Advanced Algorithms

Lecture 21: Cheeger's Inequality

Lecturer: Nathan Klein

1 Normalized Laplacian

Last time we learned about the Laplacian matrix L for a graph G. It is defined $L = D - A$, where A is the adjacency matrix and D the diagonal degree matrix in which all off-diagonal entries are 0 and diagonals $d_{ii} = d_i$ where d_i is the degree of the ith vertex of the graph.

We learned last time that $\lambda_1 = 0$ for every graph, and $\lambda_2 > 0$ if and only if G is connected. So, you could guess that λ_2 is a measure of how connected the graph is. If this intuition is correct, it should be large for a complete graph, which is maximally connected.

Fact 1.1. *For the complete graph, the second eigenvalue of the Laplacian is n.*

Proof. We have $L = n \cdot I - \mathbf{1}\mathbf{1}^T$. So:

$$Lx = (n \cdot I - \mathbf{1}\mathbf{1}^T)x = n \cdot x - \mathbf{1}\mathbf{1}^T x$$

but since the first eigenvector is $\mathbf{1}$, any eigenvector after the first is orthogonal to $\mathbf{1}$. So for any other eigenvector, we simply have $Lx = nx$. Therefore this matrix has $\lambda_1 = 0$ and $\lambda_i = n$ for all $i \geq 2$. $\square$

Great: maybe our intuition is right. But, if we want to have a concrete measure of connectedness, it's better for it not to depend on n. It would be nice to have a single number, say between 0 and 1, that measured the connectivity, where 0 was disconnected and 1 was very connected. The normalized Laplacian just about does this, although it technically can have λ_2 as large as 2.

Definition 1.2 (Normalized Adjacency Matrix and Normalized Laplacian)**.** *Given a graph G, the normalized adjacency matrix $\tilde{A}$ is:*

$$\tilde{A} = D^{-1/2} A D^{-1/2}$$

where D is the diagonal matrix of vertex degrees. In this way, $\tilde{A}_{i,j} = \frac{A_{i,j}}{\sqrt{d_i d_j}}$. The normalized Laplacian $\tilde{L}$ is then equal to

$$\tilde{L} = I - \tilde{A}$$

We will prove that for $\tilde{L}$, we have $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq 2$. But first we need a simple fact:

Fact 1.3. *If for two symmetric matrices A, B we have $A - B \succeq 0$ (i.e. $A - B$ is PSD), then we write $A \succeq B$, and this implies that the kth eigenvalue of B is less than the kth eigenvalue of A for all k.*

Proof. This is because $x^T(A - B)x \geq 0$ implies $x^T A x \geq x^T B x$. So, using the Courant-Fischer theorem, $\lambda_k(A) \geq \lambda_k(B)$ for any k. $\square$

Fact 1.4. *The eigenvalues of $\tilde{A}$ are between -1 and 1. The eigenvalues of $\tilde{L}$ are between 0 and 2.*

Proof. First, note that $\tilde{L}$ is still PSD. This is because:

$$x^T \tilde{L} x = x^T D^{-1/2} L D^{-1/2} x = \sum_{e \in E} x^T D^{-1/2} L_e D^{-1/2} x = \sum_{\{v_i, v_j\} \in E} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 \geq 0$$

In other words, we have $\tilde{L} = I - \tilde{A} \succeq 0$. So $\tilde{A} \preceq I$, so its eigenvalues are at most 1. We can similarly prove that $I + \tilde{A} \succeq 0$, which demonstrates that the eigenvalues of $\tilde{A}$ are at least -1.

So, the eigenvalues of $\tilde{A}$ are between -1 and 1, which implies that the eigenvalues of $I - \tilde{A}$ are between 0 and 2. □

Also, note that 0 is still an eigenvalue of $\tilde{L}$ because:

$$\tilde{L}(D^{1/2}\mathbf{1}) = (D^{-1/2} L D^{-1/2})(D^{1/2}\mathbf{1}) = D^{-1/2} L \mathbf{1} = \mathbf{0}$$

And it remains true that a graph is connected if and only if the second eigenvalue of $\tilde{L}$ is greater than 0. And for the complete graph, we can see that $\lambda_2 = \frac{n}{n-1}$. This is because now $\tilde{L} = \frac{n}{n-1} I - \frac{1}{n-1}\mathbf{1}\mathbf{1}^T$ since $\tilde{A} = \frac{1}{n-1}$ on all off-diagonal entries and 0 on the diagonal. Furthermore, the first eigenvector (since the degrees are all equal) is the all 1s vector. So, for any x orthogonal to that we have:

$$\tilde{L}x = \left(\frac{n}{n-1} I - \frac{1}{n-1}\mathbf{1}\mathbf{1}^T\right)x = \frac{n}{n-1} x$$

Things become particularly clean for regular graphs, in which all vertices have the same degree. Then, $\tilde{L} = \frac{1}{d} L$.

2 Cheeger's Inequality

The quantity λ_2 is measuring turns out to be similar to what's known as the *conductance* of the graph.

Definition 2.1 (Conductance)**.** *Given a set $S \subseteq V$, its conductance $\phi(S)$ is equal to:*

$$\phi(S) = \frac{|\delta(S)|}{\mathrm{vol}(S)}$$

where $\mathrm{vol}(S) = \sum_{v \in S} d_v$ is the sum of degrees of vertices in S. The conductance of a graph is then the minimum conductance over all cuts:

$$\phi(G) = \min_{S \subseteq V : \mathrm{vol}(S) \leq \frac{1}{2}\mathrm{vol}(V)} \phi(S)$$

Notice that $0 \leq \phi(G) \leq 1$. If the graph is disconnected, clearly $\phi(G) = 0$ as we can choose the cut S with $|\delta(S)| = 0$. It is also at most 1, since $|\delta(S)| \leq \sum_{v \in S} d_v = \mathrm{vol}(S)$ so the ratio is at most 1.

The conductance is a useful property because it encodes whether the graph has any bottlenecks. For example, in a social network it might correspond to a community: one which has a lot of internal connections and few external ones. A graph for which $\phi(G) \geq \Omega(1)$ is called an **expander** graph. As we will see shortly, this is equivalent to asking that $\lambda_2 \geq \Omega(1)$, where λ_2 is the second smallest eigenvalue of $\tilde{L}$. The complete graph is an easy example of an expander:

Fact 2.2. *The conductance of the complete graph is $\frac{1}{2} \cdot \frac{n}{n-1} = \frac{1}{2}\lambda_2$, where λ_2 is the second smallest eigenvalue of its normalized Laplacian.*

Proof. The conductance of a set S is:

$$\phi(S) = \frac{|\delta(S)|}{\text{vol}(S)} = \frac{|S|(n - |S|)}{(n-1) \cdot |S|} = \frac{n - |S|}{n-1}$$

To minimize this over all S of volume at most $\frac{1}{2}\text{vol}(V)$, we pick $|S| = \frac{n}{2}$, achieving

$$\phi(G) = \frac{n/2}{n-1} = \frac{1}{2}\frac{n}{n-1} = \frac{1}{2}\lambda_2$$ □

Given our framing of the above fact, one may suspect that there is a relationship between the conductance and the second smallest eigenvalue of a graph. This is true, and is formalized by Cheeger's inequality.

Theorem 2.3 (Cheeger's Inequality)**.** *For a graph G, if λ_2 is the second smallest eigenvalue of normalized Laplacian $\tilde{L}$, we have:*

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

We have seen for the complete graph, $\phi(G) = \frac{1}{2}\lambda_2$, so the first inequality is tight. It turns out the other is tight as well.

Computing the conductance of a graph turns out to be NP-Hard, and it is also difficult to approximate: there is no $O(1)$ approximation known. So, computing λ_2 is a good way to get a sense of the conductance of a graph. If $\phi(G)$ is a constant, then λ_2 is a constant as well, so it at least tells us if G is an expander.

2.1 Spectral Partitioning Algorithm

Before we prove Cheeger's inequality, let's see an important algorithm that comes out of it: the **spectral partitioning algorithm.**

1. Compute the second eigenvector x of $\tilde{L}$
2. Sort the vertices so that $x_1 \geq x_2 \geq \cdots \geq x_n$.
3. Consider all n "threshold" cuts arising from taking vertices $\{v_1, \ldots, v_i\}$ for all $1 \leq i \leq n$, and return the one which minimizes $\phi(S)$.

It turns out that this algorithm can be implemented in nearly linear time in $|E|$, since all that is required is computing v_2, λ_2, which can be done quickly using something called the "power method."

It typically works very well in practice and can be used to find sparse cuts in massive graphs.

2.2 Proof of Cheeger: Easy Direction

We will prove Cheeger for d-regular graphs. The general proof is not that different. We'll start with the "easy" direction:

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

For the d-regular case, the first eigenvector is $\mathbf{1}$ (just like for the unnormalized Laplacian). So:

$$\lambda_2 = \min_{x\in\mathbb{R}^n, x\neq 0 \langle x,\mathbf{1}\rangle=0} \frac{x^T\tilde{L}x}{x^Tx} = \frac{1}{d}\cdot \min_{x\in\mathbb{R}^n, x\neq 0 \langle x,\mathbf{1}\rangle=0} \frac{x^TLx}{x^Tx} = \frac{1}{d}\cdot \min_{x\in\mathbb{R}^n, x\neq 0 \langle x,\mathbf{1}\rangle=0} \frac{\sum_{\{i,j\}\in E}(x_i-x_j)^2}{x^Tx}$$

where we used that since G is d-regular, the normalized Laplacian is just a $\frac{1}{d}$ scaling of L.

For this direction, we need to show an upper bound on λ_2. So we just need to construct some vector x that's orthogonal to $\mathbf{1}$ and for which $\frac{\sum_{\{i,j\}\in E}(x_i-x_j)^2}{x^Tx}$ is small. I imagine most of you have a good guess for what to use here! We want to pick the cut S for which $\phi(S)=\phi(G)$ and probably let $x=\mathbf{1}_S$. We will have $|S|\le\frac{n}{2}$ since the graph is d-regular and we optimize over sets with volume at most $\frac{1}{2}\mathrm{vol}(V)$.

This is a great start, but $\mathbf{1}_S$ is not orthogonal to $\mathbf{1}$. So we should pick some scaling of $\mathbf{1}_S$ which has the property that $\langle x,\mathbf{1}\rangle = \sum_{i=1}^n x_i = 0$. We will let:

$$x_i = \begin{cases} +\frac{1}{|S|} & \text{for } i\in S \\ -\frac{1}{|V\setminus S|} & \text{otherwise}\end{cases}$$

Now, $\sum_{i=1}^n x_i = |S|\cdot\frac{1}{|S|} - |V\setminus S|\cdot\frac{1}{|V\setminus S|} = 0$, so it is orthogonal to $\mathbf{1}$. And now:

$$\lambda_2 \le \frac{1}{d}\frac{\sum_{\{i,j\}\in E}(x_i-x_j)^2}{x^Tx} = \frac{1}{d}\frac{(\frac{1}{|S|}+\frac{1}{|V\setminus S|})^2|\delta(S)|}{|S|\cdot\frac{1}{|S|^2}+|V\setminus S|\cdot\frac{1}{|V\setminus S|^2}}$$

where in the numerator we used that we only count edges where $i\in S$ and $j\notin S$ or vice versa and so the signs will match in x_i and $-x_j$. But now simplifying, this is:

$$\frac{(\frac{1}{|S|}+\frac{1}{|V\setminus S|})|\delta(S)|}{d} = \frac{|V|\cdot|\delta(S)|}{d\cdot|S|\cdot|V\setminus S|} = \frac{|V|}{|V\setminus S|}\cdot\frac{|\delta(S)|}{d\cdot|S|} \le 2\cdot\frac{|\delta(S)|}{\mathrm{vol}(S)} = 2\phi(S) = 2\phi(G)$$

where to bound $\frac{|V|}{|V\setminus S|}$ we use that $|S|\le\frac{n}{2}$.

2.3 Proof of Cheeger: Hard Direction

For the hard direction, we now will need to take the second eigenvector, call it x, with eigenvalue λ_2 and use it to *construct* a sparse cut with conductance at most $\sqrt{2\lambda_2}$. If the eigenvector is similar to what we used for the easy direction (taking one positive value and one negative value), then this is easy to do. But this may not be the case. Since G is connected, the first eigenvector is $\mathbf{1}$ and x is orthogonal to $\mathbf{1}$.

We now want to constrain x so that it has at most $\frac{n}{2}$ non-zero entries. Since we are still working only with d-regular graphs, this will ensure that any cut S containing indices that are non-zero in x will have $\mathrm{vol}(S)\le\frac{n}{2}=\frac{1}{2}\mathrm{vol}(V)$.

WLOG assume that there at least as many negative entries as positive entries (otherwise let $x = -x$, which does not change the value of $x^T \tilde{L} x$). Then let y be the entrywise maximum of x and 0, so that $y_i = x_i$ if $x_i \geq 0$ and $y_i = 0$ otherwise.

Fact 2.4.

$$\frac{x^T \tilde{L} x}{x^T x} \geq \frac{y^T \tilde{L} y}{y^T y}$$

Proof. First notice that for all i, $(\tilde{L} y)_i = y_i - \sum_{j:\{i,j\} \in E} \frac{1}{d} y_j$. This is because the ith row of $\tilde{L}$ has a 1 in position i (as $\tilde{L} = I - \tilde{A}$) and -1 on the entries $\{i, j\}$ where $\{i, j\} \in E$.

For indices where $y_i > 0$, this is at most $x_i - \sum_{j:\{i,j\} \in E} \frac{1}{d} x_j = (\tilde{L} x)_i = \lambda_2 x_i$. So:

$$y^T \tilde{L} y = \sum_{i=1}^{n} y_i (\tilde{L} y)_i \leq \sum_{i: y_i > 0} \lambda_2 x_i^2 = \sum_{i=1}^{n} \lambda_2 y_i^2$$

which implies that $\frac{y^T \tilde{L} y}{y^T y} \leq \lambda_2 = \frac{x^T \tilde{L} x}{x^T x}$. □

Scale y so that the largest entry of y is 1. Now, to find our cut, let's pick a random threshold t between 0 and 1. Let $S = \{v_i \mid y_i^2 \geq t\}$.

Theorem 2.5 ("Hard" Direction of Cheeger)**.** *There exists a cut S with $|S| \leq \frac{n}{2}$ (i.e. $vol(S) \leq vol(V)$) and $\phi(S) \leq \sqrt{2\lambda_2}$.*

Proof. First let's look at the expected number of edges cut, $\mathbb{E}\left[|\delta(S)|\right]$.

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

This is because we will cut an edge exactly when index i is less than the threshold and index j is greater than the threshold, since we assume WLOG $y_i \leq y_j$. But now, this is the probability $t \in [y_i^2, y_j^2]$ which is $y_j^2 - y_i^2$. So our expected cut size is:

$$\sum_{\{i,j\} \in E} y_j^2 - y_i^2 = \sum_{\{i,j\} \in E} (y_j - y_i)(y_j + y_i)$$

Now, think of this as the multiplication of two vectors u, v with entries in $\mathbb{R}^E$, where $u_e = (y_j - y_i)$ and $v_e = (y_j + y_i)$ for each $e = \{i, j\} \in E$. Then, by Cauchy-Schwarz:

$$\sum_{\{i,j\} \in E} (y_j - y_i)(y_j + y_i) = \langle u, v \rangle \leq \|u\|_2 \|v\|_2 = \sqrt{\sum_{\{i,j\} \in E} (y_j - y_i)} \sqrt{\sum_{\{i,j\} \in E} (y_j + y_i)}$$

The left term here is exactly $\sqrt{y^T L y}$. The RHS is at most $\sqrt{2 \sum_{\{i,j\} \in E} (y_i^2 + y_j^2)}$, as $(y_i + y_j)^2 \leq 2y_i^2 + 2y_j^2$. Then, this is $\sqrt{2d \sum_{i=1}^{n} y_i^2}$ since each index appears d times. So, overall we have:

$$\leq \sqrt{x^T L x} \sqrt{2d \sum_{i=1}^{n} y_i^2} = \sqrt{x^T L x} \sqrt{2d \sum_{i=1}^{n} y_i^2} \cdot \frac{\sqrt{d \cdot y^T y}}{\sqrt{d \cdot y^T y}} = \sqrt{2 \frac{y^T L y}{d \cdot y^T y} \cdot d \sum_{i=1}^{n} y_i^2}$$

and by our above fact, this is at most $d\sqrt{2\lambda_2}\sum_{i=1}^{n} y_i^2$.

Second, note that $\mathbb{E}\left[|S|\right] = \sum_{i=1}^{n} \mathbb{P}\left[y_i^2 \geq t\right] = \sum_{i=1}^{n} y_i^2$. Therefore:

$$\mathbb{E}\left[|\delta(S)|\right] \leq d\sqrt{2\lambda_2} \cdot \sum_{i=1}^{n} y_i^2 = d\sqrt{2\lambda_2}\mathbb{E}\left[|S|\right]$$

Therefore, there exists some cut where $|\delta(S)| \leq d\sqrt{2\lambda_2}|S|$, or equivalently $\phi(S) \leq \sqrt{2\lambda_2}$. Furthermore, $S \neq \emptyset$ since $t < 1$ with probability 1 and we scaled y so that its maximum coordinate was 1. $\square$

This implies that the spectral partitioning algorithm returns a cut with conductance at most $\sqrt{2\lambda_2}$, since it will try all possible cuts of the above type. This gives an $O(\frac{1}{\sqrt{\lambda_2}})$ approximation for the problem of finding a cut minimizing $\phi(S)$, which is known as the sparsest cut problem. The best known approximation for sparsest cut currently gives a ratio of $O(\sqrt{\log n})$, and finding a better algorithm is a major open problem.