Cheeger's Inequality:
Looking at Higher Eigenvalues in Spectral Graph Theory

Armaun Sanayei, Patin Inkaew

1 Introduction

Lecture 11 and 12 introduce Spectral Graph Theory: the study of graph structure using linear algebra. We see that many features of the underlying graph can be discovered through adjacency matrix and Laplacian matrix. We also have an opportunity to experimenting on graph partitioning in mini project 6. In this paper, we will formally prove Cheeger's inequality and see how this inequality informed us about partitioning and disconnectedness.

Most importantly, Cheeger's inequality gives a performance guarantee for Spectral Partitioning Algorithm which approximately computes sparse cut. Finding sparse cut can help designing divide and conquer algorithm. Other applications include image segmentation, community detection in networks, and data clustering. For example, in the paper "Normalized Cuts and Image Segmentation", Shi and Malik show how graph cut can be used to segment regions of image in groups. In the paper, "Improved bounds for mixing rates of Markov chains and multicommodity", Sinclair uses Cheeger's inequality to improve bound on the mixing rate of Markov chains.

After providing proof for Cheeger's inequality, we will discuss some intuitions, generalization, and modern development on Cheeger's inequality and sparse cut.

2 Preliminaries

To prove Cheeger's inequality, we will need to develop some machinery. This section will prove and discuss these preliminary results.

2.1 Adjacency matrix and Laplacian matrix

Let's revisit the definition of Adjacency matrix and Laplacian matrix.

Definition 2.1. (Adjacency matrix) Given an undirected graph G with vertices $V(G) = \{1, ..., n\}$, we define adjacency matrix $A(G)$ as an $n \times n$ matrix where element a_{ij} is defined as

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

Since graph G is undirected, we can see that $a_{ij} = a_{ji}$ for all $i, j = 1, ..., n$. Thus, adjacency matrix is symmetric. Now, let's work on some examples to see how adjacency can give information about the structure of the graph.

Example 2.2. (Adjacency matrix of a complete graph)
Recall that a complete graph K_n with n vertices is a graph where every vertex is connected to all other vertices in the graph.
The adjacency matrix $A(K_n)$ can be written as

$$A(K_n) = J - I$$

where J is an all-one matrix and I is an identity matrix.
Since J has rank 1, we can see that all-one vector is an eigenvector of J with eigenvalue n and all other eigenvectors of J correspond to eigenvalue 0. Then, any vector is an eigenvector of I with eigenvalue 1. So eigenvalues of $A(K_n)$ are one less than those of J: the eigenvalues of $A(K_n)$ are $n-1$ with multiplicity 1 and -1 with multiplicity -1.
This example shows the largest gap between the largest eigenvalue and the second largest eigenvalue.

Next, we review the definition of Laplacian

Definition 2.3. (Laplacian matrix) Given an undirected graph G with vertices $V(G) = \{1, ..., n\}$, we define Laplacian matrix $L(G)$ as

$$L(G) = D(G) - A(G)$$

where $D(G) = \text{diag}(d_1, ..., d_n)$ where d_i is the degree of vertex i. Notice that $L(G)$ is symmetric since both $D(G)$ and $A(G)$ are symmetric. Below, we present some properties of $L(G)$ which are proved in lecture 11.

Lemma 2.4. *For any undirected graph G, $L(G)$ is positive semidefinite.*

Lemma 2.5. *For any undirected graph G, 0 is an eigenvalue of $L(G)$.*

Lemma 2.6. *For any undirected graph G, the number of connected components in G equals to multiplicity of zero eigenvalues of $L(G)$*

Corollary 2.7. *An undirected graph G is connected if and only if 0 is an eigenvalue of $L(G)$ with multiplicity 1. In other word, an undirected graph G is connected if and only if the second smallest eigenvalue $\lambda_2 > 0$ where λ_i are eigenvalues of $L(G)$ and $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.*

Proof. This corollary is an directed result from Lemma 2.6.

2.2 Regular Graph

In this section, we define a regular graph to help simplify the proof of Cheeger's inequality and discuss some notations.

Definition 2.8. (d-regular graph) An undirected graph is d-regular if every vertex in G has degree d.

Example 2.9. (Laplacian matrix of d-regular graph)
Let G be a d-regular graph. By definition of d-regular graph, we can see that $D(G) = \text{diag}(d, ..., d) = dI$. So the Laplacian matrix is $L(G) = dI - A$. That is, eigenvalues of $L(G)$ are $d - \alpha_i$ where α_i are eigenvalues of adjacency matrix $A(G)$.
We can see that if G is d-regular, spectrum of $L(G)$ and $A(G)$ are equivalent. Of course, this is not a case if G is not regular.

From the above example, it is reasonable to order the eigenvalues of $L(G)$ and $A(G)$ in the opposite way. When discussing $L(G)$ or $A(G)$, we will reserve λ_i for eigenvalues of $L(G)$ and α_i for eigenvalues of $A(G)$. We will index eigenvalues λ_i of $L(G)$ in non-decreasing order,

$$\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$$

and index these eigenvalues λ_i of $L(G)$ in non-increasing order,

$$\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_n$$

2.3 Rayleigh Quotient

In this section, we develop Rayleigh Quotient which is a useful tool to relate eigenvalues with optimization problem. Lecture 11 discusses how we can intuitive view computing eigenvalues as solving optimization. Here, we will prove this viewpoint formally.

Lemma 2.10. *Let A be a real symmetric matrix. Claim $\lambda_{max} = \max_x \frac{x^\top Ax}{x^\top x}$ where λ_{max} is the largest eigenvalue of A.*

Proof. Since A is a real symmetric matrix, by spectral theorem, we can decompose $A = Q\Lambda Q^\top$ where Q is orthogonal matrix and Λ is diagonal matrix whose main diagonal contains eigenvalues of A. We also know that column vectors q_i form an orthonormal basis.
We consider for any vector x,

$$\begin{aligned} x^\top Ax &= x^\top Q\Lambda Q^\top x \\ &= (Q^\top x)^\top \Lambda (Q^\top x) \\ &= \sum_{i=1}^{n} \lambda_i (q_i^\top x)^2 \\ &\leq \lambda_{max} \sum_{i=1}^{n} (q_i^\top x)^2 \\ &= \lambda_{max}(x^\top QQ^\top x) \\ &= \lambda_{max} x^\top x \end{aligned}$$

Therefore, we have

$$\frac{x^\top Ax}{x^\top x} \leq \lambda_{max}$$

We can see that the equality is obtained when x is the eigenvector of A with eigenvalue λ_{max}. Hence, we conclude that

$$\lambda_{max} = \max_x \frac{x^\top Ax}{x^\top x}$$

as desired.

Lemma 2.11. *Let A be a real symmetric matrix. Claim $\lambda_{min} = \min_x \frac{x^\top Ax}{x^\top x}$ where λ_{min} is the smallest eigenvalue of A.*

Proof. The proof is very similar to the proof of Lemma 2.10 by switching from $\geq$ to $\leq$ and λ_{max} to λ_{min}

Now, we will generalize these two lemmas for any eigenvalues of symmetric matrix A.

Theorem 2.12. *(Rayleigh Quotient) Let A be a real symmetric matrix with eigenvalue $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ and corresponding orthonormal eigenvector $v_1, v_2, ..., v_n$. Denote T_k as the set of vectors orthogonal to $v_1, v_2, ..., v_{k-1}$. The eigenvalue λ_k is given by the optimization problem:*

$$\lambda_k = \max_{x \in T_k} \frac{x^\top Ax}{x^\top x}$$

Proof. First, note that the existence of vectors $v_1, ..., v_n$ follows from spectral theorem. We consider any vector $x \in T_k$. Since $v_1, ..., v_n$ form an orthonormal basis, we can express x as a linear combination of $v_1, ..., v_n$. We write

$$x = c_1v_1 + c_2v_2 + ... + c_nv_n$$

Then, we compute

$$\begin{aligned} x^\top Ax &= (c_1v_1 + c_2v_2 + ... + c_nv_n)^\top A(c_1v_1 + c_2v_2 + ... + c_nv_n) \\ &= (c_1v_1 + c_2v_2 + ... + c_nv_n)^\top (c_1\lambda_1v_1 + c_2\lambda_2v_2 + ... + c_n\lambda_nv_n) \\ &= \lambda_1c_1^2v_1^\top v_1 + \lambda_2c_2^2v_2^\top v_2 + ... + \lambda_nc_n^2v_n^\top v_n \\ \therefore x^\top Ax &= \sum_{i=1}^n \lambda_ic_i^2 \end{aligned}$$

In the third line, we use the fact that v_i's are orthogonal. Similarly, it's easy to show that

$$x^\top x = \sum_{i=1}^n c_i^2$$

Thus, we can write

$$\frac{x^\top Ax}{x^\top x} = \frac{\sum_{i=1}^n \lambda_ic_i^2}{\sum_{i=1}^n c_i^2}$$

Recall that the coefficient $c_i = v_i^\top x$. Since $x \in T_k$, we have that $c_1 = c_2 = ... = c_{k-1} = 0$. Thus, for vector $x \in T_k$, we have

$$\frac{x^\top Ax}{x^\top x} = \frac{\sum_{i=k}^n \lambda_i c_i^2}{\sum_{i=k}^n c_i^2} \leq \frac{\lambda_k \sum_{i=k}^n c_i^2}{\sum_{i=k}^n c_i^2} = \lambda_k$$

We see that the equality is obtained if $x = v_k$. Thus, we conclude that

$$\lambda_k = \max_{x \in T_k} \frac{x^\top Ax}{x^\top x}$$

as desired.

2.4 Courant-Fischer Theorem

Rayleigh Quotient is a very useful tool; however, it relies on knowing eigenvectors used in the constraints. Courant-Fischer Theorem extends Rayleigh Quotient to give eigenvalue characterization without knowing eigenvectors.

Theorem 2.13. *(Courant-Fisher Theorem) Let A be a real symmetric matrix with eigenvalue $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ The eigenvaule λ_k can be obtained by solving the optimization problem:*

$$\lambda_k = \max_{\substack{S \subseteq \mathbb{R}^n \\ dim(S)=k}} \min_{x \in S} \frac{x^\top Ax}{x^\top x}$$

Proof. We use similar notations used in Rayleigh Quotient. Let $v_1, ..., v_n$ be eigenvectors corresponding to eigenvalue $\lambda_1, ..., \lambda_n$ of A.
Let $S_k \in \mathbb{R}^n$ be a k-dimensional subspace spanned by $\{v_1, ..., v_k\}$. For any vector $x \in S_k$, we can write $x = \sum_{i=1}^k c_i v_k$. Then, we compute

$$\frac{x^\top Ax}{x^\top x} = \frac{\sum_{i=1}^k \lambda_i c_i^2}{\sum_{i=1}^k \lambda_i c_i^2} \geq \frac{\lambda_k \sum_{i=1}^k \lambda_i c_i^2}{\sum_{i=1}^k \lambda_i c_i^2} = \lambda_k$$

Hence, we have

$$\max_{\substack{S \subseteq \mathbb{R}^n \\ dim(S)=k}} \min_{x \in S} \frac{x^\top Ax}{x^\top x} \geq \min_{x \in S_k} \frac{x^\top Ax}{x^\top x} \geq \lambda_k$$

Next, we prove that the maximum cannot exceed eigenvalue λ_k. Observe that any k-dimensional subspace must intersect with $(n-k+1)$-dimensional subspace S'_k spanned by $\{v_k, v_{k+1}, ..., v_n\}$. For any $x \in S'_k$, we have

$$\frac{x^\top Ax}{x^\top x} = \frac{\sum_{i=k}^n \lambda_i c_i^2}{\sum_{i=k}^n c_i^2} \leq \frac{\lambda_k \sum_{i=k}^n c_i^2}{\sum_{i=k}^n c_i^2} = \lambda_k$$

Hence, we have

$$\max_{\substack{S \subseteq \mathbb{R}^n \\ dim(S)=k}} \min_{x \in S} \frac{x^\top Ax}{x^\top x} \leq \max_{\substack{S \subseteq \mathbb{R}^n \\ dim(S)=k}} \min_{x \in S \cap S'_k} \frac{x^\top Ax}{x^\top x} \leq \lambda_k$$

From both inequalities, we finally conclude that

$$\lambda_k = \max_{\substack{S \subseteq \mathbb{R}^n \\ dim(S)=k}} \min_{x \in S} \frac{x^\top Ax}{x^\top x}$$

Here, we have shown Courant-Fisher Theorem.

2.5 Graph Expansion

From Corollary 2.7, we know that graph is disconnected if and only if the second smallest eigenvalue of Laplacian matrix $\lambda_2 = 0$. We might expect that if λ_2 is small, we can say that graph is close to be disconnected. In fact, Cheeger's inequality will formally justify that λ_2 is small if and only if graph is closed to be disconnected.

In this section, we will give quantitative measure for connectedness/disconnectedness of a graph. We introduce two measures: expansion and conductance.

Definition 2.14. Given an undirected graph G, for an induced subgraph $S \subseteq V$, we denote $\delta(S)$ as the set of edge cut. That is, $\delta(S) = \{(i,j) \in E \mid i \in S \text{ and } j \in V \setminus S\}$

Definition 2.15. (Expansion/Edge-expansion) The expansion of graph G is defined as

$$\Phi(G) = \min_{\substack{S \subseteq V \\ |S| \leq |V|/2}} \Phi(S), \text{ where } \Phi(S) = \frac{|\delta(S)|}{|S|}$$

In words, expansion is the ratio of number of edges cut to the number of vertices.
In other literature, expansion is also called edge-expansion, or isoperimetric number, or Cheeger number and is also denoted by $h(G)$.

Definition 2.16. (Conductance) The conductance of graph G is defined as

$$\phi(G) = \min_{\substack{S \subseteq V \\ vol(S) \leq |E|}} \phi(S), \text{ where } \phi(S) = \frac{|\delta(S)|}{vol(S)}, \text{ and } vol(S) = \sum_{i \in S} deg(i)$$

In words, conductance is the ratio of number of edges cut to total degree of the set.

Notice that If the graph is d-regular, expansion and conductance are equivalent: $\Phi(G) = d\phi(G)$. Cheeger's inequality will relate graph conductance to the second smallest eigenvalue.

Definition 2.17. We say G is an expander graph if $\phi(G)$ is large (e.g. $\phi(G) \geq 0.1$) and S is sparse cut if $\phi(S)$ is small.

Observe that for every $S \subseteq V$, we have $0 \leq \phi(S) \leq 1$.

2.6 Spectral Partitioning Algorithm

In order to compute graph conductance $\phi(G)$, we will need to find sparse cut S with minimum $\phi(S)$. This problem is often called graph conductance problem. It has been shown that graph conductance problem is NP-hard, so we need approximate algorithm. A good heuristic to finding sparse cut in practice is the following "Spectral Partitioning Algorithm"

Algorithm 1: SpectralPartitioningAlgorithm

(1) Compute the second eigenvector x of normalized Laplacian $\mathcal{L}$ (eigenvector corresponds to the second smallest eigenvalue λ_2, we will discuss normalized Laplacian $\mathcal{L}$ in the next section)

(2) Sort the vertices in non-increaseing order: $x_1 \geq x_2 \geq ... \geq x_n$ where $n = |V|$

(3) Define $S_i = \begin{cases} \{1, ..., i\} & \text{if } i \leq n/2 \\ \{i+1, ..., n\} & \text{if } i > n/2 \end{cases}$

(4) Output $\min_i\{\phi(S_i)\}$

This algorithm is very popular in practice since it can be run in linear time. This heuristic found to be very effective in clustering and image segmentation. Cheeger's inequality will give performance guarantee for this Spectral Partitioning Algorithm.

2.7 Normalized Matrices

The last machinery to help us state Cheeger's inequality nicely is the notion of normalized matrices.

Definition 2.18. We define **normalized adjacency matrix** $\mathcal{A}(G)$ as

$$\mathcal{A}(G) = D^{-1/2}AD^{-1/2}$$

and **normalized Laplacian matrix** $\mathcal{L}(G)$ as

$$\mathcal{L}(G) = I - \mathcal{A}(G)$$

Notice that equivalently, we can write

$$\mathcal{L}(G) = D^{-1/2}LD^{-1/2}$$

These normalized matrices remove the dependency on the maximum degree of the graph.

Lemma 2.19. *Let $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_n$ be eigenvalues of $\mathcal{A}$ and $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ be eigenvalues of $\mathcal{L}$. Claim $1 = \alpha_1 \geq \alpha_n \geq -1$ and $0 = \lambda_1 \leq \lambda_n \leq 2$.*

Proof. Notice that $D^{1/2}\mathbf{1}$ is eigenvector of $\mathcal{L}$ with 0 eigenvalue because

$$\mathcal{L}D^{1/2}\mathbf{1} = D^{-1/2}LD^{-1/2}D^{1/2}\mathbf{1} = D^{-1/2}L\mathbf{1} = 0$$

Furthermore, $\mathcal{L}$ is semi-positive definite because of any vector x, we have

$$x^\top \mathcal{L}x = x^\top D^{-1/2}LD^{-1/2}x = (D^{-1/2}x)^\top L(D^{-1/2}x) \geq 0$$

We view $D^{-1/2}x$ as another vector and use the fact that Laplacian matrix L is positive semi definite. This proves that $\lambda_1 = 0$.
Now, since $\mathcal{L} = I - \mathcal{A}$, we have $\alpha_1 = 1 - \lambda_1 = 1$. This is because any vector is eigenvector of I with eigenvalue 1.
Next, for any vector x, we can write

$$\begin{aligned}
x^\top(I+\mathcal{A})x &= x^\top \mathcal{L}x + 2x^\top \mathcal{A}x \\
&= \sum_{(i,j)\in E} \left[\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 + 2\frac{x_i x_j}{\sqrt{d_i d_j}} \right] \\
&= \sum_{(i,j)\in E} \left[\left(\frac{x_i}{\sqrt{d_i}} \right)^2 + \left(\frac{x_j}{\sqrt{d_j}} \right)^2 \right] \\
&\geq 0
\end{aligned}$$

Thus, $I + \mathcal{A}$ is semi-positive definite and $\alpha_n \geq -1$. Since $\mathcal{L} = I - \mathcal{A}$, we conclude that $\lambda_n = 1 - \alpha_n \leq 2$.

3 Proof of Cheeger's Inequality

Theorem 3.1. *Cheeger's Inequality:*

$$\frac{1}{2}\lambda_2 \leq \phi(G) \leq \sqrt{2\lambda_2}$$

where λ_2 is the second smallest eigenvalue of $\mathcal{L}$, the normalized Laplacian.

For the proof Cheeger's inequality, we will assume that the graph is d-regular, which is not a significant assumption. It only helps simplify a few computations, which can be easily generalized.

Easy Direction: $\frac{1}{2}\lambda_2 \leq \phi(G)$

The lower bound is the easier direction to prove in the inequality. It is "easy" because of the nice machinery we developed in the preliminaries in relation to Rayleigh quotients. Since we showed that λ_2 is the minimum of some optimization, it is not surprising that the minimum is easy to upper bound.

We first start off with utilizing our assumption of d-regularity. The first eigenvector the of the Laplacian matrix of a d-regular graph is the all-one vector.
As the discussed in the preliminaries, we can write out λ_2 in terms of the Rayleigh quotient:

$$\lambda_2 = \min_{x \perp \vec{1}} \frac{x^T \mathcal{L} x}{x^T x} = \min_{x \perp \vec{1}} \frac{x^T L x}{d x^T x} = \min_{x \perp \vec{1}} \frac{\sum_{i,j \in E} (x_i - x_j)^2}{d \sum_{i \in V} x_i^2}$$

where the first inequality comes from the definition of the normalized laplacian and the last equality comes by writing out the matrix multiplication. This equation just codifies the unique optimization problem that λ_2 solves. This property was in fact investigate on mini-project 6.
As discussed, we now try to upper bound λ_2 by finding a vector that fulfills the constraints (i.e. $x \perp \vec{1}$) but is not the minimum. To do so, we will need to consider a subgraph $S \subset V$ of our graph G such that $\phi(G) = \phi(S)$, which exists by definition of conductance.
Now, with this subgraph, we can define a binary solution vector argument for the optimization, a vector x such that:

$$x_i = \begin{cases} \frac{1}{|S|} & \text{if } i \in S \\ -\frac{1}{|V-S|} & \text{if } i \notin S \end{cases}$$

Thus, with this vector, we know first of all that that it is $x \perp \vec{1}$ since by construction the sums cancel out. Moreover, we have that:

$$\begin{aligned} \lambda_2 &\leq \frac{\sum_{i,j \in E} (x_i - x_j)^2}{d \sum_{i \in V} x_i^2} \\ &= \frac{|\delta(S)| \left(\frac{1}{|S|} + \frac{1}{|V-S|} \right)^2}{d \left(|S| \frac{1}{|S|^2} + |V-S| \frac{1}{|V-S|^2} \right)} \\ &= \frac{|\delta(S)| \cdot |V|}{d \cdot |S| \cdot |V-S|} \\ &\leq 2\phi(S) \\ &= 2\phi(G) \end{aligned}$$

Thus, as desired we have that:

$$\frac{1}{2}\lambda_2 \leq \phi(G)$$

Hard Direction: $\phi(G) \leq \sqrt{2}\lambda_2$

Our focus here will begin with the second eigenvector, x.
Without loss of generality, assume that there are fewer positive entries than negative entries in x (this is possible since sign is not of importance in the eigenvector). By making this assumption, we are setting ourselves up to deal with a vector where less than half of it is non-zero. This will come in handy soon.
Now, making this assumption on x, we define:

$$y_i = \begin{cases} x_i & \text{if} \quad x_i \geq 0 \\ 0 & \text{if} \quad x_i < 0 \end{cases}$$

Now with these two vectors in hand, we make the following proposition:

Proposition 3.2.

$$R(y) \leq R(x)$$

where $R(x) := \frac{x^T \mathcal{L} x}{x^T x}$ *is the Rayleigh coefficient.*

Proof. We define the support of y as $\text{supp}(y) = \{i | y(i) > 0\}$
Thus, we consider for $i \in \text{supp}(y)$ that:

$$\begin{aligned}
(\mathcal{L}y)_i &= y_i - \sum_{j \in \text{neighbor}(i)} \frac{y_i}{d} \\
&\leq x_i - \sum_{j \in \text{neighbor}(i)} \frac{x_j}{d} \\
&= (Lx)_i \\
&= \lambda_2 x_i
\end{aligned}$$

Thus:

$$\begin{aligned}
y^T \mathcal{L} y &= \sum_{i \in V} y_i (Ly)_i \\
&\leq \sum_{i \in \text{supp}(y)} \lambda_2 x_i^2 \\
&= \sum_i \lambda_2 y_i^2 \\
&\Downarrow \\
\frac{y^T \mathcal{L} y}{y^T y} &\leq \lambda_2
\end{aligned}$$

Thus, we get that:

$$\frac{y^T \mathcal{L} y}{y^T y} \leq \lambda_2 = \frac{x^T \mathcal{L} x}{x^T x}$$

which is what we wished to show.

We move onto another Lemma which in combination with the proposition above will give us Cheeger's inequality.

Lemma 3.3. *Given any vector* y*, there exists a subset* $S \subset \text{supp}(y)$ *such that* $\phi(S) \leq \sqrt{2R(y)}$

Proof. First, we scale y such that $\forall i \quad 0 \leq y_i \leq 1$. This proof will involve probabilistic methods, so we will begin with defining our random variable.
We have $t \in (0, 1]$ chosen uniformly random from the interval. From this random variable, we define a random subset. More specifically, we have that:

$$S_t = \{i | y_i^2 \geq t\}$$

Notice that $S_t \subset \mathrm{supp}(y)$ by construction. Since we have random variables, we will now examine expected values of different quantities, namely $|\delta(S_t)|$, the size of the edge boundary. We have that:

$$\mathbb{E}(|\delta(S_t)|) \;=\; \sum_{i,j\in E} [\mathbb{P}(v_{\mathrm{ij}}\,\mathrm{is\,cut})]$$

This comes straight from the definition of the size of the edge boundary, where each edge adds 1 weighted by the probability that it is cut (i.e. on the boundary). Furthermore, we have that $\mathbb{P}(v_{\mathrm{ij}}\,\mathrm{is\,cut}) = \mathbb{P}(y_i^2 < t \leq y_j^2)$ since the probability the edge is a boundary edge depends on how far the threshold t kept in S_t. Thus, we can further simplify:

$$\mathbb{E}(|\delta(S_t)|) \;=\; \sum_{i,j\in E} [\mathbb{P}(y_i^2 < t \leq y_j^2)]$$

Since t is uniformly random, this probability can be easily simplified to:

$$\mathbb{E}(|\delta(S_t)|) \;=\; \sum_{i,j\in E} |y_i^2 - y_j^2| = \sum_{i,j\in E} |y_i - y_j||y_i + y_j|$$

Now, we can upper bound this by Cauchy Schwarz:

$$\begin{aligned}
\sum_{i,j\in E} |y_i - y_j||y_i + y_j| &\leq \sqrt{\sum_{i,j\in E} (y_i - y_j)^2}\sqrt{\sum_{i,j\in E} (y_i + y_j)^2} \\
&\leq \sqrt{\sum_{i,j\in E} (y_i - y_j)^2}\sqrt{2\sum_{i,j\in E} (y_i + y_j)^2} \\
&\leq \sqrt{\sum_{i,j\in E} (y_i - y_j)^2}\sqrt{2d\sum_{i,j\in E} y_i^2} \\
&= \left(d\sum_{i\in V} y_i^2\right)\sqrt{2R(y)}
\end{aligned}$$

where the assumption of d-regular comes in and our definition $R(y)$. More precisely this tells us that:

$$\mathbb{E}(|\delta(S_t)|) \leq \left(d\sum_{i\in V} y_i^2\right)\sqrt{2R(y)}$$

We can similarly perform an expectation analysis on the size of S_t, we have that:

$$\begin{aligned}
\mathbb{E}[|S_t|] &= \sum_{i\in V} \mathbb{P}[y_i^2 \geq t] \\
&= \sum_{i\in V} y_i^2
\end{aligned}$$

This comes directly from the fact that t is uniformly distributed. Thus, combining these two expectation bounds together we get that:

$$\frac{\mathbb{E}(|\delta(S_t)|)}{\mathbb{E}[|S_t|]} \leq \sqrt{2R(y)}$$

Thus, by linearity of expectation, we get that:

$$\mathbb{E}\left[|\delta(S_t)| - d|S_t|\sqrt{2R(y)}\right] \leq 0$$

Now, using the probabilistic method, we can say that since the mean is less than 0, there must exist a t such that:

$$\frac{|\delta(S_t)|}{d|S_t|} \leq \sqrt{2R(y)}$$

Now, we can combine our proposition and this lemma to get that since $\lambda_2 = R(x) \leq R(y)$, we have that:

$$\phi(G) \leq \frac{|\delta(S_t)|}{d|S_t|} \leq \sqrt{2R(y)} \leq \sqrt{2R(x)} = \sqrt{2\lambda_2}$$

Thus, we have shown the hard direction of Cheeger's inequality.

4 Discussion and Intuition

The two directions of Cheeger's inequality provide us a lot of information.
The most intuitive direction is naturally the easy direction (i.e. the lower bound). The lower bound can be interpreted in a multitude of ways depending on the graph of interest(i.e. a Markov chain, expander graph, etc.). Overall, however, we can extract some general meaning.
If there is a sparse cut (i.e. $\phi(G)$ is small), then λ_2 is small. This makes sense with our intuition. Even thinking about the extreme case where a graph is disconnected, we get that λ_2 goes to 0, forced down by $\phi(G) = 0$ since there are disconnected components. On the other hand, if λ_2 is larger, then we know that G has no sparse cut and that it is well connected. This is the intuition we usually can garner from the second eigenvalue.
The harder direction is less intuitive but the insights it provides are more rewarding.
The intuition can begin with a short discussion on why the hard direction is "hard". In the proof of the easy direction, we were able to take a "short-cut" by examining binary vectors, which allowed us to side-step the the constraint on orthogonality with the all-one vector. Because of the generosity of the optimization, it was ok to restrict our domain to binary vectors.
In the hard direction, however, we are not let off as easily. We must consider the entire domain of optimization. The entire domain, however, presents many challenges in that there could be solutions/vectors that are difficult to even begin to determine the sparse cut or visualize. This explains why we had to use the technique of the probabilistic method, which allowed us to consider characteristics of the graph in aggregate. Although the proof does not construct any tangible subgraph or have much intuitive steps, the probability and expectation save us from trying to work the specifics of what could be very ugly graphs.
The interpretation of the upper bound in Cheeger's inequality tells us that although $\phi(G)$ bounds $\frac{1}{2}\lambda_2$ from above, that λ_2 cannot be "too small". If λ_2 is too small, then that implies that we can extract a sparse cut from the graph, which in turn means that $\phi(G)$ must be

small. A lot of the power of this comes from the square root. In many algorithms that use Cheeger's inequality or Cheeger's number, the square root establishes bounds and allows efficient approximation of Cheeger's number up to a square root factor. This is very useful for clustering and other machine learning tasks and used widely.

5 Tightness of the Inequality

Cheeger's inequality is tight on both sides.
One does not have to look far for an almost-tight upper bound for the Cheeger's inequality. In fact, we partially computed in our mini-project. If we consider a cycle graph of length n. We showed in our mini-project (i.e. not the general case, but a specific example) that the second eigenvalue of a cycle graph in the general case is:

$$\lambda_2 \leq \frac{x^T \mathcal{L} x}{x^T x} = \frac{x^T L x}{d x^T x} = \frac{\sum_{i,j \in E}(x_i - x_j)^2}{2 \sum_i x_i^2} = \frac{n \cdot (1^2)}{8 \sum_{j=1}^{n/4} j^2} \leq O\left(\frac{1}{n^2}\right)$$

while we have that $\phi(C_n)$ is simply:

$$\phi(C_n) = \min_{S \subset V} \frac{|\delta(S)|}{\mathrm{vol}(S)} = \min_{1 \leq s \leq \frac{n}{2}} \frac{2}{2s} = \frac{2}{n}$$

Thus, we have that the upper bound is fulfilled:

$$\phi(C_n) \leq \sqrt{2\lambda_2} = \sqrt{2O\left(\frac{1}{n^2}\right)} \leq O\left(\frac{1}{n}\right)$$

6 Generalizations and Modern Development

Over time, the Cheeger inequality has been generalized and modified. Here we will give a quick discussion of some modern variants of Cheeger's inequality.
As shown in (Lee, Oveis Gharan, Trevisan 2011), they showed

Theorem 6.1. *Higher Order Cheeger's Inequality:*

$$\frac{\lambda_k}{2} \leq \phi_k(G) \leq O(k^2)\sqrt{\lambda_k}$$

where $\phi_k(G)$ is defined to be the k-way conductance of the graph:

$$\phi_k(G) = \min_{S_1, S_2, \ldots, S_n \subset V, S_i \cap S_j = \varnothing} \quad \max_{1 \leq i \leq k} \phi(S_i)$$

This builds off of our intuition we developed in class. Just like $\lambda_2 = 0$, indicated to us that there were two disconnected components, we have similarly that $\lambda_k = 0$ indicates that there k disconnected components. Therefore, it is clear that these eigenvalues carry important information just like λ_2, therefore, it makes sense to search and find a higher-order Cheeger's inequality, which is able squeeze out important information from as many eigenvectors as possible.

As discussed earlier, the bounds Cheeger's inequality provides are very importany for many parts of computer science. This fundamental nature of the bound, incentivizes continuous and active research on its implications, simplifications, and specification. One example of such an active place of research,which was slightly touched on earlier, is the Spectral Partition Algorithm.

We have discussed one example regading the square root approximation guarantee of the Cheeger number in clustering algorithms. Another closely related application is the Spectral Partitioning Algorithm (Cheeger, 1970) which gives a $\frac{1}{\sqrt{\phi(G)}}$ approximation for the graph conductance problem.

In the $\frac{1}{\sqrt{\phi(G)}}$ approximation however $\phi(G)$ can be as small as $\frac{1}{n^2}$ for a simple graph. So the worst case approxmination ratio can be as large as $\Omega(n)$ Leighton and Rao (1999) designed a better $O(\log n)$-approximation. The best algorithm for sparse cut problem is by Arora, Rao, and Vazirani (2009), which is $O(\sqrt{\log n})$ - approximation. These results are actively being improved upon and are important for many application from expander graphs, derandomization, to designing divide and conquer algorithms.

7 Conclusion

In CS168, we were able to explore Cheeger's inequality along with many other important theorems. Through this project, by deeply examining Cheeger's inequality, we were able to see how fundamental, intricate, and useful such a simple-looking bound is. Through these investigations, we saw the importance Cheeger's inequality, its proof, the intuition surrounding Cheegers inequality, and its modern day applications and improvements.

The spirit of CS 168 is to learn about the "modern algorithmic toolbox", and Cheeger's inequality is unambiguously in this category. For practitioners, it is incredibly important bound that can help increase performance of common algorithms, like clustering, that we use everyday. For theorists, it presents an important theoretical challenge to try and squeeze as much information from the inequality and its variants as possible. It is rare to find something that rests on this line between theory and application, and this is what makes Cheeger's inequality so fundamental.

References

1. Lecture notes from CS860 at University of Waterloo: `https://cs.uwaterloo.ca/~lapchi/cs860/notes.html`

2. Lecture notes from CS168 at Stanford University: `https://web.stanford.edu/class/cs168/index.html`

3. Lecture notes from EE263 at Stanford University: `http://ee263.stanford.edu/lectures.html`

4. "Expander Flows, Geometric Embeddings and Graph Partitioning" from https://www.cs.princeton.edu/~arora/pubs/arvfull

5. "Multi-way spectral partitioning and higher-order Cheeger inequalities" by James R. Lee, Shayan Oveis Gharan, Luca Trevisan from https://arxiv.org/abs/1111.1055

6. Expander flows, geometric embeddings and graph partitioning by S Arora, S Rao, U Vazirani from https://dl.acm.org/doi/pdf/10.1145/1502793.1502794?casa_token=Qk8-mzvhgrQAAAAA:oVL2U4jYA1sEDT7pnNwBpA19o5JJIH1GJVllSV4zic4bJFVcKYGb6iN-NoFxdpdU-f

7. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms by Leighton and Rao from https://dl.acm.org/doi/pdf/10.1145/331524.331526?casa_token=rTS1cNqG0CoAAAAA:1egfjA8BGBv_Ind2mj_XLSAn2jQXmNqBGL1BYdmGK