Spectra of Random Graphs with Planted Partitions

Sanjoy Dasgupta[1] Alexandra Kolla[2] and Konstantinos Koiliaris[2]

[1] University of California San Diego
[2] University of Illinois Urbana-Champaign

Abstract. Spectral methods for clustering are now standard, and there are many toy examples in which they can be seen to yield more sensible solutions than classical schemes like vanilla k-means. A more rigorous analysis of these methods has proved elusive, however, and has so far consisted mostly of probabilistic analyses for random inputs with planted clusterings. Such an analysis, typically calls for proving tight asymptotic bounds on the spectrum of the graph in question. In this paper, we study a considerably broad data model first introduced by Feige and Kilian [FK01]: the planted partition graph model. We prove tight bounds on the Laplacian and Adjacency spectrum of those graphs which we think will be crucial to the design and analysis of an exact algorithm for planted partition as well as semi-random graph k-clustering.

1 Introduction

Clustering is a basic primitive of statistics and machine learning. In a typical formulation, the input consists of a data set $x_1, \ldots, x_n \in \mathbb{R}^d$, along with an integer k, and the goal is to partition the data into k groups. The most popular algorithms for doing this include: k-means, which attempts to find k centers such that the average squared distance from a point to its nearest center is minimized; EM, which fits a mixture of k Gaussians to the data; and spectral clustering, the subject of this paper.

There are several heuristics that fall in the category of spectral clustering. Here is a typical one:

- Create an undirected graph with n nodes, one per data point. Put an edge between two nodes if the corresponding point are close together according to some criterion.
- Let A denote the $n \times n$ adjacency matrix of this graph, and let D be the diagonal matrix of vertex degrees; that is, $D = \mathrm{diag}(d_1, \ldots, d_n)$, where d_i is the degree of node i. Define the Laplacian of the graph to be the $n \times n$ matrix $L = D - A$.
- Find the bottom k eigenvectors of L, say $u_1, \ldots, u_k \in \mathbb{R}^n$. Associate each data point with one row in the $n \times k$ matrix whose columns are the u_i.
- Cluster these rows using k-means.

An excellent survey of such algorithms is that of von Luxburg [vL07]. Part of the reason for their popularity is that, with suitable initial graph construction, they can handle cases, such as concentric clusters, that are difficult for competitor methods.

The analysis of spectral clustering has focused primarily on random graphs with planted partitions. The graph G is assumed to be generated as follows:

- Each index $1 \leq i \leq n$ is associated with a particular cluster $c(i) \in \{1, \ldots, k\}$.
- The adjacency matrix is constructed by setting its diagonal elements to zero, and sampling each off-diagonal entry $(i, j), i < j$, independently as follows:
  - If $c(i) = c(j)$ then $A(i, j)$ is 1 with probability p (and otherwise 0).
  - If $c(i) \neq c(j)$, then $A(i, j)$ is 1 with probability q.

  Here $p > q$, and the final matrix is made symmetric by coping elements above the diagonal to their counterparts below.

When $k = 2$ and each cluster consists of exactly half the points, this is the *planted bisection* model. The pioneering work of Boppana [B87] exhibited a convex optimization algorithm, with embedded spectral procedure, that is guaranteed to recover the bisection with high probability when $p - q > \sqrt{(p \log n)/n}$. Subsequent work of McSherry [M01] showed a somewhat simpler spectral method that can achieve roughly the same result with multiple clusters $k \geq 2$. In the machine learning literature, [NJW01] proposed essentially the algorithm outlined above, and gave a basic probabilistic analysis that was later refined and improved by [BXKS11]. The latter uses matrices with a different noise model, and thus the results are not immediately comparable.

The success of spectral methods on random graphs is certainly evidence of their efficacy. And in fact, similar random graph models with planted solutions have been used to analyze heuristics for many other NP-hard problems, such as maximum clique and independent set. This is a mathematically pleasing form of analysis, but the graph models are extremely specific and unlikely to be good reflections of real data. This caveat has led several authors to investigate models with less randomness. Blum and Spencer [BS95] considered a situation where the graph is chosen adversarially, but then each edge is flipped with a certain probability. Feige and Kilian [FK01] introduced the semirandom graph model: for bisection, this would start with a planted instance as above, but then an adversary would be allowed to add edges within either half of the bisection, or to remove edges between the two halves. These changes seem to merely enhance the planted bisection, but the adversary can use them to gain control of the principal eigenvectors and thus derail naive spectral solutions. In their paper, Feige and Kilian show how to recover the planted bisection in a semi-random graph as above *exactly*, with high probability by exhibiting an algorithm which uses a combination of semi-definite programming and spectral techniques to recover the bisection. Interestingly, one of the most crucial ingredients of their algorithm is proving tight bounds on the spectrum of the underlying planted partition random instance. They, in fact, show that understanding the spectrum of the planted partition instance is sufficient and thus, there is no need to analyze the spectrum of the graph after the adversary performed her action.

Our work. In this paper, we consider a planted partition graph model as above, for any parameter k which we describe again for simplicity. There are k clusters $c(i)$ for $i = 1, \cdots, k$ of size n/k each. For each two vertices u, v in the same cluster, we add an edge between them with probability p. For each two vertices u, v in different clusters, we add an edge between them with probability $q < p$. We denote such a (random) graph with $G_{p,q}(n, k)$.

We analyze the spectrum of the adjacency and Laplacian matrices of $G_{p,q}(n, k)$ and give tight asymptotic lower and upper bounds respectively. Our main contribution is the following theorem:

Theorem 1. *Let G be a graph satisfying model $G_{p,q}(n, k)$ (explained in detail later), c be a sufficiently large constant, $a > 0$ an arbitrary constant and n sufficiently large. Let, A_G be the adjacency matrix with $\lambda_1(A_G) \geq \lambda_2(A_G) \geq \cdots \geq \lambda_n(A_G)$ its eigenvalues, $\mathcal{L}_\mathcal{G}$ the normalized Laplacian matrix with $0 = \lambda_1(\mathcal{L}_\mathcal{G}) \leq \lambda_2(\mathcal{L}_\mathcal{G}) \leq \cdots \leq \lambda_n(\mathcal{L}_\mathcal{G})$ its eigenvalues, and δ, Δ the minimum and maximum degrees of G. Finally, let s be the index of the smallest non-negative eigenvalue of A_G. Then:*

1. *For all but the first and last k eigenvalues of A_G we have:*

$$\max\left[|\lambda_k(A_G)|, |\lambda_{n-k}(A_G)|\right] \leq c\left[\sqrt{k}\sqrt{n\log\frac{n}{k}} + \sqrt{n\log n}\right]$$

   with probability at least $1 - n^{-a}$ over the choice of A_G. Which immediately implies:

$$\lambda_k(\mathcal{L}_\mathcal{G}) > 1 - \frac{1}{\delta}c\left[\sqrt{k}\sqrt{n\log\frac{n}{k}} + \sqrt{n\log n}\right]$$

   with probability at least $1 - n^{-a}$ over the choice of $\mathcal{L}_\mathcal{G}$.
2. *And for all the eigenvalues including the first and last k, we also have a different (not as tight) bound:*

   For each $1 \leq u \leq s$: $\lambda_u(\mathcal{L}_G) > 1 - \frac{1}{\delta}\max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right]$

   For each $s{+}1 \leq v \leq n$: $\lambda_v(\mathcal{L}_G) > 1{+}\frac{1}{\Delta}\max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right]$

   with probability at least $1 - n^{-a}$ over the choice of $\mathcal{L}_\mathcal{G}$.

To our knowledge, this is the first result on spectrum of $G_{p,q}(n, k)$, for $k > 2$. Similarly to [FK01], we believe that understanding the spectrum of the planted partition model is enough in order to design algorithms that will exactly recover, with high probability, the k clusters in a semi-random graph model, where the adversary can remove edges between clusters and add edges within each cluster as she wishes.

Other related work. We note that in Thm. 1 above, we have obtained much tighter bounds for the adjacency matrix and Laplacian eigenvalues other than the first (and last) k ones. The obtained bounds reflect (informally) the fact that, for $p >> q$, each of the k clusters of the planted partition model is a very good expander compared to the expansion of cuts that have endpoints at different clusters and the quality of expansion only starts to reveal itself in the eigenvalues that are higher (in order) than λ_k. This connection between the so-called *small set expansion* property of our graphs and their spectrum has been known for a long time and especially lately has been a subject of increasing interest. For instance, simple fact regarding higher eigenvalues is that for a graph G, if (and only if) the k-th Laplacian eigenvalue $\lambda_k = 0$ then G has at least k connected components.

Moreover, several works ([LOT12, LRTV11, LRTV12, OW12]) have contributed in understanding the complete spectrum of graphs and the direct connection between higher eigenvalues and expansion of small sets . Specifically, a higher order Cheeger inequality was recently obtained by Louis, Raghavendra, Tetali, and Vempala ([LRTV11, LRTV12]) and by Lee, Oveis Gharan, and Trevisan ([LOT12]). Specifically, the latter two results show that for any k, one can partition the vertex set V into $\Omega(k)$ disjoint nonempty sets S_i, each of which has conductance $\Phi[S_i] \leq O(\sqrt{\lambda k \log k})$.

Paper organization. The rest of the paper is organized as follows. In Section 2, we give some background on graph spectra and previously known results that we use. We devote Section 3 to proving our bounds on the Adjacency and Laplacian eigenvalues of the planted partition graphs (Thm. 1 above). We give some conclusions and open questions in Section 4.

2 Preliminaries

We begin this section with some necessary preliminaries on eigenvalues along with several results that we will be using later and we continue with a detailed presentation of our graph model.

2.1 Basics

For a graph G, the adjacency matrix $A = A_G$ is defined as:

$$A_G = \begin{cases} 1 \text{ if } (u,v) \in E \\ 0 \text{ if } (u,v) \notin E \end{cases}$$

If the graph has n vertices, A_G has n real eigenvalues $\lambda_1(A_G) \geq \lambda_2(A_G) \geq \cdots \geq \lambda_n(A_G)$. The eigenvectors that correspond to these eigenvalues form an orthonormal basis of $\mathbb{R}^n$. We note that if the graph is d-regular then the largest eigenvalue is equal to d and the corresponding eigenvector is the all-one's vector.

We can use the Courant-Fisher Theorem to characterize the spectrum of A. The largest eigenvalue satisfies

$$\lambda_1(A_G) = \max_{x \in \mathbb{R}^n} \frac{x^T A x}{x^T x}$$

If we denote the first eigenvector by x_1 then

$$\lambda_2(A_G) = \max_{x \in \mathbb{R}^n, x \perp x_1} \frac{x^T A x}{x^T x}$$

Similar definitions hold for the eigenvalues λ_i, $i \geq 3$.

We will also need to define the Laplacian of a graph. For a graph G, let D be the diagonal matrix with diagonal entry $D(u,u) = d_u$ equal to the degree of node u. The Laplacian of G is defined as follows:

$$L_G = D - A_G$$

If the graph has n vertices, L_G has n real eigenvalues $0 = \lambda_1(L_G) \leq \lambda_2(L_G) \leq \cdots \leq \lambda_n(L_G)$. We can always choose n eigenvectors $\gamma_1, \cdots, \gamma_n$ such that γ_i has eigenvalue λ_i which form an orthonormal basis of $\mathbb{R}^n$. We note that 0 is always an eigenvalue with corresponding unit length eigenvector the (normalized) all-one's vector. Moreover, if and only if the graph has k connected components, then L_G has k eigenvalues equal to zero. We also define the Normalized Laplacian to be the matrix:

$$\mathcal{L}_\mathcal{G} = D^{-1/2} L_G D^{-1/2}$$

2.2 Some Results we Will Use

We next present a few lemmas and theorems we will need in our proof.
Convention: We will be using $\lambda(A)$ to denote eigenvalues of adjacency matrices and $\lambda(L)$ to denote eigenvalues of Laplacian matrices.

Lemma 1. [CDHLPS] *Let G be a connected graph, let H be a proper connected subgraph of G and let L_G and L_H be their Laplacian matrices. Then $\lambda_i(L_H) < \lambda_i(L_G), \forall i \in [1,n]$.*

Theorem 2. [FK01] *We denote the eigenvalues of a matrix by $\nu_1 \geq \nu_2 \geq ... \geq \nu_n$. Let A, B be two symmetric matrices of order n. Then $\nu_n(A) + \nu_i(B) \leq \nu_i(A+B) \leq \nu_1(A) + \nu_i(B)$.*

Theorem 3. [FK01] *Let c be a sufficiently large constant, $a > 0$ an arbitrary constant and n sufficiently large. Let, also, B be the adjacency matrix of a random graph with n vertices in which each edge has probability p, where $1/n \leq p < (n-1)/n$. Then with probability at least $1 - n^{-a}$ over the choice of B, we have that:*

$$\max\left[|\lambda_2(B)|, |\lambda_n(B)|\right] \leq \max\left[c\sqrt{pn\log n},\ c\sqrt{(1-p)n\log n}\right]$$

Let C be the adjacency matrix of a random bipartite graph on $n/2 + n/2$ vertices in which each edge connecting the two parts has probability p. Then with probability at least $1 - n^{-a}$ over the choice of C, we have that:

$$\max\left[|\lambda_2(C)|, |\lambda_{n-1}(C)|\right] \leq \max\left[c\sqrt{pn\log n},\ c\sqrt{(1-p)n\log n}\right]$$

as well.

Theorem 4. [CAV01] *Let G be a graph of size n with no isolated vertices, and δ and Δ the minimum and maximum degrees (respectively) of G. Let s be such that*

$$\lambda_1(A) \geq ... \geq \lambda_s(A) \geq 0 > \lambda_{s+1}(A) \geq ... \geq \lambda_n(A).$$

Then the following statements hold.

$$\textit{For each } 1 \leq k \leq s \ : \ 1 - \frac{\lambda_k(A)}{\delta} \leq \lambda_k(\mathcal{L}) \leq 1 - \frac{\lambda_k(A)}{\Delta}$$

$$\textit{For each } s+1 \leq k \leq n \ : \ 1 - \frac{\lambda_k(A)}{\Delta} \leq \lambda_k(\mathcal{L}) \leq 1 - \frac{\lambda_k(A)}{\delta}$$

2.3 Our Model $G_{p,q}(n,k)$

We next formalize the k-planted partition graph model that was mentioned in the introduction.

We will say that a random graph G on n vertices satisfies our model $G_{p,q}(n,k)$ where $p > q > \frac{\log n}{n}$, if it is defined as follows:

- Partition the vertices into k groups of $\frac{n}{k}$ vertices each, $k \geq 2$.
- Add an edge for every two vertices in the same cluster independently at random with probability p.
- Add an edge for every two vertices in different clusters independently at random with probability q.

3 Eigenvalue Bounds

We will now show that the Laplacian eigenvalues of a graph G satisfying this model $G_{p,q}(n,k)$, are bounded with high probability.

3.1 k-partite Random Graphs

In this section, we improve Thm. 3 of Feige and Kilian by stating a more generic result. We prove that the bounds described hold not only for random and bipartite random graphs, but indeed for any k-partite random graph. In particular:

Theorem 5. *Let c be a sufficiently large constant, $a > 0$ an arbitrary constant and n sufficiently large. Let, also, C be the adjacency matrix of a k-partite random graph with n/k vertices in each cluster and where each edge connecting any two clusters has probability p, where $1/n \leq p < (n-1)/n$. Then with high probability (at least $1 - n^{-a}$) over the choice of C, we have that:*

$$\max\left[|\lambda_k(C)|, |\lambda_{n-k}(C)|\right] \leq \max\left[c\sqrt{pn\log n},\ c\sqrt{(1-p)n\log n}\right].$$

Proof. Consider a matrix C as per the theorem and assume $p \leq 1/2$ and let C_i, $i = 1..k$ denote the k different clusters. Moreover, let D be a matrix with entries $d_{ij} = 0$ whenever $i, j \in C_l$, $l = 1..k$ and 1 otherwise.

D has k non zero eigenvalues: $(k-1)$ of them equal to $-n/k$ and one equal to $n/k \cdot (k-1)$.

Let $A = C - pD$.

Claim. It holds that $\lambda_k(C) \leq \lambda_1(A)$ and $\lambda_{n-k}(C) \geq \lambda_n(A)$.

*Proof (*Claim 1*).* Recall that $A = C - pD \Rightarrow C = A + pD$ and that A and C are both symmetric and of order n. Therefore by Thm. 2:

$$\lambda_i(C) \leq \lambda_1(A) + \lambda_i(pD) \overset{i=k}{\Longrightarrow} \lambda_k(C) \leq \lambda_1(A) + \lambda_k(pD) \overset{\lambda_k(pD)=0}{\Longrightarrow} \lambda_k(C) \leq \lambda_1(A)$$

$$\lambda_n(A) + \lambda_i(pD) \leq \lambda_i(C) \overset{i=n-k}{\Longrightarrow} \lambda_n(A) + \lambda_{n-k}(pD) \leq \lambda_{n-k}(C) \overset{\lambda_{n-k}(pD)=0}{\Longrightarrow} \lambda_n(A) \leq \lambda_{n-k}(C)$$

Let m be an even integer, we will later choose $m = \Theta(\log n)$, will prove that the upper bound on $E[Tr(A^m)]$ applies as before, and hence the same technique used in the proof of Thm. 3.

The matrix A is a random matrix.

Claim. Each entry of A has expectation 0 and variance $p(1-p)$ (except for the entries in the main diagonal which are all 0).

*Proof (*Claim 2*).* Let us begin by calculating the expectation of the entries of A. In order to do that, we will look into two cases. The first will be for entries ij such that both i, j are in the same cluster $C_i, i = 1..k$ and the second for entries ij such that i, j are not in the same cluster $C_i, i = 1..k$. Then $E[A_{ij}] = E[C_{ij}] - E[pD_{ij}]$ and we have the following analysis.

- If $i, j \in C_i$ then $E[C_{ij}] = 0$ since all the entries are 0 as there are no edges between any two vertices inside a cluster and $E[pD_{ij}] = p \cdot 0 = 0$ since again all entries are 0 by the design of D. So, $E[A_{ij}] = 0 - 0 = 0$.

- If $i, j \notin C_i$ then $E[C_{ij}] = p$ since all entries are either 0 or 1 with probability p and $E[pD_{ij}] = p \cdot 1 = p$ because all entries of D are 1 and linearity of expectation. So, $E[A_{ij}] = p - p = 0$.

Hence, $E[A_{ij}] = 0$ for all $i, j = 1..n$ and therefore $E[A] = 0$.

For the calculation of the variance of the entries of A, we will again investigate the same two cases. We know that $Var(A_{ij}) = Var(C_{ij}) - Var(pD_{ij})$ and again our cases analysis is:

– If $i, j \in C_i$ then $Var(C_{ij}) = 0$ since all the entries are 0 and $Var(pD_{ij}) = p^2 \cdot Var(D_{ij}) = p^2 \cdot 0 = 0$ since again all entries are 0. So, $Var(A_{ij}) = 0 - 0 = 0$.

– If $i, j \notin C_i$ then $Var(C_{ij}) = E[C_{ij}^2] - (E[C_{ij}])^2 = p - p^2 = p(1-p)$ since all entries are either 0 or 1 with probability p and $Var(pD_{ij}) = p^2 \cdot Var(D_{ij}) = p^2 \cdot 0 = 0$ because all entries of D are 1. So, $Var(A_{ij}) = p(1-p) - 0 = p(1-p)$.

Hence, $Var(A_{ij}) = \begin{cases} 0 \text{ for } i, j \in C_i \\ p(1-p) \text{ for } i, j \notin C_i \end{cases}, i, j = i..n.$

Let $\lambda(A) = \max[|\lambda_k(A)|, |\lambda_{n-k}(A)|]$. As A is random, we shall compute the expectation $E[\lambda]$ over the choice of random A. Then $E[\lambda]^m \leq E[\lambda^m] \leq E[Tr(A^m)]$.

From this point on the proof continues along the lines of Feige and Kilian and is omitted. □

3.2 All Eigenvalues But the First and Last k

Theorem 6. *Let G be a graph satisfying model $G_{p,q}(n, k)$, let c be a sufficiently large constant, $a > 0$ an arbitrary constant and n sufficiently large. Let, A_G be the adjacency matrix of G with $\lambda_1(A_G) \geq \lambda_2(A_G) \geq \cdots \geq \lambda_n(A_G)$ its eigenvalues. Then:*

$$\max\left[|\lambda_k(A_G)|, |\lambda_{n-k}(A_G)|\right] \leq c\left[\sqrt{k}\sqrt{n\log\frac{n}{k}}\max\left[\sqrt{p}, \sqrt{1-p}\right] + \sqrt{n\log n}\max\left[\sqrt{q}, \sqrt{1-q}\right]\right]$$

with probability at least $1 - n^{-a}$ over the choice of A_G.

Proof. Let G^c be the edge complement graph of G and let the adjacency matrix of this graph be denoted A^c and the normalized Laplacian as $\mathcal{L}^{\rfloor}$.

In order to choose at random a graph G^c with the right distribution, we do the following process:

Remark 1. For simplicity, from now on, we will be identifying graphs with their adjacency matrices. Moreover, every adjacency matrix of a graph is assumed to be onto the total of n vertices, i.e., every matrix is increased to $n \times n$ order by filling the extra rows and columns with zeros.

– Choose a random k-partite graph Q_k^c with parts $P_1, P_2, ..., P_k$, over n vertices. For each pair of vertices of different parts, there exists an edge with probability $(1 - q)$.
– Choose k random graphs $R_i^c, i = 1..k$ of size $\frac{n}{k}$ over each part $P_1, P_2, ..., P_k$ of Q_k^c. In each of these graphs, for each pair of vertices, an edge exists with probability $(1 - p)$. Moreover, let $R^c = \sum_{i=1}^{k} R_i^c$.

Then, it is easy to see that

$$Q_k^c + \sum_{i=1}^{k} R_i^c = Q_k^c + R^c = A^c$$

Now, we will prove that the eigenvalues of A^c are bounded:
By applying Thm. 2 we have that:

$$\begin{aligned}
\lambda_k(A^c) &\leq \lambda_1(R^c) + \lambda_k(Q_k^c) \overbrace{\leq}^{Thm.5} \\
&\leq \lambda_1(R^c) + \max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right] \overbrace{\leq}^{Thm.2} \\
&\leq \sum_{i=1}^{k} \lambda_1(R_i^c) + \max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right] \overbrace{\leq}^{Thm.3} \\
&\leq k \cdot \max\left[c\sqrt{p\frac{n}{k}\log\frac{n}{k}}, c\sqrt{(1-p)\frac{n}{k}\log\frac{n}{k}}\right] \\
&\qquad + \max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right] \leq \\
&\leq c\left[\sqrt{k}\sqrt{n\log\frac{n}{k}}\max\left[\sqrt{p}, \sqrt{1-p}\right] + \sqrt{n\log n}\max\left[\sqrt{q}, \sqrt{1-q}\right]\right]
\end{aligned}$$

Hence,

$$\lambda_k(A^c) \leq c\left[\sqrt{k}\sqrt{n\log\frac{n}{k}}\max\left[\sqrt{p}, \sqrt{1-p}\right] + \sqrt{n\log n}\max\left[\sqrt{q}, \sqrt{1-q}\right]\right]$$

Remark 2. Notice that the same bound holds for $\lambda_k(A)$ since the switch of the edge probabilities takes place inside the max and therefore does not alter the result.

So we can directly write:

$$\lambda_k(A) \leq c\left[\sqrt{k}\sqrt{n\log\frac{n}{k}}\max\left[\sqrt{p}, \sqrt{1-p}\right] + \sqrt{n\log n}\max\left[\sqrt{q}, \sqrt{1-q}\right]\right]$$

Moreover, by applying the same proof we can reach the following bound for the negative $\lambda_{n-k}(A)$:

$$\lambda_{n-k}(A) \geq -c\left[\sqrt{k}\sqrt{n\log\frac{n}{k}}\max\left[\sqrt{p}, \sqrt{1-p}\right] + \sqrt{n\log n}\max\left[\sqrt{q}, \sqrt{1-q}\right]\right]$$

Combining these two results, we get the claimed bound. □

By using Thm. 4 we get the following immediate corollary.

Corollary 1. *Let G be a graph satisfying model $G_{p,q}(n,k)$, let c be a sufficiently large constant, $a > 0$ an arbitrary constant and n sufficiently large. Let $\mathcal{L}_\mathcal{G}$ the normalized Laplacian matrix of G with $0 = \lambda_1(\mathcal{L}_\mathcal{G}) \leq \lambda_2(\mathcal{L}_\mathcal{G}) \leq \cdots \leq \lambda_n(\mathcal{L}_\mathcal{G})$ its eigenvalues, and δ and Δ the minimum and maximum degrees (respectively) of G. Then:*

$$\lambda_k(\mathcal{L}_\mathcal{G}) > 1 - \frac{1}{\delta} c \left[\sqrt{k} \sqrt{n \log \frac{n}{k}} \max \left[\sqrt{p}, \sqrt{1-p} \right] + \sqrt{n \log n} \max \left[\sqrt{q}, \sqrt{1-q} \right] \right]$$

with probability at least $1 - n^{-a}$ over the choice of $\mathcal{L}_\mathcal{G}$.

Proof. We know that:

$$\max \left[|\lambda_k(A)|, |\lambda_{n-k}(A)| \right] \leq c \left[\sqrt{k} \sqrt{n \log \frac{n}{k}} \max \left[\sqrt{p}, \sqrt{1-p} \right] + \sqrt{n \log n} \max \left[\sqrt{q}, \sqrt{1-q} \right] \right]$$

By Thm. 4 we can bound the normalized Laplacian eigenvalues of G and get the claimed result:

$$\lambda_k(\mathcal{L}) > 1 - \frac{1}{\delta} c \left[\sqrt{k} \sqrt{n \log \frac{n}{k}} \max \left[\sqrt{p}, \sqrt{1-p} \right] + \sqrt{n \log n} \max \left[\sqrt{q}, \sqrt{1-q} \right] \right]$$

□

Remark 3. Notice that because $1/\sqrt{2} < \max \left[\sqrt{p}, \sqrt{1-p} \right] < 2$, all expressions of the form $\max \left[\sqrt{p}, \sqrt{1-p} \right]$ can be absorbed in the unspecific constant c, giving us the final form of the first section of Thm. 1.

3.3 All Eigenvalues Including the First and Last k

Theorem 7. *Let G be a graph satisfying model $G_{p,q}(n,k)$, let c be a sufficiently large constant, $a > 0$ an arbitrary constant and n sufficiently large. Let, $\mathcal{L}_\mathcal{G}$ be the normalized Laplacian matrix, and δ and Δ the minimum and maximum degrees (respectively) of G. Then:*

$$\text{For each } 1 \leq u \leq s \ : \ \lambda_u(\mathcal{L}_G) > 1 - \frac{1}{\delta} \max \left[c\sqrt{qn \log n}, \ c\sqrt{(1-q)n \log n} \right]$$

$$\text{For each } s+1 \leq v \leq n \ : \ \lambda_v(\mathcal{L}_G) > 1 + \frac{1}{\Delta} \max \left[c\sqrt{qn \log n}, \ c\sqrt{(1-q)n \log n} \right]$$

with probability at least $1 - n^{-a}$ over the choice of $\mathcal{L}_\mathcal{G}$.

Proof. In order to prove that the last k eigenvalues are bounded w.h.p. we will use our known result Thm. 3 for random graphs and Lemma 1.

Assume the following construction.

– Begin with a graph G satisfying our model $G_{p,q}(n,k)$, with C_k clusters; and
– Remove each edge inside a cluster with independent probability $(1-\frac{q}{p})$.
– Call the resulting graph G^-.

Notice that after the above procedure finishes, the event of an edge $e_C = (u,v) : u,v \in C_i$ for some $i = 1..k$, existing in G^- is:

$$\Pr[e_C \in G^-] = Pr[e_C \in G] \cdot \Pr[e_C \text{not to be deleted}] = p \cdot 1 - \left(1 - \frac{q}{p}\right) = q$$

Moreover, the event of an edge $e_X = (u',v') : u \in C_i, v \in C_j$ for $i \neq j$, existing in G^- is:

$$\Pr[e_X \in G^-] = Pr[e_X \in G] = q$$

So, for any edge e: $\Pr[e \in G^-] = q$

Hence, with this construction we have created a new graph G^- that is a proper subset of G, and both G^-, G are connected with high probability since the probability of any edge existing in any of them is at least $q > \frac{\log n}{n}$.

By Lemma 1, it follows that with high probability $\lambda(L_G) > \lambda(L_{G^-}) \Rightarrow \lambda(\mathcal{L}_G) > \lambda(\mathcal{L}_{G^-})$.

Also, notice that G^- satisfies the prerequisites of Thm. 3, and as such:

$$\max\left[|\lambda_2(A_{G^-})|, |\lambda_n(A_{G^-})|\right] \leq \max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right]$$

Assume that $\lambda_s(A_{G^-})$, where $1 \leq s \leq n$ is the smallest positive eigenvalue of A_{G^-}. We can separate the positive and the negative eigenvalues and apply Thm. 4 to get the bounds for the normalized Laplacian of G^-:

For each $1 \leq u \leq s$ for which $\lambda_u(A_{G^-}) > 0$ we get

$$\lambda_u(\mathcal{L}_G) > \lambda_u(\mathcal{L}_{G^-}) > 1 - \frac{1}{\delta}\max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right]$$

For each $s+1 \leq v \leq n$ for which $\lambda_v(A_{G^-}) \leq 0$ we get

$$\lambda_v(\mathcal{L}_G) > \lambda_v(\mathcal{L}_{G^-}) > 1 + \frac{1}{\Delta}\max\left[c\sqrt{qn\log n},\ c\sqrt{(1-q)n\log n}\right]$$

And the statement is proved. □

Remark 4. The expected degree of each node in $G_{p,q}(n,k)$ is $E[deg] = p\frac{n}{k} + q\frac{n(k-1)}{k}$. By using standard Chernoff bound arguments (see also [BOL1], [BOL2]) we can obtain that w.h.p.

$$\delta \geq p\frac{n}{k} + q\frac{n(k-1)}{k} - \Theta\left(\sqrt{p\frac{n}{k}} + \sqrt{q\frac{n(k-1)}{k}}\right)$$

and

$$\Delta \leq p\frac{n}{k} + q\frac{n(k-1)}{k} + \Theta\left(\sqrt{p\frac{n}{k}} + \sqrt{q\frac{n(k-1)}{k}}\right).$$

Here Θ denotes some absolute constant. We omit expressing the above eigenvalue bounds using the degree guarantees for the sake of presentation.

4 Conclusion and Open Questions

We believe that our bounds will prove extremely useful in semi-definite programming based exact algorithms for k-clustering of semi-random graphs. In particular, we believe that those bounds can be used to show that a Feige-Kilian type semidefinite program will recover those clusters with high probability. It would also be interesting to see similar spectral analysis for different models of semi-random graphs, and particularly those that appear in machine learning applications. Our eigenvalue bounds are also of independent interest in the area of spectral graph theory. They exhibit once more the connection between small set expansion and higher eigenvalues, as seen in previous works [LOT12, LRTV11, LRTV12, OW12].

References

[BXKS11] Balakrishnan S., Xu M., Krishnamurthy A., and Singh A.: Noise thresholds for spectral clustering. Neural Information Processing Systems. (2011)

[BS95] Blum A., Spencer J.: Coloring random and semi-random k-colorable graphs. Journal of Algorithms. 19 2 (1995) 204–234

[BOL1] Bollobas B.: Degree sequences of random graphs. Discrete Math. 33 119. (1981)

[BOL2] Bollobas B.: Random Graphs. Academic Press, London/New York. (1985)

[B87] Boppana R.: Eigenvalues and graph bisection. 28th Annual Symposium on Foundations of Computer Science. (1987) 280–285

[CAV01] Cavers M.: The Normalized Laplacian Matrix and General Randić Index of Graphs. PhD Thesis. (2010)

[FK01] Feige U., Kilian J.: Heuristics for Semirandom Graph Problems. Journal of Computing and System Sciences. 63 (2001) 639–671

[vL07] von Luxburg U.: A tutorial on spectral clustering. Statistics and Computing. 17 4 (2007) 395–416

[NJW01] Ng A., Weiss Y., Jordan M.I.: On spectral clustering: analysis and an algorithm. Neural Information Processing Systems. (2001)

[M01] McSherry F.: Spectral partitioning of random graphs. 42nd Annual Symposium on Foundations of Computer Science. (2001) 529–537

[LOT12] Lee J., Oveis Gharan S., Trevisan L.: Multi-way spectral partitioning and higher-order Cheeger inequalities. 44th Annual ACM Symposium on Theory of Computing. (2012)

[LRTV11] Louis A., Raghavendra P., Tetali P., Vempala, S.: Algorithmic extensions of Cheegers inequality to higher eigenvalues and partitions. 14th Annual International Workshop on Approximation Algorithms for Combinatorial Optimization Problems. (2011) 315–326

[LRTV12] Louis A., Raghavendra P., Tetali P., Vempala, S.: Many sparse cuts via higher eigenvalues. 44th Annual ACM Sympo- sium on Theory of Computing. (2012)

[OW12] O'Donnell R., Witmer D.: Improved small-set expansion from higher eigenvalues. CoRR. (2012)

[CDHLPS] Chen G., Davis G., Hall F., Li Z., Patel K., Stewart M.: An interlacing result on normalized Laplacians. SIAM Journal on Discrete Mathematics. 18:353–361, (2004)