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

Lecture 8: Spectral Graph Theory III

October 2, 2013

Lecturer: Ryan O'Donnell *Scribe: Jeremy Karp*

1 Recap

Last time we showed that every function $f : V \to \mathbb{R}$ is uniquely expressible as

$$f = \hat{f}(0)\phi_0 + \hat{f}(1)\phi_1 + \ldots + \hat{f}(n-1)\phi_{n-1}$$

In this representation, the ϕ_is form an orgthogonal basis of the eigenvectors of L. We can also compute the eigenvalues of L and order them as follows:

$$0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{n-1} \leq 2$$

The number of eigenvalues that are 0 is the number of connected components in the graph. Also recall that if we apply L to any function f:

$$Lf = \lambda_0\hat{f}(0)\phi_0 + \lambda_1\hat{f}(1)\phi_1 + \ldots + \lambda_{n-1}\hat{f}(n-1)\phi_{n-1}$$

It also follows that

$$\langle f,\ g\rangle = \sum_{i=0}^{n-1} \hat{f}(i)\hat{g}(i)$$

A corollary of this is

$$\xi(f) = \langle f,\ Lf\rangle = \sum_{i>0} \lambda_i \hat{f}(i)^2$$

We also have

$$||f||_2^2 = \sum_i \hat{f}(i)^2$$

and

$$\mathbf{E}[f] = \langle f,\ \mathbf{1}\rangle$$

From orthonormality, everything in the previous inner product drops out except the first term, so $\mathbf{Var}[f] = \mathbf{E}[f^2] - \mathbf{E}[f]^2 = \sum_{i>0} \hat{f}(i)^2$

2 Conductance and the second eigenvalue

Now, we can observe that

$$\min_{Var[f]=1}\{\xi(f)\} = \lambda_1$$

This comes from the observatrion that the variance is a sum of non-negative numbers that add up to 1, but $\xi(f)$ is the sum of these same numbers multiplied by the eigenvalues. Then, the smallest this result could be is the smallest of these eigenvalues, λ_1. Another way to think of this idea is in the form of the Poincare inequality:

$$\lambda_1 \mathbf{Var}[f] \leq \xi[f]$$

Note that λ_1 is referred to as the second eigenvalue, despite its index being 1.

We care about the relationship between $\mathbf{Var}[f]$ and $\xi[f]$ because it relates to conductance. Recall from last time the definition of conductance:

Definition 2.1. The conductance of $S \subseteq V$ is

$$\begin{aligned}\Phi(S) &= \mathbf{Pr}_{u\sim v}[v \notin S |\ u \in S] \\ &= \frac{\xi[\mathbf{1}_S]}{||\mathbf{1}_S||_2^2}\end{aligned}$$

We can relate these to the sparsest cut in the graph by defining the conductance of a graph:

Definition 2.2. Φ_G, the conductance of G is

$$\min_{S:\ 0<vol(S)\leq\frac{1}{2}}\{\Phi(S)\}$$

Note that a lower bound on the conductance is

$$\frac{1}{2}\min_{S:\ 0<vol(S)\leq\frac{1}{2}}\{\frac{\xi[\mathbf{1_S}]}{\mathbf{Var}[\mathbf{1}_S]}\}$$

because $\mathbf{Var}[\mathbf{1}_S] = vol(S)(1-vol(S)) \geq \frac{1}{2}vol(S) = \frac{1}{2}||\mathbf{1}_S||_2^2$.

As we saw in Section 1, we can similarly express the second eigenvalue as the following minimization problem:

$$\lambda_1 = \min_{f:\ u\to\mathbb{R},\ Var(f)\neq 0}\{\frac{\xi[f]}{\mathbf{Var}[f]}\}$$

This is a relaxation of the optimization problem that we solved to determine the conductance of G. Now, we can optimize over a larger range of functions. Interestingly, we can compute λ_1 efficiently, but it is NP-hard to determine the conductance of a graph. In the next section, we relate these two values to each other using Cheeger's inequality.

3 Cheeger's inequality

The following theorem is called Cheeger's inequality:

Theorem 3.1 (Alon, 1986)**.** $\frac{1}{2}\lambda_1 \leq \Phi_G \leq 2\sqrt{\lambda_1}$

Proof. The first inequality is a consequence of our phrasing of λ_1 as a minimization problem. Specifically, we have

$$\begin{aligned}\frac{1}{2}\lambda_1 =& \frac{1}{2}\min_{f:\, u\to\mathbb{R},\, Var(f)\neq 0}\{\frac{\xi[f]}{\mathbf{Var}[f]}\}\\ \leq& \frac{1}{2}\min_{S:\, 0<vol(S)\leq\frac{1}{2}}\{\frac{\xi[\mathbf{1}_S]}{\mathbf{Var}[\mathbf{1}_S]}\}\\ \leq& \Phi_G\end{aligned}$$

The second inequality is proved by Theorem 3.2. □

Observe that a small λ_1 means there exists a function f where $\xi[f]$ is small compared to $\mathbf{Var}[f]$. Then, a natural question to ask is whether we can convert this function to a $\{0,1\}$-valued function, which would apply to Φ_G. This process is where we lose a square-root factor in the second inequality.

Theorem 3.2. *Let f be a non-constant function such that $\xi[f] \leq \lambda_1 \mathbf{Var}[f]$. Then, $\exists S \subseteq V$ such that $0 < vol(S) \leq \frac{1}{2}$ and $\Phi(S) \leq 2\sqrt{\lambda_1}$*

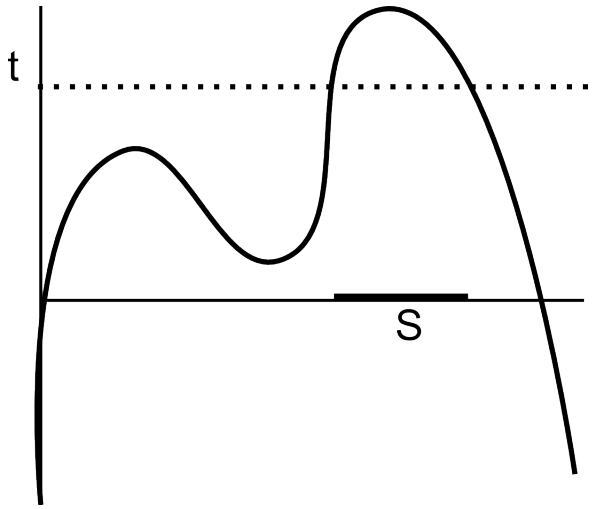

Figure 1: An illustration of how S is determined by function f and value t

In fact, there exists a set S with these proterties where $S = \{u : \ f(u) \geq t\}$ or $\{u : f(u) < t\}$ for some value t. This result allows us to approximately solve the sparsest cut problem. However, we need to do some more work before we are ready to prove Theorem 3.2.

Definition 3.3. $g: V \to \mathbb{R}$ is "convenient" if $g \geq 0$ and $vol(\{u: g(u) \neq 0\}) \leq \frac{1}{2}$

We claim without proof that we can assume f is convenient in our proof of Theorem 3.2, losing only a factor of 2 in the inequality $\xi[f] \leq \lambda_1 \mathbf{Var}[f]$.

A reasonable question to ask, regarding the cut-off value t we discussed earlier is, for a value t, what is the expectation of the set you obtain, S. This question is answered by the following Lemma:

Lemma 3.4. *[Coarea formula] For a non-negative function $g: V \to \mathbb{R}^+$*

$$\int_0^\infty (\mathbf{Pr}_{u\sim v}[(u,v) \text{ crosses } g^{>t} \text{ boundary}]dt = \mathbf{E}_{u\sim v}[|g(u) - g(v)|]$$

where $g^{>t} = \{u: g(u) > t\}$.

Proof.

$$\begin{aligned}\int_0^\infty (\mathbf{Pr}_{u\sim v}[(u,v) \text{ crosses } g^{>t} \text{ boundary})dt &= \int_0^\infty (\mathbf{Pr}_{u\sim v}[t \text{ is between } g(u),\ g(v)])dt \\ &= \mathbf{E}_{u\sim v}[\int_0^\infty \mathbf{1}^T\{t \text{ is between } g(u),\ g(v)\}dt] \\ &= \mathbf{E}_{u\sim v}[|g(u) - g(v)|]\end{aligned}$$

□

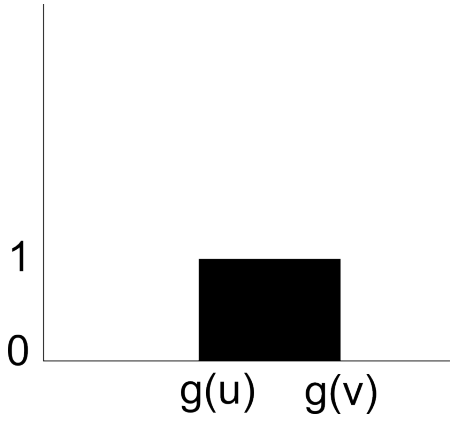

Figure 2: We see in this illustration that the proof of the Coarea formula relies on the non-negativity of g

We also have the following:

Corollary 3.5. *Let g be convenient. Then, $\mathbf{E}[|g(u) - g(v)|] \geq 2\Phi_G \mathbf{E}_{u\sim v}[g(u)]$.*

In this proof, note that $\xi[1_S] = \mathbf{Pr}_{u\sim v}[u \in S,\ v \notin S]$.

Proof.

$$\begin{aligned}
\mathbf{E}[|g(u) - g(v)|] &= \int_0^\infty 2\xi[\mathbf{1}_{g^{>t}}]dt \\
&= 2\int_0^\infty \Phi(g^{>t})vol(g^{>t})dt \\
&\geq 2\Phi_G \int_0^\infty \mathbf{Pr}_{u\sim\pi}[u \in g^{>t}]dt \\
&= 2\Phi_G \int_0^\infty \mathbf{Pr}_{u\sim\pi}[g(u) > t]dt \\
&= 2\Phi_G \mathbf{E}_{u\sim v}[g(u)]
\end{aligned}$$

□

Note that the first inequality in the previous proof relies on the fact that g is convenient. We can now use this corollary to prove Theorem 3.2.

Proof. [Proof of Theorem 3.2]

Assuming g is convenient and not always 0, we see that

$$\Phi_G \leq \frac{\mathbf{E}_{u\sim v}[|g(u) - g(v)|]}{2\mathbf{E}_{u\sim v}[g(u)]}$$

We have a convenient function f such that $\frac{\xi[f]}{\mathbf{Var}[f]} \leq 2\lambda_1$. Then, let $g = f^2$. g is still non-negative and its support is the same as the support of f, so g is also convenient.

$$\begin{aligned}
\mathbf{E}_{u\sim v}[|f(u)^2 - f(v)^2|] &= \mathbf{E}_{u\sim v}[|f(u) - f(v)| \cdot |f(u) + f(v)|] \\
&\leq \sqrt{\mathbf{E}_{u\sim v}[(f(u) - f(v))^2]}\sqrt{\mathbf{E}_{u\sim v}[(f(u) + f(v))^2]} \\
&\leq \sqrt{2\xi[f]}\sqrt{\mathbf{E}_{u\sim v}[2f(u)^2 + 2f(v)^2]} \\
&\leq \sqrt{4\lambda_1 \mathbf{E}_{u\sim v}[f(u)^2]}2\sqrt{\mathbf{E}_{u\sim v}[f(u)^2]} \\
&= 4\sqrt{\lambda_1}\mathbf{E}_{u\sim v}[f(u)^2]
\end{aligned}$$

We then combine this calculation with the previous one to prove the theorem:

$$\begin{aligned}
\Phi_G &\leq \frac{\mathbf{E}_{u\sim v}[|g(u) - g(v)|]}{2\mathbf{E}_{u\sim v}[g(u)]} \\
&= \frac{\mathbf{E}_{u\sim v}[|f(u)^2 - f(v)^2|]}{2\mathbf{E}_{u\sim v}[f(u)^2]} \\
&\leq \frac{4\sqrt{\lambda_1}\mathbf{E}_{u\sim v}[f(u)^2]}{2\mathbf{E}_{u\sim v}[f(u)^2]} \\
&= 2\sqrt{\lambda_1}
\end{aligned}$$

□

Before moving on, let's consider more specifically how we can use Theorem 3.1 to compute an approximately-minimal conductance set S in G efficiently.

Given G, it is possible to compute a value $\widetilde{\lambda}_1$ and a function $f : V \to \mathbb{R}$ such that:

a)

$$|\widetilde{\lambda}_1 - \lambda_1| \leq \epsilon$$

,

b)

$$\mathcal{E}[f] \leq \widetilde{\lambda}_1 \mathbf{Var}[f]$$

in time either:

i)

$$\widetilde{O}(n) \cdot \frac{1}{\epsilon}$$

using "The Power Method", or

ii)

$$poly(n) \cdot \log \frac{1}{\epsilon}$$

using "more traditional linear algebra methods".

Note that in some sense, (i) above is not a polynomial time algorithm, since the input length of ϵ is roughly $\log \frac{1}{\epsilon}$. That said, since the "Cheeger algorithm" only gives an approximation to the sparsest cut anyway, maybe you don't mind if ϵ is rather big. (ii) is actually a polynomial time algorithm, but the polynomial on the n isn't very good (perhaps n^4) so in practice it might not be a great algorithm to run.

4 Convergence to the stationary distribution

So far, we have proved Cheeger's inequality, which provides us with a method to find the sparsest cut in a graph up to a square root factor. Cheeger's inequality also implies that λ_1 is "large" if and only if there does not exists a set with small conductance. This observation relates to the speed of convergence to π, the stationary distribution. By this we mean, if we are to start at any vertex in the graph and take a random walk, how long does it take for the probability distribution on our location to converge to π? We will see that if λ_1 is large, then convergence occurs quickly. These graphs where λ_1 is large are called expander graphs. If we want to determine the speed of convergence, we might ask what it means for two distributions to be "close". We first observe that

$$(I - L)\phi_i = \phi_i - \lambda_i \phi_i = (1 - \lambda_i)\phi_i$$

Then, let $\kappa_i = (1 - \lambda_i)$ which gives us the ordering $1 = \kappa_0 \geq \kappa_1 \geq \ldots \geq \kappa_{n-1} - 1$. Then, if λ_1 is large, κ_1 has a large gap from 1.

4.1 Lazy graphs

Unfortunately, when a graph is bipartite we never converge to the stationary distribution because the number of steps we have taken determines which side of the bipartition we are necessarily on. Fortunately, we can fix this issue by introducing "lazy" graphs.

Definition 4.1. The lazy version of graph $G = (V, E)$ is $G_L = (V, E \cup S)$ where S is a set of self-loops such that each degree d node has d self-loops in set S.

Lazy versions of graphs have the property that at each step of the random walk, with probability $\frac{1}{2}$ we stay at the same vertex and with probability $\frac{1}{2}$ we take a random step according to the same probability distribution as in the original graph. This fixes the issue we had with bipartite graphs, while keeping many of the important properties of the original graph.

Now, in the lazy version of our graph, we have the operator $(\frac{1}{2}I + \frac{1}{2}\kappa)$ and if we apply this operator to ϕ_i we get

$$\begin{aligned}(\frac{1}{2}I + \frac{1}{2}\kappa)\phi_i &= \frac{1}{2}\phi_i + \frac{1}{2}\kappa_i\phi_i \\ &= (\frac{1}{2} + \frac{1}{2}\kappa_i)\phi_i\end{aligned}$$

In the lazy version of the graph, $\phi_0, \ldots, \phi_{n-1}$ are still eigenvectors, but each eigenvalue κ_i now becomes $\frac{1}{2} + \frac{1}{2}\kappa_i$ which is between 0 and 1. Similarly, the eigenvalues of the new Laplacian are also between 0 and 1.

4.2 Closeness of two distributions

We now return to the question of how to measure closeness of distributions.

Definition 4.2. A probability density $\psi : V \to \mathbb{R}$ is such that $\psi \geq 0$ and $\mathbf{E}_{u\sim\pi}[\psi(u)] = 1$

This probability density is sort of like the relative density with respect to π. More formally, a probability density function corresponds to a probability distribution $\mathcal{D}_\psi$ where $\mathbf{Pr}_{u\sim\mathcal{D}_\psi} := \psi(u)\pi(u)$. For example, if $\psi \equiv 1$ then $\mathcal{D}_\psi = \pi$.

Fact 4.3.

$$\begin{aligned}\langle \psi,\ f\rangle &= \mathbf{E}_{u\sim\pi}[\psi(u)f(u)] \\ &= \sum_u \pi(u)\psi(u)f(u) \\ &= \sum_u \mathcal{D}_\psi(u)f(u) \\ &= \mathbf{E}_{u\sim\mathcal{D}_\psi}[f(u)]\end{aligned}$$

Additionaly, the density corresponding to the probability distribution placing all probability of vertex u is $\psi_u = \frac{1}{\pi(u)} 1_u$.

We can now ask is $\mathcal{D}_\psi$ "close" to π?

Definition 4.4. The χ^2-divergence of $\mathcal{D}_\psi$ from π is

$$\begin{aligned} d_{\chi^2}(\mathcal{D}_\psi, \pi) &= \mathbf{Var}_\pi[\psi] \\ &= \mathbf{E}_{u\sim\pi}[(\psi(u) - 1)^2] \\ &= ||\psi||_2^2 - 1 \\ &= \sum_i \hat{\psi}(i)^2 - \hat{\psi}(0)^2 \\ &= \sum_{i>0} \hat{\psi}(i)^2 \end{aligned}$$

This definition makes some intuitive sense because the distance ψ is from 1 at each point is an indicator of how similar the $\mathcal{D}_\psi$ is to π. Goin forward, we will use χ^2-divergence as our measure of closeness, even if it has some strange properties. One unusual feature, in particular, is that this measure of closeness is not symmetric. We will also present an alternative closeness metric, total variation distance.

Definition 4.5. The total variation distance between $\mathcal{D}_\psi$ and π is

$$d_{TV}(\mathcal{D}_\psi, \pi) = \frac{1}{2} \sum_{u\in V} |\mathcal{D}_\psi(u) - \pi(u)|$$

Note that total variation distance is symmetric and between 0 and 1. The following fact illustrates that the previous two measures of closeness are related

Fact 4.6.

$$\begin{aligned} d_{TV}(\mathcal{D}_\psi, \pi) &= \frac{1}{2} \sum_{u\in V} |\mathcal{D}_\psi(u) - \pi(u)| \\ &= \frac{1}{2} \sum_{u\in V} \pi(u) |\frac{\mathcal{D}_\psi(u)}{\pi(u)} - 1| \\ &= \frac{1}{2} \mathbf{E}_{u\sim\pi}[|\psi(u) - 1|] \\ &\leq \frac{1}{2} \sqrt{\mathbf{E}_{u\sim\pi}[(\psi(u) - 1)^2]} \\ &= \frac{1}{2} \sqrt{d_{\chi^2}(\mathcal{D}_\psi, \pi)} \end{aligned}$$

We are now interested in the question of how close are we to π if we take a random walk for t steps from some pre-determined starting vertex u. To answer this, suppose ψ is a density. Then $K\psi$ is some function.

Fact 4.7.

$$\begin{aligned}\langle K\psi,\ 1\rangle &= \langle \psi, K1\rangle \\ &= \langle \psi, 1\rangle \\ &= \mathbf{E}[\psi] \\ &= 1\end{aligned}$$

Then, $K\psi$ is also a density because it's expectation is 1 and it's non-negative.

Theorem 4.8. *$\mathcal{D}_{K\psi}$ is equivalent to the probability distribution where we drew a vertex from $\mathcal{D}_\psi$, then take one random step.*

Proof. [Proof sketch] We use the fact about $\langle \psi, f\rangle_\pi$:

$$\begin{aligned}\langle K\psi, f\rangle &= \langle \psi, Kf\rangle \\ &= \mathbf{E}_{u\sim\psi}[(Kf)(u)] \\ &= \mathbf{E}_{u\sim\mathcal{D}_\chi}[\mathbf{E}_{v\sim u}[f(u)]] \\ &= \mathbf{E}_{u\sim\mathcal{D}_{K\psi}}[f(u)]\end{aligned}$$

□

It is also true that $\mathcal{D}_{K^t\psi}$ corresponds to taking t random steps.

Note also that the operator K^t satisfies $K^t\phi_i = \kappa_i^t\phi_i$; i.e., the eigenvectors of K^t are still $1 = \phi_0, \phi_1, \ldots, \phi_n$ and the associated eigenvalues are κ_i^t. Let us henceforth assume that G is "made lazy" so that $\kappa_i \geq 0$ for all i. Thus we have $1 = \kappa_0^t \geq \kappa_1^t \geq \cdots \geq \kappa_{n-1}^t$. (Note that if we hadn't done that then maybe $\kappa_{n-1} = -1$ and so $\kappa_{n-1}^t = 1$!)

We also saw that the "χ^2-divergence" of $\mathcal{D}_\psi$ from π was $\mathbf{Var}[\psi]$, and that this was also an upper bound on $\sqrt{2d_{TV}(\mathcal{D}_\psi, \pi)}$; i.e., it's indeed a reasonable way to measure closeness to π.

Now we can do a simple calculation:

$$\begin{aligned}d_{\chi^2}(\mathcal{D}_{K^\psi}, \pi) &= \mathbf{Var}[K^t\psi] \\ &= \sum_{i>0} \widehat{K^t\psi}(i)^2 \\ &= \sum_{i>0} \kappa_i^t \cdot \widehat{\psi}(i)^2 \\ &\leq \max_{i>0}\{\kappa_i^t\} \sum_{i>0} \widehat{\psi}(i)^2 \\ &= \kappa_1^t \mathbf{Var}[\psi] \\ &= \kappa_1^t d_{\chi^2}(\mathcal{D}_{K\psi}, \pi)\end{aligned}$$

Thus if we do a random walk in G with the starting vertex chosen according to $\mathcal{D}_\psi$, the χ^2-divergence from π of our distribution at time t literally just goes down exponentially in

t, with the base of the exponent being κ_1. In particular, if κ_1 has a "big gap" from 1, say $\kappa_1 = 1 - \lambda_1$ where λ_1 (the second eigenvalue of L) is "large", then the exponential decay is $(1-\lambda_1)^t d_{\chi^2}(\mathcal{D}_{K^\psi}, \pi)$. We may also upper-bound this more simply as $\exp(-\lambda_1)^t d_{\chi^2}(\mathcal{D}_{K\psi}, \pi) = \exp(-\lambda_1 t) d_{\chi^2}(\mathcal{D}_{K\psi}, \pi)$.

Finally, what is the "worst possible" starting distribution? It seems intuitively clear that it should be a distribution concentrated on a single vertex, some u. As we saw, this distribution has density $\psi_u = \frac{1}{\pi(u)}\mathbf{1}_u$. It satisfies

$$d_{\chi^2}(\psi_u, \pi) = \mathbf{Var}[\psi_u] \leq \mathbf{E}[\psi_u^2] = \pi(u) \cdot \frac{1}{\pi(u)^2} = \frac{1}{\pi(u)}$$

.

Note that in the case of a regular graph G, all of the $\pi(u)$'s are $1/n$, so this distance is upper-bounded by n. In general, we'll just say it's at most $\frac{1}{\pi^*}$, where π^* is defined to be $\min_u\{\pi(u)\}$.

Finally, it is an exercise (use convexity!) to show that divergence from π after t steps starting from any distribution is no bigger than the maximum divergence over all single-vertex starting points.

We conclude with a theorem.

Theorem 4.9. *For a lazy G, if we start a t-step random walk from any starting distribution, the resulting distribution $\mathcal{D}$ on vertices satisfies*

$$d_{\chi^2}(\mathcal{D}, \pi) \leq \exp(-\lambda_1 t)\frac{1}{\pi^*}$$

where $\pi^* = \min_u\{\pi(u)\}$, which is $1/n$ in case G is regular.

Inverting this, we may say that we get ϵ-close to π (in χ^2-divergence) after $t = \frac{\ln(n/\epsilon)}{\lambda_1}$ steps when G is regular. Typically, we might have λ_1 equal to some small absolute constant like .01, and we might choose $\epsilon = 1/n^{100}$. In this case, we get super-duper-close to π after just $O(\log n)$ steps!