**Stat260/CS294: Spectral Graph Methods** **Lecture 20 - 04/07/2015**

# Lecture: Local Spectral Methods (3 of 4)

*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.*

## 20 An optimization perspective on local spectral methods

Last time, we considered local spectral methods that involve short random walks started at a small set of localized seed nodes. Several things are worth noting about this.

- The basic idea is that these random walks tend to get trapped in good conductance clusters, if there is a good conductance cluster around the seed node. A similar statement holds for approximate localized random walks, e.g., the ACL push procedure—meaning, in particular, that one can implement them "quickly," e.g., with the push algorithm, without even touching all of the nodes in $G$.
- The exact statement of the theorems that can be proven about how these procedures can be used to find good locally-biased clusters is quite technically complicated—since, e.g., one could step outside of the initial set of nodes if one starts near the boundary—certainly the statement is much more complicated than that for the vanilla global spectral method.
- The global spectral method is on the one hand a fairly straightforward algorithm (compute an eigenvector or some other related vector and then perform a sweep cut with it) and on the other hand a fairly straightforward objective (optimize the Rayleigh quotient variance subject to a few reasonable constraints).
- Global spectral methods often do very well in practice.
- Local spectral methods often do very well in practice.
- A natural question is: what objective do local spectral methods optimize—exactly, not approximately? Or, relatedly, can one construct an objective that is quickly-solvable and that also comes with similar locally-biased Cheeger-like guarantees?

To this end, today we will present a local spectral ansatz that will have several appealing properties:

- It can be computed fairly quickly, as a PPR.
- It comes with locally-biased Cheeger-like guarantees.
- It has the same form as several of the semi-supervised objectives we discussed.
- Its solution touches all the nodes of the input graph (and thus it is not as quick to compute as the push procedure which does not).

- The strongly local spectral methods that don't touch all of the nodes of the input graph are basically $\ell_1$-regularized variants of it.

## 20.1 A locally-biased spectral ansatz

Here is what we would like to do.

- We would like to introduce an ansatz for an objective function for locally-biased spectral graph partitioning. This objective function should be a locally-biased version of the usual global spectral partitioning objective; its optimum should be relatively-quickly computable; it should be useful to highlight locally-interesting properties in large data graphs; and it should have some connection to the local spectral algorithms that we have been discussing.

In addition to having an optimization formulation of locally-biased spectral partitioning methods, there are at least two reasons one would be interested in such an objective.

- A small sparse cut might be poorly correlated with the second (or even all) global eigenvectors of $L$, and so it might be invisible to global spectral methods.
- We might have exogenous information about a specific region of a large graph in which we are most interested, and so we might want a method that finds clusters near that region, e.g., to do exploratory data analysis.

Here is the approach we will take.

- We will start with the usual global spectral partitioning objective function and add to it a certain locality constraint.
- This program will be a non-convex problem (as is the global spectral partitioning problem), but its solution will be computable as a linear equation that is a generalization of the PR spectral ranking method.
- In addition, we will show that it can be used to find locally biased partitions near an input seed node, it has connections with the ACL push-based local spectral method, etc.

Let's set notation. The Laplacian is $L = D - A$; and the normalized Laplacian is $\mathcal{L} = D^{-1/2} L D^{-1/2}$. The degree-weighted inner product is given by $x^T D y = \sum_{i=1}^n x_i y_i d_i$. In this case, the weighted complete graph is given by

$$A_{K_n} = \frac{1}{\mathrm{Vol}(G)} D 1 1^T D,$$

in which case $D_{K_n} = D_G$ and thus

$$L_{K_n} = D_{K_n} - A_{K_n} = D_G - \frac{1}{\mathrm{Vol}(G)} D 1 1^T D.$$

Given this notation, see the left panel of Figure 1 for the usual spectral program $\mathsf{Spectral}(G)$, and see the right panel of Figure 1 for $\mathsf{LocalSpectral}(G, s, \kappa)$, a locally-biased spectral program. (Later, we'll call the $\mathsf{LocalSpectral}$ objective "MOV," to compare and contrast it with the "ACL" push procedure.)

$$\begin{aligned} \min \quad & x^T L_G x \\ \text{s.t.} \quad & x^T D_G x = 1 \\ & (x^T D_G 1)^2 = 0 \\ & x \in \mathbb{R}^V \end{aligned} \qquad \begin{aligned} \min \quad & x^T L_G x \\ \text{s.t.} \quad & x^T D_G x = 1 \\ & (x^T D_G 1)^2 = 0 \\ & (x^T D_G s)^2 \geq \kappa \\ & x \in \mathbb{R}^V \end{aligned}$$

Figure 1: Global and local spectral optimization programs. Left: The usual spectral program $\mathsf{Spectral}(G)$. Right: A locally-biased spectral program $\mathsf{LocalSpectral}(G, s, \kappa)$. In both cases, the optimization variable is the vector $x \in \mathbb{R}^n$.

In the above, we assume WLOG that

$$s \text{ is such that } \begin{cases} s^T D s = 1 \\ s^T D 1 = 0 \end{cases} .$$

This "WLOG" just says that one can subtract off the part of $s$ along the all-ones vector; we could have parameterized the problem to include this component and gotten similar results to what we will present below, had we not done this. Note that $s$ can actually be any vector (that isn't in the span of the all-ones vector); but it is convenient to think of it as an indicator vector of a small "seed set" of nodes $S \subset V$.

The constraint $\left(x^T D s\right)^2 \geq \kappa$ says that the projection of the solution $x$ is at least $\sqrt{\kappa}$ in absolute value, where $\kappa \in (0, 1)$. Here is the interpretation of this constraint.

- The vector $x$ must be in a spherical cap centered at $s$ with angle at most $\arccos\left(\sqrt{\kappa}\right)$ from $s$.
- Higher values of $\kappa$ correspond to finding a vector that is more well-correlated with the seed vector. While the technical details are very different than with strongly local spectral methods such as ACL, informally one should think of this as corresponding to shorter random walks or, relatedly, higher values of the teleportation parameter that teleports the walk back to the original seed set of nodes.
- If $\kappa = 0$, then there is no correlation constraint, in which case we recover $\mathsf{Spectral}(G)$.

## 20.2 A geometric notion of correlation

Although LocalSpectral is just an objective function and no geometry is explicitly imposed, there is a geometric interpretation of this in terms of a geometric notion of correlation between cuts in $G$. Let's make explicit the geometric notion of correlation between cuts (or, equivalently, between partitions, or sets of nodes) that is used by LocalSpectral.

Given a cut $(T, \bar{T})$ in a graph $G = (V, E)$, a natural vector in $\mathbb{R}^n$ to associate with it is its indicator/characteristic vector, in which case the correlation between a cut $(T, \bar{T})$ and another cut $(U, \bar{U})$ can be captured by the inner product of the characteristic vectors of the two cuts. Since we

are working on the space orthogonal to the degree-weighted all-ones vector, we'll do this after we remove from the characteristic vector its projection along the all-ones vector. In that case, again, a notion of correlation is related to the inner product of two such vectors for two cuts. More precisely, given a set of nodes $T \subseteq V$, or equivalently a cut $(T, \bar{T})$, one can define the unit vector $s_T$ as

$$s_T \stackrel{\text{def}}{=} \sqrt{\frac{\text{vol}(T)\text{vol}(\bar{T})}{2m}} \left( \frac{1_T}{\text{vol}(T)} - \frac{1_{\bar{T}}}{\text{vol}(\bar{T})} \right),$$

in which case

$$s_T(i) = \begin{cases} \sqrt{\frac{\text{vol}(T)\text{vol}(\bar{T})}{2m}} \cdot \frac{1}{\text{vol}(T)} & \text{if } i \in T \\ -\sqrt{\frac{\text{vol}(T)\text{vol}(\bar{T})}{2m}} \cdot \frac{1}{\text{vol}(\bar{T})} & \text{if } i \in \bar{T} \end{cases}.$$

Several observations are immediate from this definition.

- One can replace $s_T$ by $s_{\bar{T}}$ and the correlation remains the same with any other set, and so this is well-defined. Also, $s_T = -s_{\bar{T}}$; but since here we only consider quadratic functions of $s_T$, we can consider both $s_T$ and $s_{\bar{T}}$ to be representative vectors for the cut $(T, \bar{T})$.
- Defined this way, it immediately follows that $s_T^T D_G 1 = 0$ and that $s_T^T D_G s_T = 1$. Thus, $s_T \in \mathcal{S}_D$ for $T \subseteq V$, where we denote by $\mathcal{S}_D$ the set of vectors $\{x \in \mathbb{R}^V : x^T D_G 1 = 0\}$; and $s_T$ can be seen as an appropriately normalized version of the vector consisting of the uniform distribution over $T$ minus the uniform distribution over $\bar{T}$.
- One can introduce the following measure of correlation between two sets of nodes, or equivalently between two cuts, say a cut $(T, \bar{T})$ and a cut $(U, \bar{U})$:

  $$K(T, U) \stackrel{\text{def}}{=} (s_T D_G s_U)^2.$$

  Then it is easy to show that: $K(T, U) \in [0, 1]$; $K(T, U) = 1$ if and only if $T = U$ or $\bar{T} = U$; $K(T, U) = K(\bar{T}, U)$; and $K(T, U) = K(T, \bar{U})$.
- Although we have described this notion of geometric correlation in terms of vectors of the form $s_T \in \mathcal{S}_D$ that represent partitions $(T, \bar{T})$, this correlation is clearly well-defined for other vectors $s \in \mathcal{S}_D$ for which there is not such a simple interpretation in terms of cuts.

Below we will show that the solution to LocalSpectral can be characterized in terms of a PPR vector. If we were interested in objectives that had solutions of different forms, e.g., the form of a heat kernel, then this would correspond to an objective function with a different constraint, and this would then imply a different form of correlation.

## 20.3 Solution of LocalSpectral

Here is the basic theorem characterizing the form of the solution of LocalSpectral.

**Theorem 1** (Solution Characterization)**.** *Let $s \in \mathbb{R}^n$ be a seed vector such that $s^T D_G 1 = 0$, $s^T D_G s = 1$, and $s^T D_G v_2 \neq 0$, where $v_2$ is the second generalized eigenvector of $L_G$ with respect to $D_G$. In addition, let $1 > \kappa \geq 0$ be a correlation parameter, and let $x^\star$ be an optimal solution to LocalSpectral$(G, s, \kappa)$. Then, there exists some $\gamma \in (-\infty, \lambda_2(G))$ and a $c \in [0, \infty]$ such that*

$$x^\star = c(L_G - \gamma D_G)^+ D_G s. \tag{1}$$

Before presenting the proof of this theorem, here are several things to note.

- $s$ and $\kappa$ are the parameters of the program; $c$ is a normalization factor that rescales the norm of the solution vector to be 1 (and that can be computed in linear time, given the solution vector); and $\gamma$ is implicitly defined by $\kappa$, $G$, and $s$.
- The correct setting of $\gamma$ ensures that $(s^T D_G x^\star)^2 = \kappa$, i.e., that $x^\star$ is found exactly on the boundary of the feasible region.
- $x^\star$ and $\gamma$ change as $\kappa$ changes. In particular, as $\kappa$ goes to 1, $\gamma$ tends to $-\infty$ and $x^\star$ approaches $s$; conversely, as $\kappa$ goes to 0, $\gamma$ goes to $\lambda_2(G)$ and $x^\star$ tends towards $v_2$, the global eigenvector.
- For a fixed choice of $G$, $s$, and $\kappa$, an $\epsilon$-approximate solution to LocalSpectral can be computed in time $\tilde{O}\left(\frac{m}{\sqrt{\lambda_2(G)}} \cdot \log(\frac{1}{\epsilon})\right)$ using the Conjugate Gradient Method; or in time $\tilde{O}\left(m \log(\frac{1}{\epsilon})\right)$ using the Spielman-Teng linear-equation solver (that we will discuss in a few weeks), where the $\tilde{O}$ notation hides $\log\log(n)$ factors. This is true for a fixed value of $\gamma$, and the correct setting of $\gamma$ can be found by binary search.

  While that is theoretically true, and while there is a lot of work recently on developing practically-fast nearly-linear-time Laplacian-based solvers, this approach might not be appropriate in certain applications. For example, in many applications, one has precomputed an eigenvector decomposition of $L_G$, and then one can use those vectors and obtain an approximate solution with a small number of inner products. This can often be much faster in practice.

In particular, solving LocalSpectral is *not* "fast" in the sense of the original local spectral methods, i.e., in that the running time of those methods depends on the size of the output and doesn't depend on the size of the graph. But the running time to solve LocalSpectral *is* fast, in that its solution depends essentially on computing a leading eigenvector of a Laplacian $L$ and/or can be solved with "nearly linear time" solvers that we will discuss in a few weeks.

While Eqn. (1) is written in the form of a linear equation, there is a close connection between the solution vector $x^\star$ and the Personalized PageRank (PPR) spectral ranking procedure.

- Given a vector $s \in \mathbb{R}^n$ and a *teleportation* constant $\alpha > 0$, the PPR vector can be written as

$$\mathrm{pr}_{\alpha,s} = \left(L_G + \frac{1-\alpha}{\alpha} D_G\right)^{-1} D_G s.$$

  By setting $\gamma = -\frac{1-\alpha}{\alpha}$, one can see that the optimal solution to LocalSpectral is proved to be a generalization PPR.
- In particular, this means that for high values of the correlation parameter $\kappa$ for which the corresponding $\gamma$ satisfies $\gamma < 0$, the optimal solution to LocalSpectral takes the form of a PPR vector. On the other hand, when $\gamma \geq 0$, the optimal solution to LocalSpectral provides a smooth way of transitioning from the PPR vector to the global second eigenvector $v_2$.
- Another way to interpret this is to say that for values of $\kappa$ such that $\gamma < 0$, then one could compute the solution to LocalSpectral with a random walk or by solving a linear equation, while for values of $\kappa$ for which $\gamma > 0$, one can only compute the solution by solving a linear equation and not by performing a random walk.

$$\begin{aligned} \text{minimize} \quad & L_G \circ X \\ \text{s.t.} \quad & L_{K_n} \circ X = 1 \\ & (D_G s)(D_G s)^T \circ X \geq \kappa \\ & X \succeq 0 \end{aligned} \qquad \begin{aligned} \text{maximize} \quad & \alpha + \kappa\beta \\ \text{s.t.} \quad & L_G \succeq \alpha L_{K_n} + \beta (D_G s)(D_G s)^T \\ & \beta \geq 0 \\ & \alpha \in \mathbb{R} \end{aligned}$$

Figure 2: Left: Primal SDP relaxation of $\mathsf{LocalSpectral}(G, s, \kappa)$: $\mathsf{SDP}_p(G, s, \kappa)$. For this primal, the optimization variable is $X \in \mathbb{R}^{n \times n}$ such that $X$ is SPSD. Right: Dual SDP relaxation of $\mathsf{LocalSpectral}(G, s, \kappa)$: $\mathsf{SDP}_d(G, s, \kappa)$. For this dual, the optimization variables are $\alpha, \beta \in \mathbb{R}$.

About the last point, we have talked about how random walks compute regularized or robust versions of the leading nontrivial eigenvector of $L$—it would be interesting to characterize an algorithmic/statistical tradeoff here, e.g., if/how in this context certain classes of random walk based algorithms are less powerful algorithmically than related classes of linear equation based algorithms but that they implicitly compute regularized solutions more quickly for the parameter values for which they are able to compute solutions.

## 20.4 Proof of Theorem 1

Here is an outline of the proof, which essentially involves "lifting" a rank-one constraint to obtain an SDP in order to get strong duality to apply.

- Although LocalSpectral is not a convex optimization problem, it can be relaxed to an SDP that is convex.
- From strong duality and complementary slackenss, the solution to the SDP is rank one.
- Thus, the vector making up the rank-one component of this rank-one solution is the solution to LocalSpectral.
- The form of this vector is of the form of a PPR.

Here are some more details. Consider the primal $\mathsf{SDP}_p$ and dual $\mathsf{SDP}_d$ SDPs, given in the leftt panel and right panel, respectively, of Figure 2.

Here are a sequence of claims.

**Claim 1.** *The primal SDP, $\mathsf{SDP}_p$ is a relaxation of LocalSpectral.*

*Proof.* Consider $x \in \mathbb{R}^n$, a feasible vector for LocalSpectral. Then, the SPSD matrix $X = xx^T$ is feasible for $\mathsf{SDP}_p$. □

**Claim 2.** *Strong duality holds between $\mathsf{SDP}_p$ and $\mathsf{SDP}_d$.*

*Proof.* The program $\mathsf{SDP}_p$ is convex, and so it suffices to check that Slater's constraint qualification conditions hold for $\mathsf{SDP}_p$. To do so, consider $X = ss^T$. Then,

$$(D_G s)\,(D_G s)^T \circ ss^T = \left(s^T D_G s\right)^2 = 1 > \kappa.$$

□

**Claim 3.** *The following fewsibility and complementary slackness conditions are sufficient for a primal-dual pair $X^*$, $\alpha^*$, $\beta^*$ to be an optimal solution. The feasibility conditions are:*

$$\begin{aligned} L_{K_n} \circ X^\star &= 1, \\ (D_G s)(D_G s)^T \circ X^\star &\geq \kappa, \\ L_G - \alpha^\star L_{K_n} - \beta^\star (D_G s)(D_G s)^T &\succeq 0, \text{ and} \\ \beta^\star &\geq 0, \end{aligned} \tag{2}$$

*and the complementary slackness conditions are:*

$$\alpha^\star (L_{K_n} \circ X^\star - 1) = 0, \tag{3}$$

$$\beta^\star ((D_G s)(D_G s)^T \circ X^\star - \kappa) = 0, \text{ and} \tag{3}$$

$$X^\star \circ (L_G - \alpha^\star L_{K_n} - \beta^\star (D_G s)(D_G s)^T) = 0. \tag{4}$$

*Proof.* This follows from the convexity of $\mathsf{SDP}_p$ and Slater's condition. □

**Claim 4.** *The feasibility and complementary slackness conditions, coupled with the assumptions of the theorem, imply that $X^*$ is rank one and that $\beta^* \geq 0$.*

*Proof.* If we plug $v_2$ in Eqn. (2), then we obtain that $v_2^T L_G v_2 - \alpha^\star - \beta^\star (v_2^T D_G s)^2 \geq 0$.

But $v_2^T L_G v_2 = \lambda_2(G)$ and $\beta^\star \geq 0$. Hence, $\lambda_2(G) \geq \alpha^\star$. Suppose $\alpha^\star = \lambda_2(G)$. As $s^T D_G v_2 \neq 0$, it must be the case that $\beta^\star = 0$. Hence, by Equation (4), we must have $X^\star \circ L(G) = \lambda_2(G)$, which implies that $X^\star = v_2 v_2^T$, *i.e.*, the optimum for $\mathsf{LocalSpectral}$ is the global eigenvector $v_2$. This corresponds to a choice of $\gamma = \lambda_2(G)$ and $c$ tending to infinity.

Otherwise, we may assume that $\alpha^\star < \lambda_2(G)$. Hence, since $G$ is connected and $\alpha^\star < \lambda_2(G)$, $L_G - \alpha^\star L_{K_n}$ has rank exactly $n-1$ and kernel parallel to the vector 1.

From the complementary slackness condition (4) we can deduce that the image of $X^\star$ is in the kernel of $L_G - \alpha^\star L_{K_n} - \beta^\star (D_G s)(D_G s)^T$.

If $\beta^\star > 0$, we have that $\beta^\star (D_G s)(D_G s)^T$ is a rank one matrix and, since $s^T D_G 1 = 0$, it reduces the rank of $L_G - \alpha^\star L_{K_n}$ by one precisely. If $\beta^\star = 0$ then $X^\star$ must be 0 which is not possible if $\mathsf{SDP}_p(G, s, \kappa)$ is feasible.

Hence, the rank of $L_G - \alpha^\star L_{K_n} - \beta^\star (D_G s)(D_G s)^T$ must be exactly $n-2$. As we may assume that 1 is in the kernel of $X^\star$, $X^\star$ must be of rank one. This proves the claim. □

**Remark.** It would be nice to have a cleaner proof of this that is more intuitive and that doesn't rely on "boundary condition" arguments as much.

Now we complete the proof of the theorem. From the claim it follows that, $X^\star = x^\star x^{\star T}$ where $x^\star$ satisfies the equation

$$(L_G - \alpha^\star L_{K_n} - \beta^\star (D_G s)(D_G s)^T) x^\star = 0.$$

From the second complementary slackness condition, Equation (3), and the fact that $\beta^\star > 0$, we obtain that $(x^\star)^T D_G s = \pm\sqrt{\kappa}$. Thus, $x^\star = \pm\beta^\star\sqrt{\kappa}(L_G - \alpha^\star L_{K_n})^+ D_G s$, as required.

## 20.5 Additional comments on the LocalSpectral optimization program

Here, we provide some additional discussion for this locally-biased spectral partitioning objective. Recall that the proof we provided for Cheeger's Inequality showed that in some sense the usual global spectral methods "embed" the input graph $G$ into a complete graph; we would like to say something similar here.

To do so, observe that the dual of LocalSpectral is given by the following.

$$\begin{aligned} \text{maximize} \quad & \alpha + \beta\kappa \\ \text{s.t.} \quad & L_G \succeq \alpha L_{K_n} + \beta\Omega_T \\ & \beta \geq 0, \end{aligned}$$

where $\Omega_T = D_G s_T s_T^T D_G$. Alternatively, by subtracting the second constraint of LocalSpectral from the first constraint, it follows that

$$x^T \left(L_{K_n} - L_{K_n} s_T s_T^T L_{K_n}\right) x \leq 1 - \kappa.$$

Then it can be shown that

$$L_{K_n} - L_{K_n} s_T s_T^T L_{K_n} = \frac{L_{K_T}}{\text{vol}(\bar{T})} + \frac{L_{K_{\bar{T}}}}{\text{vol}(T)},$$

where $L_{K_T}$ is the $D_G$-weighted complete graph on the vertex set $T$. Thus, LocalSpectral is equivalent to

$$\begin{aligned} \text{minimize} \quad & x^T L_G x \\ \text{s.t.} \quad & x^T L_{K_n} x = 1 \\ & x^T \left(\frac{L_{K_T}}{\text{vol}(\bar{T})} + \frac{L_{K_{\bar{T}}}}{\text{vol}(T)}\right) x \leq 1 - \kappa. \end{aligned}$$

The dual of this program is given by the following.

$$\begin{aligned} \text{maximize} \quad & \alpha - \beta(1 - \kappa) \\ \text{s.t.} \quad & L_G \succeq \alpha L_{K_n} - \beta \left(\frac{L_{K_T}}{\text{vol}(\bar{T})} + \frac{L_{K_{\bar{T}}}}{\text{vol}(T)}\right) \\ & \beta \geq 0. \end{aligned}$$

Thus, from the perspective of this dual, LocalSpectral can be viewed as "embedding" a combination of a complete graph $K_n$ and a weighted combination of complete graphs on the sets $T$ and $\bar{T}$, i.e., $K_T$ and $K_{\bar{T}}$. Depending on the value of $\beta$, the latter terms clearly discourage cuts that substantially cut into $T$ or $\bar{T}$, thus encouraging partitions that are well-correlated with the input cut $(T, \bar{T})$.

If we can establish a precise connection between the optimization-based LocalSpectral procedure and operational diffusion-based procedures such as the ACL push procedure, then this would provide additional insight as to "why" the short local random walks get stuck in small seed sets of nodes. This will be one of the topics for next time.