CSE 521: Design and Analysis of Algorithms I **Winter 2017**

Lecture 12: Cheeger's Inequality cont., Spectral Clustering, Power Method

Lecturer: Shayan Oveis Gharan *02/22/17* *Scribe: Elizabeth Clark*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

12.1 Cheeger's Inequality (continued)

12.1.1 Review from last class

Definition 12.1 (Conductance)**.** *Given a graph $G = (V, E)$ with V partitioned into S and $\bar{S}$, the conductance of S is defined as:*

$$\phi(S) = \frac{|E(S, \bar{S})|}{\text{vol}(S)}$$

The conductance of G is defined as:

$$\phi(G) = \min_{\text{vol}(S) \leq \frac{\text{vol}(V)}{2}} \phi(S)$$

Question: Why aren't we looking at the mincut as a measure of conductance? Why are we normalizing by the size/volume of S in $\frac{|E(S,\bar{S})|}{\text{vol}(S)}$?

Answer: Consider a network of roads. One road in this network is a highway that connects two major cities. Another is your driveway that connects your house to the rest of the network. If cut either of these roads, it will divide the network into two disconnected sub-networks, so these are two mincuts of our graph each with size 1. However, the two roads have very different levels of conductance. The numbers of drivers inconvenienced by the shutdown of a highway is much greater than those inconvenienced by the shutdown of your driveway.

12.1.2 Cheeger's Inequality

The following theorem is one of the fundamental inequalities in spectral graph theory.

Theorem 12.2 (Cheeger's Inequality)**.** *For any graph G,*

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

where λ_2 is the 2nd smallest eigenvalue of $\tilde{L}$.

Cheeger's inequality relates the combinatorial property of conductance to a spectral property, the 2nd smallest eigenvalue. Observe that in the extreme case where $\lambda_2 = 0$, we also have $\phi(G) = 0$ and vice versa.

A crucial fact about the above inequality is that it does not depend on the size of G, n. It implies that if G has a "small" 2nd eigenvalue, then it is partitionable, whereas if the 2nd eigenvalue is "large" the graph is similar to a complete graph and it is not partitionable.

The proof of the right side of Cheeger's inequality, $\phi(G) \leq \sqrt{2\lambda_2}$ is constructive, and it shows that the spectral partitioning algorithm always returns a set S such that $\text{vol}(S) \leq \text{vol}(V)/2$ and

$$\phi(S) \leq \sqrt{2\lambda_2} \leq \sqrt{4\phi(G)}.$$

We now discuss several consequences of the above theorem for a special family of graphs.

Definition 12.3 (Expander Graphs). *Expander graphs are sparse highly connected graphs with large 2nd eigenvalues, i.e., $\lambda_2 \geq \Omega(1)$. So, the can be seen as a sparse complete graphs which have $\lambda_2 = 1$. It turns out that most of the graphs are expanders, because a random d-regular graph satisfies $\lambda_2 \geq 1 - \frac{2}{\sqrt{d}}$*

Expander graphs are the easiest instances to use for many of the optimization problems (see PS4 for applications to the max cut problem). They are frequently used in coding theory (see PS4) and in psuedorandom number generators. In Problem Set 4, we will discuss the expander mixing lemma, which states that Expander Graphs are approximately the same as random graph. In words, in a d-regular expander graphs, for every disjoint large sets S, T, $|E(S,T)|$ is very close to the expected number of edges between S, T in a random $G(n, d/n)$ graph.

Definition 12.4 (Planar Graphs). *A planar graph is one where all vertices can be projected onto a plane with no crossing edges.*

We know a lot about planar graphs. For example, we know that they tend to be sparse with average degrees at most 5.

It turns out that the 2nd eigenvalue of $\tilde{L}$ of any planar graph is at most $O(1/n)$.

Theorem 12.5. *If G is a bounded degree planar graph, the*

$$\lambda_2 \leq O(\frac{1}{n}).$$

Using the Cheeger's inequality, we can show that for every bounded degree planar graph G, $\phi(G) \leq O(1/\sqrt{n})$. In fact, by repeatedly peeling off sets of small conductance in G, we can show that every planar graph with bounded degree has a sparse bisection, i.e., a set $S \subseteq V$ such that $\text{vol}(V)/3 \leq \text{vol}(S) \leq 2\,\text{vol}(V)/3$ and $\phi(S) \leq O(1/\sqrt{n})$. This means that it is very easy to break a bounded degree planar graph into two sets such that there is very small number of connections between the two. This makes planar graphs ideal candidate for divide and conquer algorithms. We can recursively solve our problem on the two sides $S, \overline{S}$ and then merge the solutions. Since there are only $O(\sqrt{n})$ edges between $S, \overline{S}$, the merge operation can be done very efficiently

Finally, we also note that we can generalize the task of partitioning a graph into two sets, into partitioning a graph into k sets. Doing so, we define the k-way conductance of a graph to be:

Definition 12.6 (k-way conductance). *For an integer $k > 1$ and a graph G, let*

$$\phi_k(G) = \min_{disjoint S_1,\ldots,S_k} \max_{1 \leq i \leq k} \phi(S)$$

where the min is over all k disjoint sets $S_1, \ldots, S_k$ in G.

In other words, we are interested in finding k disjoint sets such that their maximum conductance is as small as possible. With this definition, $\phi(G)$ corresponds to the case $k = 2$.

It turns out that there is a natural generalization of Cheeger's inequality for larger values of k.

Theorem 12.7. *For any graph G and an integer $k > 2$,*

$$\begin{aligned} \lambda_k/2 \leq \phi_k(G) &\leq O(\sqrt{\lambda_k} \cdot k^2) \\ \phi_k(G) &\leq O(\sqrt{\log k \cdot \lambda_{2k}}). \end{aligned}$$

Note that the second inequality is stronger than the first one only when λ_{2k} is not much larger than λ_k. Similar to Cheeger's inequality, the proof of the right side of this inequality is constructive and provides an algorithm to k disjoint sets with small conductance.

12.1.3 Proof of "easier side" of Cheeger's Inequality

In this lecture we prove the easy direction of Cheeger's inequality, i.e., we show that, for any graph G,

$$\frac{\lambda_2}{4} \leq \phi(G). \tag{12.1}$$

Recall that the normalized Laplacian matrix is defined as $\tilde{L} = D^{-1/2}LD^{-1/2}$, So, the first eigenvector of $\tilde{L}$ is $D^{1/2}\mathbf{1}$ with eigenvalue 0. This is because,

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

By Rayleigh quotient,

$$\begin{aligned} \lambda_2 &= \min_{x:x\perp D^{1/2}\mathbf{1}} \frac{x^T\tilde{L}x}{x^Tx} \\ &= \min_{x:x\perp D^{1/2}\mathbf{1}} \frac{x^TD^{-1/2}LD^{-1/2}x}{xD^{-1/2}DD^{-1/2}x} \\ &= \min_{x:\substack{x\perp D^{1/2}\mathbf{1} \\ y=D^{-1/2}x}} \frac{y^TLy}{yDy} \\ &= \min_{x:\substack{x\perp D^{1/2}\mathbf{1} \\ y=D^-1/2x}} \frac{\sum_{i\sim j}(y_i-y_j)^2}{\sum_i d_iy_i^2}. \end{aligned} \tag{12.2}$$

To prove (12.1), we need to relate this value to

$$\phi(G) = \min_{S:\mathrm{vol}(S)\leq\mathrm{vol}(V)/2} \phi(S).$$

Let S be the best set in the RHS of above, i.e., assume $\phi(S) = \phi(G)$ and $\mathrm{vol}(S) \leq \mathrm{vol}(V)/2$. We can write,

$$\begin{aligned} \phi(S) &= \frac{|E(S,\overline{S}|}{\mathrm{vol}(S)} \\ &= \frac{\sum_{i\sim j}|\mathbb{I}[i\in S] - \mathbb{I}[j\in S]|}{\sum_{i\in S}d_i} \\ &= \frac{\sum_{i\sim j}|\mathbb{I}[i\in S] - \mathbb{I}[j\in S]|^2}{\sum_i \mathbb{I}[i\in S]^2} \end{aligned}$$

To see the last identity note that the absolute value of the difference of two indicator functions is either 0 or 1, so $|\mathbb{I}[i \in S] - \mathbb{I}[j \in S]| = |\mathbb{I}[i \in S] - \mathbb{I}[j \in S]|^2$. As usual let $1_i^S = \begin{cases} 1 & i \in S \\ 0 & \text{otherwise.} \end{cases}$ Note that the above equation is very similar to (12.2). Roughly speaking, in the above we are looking at the Rayleigh quotient for a specific vector $\mathbf{1}^S$ whereas the RHS of (12.2) is the minimum possible value of Rayleigh quotient over all vectors in $\mathbb{R}^n$. So, it seems that we should get $\phi(G) = \phi(S) \geq \lambda_2$.

This however is not quite right. First let us define a new inner-product operator. For two vectors x, y we write

$$\langle x, y\rangle_D = \sum_i x_i y_i d_i.$$

So, we scale the product of x_i, y_i with the degree of vertex i. Also, note that $\|x\|_D = \sqrt{\langle x, x\rangle_D}$. Recall that the minimum in (12.2) is taken over all y where $y = D^{-1/2}x$ and $x \perp D^{1/2}\mathbf{1}$; in other words, all y where

$$\langle y, \mathbf{1}\rangle_D = \langle D^{-1/2}x, \mathbf{1}\rangle_D = \langle DD^{-1/2}x, \mathbf{1}\rangle = \langle x, D^{1/2}\mathbf{1}\rangle = 0$$

Observe that

$$\langle 1^S, \mathbf{1}\rangle_D = \sum_i 1_i^S d_i = \sum_{i \in S} d_i = \text{vol}(S). \tag{12.3}$$

So, this inner product is not zero.

So, we have to perturb the 1^S vector and make it orthogonal to the all-ones vector. In general, if we to make a given vector x orthogonal to a vector v all we need to do is to take off the projection of x along the v *direction* from x.

$$x = x - \left\langle x, \frac{v}{\|v\|} \right\rangle \frac{v}{\|v\|}.$$

Note that it is crucial in the above to normalize v; $v/\|v\|$ is the unit norm vector in the direction of v.

So, we will do the same operation for the $\mathbf{1}^S$ vector. Note that we have to do all calculations with respect to the inner product space $\langle ., . \rangle_D$.

$$z := \mathbf{1}^S - \left\langle \mathbf{1}^S, \frac{\mathbf{1}}{\|\mathbf{1}\|_D} \right\rangle_D \frac{\mathbf{1}}{\|\mathbf{1}\|_D} = \mathbf{1}^S - \frac{\langle \mathbf{1}^S, \mathbf{1}\rangle_D}{\|\mathbf{1}\|_D^2}\mathbf{1}.$$

So, to prove (12.1), we need to show that

$$\frac{\sum_{i\sim j} |\mathbf{1}_i^S - \mathbf{1}_j^S|^2}{\sum_i \mathbf{1}_i^{S^2} d_i} \geq \frac{1}{2} \frac{\sum_{i\sim j}(z_i - z_j)^2}{\|z\|_D^2} \tag{12.4}$$

where z as defined above. Note that the numerator of (12.4) is shift-invariant, i.e.,

$$\sum_{i\sim j}(z_i - z_j)^2 = \sum_{i\sim j}((z_i - c) - (z_j - c))^2$$

for any $c \in \mathbb{R}$. Since z is just a shift of $\mathbf{1}^S$ because we are just taking off projection along the all-ones vector, the numerators of the left and right hand sides of (12.4) are equal. So, it is enough to show that

$$\|z\|_D^2 \geq \frac{1}{2}\,\text{vol}(S). \tag{12.5}$$

Note that the above inequality is not true if $S = V$. In fact, this is the only place in the proof that we use that $\text{vol}(S) \leq \text{vol}(V)/2$, Therefore,

$$\begin{aligned}
\|z\|_D^2 &= \left\langle \mathbf{1}^S - \frac{\langle \mathbf{1}^S, \mathbf{1}\rangle_D}{\|\mathbf{1}\|_D^2}\mathbf{1}, \mathbf{1}^S - \frac{\langle \mathbf{1}^S, \mathbf{1}\rangle_D}{\|\mathbf{1}\|_D^2}\mathbf{1} \right\rangle_D \\
&= \langle \mathbf{1}^S, \mathbf{1}^S\rangle_D + \frac{\langle \mathbf{1}^S, \mathbf{1}\rangle_D^2}{\|\mathbf{1}\|_D^4}\langle \mathbf{1}, \mathbf{1}\rangle_D - 2\frac{\langle \mathbf{1}^S, \mathbf{1}\rangle_D^2}{\|\mathbf{1}\|_D^2} \\
&= \langle \mathbf{1}^S, \mathbf{1}^S\rangle_D - \frac{\langle \mathbf{1}^S, \mathbf{1}\rangle_D^2}{\|\mathbf{1}\|_D^2} \\
&= \text{vol}(S) - \frac{\text{vol}(S)^2}{\text{vol}(V)}
\end{aligned}$$

where in the last equality we used $\langle \mathbf{1}^S, \mathbf{1}^S\rangle = \sum_{i\in S} d_i = \text{vol}(S)$, $\langle \mathbf{1}^S, \mathbf{1}\rangle = \text{vol}(S)$ by (12.3) and that $\|\mathbf{1}\|_D^2 = \langle \mathbf{1}, \mathbf{1}\rangle_D = \sum_i d_i = \text{vol}(V)$. Since $\text{vol}(S) \leq \text{vol}(V)/2$, we have

$$\|z\|_D^2 = \text{vol}(S)(1 - \text{vol}(S)/\,\text{vol}(V)) \geq \text{vol}(S)/2.$$

This proves (12.5) which implies (12.4) as desired.

We do not prove the following lemma; interested reader can see lecture notes of more advanced courses linked in the course website for the proof of the harder direction of Cheeger's inequality.

Lemma 12.8. *For all y such that $\langle y, 1\rangle_D = 0$, the spectral partitioning algorithm returns S such that $\phi(S) \leq 2\sqrt{\frac{y^TLy}{\|y\|_D^2}}$.*

The importance of the above lemma is that we don't need to find the actual eigenvector of λ_2 to use the spectral partitioning algorithm. As long as we can approximately minimize the Rayleigh quotient, $\frac{y^TLy}{yDy}$, we can run the spectral partitioning algorithm on the approximate vector to obtain a set S of small conductance. In the next lecture we will see how to find an approximate second eigenvector of the $\tilde{L}$ in almost linear time.

12.1.4 A Bad example for Spectral Partitioning Algorithm

Spectral Partitioning Algorithm does not always return the optimal solution, in fact it may return a set of a significantly larger conductance than the optimum. Consider the following example.

Suppose we have the graph shown in Figure 12.1. Consider 2 possible cuts of this graph. Cut 1 (shown in red) will give a conductance value $\frac{4}{2n}$, or $O(\frac{1}{n})$. Cut 2 (shown in green) will give a conductance value of $\frac{n\frac{50}{n^2}}{2n}$, or $O(\frac{1}{n^2})$. While Cut 2 is much better than Cut 1, SPA will return Cut 1. This is because the 2nd smallest eigenvector of this graph is the same as the 2nd smallest eigenvector of a cycle, i.e., it maps the endpoints of each dashed edge to the same value. Because of that the algorithm indeed returns a cut whose conductance is n times the optimum.

12.2 Spectral Clustering Algorithm

This is a brief discussion of Ng, Jordan, and Weiss [NJW02] paper on spectral clustering.

Motivating Example: Suppose you want to cluster a set of points, but your points look something like those depicted in Figure 12.2a. In this case, you want to find the green and black clusters. If you run k-means on this data, you won't find these clusters.

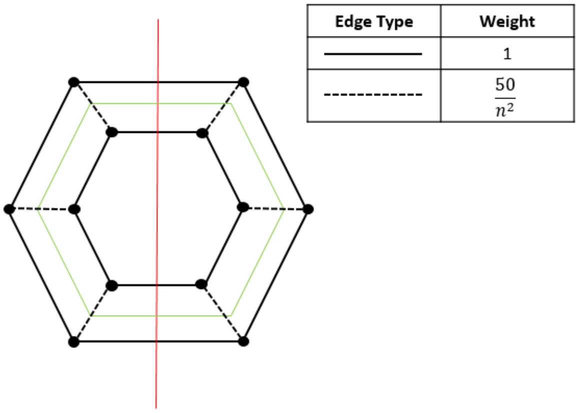
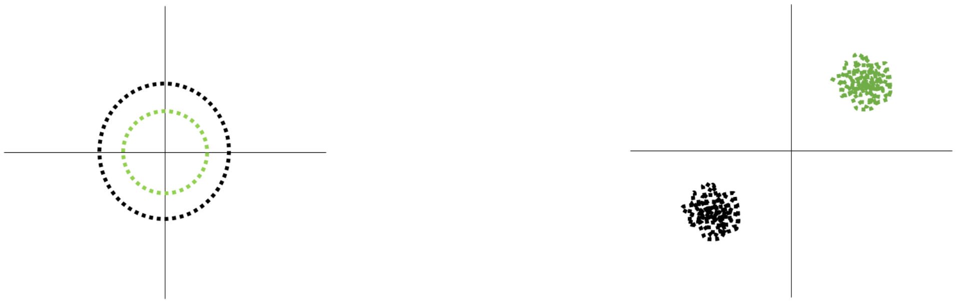


Figure 12.1: A weighted graph comprised of two cycles. The conductance of the red cut is n times the conductance of the green cut, but the spectral partitioning algorithm returns the red cut.



(a) Original data

(b) Data after spectral clustering

Figure 12.2: Spectral clustering: before and after

Instead, we can use SPA by creating a graph from this data by connecting points with an edge of weight

$$e^{\frac{-\|x_i - x_j\|^2}{\sigma^2}},$$

where x_i, x_j represents any two datapoints in $\mathbb{R}^d$. Note that the above Gaussian kernel is maximized if x_i is very close to x_j. The parameter σ must be tuned based on the particular application in mind.

After constructing this graph, we compute the normalized Laplacian matrix and the first k eigenvectors $v_1, v_2, \ldots, v_k$ of the matrix (since we want a k-partition of the graph).

Then we build the spectral embedding of graph, i.e., a matrix

$$F = \begin{bmatrix} D^{-\frac{1}{2}} v_1 \\ \vdots \\ D^{-\frac{1}{2}} v_k \end{bmatrix} \in \mathbb{R}^{k \times n},$$

which has a column for every vertex in the graph. Now, we map each vertex of graph (or each data point) i to a point in k dimensions corresponding to the i-th column of the above matrix. It turns out that in this new mapping the each cluster of points will be mapped close to one another, see Figure 12.2b and we can use k-means to find the k partition. In **??** we give a rigorous analysis of (a varaint of) this algorithm; we show that for any graph G we can find k disjoint sets $S_1, \ldots, S_k$ each of conductance $O(\sqrt{lambda_k} k^2)$. In other words, this shows that if the graph that we construct from the data points has k small eigenvalues then we can use k means to find a k partitioning of the graph. Also, conversely, if the first k eigenvalues of G are not small, then there is no “good” k partitionings of G.

References

[NJW02] A. Ng, M. Jordan, and Y. Weiss. “On spectral clustering: Analysis and an algorithm”. In: *NIPS*. 2002 (cit. on p. 12-5).