Stat260/CS294: Spectral Graph Methods **Lecture 7 - 02/12/2015**

Lecture: Spectral Methods for Partitioning Graphs (2 of 2)

Lecturer: Michael Mahoney *Scribe: Michael Mahoney*

Warning: these notes are still very rough. They provide more details on what we discussed in class, but there may still be some errors, incomplete/imprecise statements, etc. in them.

7 Proof of Cheeger's Inequality

Here, we will prove the easy direction and the hard direction of Cheeger's Inequality. Recall that what we want to show is that

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

7.1 Proof of the easy direction of Cheeger's Inequality

For the easy direction, recall that what we want to prove is that

$$\lambda_2 \leq \sigma(G) \leq 2\phi(G).$$

To do this, we will show that the Rayleigh quotient is a relaxation of the sparsest cut problem.

Let's start by restating the sparsest cut problem:

$$\begin{aligned}
\sigma(G) &= \min_{S\subset V: S\neq 0, S\neq V} \frac{E\left(S,\bar{S}\right)}{\frac{d}{|V|}|S|\cdot|\bar{S}|} \\
&= \min_{x\in\{0,1\}^n\setminus\{\vec{0},\vec{1}\}} \frac{\sum_{\{u,v\}\in E}|x_u - x_v|}{\frac{d}{n}\sum_{\{u,v\}\in V\times V}|x_u - x_v|} \\
&= \min_{x\in\{0,1\}^n\setminus\{\vec{0},\vec{1}\}} \frac{\sum_{\{u,v\}\in E}|x_u - x_v|^2}{\frac{d}{n}\sum_{\{u,v\}\in V\times V}|x_u - x_v|^2}, \qquad (1)
\end{aligned}$$

where the last equality follows since x_u and x_v are Boolean values, which means that $|x_u - x_v|$ is also a Boolean value.

Next, recall that

$$\lambda_2 = \min_{x\in\mathbb{R}^n\setminus\{\vec{0}\}, x\perp\vec{1}} \frac{\sum_{\{u,v\}\in E}|x_u - x_v|^2}{d\sum_v x_v^2}. \qquad (2)$$

Given that, we claim the following.

Claim 1.

$$\lambda_2 = \min_{x\in\mathbb{R}^n\setminus Span\{\vec{1}\}} \frac{\sum_{\{u,v\}\in E}|x_u - x_v|^2}{\frac{d}{n}\sum_{\{u,v\}}|x_u - x_v|^2}. \qquad (3)$$

Proof: Note that

$$\begin{aligned}\sum_{u,v} |x_u - x_v|^2 &= \sum_{uv} x_u^2 + \sum_{uv} x_v^2 - 2\sum_{uv} x_u x_v \\ &= 2n\sum_v x_v^2 - 2\left(\sum_v x_v\right)^2.\end{aligned}$$

Note that for all $x \in \mathbb{R}^n \setminus \{\vec{0}\}$ s.t. $x \perp \vec{1}$, we have that $\sum_v x_v = 0$, so

$$\begin{aligned}\sum_v x_v^2 &= \frac{1}{2n}\sum_{u,v} |x_u - x_v|^2 \\ &= \frac{1}{n}\sum_{\{u,v\}} |x_u - x_v|^2,\end{aligned}$$

where the first sum is over unordered pairs u, v of nodes, and where the second sum of over ordered pairs $\{u, v\}$ (i.e. we double count (u, v) and (v, u) in first sum, but not in second sum). So,

$$\min_{x \in \mathbb{R}^n \setminus \{\vec{0}\}, x \perp \vec{1}} \frac{\sum_{\{u,v\} \in E} |x_u - x_v|^2}{d\sum_v x_v^2} = \min_{x \in \mathbb{R}^n \setminus \{0\}, x \perp \vec{1}} \frac{\sum_{\{u,v\} \in E} |x_u - x_v|^2}{\frac{d}{n}\sum_{\{u,v\}} |x_u - x_v|^2}.$$

Next, we need to remove the part along the all-ones vector, since the claim doesn't have that.

To do so, let's choose an x^* that maximizes Eqn. (3). Observe the following. If we shift every coordinate of that vector x^* by the same constant, then we obtain another optimal solution, since the shift will cancel in all the expressions in the numerator and denominator. (This works for any shift, and we will choose a particular shift to get what we want.)

So, we can define x' s.t. $x'_v = x_v - \frac{1}{n}\sum_u x_u$, and note that the entries of x' sum to zero. Thus $x' \perp \vec{1}$. Note we need $x \notin Span(\vec{1})$ to have $x' \neq \vec{0}$ So,

$$\min_{x \in \mathbb{R}^n \setminus \{0\}, x \perp \vec{1}} \frac{\sum_{\{u,v\} \in E} |x_u - x_v|^2}{\frac{d}{n}\sum_{\{u,v\}} |x_u - x_v|^2} = \min_{x \in \mathbb{R}^n \setminus Span\{\vec{1}\}} \frac{\sum_{\{u,v\} \in E} |x_u - x_v|^2}{\frac{d}{n}\sum_{\{u,v\}} |x_u - x_v|^2}.$$

This establishes the claim.

◇

So, from Eqn. (1) and Eqn. (3), it follows that λ is a continuous relaxation of $\sigma(G)$, and so $\lambda_2 \leq \sigma(G)$, from which the easy direction of Cheeger's Inequality follows.

7.2 Some additional comments

Here are some things to note.

- There is nothing required or forced on us about the use of this relaxation, and in fact there are other relaxations. We will get to them later. Some of them lead to traditional algorithms, and one of them provides the basis for flow-based graph partitioning algorithms.

- Informally, this relaxation says that we can replace $x \in \{0,1\}^n$ or $x \in \{-1,1\}^n$ constraint with the constraint that x satisfies this "on average." By that, we mean that x in the relaxed problem is on the unit ball, but any particular value of x might get distorted a lot, relative to its "original" $\{0,1\}$ or $\{-1,1\}$ value. In particular, note that this is a very "global" constraint. As we will see, that has some good features, e.g., many of the well-known good statistical properties; but, as we will see, it has the consequence that any particular local pairwise metric information gets distorted, and thus it doesn't lead to the usual worst-case bounds that are given only in terms of n the size of the graph (that are popular in TCS).

- While providing the "easy" direction, this lemma gives a quick low-degree polynomial time (whatever time it takes to compute an exact or approximate leading nonrtivial eigenvector) certificate that a given graph is expander-like, in the sense that for all cuts, at least a certain number of edges cross it.

- There has been a lot of work in recent years using approaches like this one; I don't know the exact history in terms of who did it first, but it was explained by Trevisan very cleanly in course notes he has had, and this and the proof of the other direction is taken from that. In particular, he describes the randomized rounding method for the other direction. Spielman has slightly different proofs. These proofs here are a combination of results from them.

- We could have proven this "easy direction" by just providing a test vector. E.g., a test vector that is related to an indicator vector or a partition. We went with this approach to highlight similarities and differences with flow-based methods in a week or two.

- The other reason to describe λ_2 as a relaxation of $\sigma(G)$ is that the proof of the other direction that $\phi(G) \leq \sqrt{2\lambda_2}$ can be structured as a randomized rounding algorithm, i.e., given a real-valued solution to Eqn. (3), find a similarly good solution to Eqn. (1). This is what we will do next time.

7.3 A more general result for the hard direction

For the hard direction, recall that what we want to prove is that

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

Here, we will state—and then we will prove—a more general result. For the proof, we will use the randomized rounding method. The proof of this result is algorithmic/constructive, and it can be seen as an analysis for the following algorithm.

VanillaSpectralPartitioning. Given as input a graph $G = (V, E)$, a vector $x \in \mathbb{R}^n$,

1. Sort the vertices of V in non-decreasing order of values of entries of x, i.e., let $V = \{v_1, \cdots, v_n\}$, where $x_{v_1} \leq \cdots \leq x_{v_n}$.

2. Let $i \in [n-1]$ be s.t.

$$\max\{\phi(\{v_1, \cdots, v_i\}), \phi(\{v_{i+1}, \cdots, v_n\})\},$$

   is minimal.

3. Output $S = \{v_1, \ldots, v_i\}$ and $\bar{S} = \{v_{i+1}, \ldots v_n\}$.

This is called a "sweep cut," since it involves sweeping over the input vector and looking at n (rather than 2^n partitions) to find a good partition.

We have formulated this algorithm as taking as input a graph G and any vector x. You might be more familiar with the version that takes as input a graph G that first compute the leading nontrivial eigenvector and then performs a sweep cut. We have formulated it the way we did for two reasons.

- We will want to separate out the spectral partitioning question from the question about how to compute the leading eigenvector or some approximation to it. For example, say that we don't run an iteration "forever," i.e., to the asymptotic state to get an "exact" answer to machine precision. Then we have a vector that only approximates the leading nontrivial eigenvector. Can we still use that vector and get nontrivial results? There are several interesting results here, and we will get back to this.
- We will want to separate out the issue of global eigenvector to something about the structure of the relaxation. We will see that we can use this result to get local and locally-biased partitions, using both optimization and random walk based idea. In particular, we will use this to generalize to locally-biased spectral methods.

So, establishing the following lemma is sufficient for what we want.

Lemma 1. *Let $G = (V, E)$ be a d-regular graph, and let $x \in \mathbb{R}^n$ be s.t. $x \perp \vec{1}$. Define*

$$R(x) = \frac{\sum_{\{u,v\}\in E} |x_u - x_v|^2}{d \sum_v x_v^2}$$

and let S be the side with less than $|V|/2$ nodes of the output cut of VanillaSpectralPartitioning*. Then,*

$$\phi(S) \leq \sqrt{2R(x)}$$

Before proving this lemma, here are several things to note.

- If we apply this lemma to a vector x that is an eigenvector of λ_2, then $R(x) = \lambda_2$, and so we have that $\phi(G) \leq \phi(S) \leq \sqrt{2\lambda_2}$, i.e., we have the difficult direction of Cheeger's Inequality.
- On the other hand, any vector whose Rayleigh quotient is close to that of λ_2 also gives a good solution. This "rotational ambiguity" is the usual thing with eigenvectors, and it is different than any approximation of the relatation to the original expansion IP. We get "goodness" results for such a broad class of vectors to sweep over since we are measuring goodness rather modestly: only in terms of objective function value. Clearly, the actual clusters might change a lot and in general will be very different if we sweep over two vectors that are orthogonal to each other.
- This result actually holds for vectors x more generally, i.e., vectors that have nothing to do with the leading eigenvector/eigenvalue, as we will see below with locally-biased spectral methods, where we will use it to get upper bounds on locally-biased variants of Cheeger's Inequality.
- In this case, in "eigenvector time," we have found a set S with expansion s.t. $\phi(S) \leq \sqrt{\lambda_2} \leq 2\sqrt{\phi(G)}$.

- This is *not* a constant-factor approximation, or any nontrivial approximation factor in terms of n; and it is incomparable with other methods (e.g., flow-based methods) that do provide such an approximation factor. It is, however, nontrivial in terms of an important structural parameter of the graph. Moreover, it is efficient and useful in many machine learning and data analysis applications.

- The above algorithm can be implemented in roughly $O(|V| \log |V| + |E|)$ time, assuming arithmetic operations and comparisons take constant time. This is since once we have computed
$$E(\{v_1, \ldots, v_i\}, \{v_{i+1}, \ldots, v_n\}),$$
it only takes $O(\text{degree}(v_{i+1}))$ time to compute
$$E(\{v_1, \ldots, v_{i+1}\}, \{v_{i+2}, \ldots, v_n\}).$$

- As a theoretical point, there exists nearly linear time algorithm to find a vector x such that $R(x) \approx \lambda_2$, and so by coupling this algorithm with the above algorithm we can find a cut with expansion $O\left(\sqrt{\phi(G)}\right)$ in nearly-linear time. Not surprisingly, there is a lot of work on providing good implementations for these nearly linear time algorithms. We will return to some of these later.

- This quadratic factor is "tight," in that there are graphs that are that bad; we will get to these (rings or Guattery-Miller cockroach, depending on exactly how you ask this question) graphs below.

7.4 Proof of the more general lemma implying the hard direction of Cheeger's Inequality

Note that λ_2 is a relaxation of $\sigma(G)$ and the lemma provides a rounding algorithtm for real vectors that are a solution of the relaxation. So, we will view this in terms of a method from TCS known as randomized rounding. This is a useful thing to know, and other methods, e.g., flow-based methods that we will discuss soon, can also be analyzed in a similar manner.

For those who don't know, here is the one-minute summary of randomized rounding.

- It is a method for designing and analyzing the quality of approximation algorithms.

- The idea is to use the probabilistic method to convert the optimal solution of a relaxation of a problem into an approximately optimal solution of the original problem. (The probabilistic method is a method from combinatorics to prove the existence of objects. It involves randomly choosing objects from some specified class in some manner, i.e., according to some probability distribution, and showing that the objects can be found with probability > 0, which implies that the object exists. Note that it is an existential/non-constructive and not algorithmic/constructive method.)

- The usual approach to use randomized rounding is the following.
  - Formulate a problem as an integer program or integer linear program (IP/ILP).
  - Compute the optimal fractional solution x to the LP relaxation of this IP.

– Round the fractional solution x of the LP to an integral solution x' of the IP.

- Clearly, if the objective is a min, then $\text{cost}(x) \leq \text{cost}(x')$. The goal is to show that $\text{cost}(x')$ is not much more that $\text{cost}(x)$.
- Generally, the method involves showing that, given any fractional solution x of the LP, w.p. > 0 the randomized rounding procedure produces an integral solution x' that approximated x to some factor.
- Then, to be computationally efficient, one must show that $x' \approx x$ w.h.p. (in which case the algorithm can stay randomized) or one must use a method like the method of conditional probabilities (to derandomize it).

Let's simplify notation: let $V = \{1, \ldots, n\}$; and so $x_1 \leq x_2 \leq \cdots x_n$. In this case, the goal is to show that there exists $i \in [n]$ w.t.

$$\phi\left(\{1, \ldots, i\}\right) \leq \sqrt{2R(x)} \quad \text{and} \quad \phi\left(\{i+1, \ldots, n\}\right) \leq \sqrt{2R(x)}.$$

We will prove the lemma by showing that there exists a distribution D over sets S of the form $\{1, \ldots, i\}$ s.t.

$$\frac{\mathbb{E}_{S \sim D}\left\{E(S, \bar{S})\right\}}{\mathbb{E}_{S \sim D}\left\{d \min\{|S|, |\bar{S}|\}\right\}} \leq \sqrt{2R(x)}. \tag{4}$$

Before establishing this, note that Eqn. (4) does *not* imply the lemma. Why? In general, it is the case that $\mathbb{E}\left\{\frac{X}{Y}\right\} \neq \frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}$, but it suffices to establish something similar.

Fact. For random variables X and Y over the sample space, even though $\mathbb{E}\left\{\frac{X}{Y}\right\} \neq \frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}$, it is the case that

$$\mathbb{P}\left\{\frac{X}{Y} \leq \frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}\right\} > 0,$$

provided that $Y > 0$ over the entire sample space.

But, by linearity of expectation, from Eqn. (4) it follows that

$$\mathbb{E}_{S \sim D}\left[E(S, \bar{S}) - d\sqrt{2R(x)} \min\{|S|, |\bar{S}|\}\right] \leq 0.$$

So, there exists a set S in the sample space s.t.

$$E(S, \bar{S}) - d\sqrt{2R(x)} \min\{|S|, |\bar{S}|\} \leq 0.$$

That is, for S and $\bar{S}$, at least on of which has size $\leq \frac{n}{2}$,

$$\phi(S) \leq \sqrt{2R(x)},$$

from which the lemma will follow.

So, because of this, it will suffice to establish Eqn. (4). So, let's do that.

Assume, WLOG, that $x_{\lceil \frac{n}{2} \rceil} = 0$, i.e., the median of the entires of x equals 0; and $x_1^2 + x_n^2 = 1$. This is WLOG since, if $x \perp \vec{1}$, then adding a fixed constant c to all entries of x can only decrease the

Rayleigh quotient:

$$\begin{aligned} R\left(x+(c,\ldots,c)\right) &= \frac{\sum_{\{(u,v)\}\in E}\left|(x_u+c)-(x_v+c)\right|^2}{d\sum_v (x_v+c)^2} \\ &= \frac{\sum_{\{(u,v)\}\in E}\left|x_u-x_v\right|^2}{d\sum_v x_v^2 - 2dc\sum_v x_v + nc^2} \\ &= \frac{\sum_{\{(u,v)\}\in E}\left|x_u-x_v\right|^2}{d\sum_v x_v^2 + nc^2} \\ &\leq R(x). \end{aligned}$$

Also, multiplying all entries by fixed constant does *not* change the value of $R(x)$, nor does it change the property that $x_1 \leq \cdots \leq x_n$.

We have made these choices since they will allow us to define a distribution D over sets S s.t.

$$\mathbb{E}_{S\sim D} \min\left\{|S|, |\bar{S}|\right\} = \sum_i x_i^2 \tag{5}$$

Define a distribution D over sets $\{1,\ldots,i\}$, $1 \leq i \leq n-1$, as the outcome of the following probabilistic process.

1. Choose a $t \in [x_1, x_n] \in \mathbb{R}$ with probability density function equal to $f(t) = 2|t|$, i.e., for $x_1 \leq a \leq b \leq x_n$, let

$$\mathbb{P}\left[a \leq t \leq b\right] = \int_a^b 2|t|dt = \begin{cases} |a^2-b^2| & \text{if } a \text{ and } b \text{ have the same sign} \\ a^2+b^2 & \text{if } a \text{ and } b \text{ have different signs} \end{cases},$$

2. Let $S = \{u : x_i \leq t\}$

From this definition

- The probability that an element $i \leq \frac{n}{2}$ belongs to the smaller of the sets $S, \bar{S}$ equals the probability of $i \in S$ and $|S| \leq |\bar{S}|$, which equals the probability that the threshold t is in the range $[x_i, 0]$, which equals x_i^2.
- The probability that an element $i > \frac{n}{2}$ belongs to the smaller of the sets $S, \bar{S}$ equals the probability of $i \in \bar{S}$ and $|S| \geq |\bar{S}|$, which equals the probability that the threshold t is in the range $[0, x_i]$, which equals x_i^2.

So, Eqn. (5) follows from linearity of expectation.

Next, we want to estimate the expected number of edges between S and $\bar{S}$, i.e.,

$$\mathbb{E}\left[E\left(S,\bar{S}\right)\right] = \sum_{(i,j)\in E} \mathbb{P}\left[\text{edge } (i,j) \text{ is cut by } (S,\bar{S})\right].$$

To estimate this, note that the event that the edge (i,j) is cut by the partition $(S,\bar{S})$ happens when t falls in between x_i and x_j. So,

- if x_i and x_j have the same sign, then

$$\mathbb{P}\left[\text{edge } (i,j) \text{ is cut by } (S,\bar{S})\right] = |x_i^2 - x_j^2|$$

- if x_i and x_j have the different signs, then

$$\mathbb{P}\left[\text{edge } (i,j) \text{ is cut by } (S,\bar{S})\right] = x_i^2 + x_j^2$$

The following expression is an upper bound that covers both cases:

$$\mathbb{P}\left[\text{edge } (i,j) \text{ is cut by } (S,\bar{S})\right] \leq |x_i - x_j| \cdot (|x_i| + |x_j|)\,.$$

Plugging into the expressions for the expected number of cut edges, and applying the Cauchy-Schwatrz inequality gives

$$\begin{aligned}
\mathbb{E}E\left(S,\bar{S}\right) &\leq \sum_{(i,j)\in E} |x_i - x_j|\left(|x_i| + |x_j|\right) \\
&\leq \sqrt{\sum_{(i,j)\in E} (x_i - x_j)^2}\sqrt{\sum_{(i,j)\in E} \left(|x_i| + |x_j|\right)^2}
\end{aligned}$$

Finally, to deal with the expression $\sum_{(ij)\in E}\left(|x_i| + |x_j|\right)^2$, recall that $(a+b)^2 \leq 2a^2 + 2b^2$. Thus,

$$\sum_{(ij)\in E} \left(|x_i| + |x_j|\right)^2 \leq \sum_{(ij)\in E} 2x_i^2 + 2x_j^2 = 2d\sum_i x_i^2.$$

Putting all of the pieces together, we have that

$$\frac{\mathbb{E}\left[E\left(S,\bar{S}\right)\right]}{\mathbb{E}\left[d\min\{|S|,|\bar{S}|\}\right]} \leq \frac{\sqrt{\sum_{(ij)\in E}(x_i - x_j)^2}\sqrt{2d\sum_i x_i^2}}{d\sum_i x_i^2} = \sqrt{2R(x)},$$

from which the result follows.