			**JacobiConv**		
	None	**Uniform EC**	**DEC**	**None**	**Uniform EC**	**DEC**	**None**	**Uniform EC**	**DEC**
Actor	38.132	38.271	38.189	38.113	38.366	37.878	36.384	37.897	37.935
Chameleon	64.858	72.166	73.217	56.827	71.816	73.239	73.807	74.595	74.333
Citeseer	73.547	71.869	73.247	72.066	60.900	73.001	71.460	61.896	71.555
Computers	86.772	85.392	87.925	86.649	84.396	88.913	89.585	85.172	89.490
Cora	87.373	75.238	87.816	87.009	83.000	87.406	87.071	83.173	88.062
Cornell	66.170	69.149	74.255	61.277	53.191	77.660	61.064	61.277	59.362
Photo	93.573	92.763	94.059	93.730	93.750	94.909	95.425	92.222	95.552
PubMed	84.763	83.787	85.237	85.427	83.612	86.611	85.366	79.229	85.336
Squirrel	41.940	55.879	58.761	32.113	54.150	54.025	59.183	54.140	59.299
Texas	83.443	83.607	86.066	81.148	84.426	87.705	78.033	86.721	87.213

Table 3: Final evaluated accuracies (%) on real-world datasets for all filters and eigenvalue correction strategies. For a table of corresponding confidence intervals, refer to Table 4.

On average, the DEC methodology yields 4.14% better accuracy than the EC methodology and 7.10% better accuracy than no correction methodology for BernNet. Similarly, DEC outperforms EC and no correction by 7.97% and 13.64% when applied to GPRGNN, and 4.46% and 1.47% when applied to JacobiConv, respectively. This difference between three filters tested is symptomatic of how each filter propagates information between layers, similar to the differences in model performance when varying β in Figure 8. As GPRGNN relies on the centrality of individual nodes to handle propagation, it is highly dependent on having a maximally informative eigenvalue distribution with which to perform the signal filtering process. Over smoothing of the distribution, either in magnitude or the interval over which the smoothing occurs, fundamentally distorts the centrality of nodes, thus impacting model performance. The impact of correction on GPRGNN can thus be interpreted as a trade-off between the expressive range of the filter and its ability to accurately generate an output signal. Looking specifically at the dense Chameleon and Squirrel datasets, we see that the ranges of correction as defined in (5.18) as a function of $\sqrt{Np}$ are approximately $(0.619, 1.380)$ and $(0.771, 1.229)$, respectively. This is a much tighter interval than the $[0, 2]$ range used for the EC correction. The value of this more precise correction range is supported by the similarly low values of β for these two datasets in Table 2, indicating that the corrected $\boldsymbol{\nu}$ is on balance more informative than the empirical spectra. Finally, looking at the accuracy and uncertainties of these correction methods using GPRGNN in Figure 9, we see that both EC and DEC outperform no correction in both Chameleon and Squirrel. However, even with this tightened correction interval, while the mean accuracy of DEC is slightly higher, we see that the two methods have overlapping 95% confidence intervals, indicating that in the dense setting, eigenvalue multiplicity reduction is more valuable than retaining the shape of the empirical spectra. In contrast, in sparse regimes, such as Cora and Citeseer, we see that tuning of β yields values near 1 for GPRGNN, indicating that the model is predominately using the empirical eigenvalues, hence the accuracy of DEC being similar to that of having no correction applied. Here, it is thus more important to retain the shape of the underlying distribution than to reduce multiplicity.

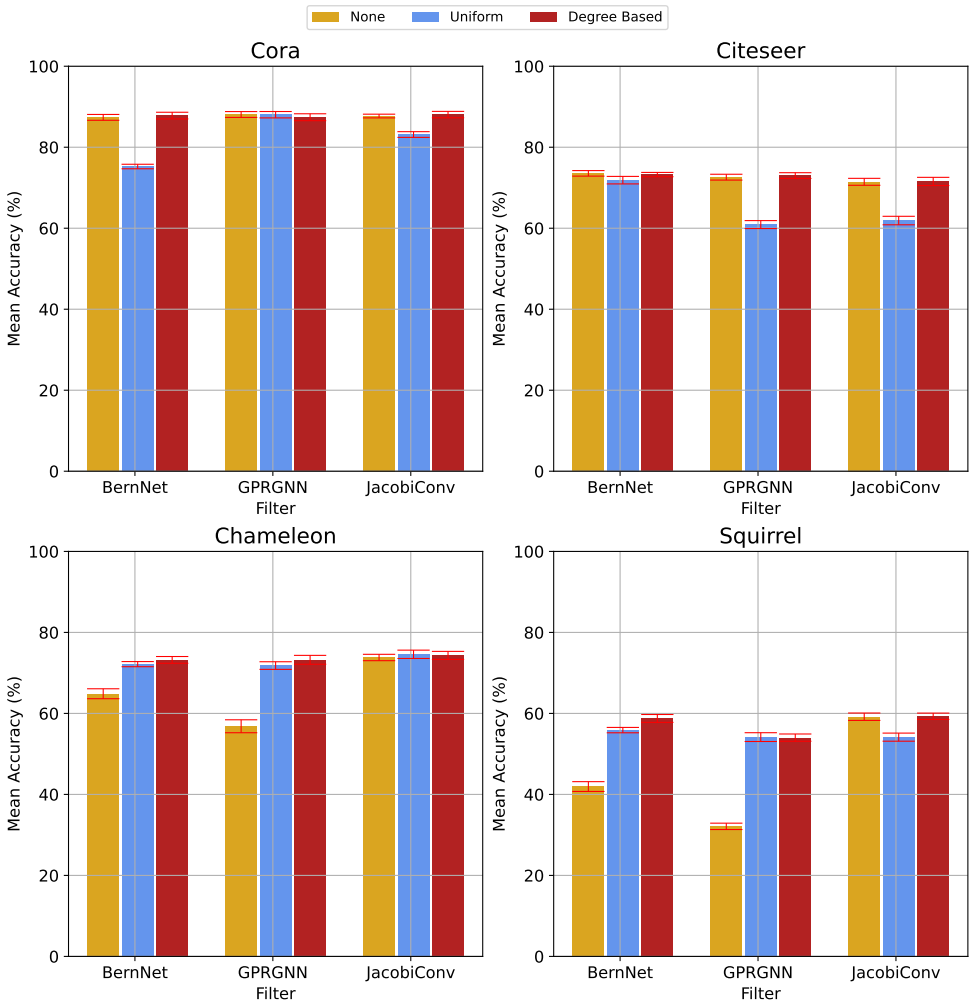
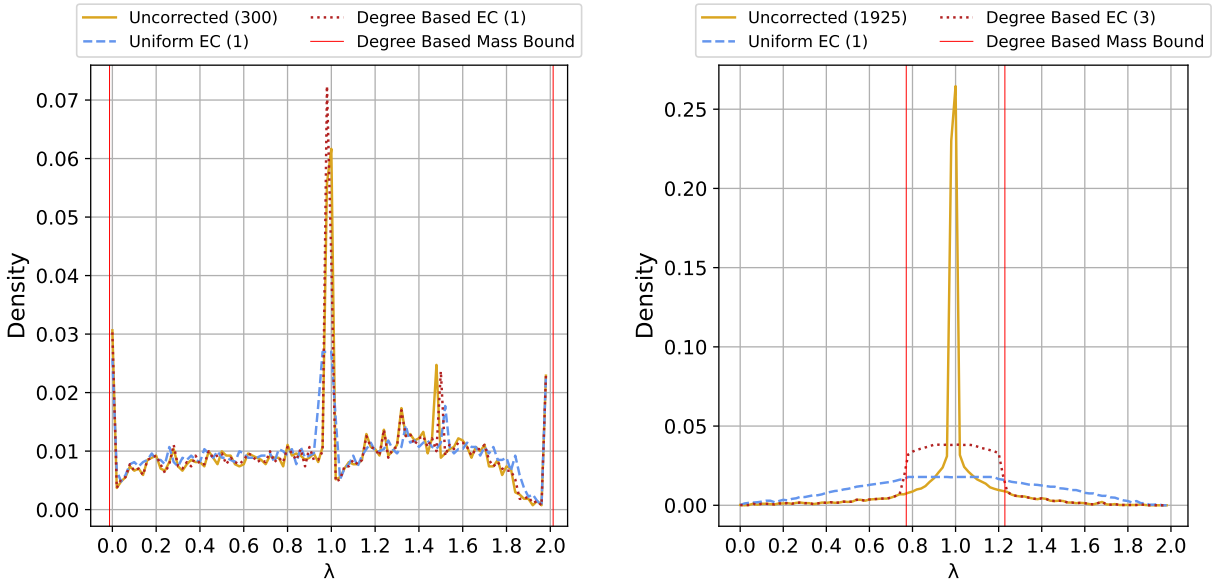


Figure 9: Final model mean accuracies, compared across the Cora (S), Citeseer (S), Chameleon (D), and Squirrel (D) datasets. Error bars (red) are the 95% confidence intervals over 10 runs for each corresponding combination of filter, correction strategy, and dataset.

The polynomial basis filters, BernNet and JacobiConv, see smaller differences in performance, as these methods aim to approximate the spectra as a polynomial function, leading to increased retention in the shape of the eigenvalue distribution. This contrasts with the emphasis on retaining individual node behavior in GPRGNN. Starting with BernNet's application to the dense datasets, we see that the broad changes in performance are similar to that of PageRank motivated methods, as the high eigenvalue multiplicity of datasets like Chameleon and the subsequent reduced expressiveness of applied filters dominate the performance of uncorrected methods. However, the gap by which DEC outperforms EC is notably wider in the densest graph, Squirrel, where the selected β is also 0.05. For a polynomial approximation using Bernstein polynomials, using a corrected spectra with a narrow correction range of $(0.771, 1.229)$ implies that much of the range is untouched, and that the range of correction is highly focused on minimally disruptive correction of a single mass. This further promotes having a small β value, as the difference across the range between λ and $\mathbf{u}$ is thus minimal when assessing the range of eigenvalues affected. Figure 10 visualizes this, with the over flattening apparent for the uniform EC, whereas DEC holds some of the initial structure even after correction by narrowing the range at which equidistant spacing is spliced in.



(a) Cora (Sparse)

(b) Squirrel (Dense)

Figure 10: Comparison of eigenvalue correction methodologies for the Squirrel dataset ($\beta = 0.44$ for EC, $\beta = 0.05$ for DEC) and for the Cora dataset ($\beta = 0.63$ for EC, $\beta = 0.95$ for DEC).

Eigenvalue correction for sparse datasets yields different results as the underlying shape is inherently difficult to approximate, as can be seen in Figure 10 for Cora, where at $\beta = 0.95$ the correction done is minimal. This minimal correction reduces eigenvalue muliplicity but ensures as much shape retention as possible, a trend that can similarly be seen for other sparse datasets in Table 2. For Cora, this minimal correction yields an improvement in the accuracy of BernNet from 87.373% with no correction to 87.816% with DEC applied. Notably, EC uses a significantly lower $\beta = 0.63$ and produces less accurate predictions, indicating that EC has lost an excessive amount of information from the distribution's shape by over-correcting. JacobiConv outperforms BernNet in dense settings as its use of polynomial coefficient decomposition and orthogonality of the Jacobi bases allows for faster solution finding [46], but this gain is reduced in sparse settings as it is a more difficult problem to approximate an extremely diffuse spectra. As such, in the spare regime, we see that any correction strategy yields similar performance between BernNet and JacobiConv. DEC thus uses the average degree of the underlying distribution to produce a more informed balance between multiplicity reduction and shape retention in approximating filter functions, with improved accuracy over both uncorrected and EC methods in dense settings. While the improvement from any correction method is reduced in sparse settings, DEC is able to perform at least as well as the uncorrected polynomial approximations via tuning of β.

Chapter 6
Discussion

This thesis aimed to fill a gap in the literature of Erdös-Rényi Random Graphs in the inhomogeneous regime by proving that the empirical spectral distribution of the normalized Laplacian converges to the deterministic limiting measure, μ_f, and applying this result and its methods to a statistical application in the approximation of PageRank, as well as a machine learning application via eigenvalue correction in GNNs. Throughout the theoretical parts of this work, we have relied on the assumption that we are operating in the dense regime with a well-connected graph, that is, where $N\varepsilon_N \gg \log(N)$.

First, we proved that the empirical spectral distribution of an appropriately centered and scaled normalized Laplacian matrix of an inhomogeneous random graph where the connection probability function can be decomposed multiplicatively indeed converges to the limiting measure, μ_f. This deterministic measure provides an extension of the semicircular law that governs the limiting behavior of the normalized Laplacian in the homogeneous regime, as proven in [29]. Additionally, in the proof, we have borrowed methods for perturbation and Gaussianization for inhomogeneous random graphs from [7], which focused on the adjacency and Laplacian matrices. The formal proof of this convergence to a deterministic limiting measure fills a gap in the literature of random graphs, extending the field's understanding of the asymptotic behaviour of $\mathbf{L}_N$ in more generalizable settings. This result and its constituent methods will enable downstream statistical methods on graphs to be applied to more diverse sets of graphs with appropriate mathematical motivation, particularly in real-world settings where inhomogeneity is commonplace.

We explore one of these methods in our exploration of approximations of PageRank in the same inhomogeneous setting, where the PageRank vector, which as a centrality measure can also be viewed as the stationary distribution of a stochastic matrix, is instead represented by the volume normalized degree vector. In proving that this approximation does indeed hold for a general class of inhomogeneous graphs, and further by providing the rate of convergence in total variation distance between the two probability vectors, $\frac{K}{\sqrt{N\varepsilon_N}}$, we have extended upon the initial work done in [2], which proved this approximation in the more specific Chung-Lu and Stochastic Block Model regimes. This proof builds directly upon the prior result regarding the convergence of the spectral measure of $\mathbf{L}_N$ to μ_f, as it similarly requires the degree behaviour of the graph to be reasonably well-behaved with the maximum degree of the graph being finite and the minimum degree of the graph being non-zero in order for a decoupling of the empirical and expected degree matrices. Additionally, we have extended the result concerning counting across maximal walk contributions as a representation of the maximum eigenvalue of $\mathbf{C}_N$ to the general inhomogeneous case, building upon the initial result from [10]. We believe that these results can be applied to statistical methods and algorithms that rely on calculation of PageRank at either the global graph level or for smaller sub-graphs, specifically for graphs that do not adhere to more common inhomogeneous settings for which prior results exist.

Finally, we applied this understanding of dependence on the behavior of the spectral measure on the degree of the underlying graph to GNNs, specifically in the context of eigenvalue correction of the normalized Laplacian as it is used for propagation between layers. This correction by imposing uniform spacing between eigenvalues improves the diversity of output values for the underlying filter functions, thereby improving the expressiveness of the overall GNN. Our contribution was in leveraging that the degree relative to the size of the graph controls the accumulation of central mass in its spectra, and so by restricting the range of eigenvalue correction as a function of the degree we arrive at a final corrected spectra which adequately reduces the multiplicity of the individual eigenvalues with minimal loss of information. This was found to be particularly effective in dense graphs, where $Np > \log(N)$, such as the Squirrel or Computers datasets, where our proposed degree based eigenvalue correction methodology outperformed the uniform methodology proposed by [33] which corrects across the entire interval of eigenvalues, $[0, 2]$. Furthermore, assessing the performance difference across a set of real-world datasets demonstrated that DEC outperforms the uniform correction methodology, on average. However, the efficacy of both our results and that of the uniform correction depend heavily on an appropriate selection of $\beta \in [0, 1)$ in order to adequately balance the original distribution and the corrected, even when the correction interval is narrower than $[0, 2]$. While this methodology yields relatively stable results in the dense regime for any values of β, sparse graphs require very careful tuning, as can be seen in Figure 8, as final model accuracy

is highly variable at lower values of β when a relatively flat distribution of evenly spaced eigenvalues is unable to accurately model an atom heavy distribution.

While this work has made significant progress towards the understanding and application of the spectral behaviour of $\mathbf{L}_N$ for inhomogeneous random graphs, there are several open questions that remain. These open questions arise from changing the underlying assumptions of Theorem 3.2.1 in order to extend these results to different settings or to evaluate different quantities.

- Theorem 3.2.1 relies on the assumption that f can be decomposed multiplicatively into $f(x,y) = r(x)r(y)$. However, we believe that the convergence of $\mathbf{L}_N$ to μ_f holds for any general $f : [0,\infty)^2 \to [0,\infty)$, so long as it remains bounded and Riemann integrable.
- In this thesis, we analyze ESD $(\mathbf{L}_N)$ and its convergence to μ_f. However, the asymptotic behaviour of the empirical spectral distribution of $\mathbf{Q}_N \coloneqq \mathbf{D}_N^{-1/2}\mathbf{A}_N\mathbf{D}_N^{-1/2}$ in the sparse regime remains unknown.
- Throughout this work, we assume a dense graph, where $N\varepsilon_N \gg \log(N)$. This analysis and its methods do not apply in the context of sparse graphs, where either $N\varepsilon_N \ll \log(N)$ or $N\varepsilon_N \to \lambda \in (0,\infty)$, and instead we refer readers to [5] for the analysis of the behaviour of atoms in that setting for the spectrum of adjacency matrices. Thus, it remains an open problem to characterize the behaviour of the spectra of the Laplacian and normalized Laplacian matrices in the sparse regime.
- We have considered a connection probability between vertices of the form $p_{ij} = \varepsilon_N f\left(\frac{i}{N}, \frac{j}{N}\right)$. However, there is a more general form for inhomogeneous connection probabilities, known as Kernel-Based Random Graphs (KBRGs). For a vertex set $\mathbf{V}_N \coloneqq \{1,\dots,N\}$, a weight set $(W_i)_{i\in\mathbf{V}_N}$ where $W_i \overset{iid}{\sim} \text{Pareto}(\tau), \tau > 0$, and a kernel function $\kappa_\sigma : [0,\infty)^2 \to [0,\infty), \sigma \geq 0$, the connection probability is instead,

  $$p_{ij} \coloneqq \frac{\kappa_\sigma(W_i, W_j)}{\|i-j\|^\alpha} \wedge 1,$$

  for α being a long range tuning parameter for the influence of distance between vertices. The limiting measure of the spectra of the adjacency and graph Laplacian matrices have been defined in [12] and [25], respectively, but the limiting measure of the normalized Laplacian in this setting still remains an open problem.

Theorem 4.2.1 also poses some open questions that extend upon the results discussed in this thesis, also surrounding the generalizability of approximations of PageRank to other inhomogeneous graphs.

- In this work, we have proved the validity of the approximation of PageRank in a rank-one setting. However, this is more broadly generalizable to any finite-rank setting where,

  $$f(x,y) = \sum_{i=1}^{\alpha} \theta_i r_i(x) r_i(y),$$

  for r_i being a bounded, continuous function, and $\alpha > 1$. As such, validating this degree based approximation for all finite-rank random graphs remains an open problem.
- Theorem 4.2.1 and its proof further restrict the earlier assumption of a dense graph, necessitating that for $\xi > 6$, $N\varepsilon_N \gg \log^\xi(N)$, thus yielding a very sharp bound. However, we believe that this approximation still holds in scenarios where this requirement is relaxed such that $\xi \geq 1$, but our methodology does not yield this extension.

Finally, the results in this thesis surrounding degree based eigenvalue correction for GNNs offer some potential points of further analysis.

- The analysis run in this thesis on real-world datasets relies on a set of fixed hyperparameters, aside from β, to remain in line with the methodology used by [33]. While this removes the need for tuning across a wide range of hyperparameters, such as layer specific learning rates, and improves comparability of results by holding all else constant beyond the correction methodology, a more in-depth analysis involving retuning of all hyperparameters could yield valuable results concerning the interplay between filter specific hyperparameters and β.
- The proposed DEC strategy uses a heuristic approach based on results from analysis of homogeneous random graphs by using $\frac{2}{\sqrt{Np}}$ as the interval upon which to execute correction. While this is computationally straightforward, this poses the open question of it is possible to extricate the empirical maximum eigenvalue from the data and integrate into an eigenvalue correction methodology that yields a more precise interval of correction.

References

[1] Konstantin Avrachenkov and Maximilien Dreveton. *Statistical Analysis of Networks*. Now Publishers, 2022.

[2] Konstantin Avrachenkov, Arun Kadavankandy, Liudmila Ostroumova Prokhorenkova, and Andrei Raigorodskii. "PageRank in undirected random graphs". In: *International Workshop on Algorithms and Models for the Web-Graph*. Springer. 2015, pp. 151–163.

[3] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Vol. 20. Springer, 2010.

[4] Patrick Billingsley. *Probability and Measure*. 3. ed. John Wiley & Sons, 1995. ISBN: 0471007102.

[5] Charles Bordenave, Arnab Sen, and Bálint Virág. "Mean quantum percolation". In: *Journal of the European Mathematical Society* 19.12 (2017), pp. 3679–3707.

[6] Arijit Chakrabarty. "The Hadamard product and the free convolutions". In: *Statistics & Probability Letters* 127 (2017), pp. 150–157.

[7] Arijit Chakrabarty, Rajat Subhra Hazra, Frank Den Hollander, and Matteo Sfragara. "Spectra of adjacency and Laplacian matrices of inhomogeneous Erdős–Rényi random graphs". In: *Random matrices: Theory and applications* 10.01 (2021), p. 2150009.

[8] Sourav Chatterjee. "A simple invariance theorem". In: *arXiv preprint math/0508213* (2005).

[9] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. "Adaptive universal generalized pagerank graph neural network". In: *arXiv preprint arXiv:2006.07988* (2020).

[10] Fan Chung, Linyuan Lu, and Van Vu. "Spectra of random graphs with given expected degrees". In: *Proceedings of the National Academy of Sciences* 100.11 (2003), pp. 6313–6318.

[11] Fan RK Chung. *Spectral graph theory*. Vol. 92. American Mathematical Soc., 1997.

[12] Alessandra Cipriani, Rajat Subhra Hazra, Nandan Malhotra, and Michele Salvi. "The spectrum of dense kernel-based random graphs". In: *arXiv preprint arXiv:2502.09415* (2025).

[13] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. "Convolutional neural networks on graphs with fast localized spectral filtering". In: *Advances in neural information processing systems* 29 (2016).

[14] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Vol. 38. Springer Science & Business Media, 2009.

[15] Xue Ding and Tiefeng Jiang. "Spectral distributions of adjacency and Laplacian matrices of random graphs". In: *The annals of applied probability* (2010), pp. 2086–2117.

[16] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

[17] Paul Erdös and Alfréd Rényi. "On random graphs I". In: *Publ. math. debrecen* 6.290-297 (1959), p. 18.

[18] Rida T Farouki. "The Bernstein polynomial basis: A centennial retrospective". In: *Computer Aided Geometric Design* 29.6 (2012), pp. 379–419.

[19] William Feller. *An introduction to probability theory and its applications, Volume 2*. Vol. 2. John Wiley & Sons, 1991.

[20] Zoltán Füredi and János Komlós. “The eigenvalues of random symmetric matrices”. In: *Combinatorica* 1 (1981), pp. 233–241.

[21] Diego Garlaschelli and Maria I Loffredo. “Structure and evolution of the world trade network”. In: *Physica A: Statistical Mechanics and its Applications* 355.1 (2005), pp. 138–144.

[22] Marco Gori, Gabriele Monfardini, and Franco Scarselli. “A new model for learning in graph domains”. In: *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.* Vol. 2. IEEE. 2005, pp. 729–734.

[23] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. “Wavelets on graphs via spectral graph theory”. In: *Applied and Computational Harmonic Analysis* 30.2 (2011), pp. 129–150.

[24] Taher H Haveliwala. “Topic-sensitive pagerank”. In: *Proceedings of the 11th international conference on World Wide Web*. 2002, pp. 517–526.

[25] Rajat Subhra Hazra and Nandan Malhotra. “Spectral properties of the Laplacian of Scale-Free Percolation models”. In: *arXiv preprint arXiv:2504.17552* (2025).

[26] Mingguo He, Zhewei Wei, Zengfeng Huang, and Hongteng Xu. “Bernnet: Learning arbitrary graph spectral filters via bernstein approximation”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14239–14251.

[27] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. “Stochastic blockmodels: First steps”. In: *Social networks* 5.2 (1983), pp. 109–137.

[28] Glen Jeh and Jennifer Widom. “Scaling personalized web search”. In: *Proceedings of the 12th international conference on World Wide Web*. 2003, pp. 271–279.

[29] Tiefeng Jiang. “Empirical distributions of Laplacian matrices of large dilute random graphs”. In: *Random Matrices: Theory and Applications* 1.03 (2012), p. 1250004.

[30] Thomas N Kipf and Max Welling. “Semi-supervised classification with graph convolutional networks”. In: *arXiv preprint arXiv:1609.02907* (2016).

[31] Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

[32] Pan Li, I Chien, and Olgica Milenkovic. “Optimizing generalized pagerank methods for seed-expansion community detection”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[33] Kangkang Lu, Yanhua Yu, Hao Fei, Xuan Li, Zixuan Yang, Zirui Guo, Meiyu Liang, Mengran Yin, and Tat-Seng Chua. “Improving expressive power of spectral graph neural networks with eigenvalue correction”. In: *Foo* 38.13 (2024), pp. 14158–14166.

[34] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web*. Tech. rep. Stanford infolab, 1999.

[35] Leonid Andreevich Pastur and Mariya Shcherbina. *Eigenvalue distribution of large random matrices*. 171. American Mathematical Soc., 2011.

[36] A Paszke. “Pytorch: An imperative style, high-performance deep learning library”. In: *arXiv preprint arXiv:1912.01703* (2019).

[37] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. “Geom-gcn: Geometric graph convolutional networks”. In: *arXiv preprint arXiv:2002.05287* (2020).

[38] Nicola Perra and Santo Fortunato. “Spectral centrality measures in complex networks”. In: *Physical Review E* 78.3 (Sept. 2008). ISSN: 1550-2376. DOI: 10.1103/physreve.78.036107. URL: http://dx.doi.org/10.1103/PhysRevE.78.036107.

[39] Victoria Powers and Bruce Reznick. “Polynomials that are positive on an interval”. In: *Transactions of the American Mathematical Society* 352.10 (2000), pp. 4677–4692.

[40] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. “Multi-scale attributed node embedding”. In: *Journal of Complex Networks* 9.2 (2021), cnab014.

[41] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. “Pitfalls of graph neural network evaluation”. In: *arXiv preprint arXiv:1811.05868* (2018).

[42] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains”. In: *IEEE signal processing magazine* 30.3 (2013), pp. 83–98.

[43] Roland Speicher. *Lecture Notes on "Random Matrices"*. 2020. arXiv: 2009.05157 [math.PR]. URL: https://arxiv.org/abs/2009.05157.

[44] Daniel Spielman. “Spectral graph theory”. In: *Combinatorial scientific computing* 18.18 (2012).

[45] Remco Van Der Hofstad. *Random graphs and complex networks*. Vol. 2. Cambridge university press, 2024.

[46] Xiyuan Wang and Muhan Zhang. “How powerful are spectral graph neural networks”. In: *International conference on machine learning*. PMLR. 2022, pp. 23341–23362.

[47] Gian-Carlo Wick. “The evaluation of the collision matrix”. In: *Physical review* 80.2 (1950), p. 268.

[48] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. “Revisiting semi-supervised learning with graph embeddings”. In: *International conference on machine learning*. PMLR. 2016, pp. 40–48.

[49] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. “Learning with local and global consistency”. In: *Advances in neural information processing systems* 16 (2003).

Appendix

A.1 Confidence Intervals on Corrected Filter Performance

Dataset	BernNet			GPRGNN			JacobiConv		
	None	Uniform EC	DEC	None	Uniform EC	DEC	None	Uniform EC	DEC
Actor	0.538	0.893	0.855	0.627	0.570	0.792	0.874	0.722	1.020
Chameleon	1.226	0.635	0.832	1.598	0.941	1.094	0.788	1.029	0.985
Citeseer	0.682	0.928	0.532	0.723	0.982	0.709	0.859	1.050	1.010
Computers	0.481	0.363	0.670	0.703	0.411	0.979	0.454	0.234	0.405
Cora	0.723	0.558	0.821	0.723	0.788	0.854	0.460	0.706	0.789
Cornell	4.894	5.750	3.830	6.596	11.069	1.702	7.878	5.532	8.723
Photo	0.359	0.248	0.213	0.410	0.309	0.273	0.354	0.374	0.278
Pubmed	0.254	0.398	0.238	0.325	0.254	0.418	0.469	0.398	0.411
Squirrel	1.210	0.663	0.970	0.788	1.085	0.894	0.913	1.000	0.778
Texas	3.279	2.623	3.119	3.443	3.279	2.295	8.361	2.955	3.443

Table 4: Final evaluated accuracy confidence intervals (%) on real-world datasets for all filters and eigenvalue correction strategies.

A.2 Final Accuracy Comparison Plots For All Datasets

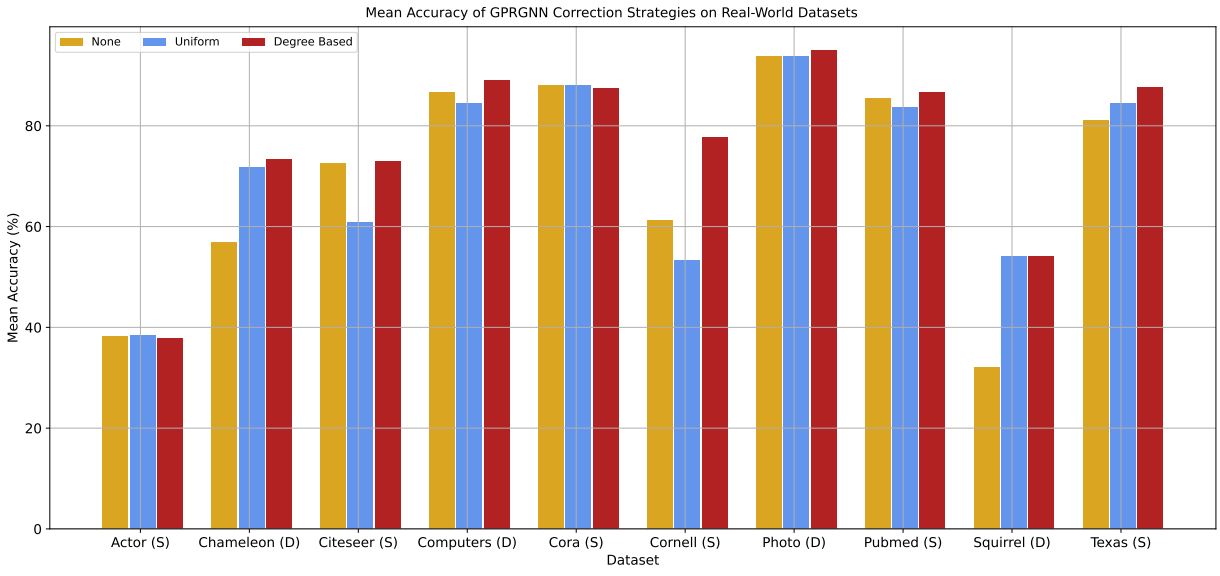


Figure 11: Final model mean accuracies for GPRGNN, compared across all real-world test datasets. Error bars (red) are the 95% confidence intervals over 10 runs for each corresponding combination of filter, correction strategy, and dataset.

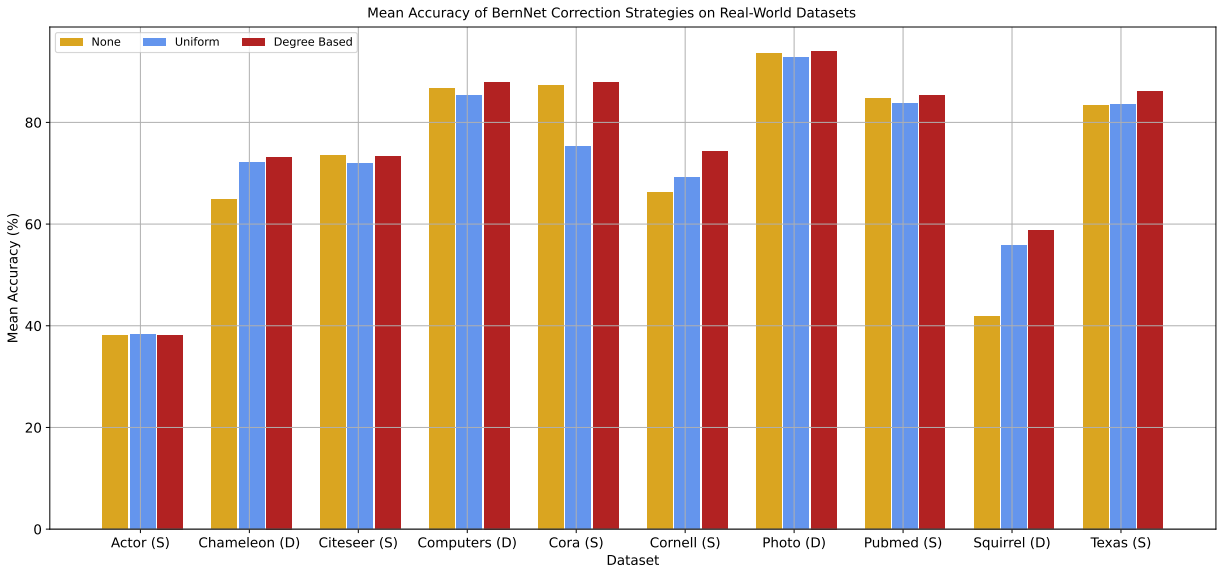


Figure 12: Final model mean accuracies for BernNet, compared across all real-world test datasets. Error bars (red) are the 95% confidence intervals over 10 runs for each corresponding combination of filter, correction strategy, and dataset.

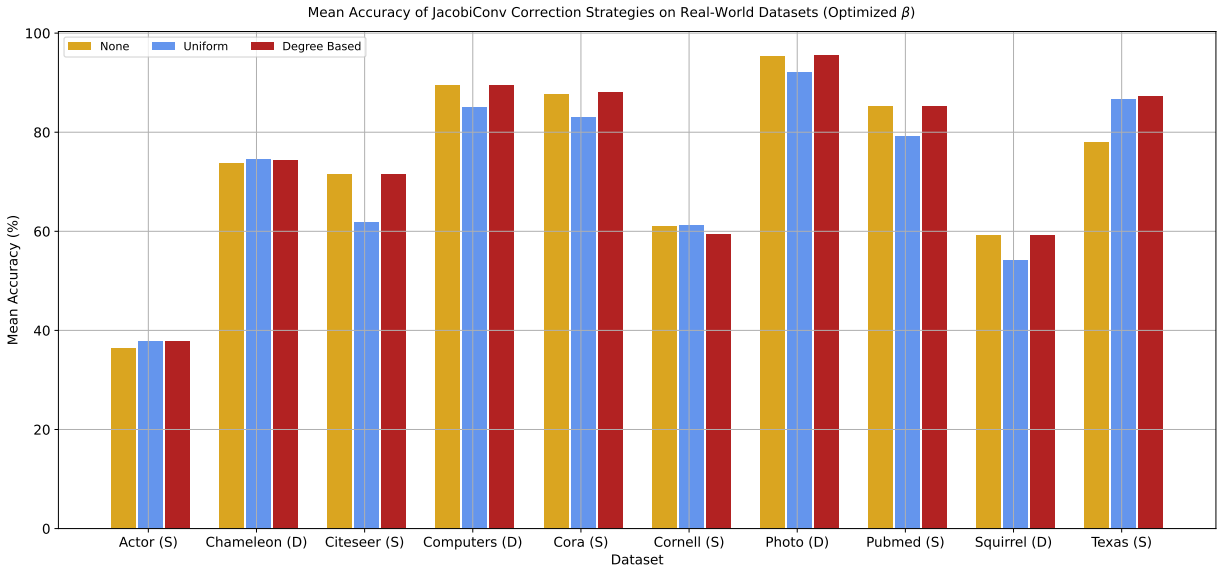


Figure 13: Final model mean accuracies for JacobiConv, compared across all real-world test datasets. Error bars (red) are the 95% confidence intervals over 10 runs for each corresponding combination of filter, correction strategy, and dataset.

Acknowledgments

I would like to thank Dr. Rajat Hazra and Dr. Odysseas Kanavetas for their guidance, encouragement, and patience during the process of completing this thesis. This past year has been an incredibly positive experience, and I would particularly like to thank Rajat for his support both in the context of my research and in my broader academic journey. Additionally, I would like to thank Kangkang Lu for providing me their code as an invaluable starting point for recreating and building upon their results. Finally, I would like to thank Chloe Cheng and Koe for useful conversations during particularly difficult problems.