Stat260/CS294: Spectral Graph Methods **Lecture 6 - 02/10/2015**

Lecture: Spectral Methods for Partitioning Graphs (1 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.

6 Introduction to spectral partitioning and Cheeger's Inequality

Today and next time, we will cover what is known as *spectral graph partitioning*, and in particular we will discuss and prove Cheeger's Inequality. This result is central to all of spectral graph theory as well as a wide range of other related spectral graph methods. (For example, the isoperimetric "capacity control" that it provides underlies a lot of classification, etc. methods in machine learning that are not explicitly formulated as partitioning problem.) Cheeger's Inequality relates the quality of the cluster found with spectral graph partitioning to the best possible (but intractable to compute) cluster, formulated in terms of the combinatorial objectives of expansion/conductance. Before describing it, we will cover a few things to relate what we have done in the last few classes with how similar results are sometimes presented elsewhere.

6.1 Other ways to define the Laplacian

Recall that $L = D - A$ is the graph Laplacian, or we could work with the normalized Laplacian, in which case $L = I - D^{-1/2}AD^{-1/2}$. While these definition might not make it obvious, the Laplacian actually has several very intuitive properties (that could alternatively be used as definitions). Here, we go over two of these.

6.1.1 As a sum of simpler Laplacians

Again, let's consider d-regular graphs. (Much of the theory is easier for this case, and expanders are more extremal in this case; but the theory goes through to degree-heterogeneous graphs, and this will be more natural in many applications, and so we will get back to this later.)

Recall the definition of the Adjacency Matrix of an unweighted graph $G = (V, E)$:

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

In this case, we can define the Laplacian as $L = D - A$ or the normalized Laplacian as $L = I - \frac{1}{d}A$.

Here is an alternate definition for the Laplacian $L = D - A$. Let G_{12} be a graph on two vertices with one edge between those two vertices, and define

$$L_{G_{12}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Then, given a graph on n vertices with just one edge between vertex u and v, we can define L to be the all-zeros matrix, except for the intersection between the u^{th} and v^{th} column and row, where we define that intersection to be

$$L_{G_{uv}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Then, for a general graph $G = (V, E)$, one we can define

$$L_G = \sum_{(u,v)\in E} L_{G_{uv}}.$$

This provides a simpler way to think about the Laplacian and in particular changes in the Laplacian, e.g., when one adds or removes edges. In addition, note also that this generalizes in a natural way to $\sum_{(u,v)\in E} w_{uv} L_{G_{uv}}$ if the graph $G = (V, E, W)$ is weighted.

Fact. This is identical to the definition $L = D - A$. It is simple to prove this.

From this characterization, several things follow easily. For example,

$$x^T L x = \sum_{(u,v)\in E} w_{uv} \left(x_u - x_v\right)^2,$$

from which it follows that if v is an eigenvector of L with eigenvalue λ, then $v^T L v = \lambda v^T v \geq 0$. This means that every eigenvalue is nonnegative, i.e., L is SPSD.

6.1.2 In terms of discrete derivatives

Here are some notes that I didn't cover in class that relate the Laplacian matrix to a discrete notion of a derivative.

In classical vector analysis, the Laplace operator is a differential operator given by the divergence of the gradient of a function in Euclidean space. It is denoted:

$$\nabla \cdot \nabla \text{ or } \nabla^2 \text{ or } \triangle$$

In the cartesian coordinate system, it takes the form:

$$\nabla = \left(\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_n}\right),$$

and so

$$\triangle f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$

This expression arises in the analysis of differential equations of many physical phenomena, e.g., electromagnetic/gravitational potentials, diffusion equations for heat/fluid flow, wave propagation, quantum mechanics, etc.

The *discrete Laplacian* is defined in an analogous manner. To do so, somewhat more pedantically, let's introduce a discrete analogue of the gradient and divergence operators in graphs.

Given an undirected graph $G = (V, E)$ (which for simplicity we take as unweighted), fix an *arbitrary* orientation of the edges. Then, let $K \in \mathbb{R}^{V \times E}$ be the edge-incidence matrix of G, defined as

$$K_{ue} = \begin{cases} +1 & \text{if edge } e \text{ exits vertex } u \\ -1 & \text{if edge } e \text{ enters vertex } u \\ 0 & \text{otherwise} \end{cases}.$$

Then,

- define the *gradient* as follows: let $f : V \to \mathbb{R}$ be a function on vertices, viewed as a row vector indexed by V; then K maps $f \to fK$, a vector indexed by E, measures the change of f along edges of the graph; and if e is an edge from u to v, then $(fK)_e = f_u - f_v$.
- define the *divergence* as follows: let $g : E \to \mathbb{R}$ be a function on edges, viewed as a column vector indexed by E; then K maps $g \to Kg$, a vector indexed by V; if we think of g as describing flow, then its divergence at vertex is the net outbound flow: $(Kg)_v = \sum_e \text{exits } {}_v\, g_e - \sum_e \text{enters } {}_v\, g_v$
- define the *Laplacian* as follows: it should map f to $KK^T f$, where $f : V \to \mathbb{R}$. So, $L = L_G = KK^T$ is the discrete Laplacian.

Note that it is easy to show that

$$L_{uv} = \begin{cases} -1 & \text{if } (u, v) \in E \\ \deg(u) & \text{if } u = v \\ 0 & \text{otherwise} \end{cases},$$

which is in agreement with the previous definition. Note also that

$$fLf^T = fKK^T f = \|fK\|_2^2 = \sum_{(u,v)\in E} (f_u - f_v)^2,$$

which we will later interpret as a smoothness condition for functions on the vertices of the graph.

6.2 Characterizing graph connectivity

Here, we provide a characterization in terms of eigenvalues of the Laplacian of whether or not a graph is connected. Cheeger's Inequality may be viewed as a "soft" version of this result.

6.2.1 A Perron-Frobenius style result for the Laplacian

What does the Laplacian tell us about the graph? A lot of things. Here is a start. This is a Perron-Frobenius style result for the Laplacian.

Theorem 1. *Let G be a d-regular undirected graph, let $L = I - \frac{1}{d}A$ be the normalized Laplacian; and let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the real eigenvalues, including multiplicity. Then:*

1. *$\lambda_1 = 0$, and the associated eigenvector $x_1 = \frac{\vec{1}}{\sqrt{n}} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$.*

2. $\lambda_2 \leq 2$.

3. $\lambda_k = 0$ *iff* G *has at least* k *connected components. (In particular,* $\lambda_2 > 0$ *iff* G *is connected.)*

4. $\lambda_n = 2$ *iff at least one connected component is bipartite.*

Proof: Note that if $x \in \mathbb{R}^n$, then $x^T L x = \frac{1}{d} \sum_{(u,v)\in E} (x_u - x_v)^2$ and also

$$\lambda_1 = \min_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^T L x}{x^T x} \geq 0.$$

Take $\vec{1} = (1, \ldots, 1)$, in which case $\vec{1}^T L \vec{1} = 0$, and so 0 is the smallest eigenvalue, and $\vec{1}$ is one of the eigenvectors in the eigenspace of this eigenvalue. This proves part 1.

We also have the following formulation of λ_k by Courant-Fischer:

$$\lambda_k = \min_{\substack{S \subseteq \mathbb{R}^n \\ dim(S)=k}} \max_{x \in S \setminus \{\vec{0}\}} \frac{x^T A x}{x^T x} \frac{\sum_{(u,v)\in E} (x_u - x_v)^2}{d \sum_u x_u^2}$$

So, if $\lambda_k = 0$, then $\exists$ a k-dimensional subspace S such that $\forall x \in S$, we have $\sum_{(u,v)\in E} (x_u - x_v)^2 = 0$. But this means that $\forall x \in S$, we have $x_u = x_v \quad \forall$ edges $(u, v) \in E$ with positive weight, and so $x_u = x_v$, for any u, v in the same connected component. This means that $x \in S$ is constant within each connected component of G. So, $k = \dim(S) \leq \Xi$, where Ξ is the number of connected components.

Conversely, if G has $\geq k$ connected components, then we can let S be the space of vectors that are constant on each component; and this S has dimension $\geq k$. Furthermore, $\forall x \in S$, we have that $\sum_{(u,v)\in E} (x_u - x_v)^2 = 0$. Thus $\max_{x \in S_k \setminus \{\vec{0}\}} \frac{x^T A x}{x^T x} = 0$ for any dimension k subspace S_k of the S we choose. Then it is clear from Courant-Fischer $\lambda_k = 0$ as any S_k provides an upperbound.

Finally, to study $\lambda_n = 2$, note that

$$\lambda_n = \max_{x \in \mathbb{R}^n \setminus \{\vec{0}\}} \frac{x^T L x}{x^T x}$$

This follows by using the variational characterization of the eigenvalues of $-L$ and noting that $-\lambda_n$ is the smallest eigenvalue of $-L$. Then, observe that $\forall x \in \mathbb{R}^n$, we have that

$$2 - \frac{x^T L x}{x^T x} = \frac{\sum_{(u,v)\in E} (x_u + x_v)^2}{d \sum_u x_u^2} \geq 0,$$

from which it follows that $\lambda_n \leq 2$ (also $\lambda_k \leq 2$ for all $k = 2, \ldots, n$).

In addition, if $\lambda_n = 2$, then $\exists x \neq 0$ s.t. $\sum_{(u,v)\in E} (x_u + x_v)^2 = 0$. This means that $x_u = -x_v$, for all edges $(u, v) \in E$.

Let v be a vertex s.t. $x_v = a \neq 0$. Define sets

$$\begin{aligned} A &= \{v : x_v = a\} \\ B &= \{v : x_v = -a\} \\ R &= \{v : x_v \neq \pm a\}. \end{aligned}$$

Then, the set $A \cup B$ is disconnected from the rest of the graph, since otherwise an edge with an endpoint in $A \cup B$ and the other endpoint in R would give a positive contribution to $\sum_{ij} A_{ij} (x_i + x_j)^2$. Also, every edge incident on A has other endpoint in B, and vice versa. So $A \cup B$ is a bipartite connected component (or a collection of connected components) of G, with bipartition A, B.

$\diamond$

(Here is an aside. That proof was from Trevisan; Spielman has a somewhat easier proof, but it is only for two components. I need to decide how much I want to emphasize the possibility of using k eigenvectors for soft partitioning—I'm leaning toward it, since several students asked about it—and if I do I should probably go with the version of here that mentions k components.) As an FYI, here is Spielman's proof of $\lambda_2 = 0$ iff G is disconnected; or, equivalently, that

$$\lambda_2 > 0 \Leftrightarrow G \text{ is connected.}$$

Start with proving the first direction: if G is disconnected, then $\lambda_2 = 0$. If G is disconnected, then G is the union of (at least) 2 graphs, call then G_1 and G_2. Then, we can renumber the vertces so that we can write the Laplacian of G as

$$L_G = \begin{pmatrix} L_{G_1} & 0 \\ 0 & L_{G_2} \end{pmatrix}$$

So, L_G has at least 2 orthogonal eigenvectors with eigenvalue 0, i.e., $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, where the two vectors are given with the same renumbering as in the Laplacians. Conversely, if G is connected and x is an eigenvector such that $L_G x = 0x$, then, $L_G x = 0$, and $x^T L_G x = \sum_{(ij) \in E} (x_i - x_j)^2 = 0$. So, for all (u, v) connected by an edge, we have that $x_u = x_u$. Apply this iteratively, from which it follows that x is a constant vector, i.e., $x_u = x_v$, forall u, v. So, the eigenspace of eigenvalue 0 has dimension 1. This is the end of the aside.)

6.2.2 Relationship with previous Perron-Frobenius results

Theorem 1 is an important result, and it has several important extensions and variations. In particular, the "$\lambda_2 > 0$ iff G is connected" result is a "hard" connectivity statement. We will be interested in how this result can be extended to a "soft" connectivity, e.g., "λ_2 is far from 0 iff the graph is well-connected," and the associated Cheeger Inequality. That will come soon enough. First, however, we will describe how this result relates to the previous things we discussed in the last several weeks, e.g., to the Perron-Frobenius result which was formulated in terms of non-negative matrices.

To do so, here is a similar result, formulated slightly differently.

Lemma 1. *Let A_G be the Adjacency Matrix of a d regular graph, and recall that it has n real eigenvalues $\alpha_1 \geq \cdots \geq \alpha_n$ and n associated orthogonal eigenvectors v_i s.t. $Av_i = \lambda_i v_i$. Then,*

- $\alpha_1 = d$, *with* $v_1 = \frac{1}{\sqrt{n}} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$.
- $\alpha_n \geq -d$.
- *The graph is connected iff* $\alpha_1 > \alpha_2$.

- *The graph is bipartite iff $\alpha_1 = -\alpha_n$, i.e., if $\alpha_n = -d$.*

Lemma 1 has two changes, relative to Theorem 1.

- The first is that it is a statement about the Adjacency Matrix, rather than the Laplacian.
- The second is that it is stated in terms of a "scale," i.e., the eigenvalues depend on d.

When we are dealing with degree-regular graphs, then A is trivially related to $L = D - A$ (we will see this below) and also trivially related to $L = I - \frac{1}{d}A$ (since this just rescales the previous L by $1/d$). We could have removed the scale from Lemma 1 by multiplying the Adjacency Matrix by $1/d$ (in which case, e.g., the eigenvalues would be in $[-1, 1]$, rather than $[-d, d]$), but it is more common to remove the scale from the Laplacian. Indeed, if we had worked with $L = D - A$, then we would have had the scale there too; we will see that below.

(When we are dealing with degree-heterogeneous graphs, the situation is more complicated. The reason is basically since the eigenvectors of the Adjacency matrix and unnormalized Laplacian don't have to be related to the diagonal degree matrix D, which defined the weighted norm which relates the normalized and unnormalized Laplacian. In the degree-heterogeneous case, working with the normalized Laplacian will be more natural due to connections with random walks. That can be interpreted as working with an unnormalized Laplacian, with an appropriate degree-weighted norm, but then the trivial connection with the eigen-information of the Adjacency matrix is lost. We will revisit this below too.)

In the above, $A \in \mathbb{R}^{n \times n}$ is the Adjacency Matrix of an undirected graph $G = (V, E)$. This will provide the most direct connection with the Perro-Frobenius results we talked about last week. Here are a few questions about the Adjacency Matrix.

- Question: Is it symmetric? Answer: Yes, so there are real eigenvalues and a full set of orthonormal eigenvectors.
- Question: Is it positive? Answer: No, unless it is a complete graph. In the weighted case, it could be positive, if there were all the edges but they had different weights; but in general it is not positive, since some edges might be missing.
- Question: Is it nonnegative? Answer: Yes.
- Question: Is it irreducible? Answer: If no, i.e., if it is reducible, then

  $$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

  must also have $A_{12} = 0$ by symmetry, meaning that the graph is disconnected, in which case we should think of it as two graphs. So, if the graph is connected then it is irreducible.
- Question: Is is aperiodic? Answer: If no, then since it must be symmetric, and so it must look like

  $$A = \begin{pmatrix} 0 & A_{12} \\ A_{12}^T & 0 \end{pmatrix},$$

  meaning that it is period equal to 2, and so the "second" large eigenvalue, i.e., the one on the complex circle equal to a root of unity, is real and equal to -1.

How do we know that the trivial eigenvector is uniform? Well, we know that there is only one all-positive eigenvector. Let's try the all-ones vector $\vec{1}$. In this case, we get

$$A\vec{1} = d\vec{1},$$

which means that $\alpha_1 = d$ and $v_1 = \frac{\vec{1}}{\sqrt{n}} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$. So, the graph is connected if $\alpha_1 > \alpha_2$, and the graph is bipartite if $\alpha_1 = -\alpha_n$.

For the Laplacian $L = D - A$, there exists a close relationship between the spectrum of A and L. (Recall, we are still considering the d-regular case.) To see this, let $d = \alpha_1 \geq \ldots \alpha_n$ be the eigenvalues of A with associated orthonormal eigenvectors $v_1, \ldots, v_n$. (We know they are orthonormal, since A is symmetric.) In addition, let $0 \leq \lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of L. (We know they are all real and in fact all positive from the above alternative definition.) Then,

$$\alpha_i = d - \lambda_i$$

and

$$A_G v_i = (dI - L_G)\, v_i = (d - \lambda_i)\, v_i.$$

So, L "inherits" eigen-stuff from A. So, even though L isn't positive or non-negative, we get Perron-Frobenius style results for it, in addition to the results we get for it since it is a symmetric matrix. In addition, if $L \to D^{-1/2} L D^{-1/2}$, then the eigenvalues of L become in $[0, 2]$, and so on. This can be viewed as changing variables $y \leftarrow D^{-1/2} x$, and then defining Laplacian (above) and the Rayleigh quotient in the degree-weighted dot product. (So, many of the results we will discuss today and next time go through to degree-heterogeneous graphs, for this reason. But some of the results, in particular the result having to do with expanders being least like low-dimensional Euclidean space, do not.)

6.3 Statement of the basic Cheeger Inequality

We know the λ_2 captures a "hard" notion of connectivity, since the above result in Theorem 1 states that $\lambda_2 = 0 \Leftrightarrow G$ is disconnected. Can we get a "soft" version of this?

To do so, let's go back to d-regular graphs, and recall the definition.

Definition 1. *Let $G = (V, E)$ be a d-regular graph, and let $\left(S, \bar{S}\right)$ be a cut, i.e., a partition of the vertex set. Then,*

- *the* sparsity of S *is:* $\sigma(S) = \frac{E\left(S,\bar{S}\right)}{\frac{d}{|V|}|S|\cdot|\bar{S}|}$
- *the* edge expansion of S *is:* $\phi(S) = \frac{E\left(S,\bar{S}\right)}{d|S|}$

This definition holds for sets of nodes $S \subset V$, and we can extend them to hold for the graph G.

Definition 2. *Let $G = (V, E)$ be a d-regular graph. Then,*

- *the* sparsity of G *is:* $\sigma(G) = \min_{S \subset V : S \neq 0, S \neq V} \sigma(S)$.
- *the* edge expansion of G *is:* $\phi(G) = \min_{S \subset V : |S| \leq \frac{|V|}{2}} \phi(S)$.

For d-regular graphs, the graph partitioning problem is to find the sparsity or edge expansion of G. Note that this means finding a number, i.e., the value of the objective function at the optimum, but people often want to find the corresponding set of nodes, and algorithms can do that, but the "quality of approximation" is that number.

Fact. For all d regular graphs G, and for all $S \subset V$ s.t. $|S| \leq \frac{|V|}{2}$, we have that

$$\frac{1}{2}\sigma(S) \leq \phi(S) \leq \sigma(S).$$

Thus, since $\sigma(S) = \sigma\left(\bar{S}\right)$, we have that

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

BTW, this is what we mean when we say that these two objectives are "equivalent" or "almost equivalent," since that factor of 2 "doesn't matter." By this we mean:

- If one is interested in theory, then this factor of 2 is well below the guidance that theory can provide. That is, this objective is intractable to compute exactly, and the only approximation algorithms give quadratic or logarithmic (or square root of log) approximations. If they could provide $1 \pm \epsilon$ approximations, then this would matter, but they can't and they are much coarser than this factor of 2.
- If one is interested in practice, then we can often do much better than this factor-of-2 improvement with various local improvement heuristics.
- In many cases, people actually write one and optimize the other.
- Typically in theory one is most interested in the number, i.e., the value of the objective, and so we are ok by the above comment. On the other hand, typically in practice, one is interested in using that vector to do things, e.g., make statements that the two clusters are close; but that requires stronger assumptions to say nontrivial about the actual cluster.

Given all that, here is the basic statement of Cheeger's inequality.

Theorem 2 (Cheeger's Inequality)**.** *Recall that*

$$\lambda_2 = \min_{x:x\perp\vec{1}} \max_{x:x\perp\vec{1}} \frac{x^T L x}{x^T x}$$

where $L = I - \frac{1}{d}A$*. Then,*

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

6.4 Comments on the basic Cheeger Inequality

Here are some notes about the basic Cheeter Inequality.

- This result "sandwiches" λ_2 and ϕ close to each other on both sides. Clearly, from this result it immediatly follows that

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

- Later, we will see that $\phi(G)$ is large, i.e., is bounded away from 0, if the graph is well-connected. In addition, other related things, e.g., that random walks will mix rapidly, will also hold. So, this result says that λ_2 is large if the graph is well-connected and small if the graph is not well-connected. So, it is a soft version of the hard connectivity statement that we had before.

- The inequality $\frac{\lambda_2}{2} \leq \phi(G)$ is sometimes known as the "easy direction" of Cheeger's Inequality. The reason is that the proof is more straightforward and boils down to showing one of two related things: that you can present a test vector, which is roughly the indicator vector for a set of interest, and since λ_2 is a min of a Rayleigh quotient, then it is lower than the Rayleigh quotient of the test vector; or that the Rayleigh quotient is a relaxation of the sparsest cut problem, i.e., it is minimizing the same objective over a larger set.

- The inequality $\phi(G) \leq \sqrt{2\lambda_2}$ is sometimes known as the "hard direction" of Cheeger's Inequality. The reason is that the proof is constructive and is basically a vanilla spectral partitioning algorithm. Again, there are two related proofs for the "hard" direction of Cheeger. One way uses a notion of inequalities over graphs; the other way can be formulated as a randomized rounding argument.

- Before dismissing the easy direction, note that it gives a polynomial-time certificate that a graph is expander-like, *i.e.*, that $\forall$ cuts (and there are 2^n of them to check) at least a certain number of edges cross that cut. (So the fact that is holds is actually pretty strong—we have a polynomial-time computable certificate of having no sparse cuts, which you can imagine is of interest since the naive way to check is to check everything.)

Before proceeding, a question came up in the class about whether the upper or lower bound was more interesting or useful in applications. It really depend on on what you want.

- For example, if you are in a networking application where you are routing bits and you are interested in making sure that your network is well-connected, then you are most interested in the easy direction, since that gives you a quick-to-compute certificate that the graph is well-connected and that your bits won't get stuck in a bottleneck.

- Alternatively, if you want to run a divide and conquer algorithm or you want to do some sort of statistical inference, both of which might require showing that you have clusters in your graph that are well-separated from the rest of the data, then you might be more interested in the hard direction which provides an upper bound on the intractable-to-compute expansion and so is a certificate that there are some well-separated clusters.