A Theorist's Toolkit **(CMU 18-859T, Fall 2013)**

Lecture 6: Spectral Graph Theory I

September 25, 2013

Lecturer: Ryan O'Donnell *Scribe: Jennifer Iglesias*

1 Graph Theory

For this course we will be working on an undirected graph $G = (V, E)$ where V is the vertex set and E is the edge set. We know the following about our graph:

- The graph is finite.
- There are no isolated vertices. An *isolated vertex* is a vertex of degree zero.
- Parallel edges and self-loops are allowed. *Parallel edges* are edges with the same endpoints. A *self-loop* is an edge where both endpoints are the same vertex; a self-loop contributes 1 to the degree of the vertex it is adjacent to and can be thought of as half an edge
- The graphs we work with need not be connected.
- We will assume the graph is unweighted as we can deal with any rational weights by using parallel edges.

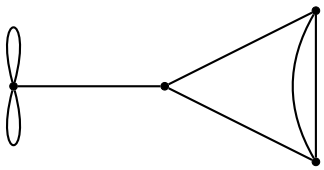


Figure 1: An example of a 3-regular graph with parallel edges and self-loops

If you are unfamiliar with graph theory, you can simplify the results by assuming the graphs we are working with are regular. A *regular* graph has the same degree at every vertex. If you have more familiarity with this subject, note that you can generalize most of the results to directed graphs and reversible Markov chains.

Often times we will consider labeling the vertices with a function $f : V \to \mathbb{R}$. This function often represents something for the vertices e.g. temperatures, voltages, coordinates for an embedding, 0/1 indicator for a subset $S \subseteq V$. We can think of f as a vector where

the ith coordinate corresponds to the value of $f(v_i)$.

$$f : V \to \mathbb{R} \equiv \begin{bmatrix} f(v_1) \\ f(v_2) \\ \vdots \\ f(v_n) \end{bmatrix}$$

We can add functions of the vertice and scale as expected:

$$(f + g)(x) := f(x) + g(x)$$
$$(c \cdot f)(x) := c \cdot (f(x))$$

This gives us linearity of the functions. In fact, the functions on V form a vectorspace and act the same as $\mathbb{R}^n$.

2 Local Variance

This is the main idea of spectral graph theory. The "Local Variance" is defined to be:

$$\mathcal{E}(f) := \frac{1}{2} \mathop{\mathbf{E}}_{\boldsymbol{u} \sim \boldsymbol{v}} \left[(f(\boldsymbol{u}) - f(\boldsymbol{v}))^2 \right]$$

The formal name for $\mathcal{E}$ is the Dirichlet form. In this class, it will often be referred to as "Local Variance", and sometimes as "Analytic boundary size" or "quadratic form of the Laplacian".

We will use $\boldsymbol{u} \sim \boldsymbol{v}$ to be a probability distribution over the edges. We can think of it as first picking an edge uniformly at random, and then pick a direction for it with each direction being equally likely. We can also think of it as every undirected edge now becomes 2 antiparallel directed edges (here a self-loop just becomes 1 directed edge). Now we can choose a directed edge uniformly at random.

Some extremely simple facts about the Local Variance:

- $\mathcal{E}(f) \geq 0$
- $\mathcal{E}[c \cdot f] = c^2 \mathcal{E}[f]$
- $\mathcal{E}[f + c] = \mathcal{E}[f]$

These are all properties which variance also has. The intuition to gain from "Local Variance" is that there is a "small" value for $\mathcal{E}[f]$ if and only if the function is f is "smooth" on the edges. Now we will go through a small example of computing $\mathcal{E}$.

Example 2.1. *Let $S \subseteq V$ and $f = 1_S$ (the indicator function for S):*

$$f(v) = \begin{cases} 1, & \text{if } v \in S \\ 0, & \text{else} \end{cases}$$

Now computing the local variance we get:

$$\begin{aligned} \mathcal{E} &= \frac{1}{2} \mathop{\mathbf{E}}_{\boldsymbol{u} \sim \boldsymbol{v}} \left[(1_S(\boldsymbol{u}) - 1_S(\boldsymbol{v}))^2 \right] \\ &= \frac{1}{2} \mathop{\mathbf{E}}_{\boldsymbol{u} \sim \boldsymbol{v}} \left[1((\boldsymbol{u}, \boldsymbol{v}) \text{ "crosses" } S) \right] \\ &= \frac{1}{2} \{ \textit{fraction of edges on } \delta S \} \\ &= \mathop{\mathbf{Pr}}_{\boldsymbol{u} \sim \boldsymbol{v}} [\boldsymbol{u} \to \boldsymbol{v} \textit{ is "stepping" out of } S] \end{aligned}$$

3 Random Vertices and Random Walks

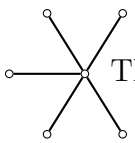
To choose a random vertex we will use the following procedure:

1. Choose $\boldsymbol{u} \sim \boldsymbol{v}$ a random edge
2. Output $\boldsymbol{u}$ (identically we could put Output $\boldsymbol{v}$ here)

This is in fact the right distribution on vertices to have, and from here on we will denote this distribution as π. This gives weights or importance to vertices based on the number of edges adjacent to them.

This distribution does not make the probability of choosing any vertex equally likely. That only happens in the case when the graph is regular.

Now consider the graph below, We are much more likely to choose the vertex in the



center than any other vertex. This seems to make sense as it is involved in all the edges in the graph and therefore should be chosen with higher probability.

Fact 3.1. *$\pi[u]$ is proportional to* $\deg(u)$.

Proof. The probability of picking u is picking an edge adjacent to u and then picking u as the endpoint. This gives:

$$\pi[u] = \frac{\deg(u)}{2|E|}$$

□

Fact 3.2. *The process of picking $\boldsymbol{u}$ from π and then picking $\boldsymbol{v}$ as a uniformly random neighbor of $\boldsymbol{u}$ is the same as drawing an edge uniformly at random $\boldsymbol{u} \sim \boldsymbol{v}$.*

Proof. The probability of picking an edge uv by the first process is $\frac{\deg u}{2|E|}\frac{1}{\deg(u)} = \frac{1}{2|E|}$. This is the same as drawing an edge uniformly at random. □

Corollary 3.3. *Let $t \in \mathbb{N}$. Now pick $\boldsymbol{u} \sim \pi$. Now do a random walk starting at $\boldsymbol{u}$ taking t steps. Then the distribution of $\boldsymbol{v}$, the endpoint of the walk, is π.*

The previous fact showed for a 1 step random walk, and by repeated iterations, we get the same result for any fixed number of steps.

Definition 3.4. π is the *stationary distribution* on vertices. Also called the *limiting distribution* or the *invariant distribution* where invariant refers to the distribution being invariant under random walks.

Now what if $u_0 \in V$ is not random and we start a random walk from u_0? As t, the number of steps, goes to infinity does the distribution of v, the current node we are at in the walk, go to π?

- No, if G is disconnected (clearly there are vertices you can never get to).
- No, if G is bipartite, the parity of t restricts which part of the graph you are on
- Yes, otherwise.

How long does it take to get close to π by doing a random walk if G is connected and not bipartite?

Consider we have something as drawn below:

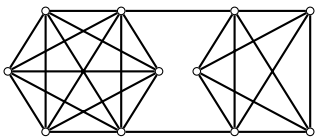


Figure 2: Two cliques with very few edges connecting them

In the figure above, we see that the graph is connected and not bipartite. We expect it to take a while to get close to π though as the walk will probably not cross between the two cliques quickly. In this example, $\mathcal{E}[1_S]$ where S is one of the cliques is small. So we expect that a "fast" convergence to π from a random walk corresponds to $\mathcal{E}[f]$ being "large" for all f. Here "large" is a relative term as if we scale f by 10 then it scales $\mathcal{E}[f]$ by 100.

Definition 3.5. Let $f : V \to \mathbb{R}$ and $\boldsymbol{u} \sim \pi$. Then $f(\boldsymbol{u})$ is a random variable. We will denote the *mean* as $\mathbf{E}[f] := \mathbf{E}_{\boldsymbol{u}\sim\pi}[f]$. We will denote the *variance* as $\mathbf{Var}[f] = \mathbf{E}_{\boldsymbol{u}\sim\boldsymbol{v}}[(f(\boldsymbol{u}) - \mu)^2]$. We can rewrite this as: $\mathbf{E}_{\boldsymbol{u}\sim\pi}[f^2(u)] - \mathbf{E}[f]^2$ and as $\frac{1}{2}\mathbf{E}_{\boldsymbol{u}\sim\pi,\boldsymbol{v}\sim\pi}[(f(\boldsymbol{u}) - f(\boldsymbol{v}))^2]$. This leads us to refer to this as the "global variance" as it looks at all pairs $\boldsymbol{u}, \boldsymbol{v}$ and not just edges like $\mathcal{E}$ does.

Example 3.6. *If we let $f = 1_S$ then we get* $\mathbf{E}[f] = \mathbf{Pr}_{\boldsymbol{u}\sim\pi}[u \in S] = vol(S)$.

Spectral graph theory is about comparing the local variance and the global variance. For example, if for all f, $\mathcal{E}[f]$ is always "large" compared to $\mathbf{Var}[f]$ this corresponds to G being an expander graph. The intuition behind this is that you mix quickly. If f is an indicator function, then $\mathcal{E}[f]$ is the edges crossing the cut and $\mathbf{Var}[f]$ is the volume. That $\mathbf{Var}[f]$ corresponds to volume is seen by:

$$\begin{aligned}\mathbf{E}[1_S^2(\boldsymbol{u})] &= \mathbf{E}[1_S(\boldsymbol{u})] = vol(S)\\ \mathbf{Var}[1_S] &= E[1_S^2(\boldsymbol{u})] - \mu^2\\ &= vol(S) - vol(S)^2\\ &= vol(S)(1 - vol(S))\\ &= vol(S)(vol(S^C))\end{aligned}$$

When S is small then $\epsilon(1-\epsilon) \approx \epsilon$. So we get that $\mathbf{E}[1_S^2] \approx \min(vol(S), vol(S^C))$, and without loss of generality we can assume that $vol(S) \leq \frac{1}{2}$. We know that the local variance is $\mathcal{E}[1_S] = \mathbf{Pr}_{\boldsymbol{u}\sim\boldsymbol{v}}[\boldsymbol{u} \in S, \boldsymbol{v} \notin S]$. Looking at the ratio we get that:

$$\frac{\mathcal{E}[1_S]}{\mathbf{E}[1_S]} = \frac{\mathbf{Pr}[\boldsymbol{u} \in S, \boldsymbol{v} \notin S]}{\mathbf{Pr}[\boldsymbol{u} \in S]} = \mathbf{Pr}[v \notin S | \boldsymbol{u} \in S]$$

This is the conductance of S and is denoted $\Phi(S)$.

4 Inner Products

Definition 4.1. Let $f, g : V \to \mathbb{R}$ we define $\langle f, g\rangle_\pi = \mathbf{E}_{\boldsymbol{u}\sim\pi}[f(u)g(u)]$. It has nothing to do with edges (except in deciding π). From here on, we will drop the π from the subscripts.

This measures how similar the functions are. It is close to what we would expect if we did the inner product of the corresponding vectors:

$$\left\langle \begin{bmatrix} f(v_1)\\ f(v_2)\\ \vdots\\ f(v_n)\end{bmatrix}, \begin{bmatrix} g(v_1)\\ g(v_2)\\ \vdots\\ g(v_n)\end{bmatrix} \right\rangle$$

This is not quite the same though as the terms do get scaled by π.

Note: In general Computer Science, $D = diag(\deg)$ and the notation will use lots of $D^{1/2}$. We will avoid this notation by using standard inner product notation.

The, $\langle \cdot, \cdot \rangle_\pi$ is in fact an inner product on the vector space defined by functions from V to $\mathbb{R}$.

- Associativity is clear: $\langle f, g\rangle = \langle g, f\rangle$.

- We also have linearity, which follows from linearity of expectation. $\langle a \cdot f + g, h\rangle = a\langle f, h\rangle + \langle g, h\rangle$.
- We also have non-negativity, $\langle f, f\rangle \geq 0$ with equality if and only if f is 0.

This also justifies the notation of $\langle f, f\rangle = \|f\|_2^2$.

What range does $\mathcal{E}[f]$ lie in? Clearly, $\mathcal{E}[f] \geq 0$ and equality holds if $f = 0$ and f is constant.

Proposition 4.2. *$\mathcal{E}[f] = 0$ if and only if f is constant on all connected components of G.*

Proof. All the terms in $\mathcal{E}[f]$ must be zero since all the terms are non-negative. So, every edge must have the same value on both ends. Therefore any path must have the same value along the path. In a connected component, every two points are connected by a path. So, every connected component must have f constant. Clearly, if f is constant on all connected components, then it is constant on every edge. Hence $\mathcal{E}[f] = 0$. □

Proposition 4.3. *The number of connected components is the same as the independent f's with $\mathcal{E}[f] = 0$.*

If $S_1, S_2, \ldots S_\ell$ are connected components then $1_{S_1}, 1_{S_2}, \ldots 1_{S_\ell}$ are linearly independent.

The number of linearly independent f which make $\mathcal{E}[f]$ small corresponds to clusters.

What is an upper bound of $\mathcal{E}[f]$ subject to $\mathbf{Var}[f] = 1$ and $\mathbf{E}[f^2] = \|f\|_2^2$.

These two constraints are in fact the same. We can first shift f so that $\mathbf{E}[f] = 0$. So we get $\mathbf{Var}[f] = \mathbf{E}[f^2] - \mathbf{E}[f]^2 = \mathbf{E}[f^2]$.

Now we want to minimize $\mathbf{E}[f^2]$ such that $\mathcal{E}[f] \geq 1$ ($\mathcal{E}$ is constant under shifts). We could also maximizes $\mathcal{E}[f]$ such that $\|f\|_2^2 = \mathbf{E}[f^2] = 1$.

Intuitively, we are trying to embed G into $\mathbb{R}$ so that edges are "long".

For what G will you be most successful? We will be successful when G is bipartite with $V = (V_1, V_2)$. Let $f = 1_{V_1} - 1_{V_2}$. Now we have:

$$f(v) = \begin{cases} +1, & \text{if } v \in V_1 \\ -1, & \text{if } v \in V_2 \end{cases}$$

Now we have $\mathbf{E}[f(u)^2] = 1$ and $\mathcal{E}[f] = \frac{1}{2}4 = 2$. This is the largest possible.

Proposition 4.4. *For all G and for all f, then $\mathcal{E}[f] \leq 2\|f\|_2^2$.*

Proof. The proof is just simply manipulation:

$$\begin{aligned} \frac{1}{2}\mathop{\mathbf{E}}_{\boldsymbol{u}\sim\boldsymbol{v}}[(f(\boldsymbol{u}) - f(\boldsymbol{v}))^2] &= \frac{1}{2}\mathbf{E}[f(\boldsymbol{u})^2] - \frac{1}{2}2\,\mathbf{E}[f(\boldsymbol{u})f(\boldsymbol{v})] + \frac{1}{2}E[f(\boldsymbol{v})^2] \\ &= \mathbf{E}[f^2] - \mathbf{E}[f(\boldsymbol{u})f(\boldsymbol{v})] \\ &\leq \mathbf{E}[f^2] + \sqrt{\mathbf{E}[f(\boldsymbol{u})^2]\,\mathbf{E}[f(\boldsymbol{v})^2]} \\ &= 2\,\mathbf{E}[f^2] \end{aligned}$$

□

This is the desired result. A good thing to note is that $\mathbf{E}[XY] \leq \sqrt{\mathbf{E}[X^2]\,\mathbf{E}[Y^2]}$, which follows from Cauchy-Schwartz.

An exercise to try at home: Show that $\mathcal{E}[f] = 2\,\mathbf{E}[f^2]$ is possible if and only if G is bipartite. If $\mathcal{E}[f]$ is close to $2\,\mathbf{E}[f^2]$ then G is close to bipartite.