Stat260/CS294: Spectral Graph Methods **Lecture 18 - 03/31/2015**

Lecture: Local Spectral Methods (1 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.

18 Introduction and overview

Last time, we showed that certain random walks and diffusion-based methods, *when not run to the asymptotic limit*, exactly solve regularized versions of the Rayleigh quotient objective (in addition to approximating the Rayleigh quotient, in a manner that depends on the specific random walk and how the spectrum of L decays). There are two ways to think about these results.

- One way to think about this is that one runs almost to the asymptotic state and then one gets a vector that is "close" to the leading eigenvector of L. Note, however, that the statement of implicit regularization from last time does *not* depend on the initial condition or how long the walk was run. (The value of the regularization parameter, etc., does, but the form of the statement does not.) Thus ...
- Another way to think about this is that one starts at any node, say a localized "seed set" of nodes, e.g., in which all of the initial probability mass is on one node or a small number of nodes that are nearby each other in the graph topology, and then one runs only a small number of steps of the random walk or diffusion. In this case, it might be more natural/useful to try to quantify the idea that: if one starts the random walk on the small side of a bottleneck to mixing, and if one runs only a few steps of a random walk, then one might get stuck in that small set.

The latter is the basic idea of so-called *local spectral methods*, which are a class of algorithms that have received a lot of attention recently. Basically, they try to extend the ideas of global spectral methods, where we compute eigenvectors, random walks, etc., that reveal structure about the entire graph, e.g., that find a partition that is quadratically-good in the sense of Cheeger's Inequality to the best conductance/expansion cut in the graph, to methods that reveal interesting structure in locally-biased parts of the graph. Not only do these provide locally-biased versions of global spectral methods, but since spectral methods are often used to provide a ranking for the nodes in a graph and/or to solve other machine learning problems, these also can be used to provide a locally-biased or personalized version of a ranking function and/or to solve other machine learning problems in a locally-biased manner.

18.1 Overview of local spectral methods and spectral ranking

Here is a brief history of local and locally-biased spectral methods.

- LS: developed a basic locally-biased mixing result in the context of mixing of Markov chains in convex bodies. They basically show a partial converse to the easy direction of Cheeger's Inequality—namely, that if the conductance $\phi(G)$ of the graph G is big, then *every* random walk must converge quickly—and from this they also show that if the random walk fails to converge quickly, then by examining the probability distribution that arises after a few steps one can find a cut of small conductance.

- ST: used the LS result to get an algorithm for local spectral graph partitioning that used truncated random walks. They used this to find good well-balanced graph partitions in nearly-linear time, which they then used as a subroutine in their efforts to develop nearly linear time solvers for Laplacian-based linear systems (a topic to which we will return briefly at the end of the semester).

- ACL/AC: improved the ST result by computing a personalized PageRank vector. This improves the fast algorithms for Laplacian-based linear solvers, and it is of interest in its own right, so we will spend some time on it.

- C: showed that similar results can be obtained by doing heat kernel computations.

- AP: showed that similar results can be obtained with an evolving set method (that we won't discuss in detail).

- MOV: provided an optimization perspective on these local spectral methods. That is, this is a locally-biased optimization objective that, if optimized exactly, leads to similar locally-biased Cheeger-like bounds.

- GM: characterized the connection between the strongly-local ACL and the weakly-local MOV in terms of ℓ_1 regularization (i.e., a popular form of sparsity-inducing regularizations) of ℓ_2 regression problems.

There are several reasons why one might be interested in these methods.

- Develop faster algorithms. This is of particular interest if we can compute locally-biased partitions without even touching all of the graph. This is the basis for a lot of work on nearly linear time solvers for Laplacian-based linear systems.

- Improved statistical properties. If we can compute locally-biased things, e.g., locally-biased partitions, without even touching the entire graph, then that certainly implies that we are robust to things that go on on the other side of the graph. That is, we have essentially engineered some sort of regularization into the approximation algorithm; and it might be of interest to quantify this.

- Locally exploring graphs. One might be interested in finding small clusters or partitions that are of interest in a small part of a graph, e.g., a given individual in a large social graph, in situations when those locally-biased clusters are not well-correlated with the leading or with any global eigenvector.

We will touch on all these themes over the next four classes.

For now, let $G = (V, E)$, and recall that $\text{Vol}(G) = \sum_{v \in T} d_v$ (so, in particular, we have that $\text{Vol}(G) = 2|E| = 2m$. Also, A is the Adjacency Matrix, $W = D^{-1}A$, and $\mathcal{L} = I - D^{-1}A$ is the random walk normalized Laplacian. For a vector $x \in \mathbb{R}^n$, let's define its support as

$$\text{Supp}(v) = \{i \in V = [n] : v_i \neq 0\}.$$

Then, here is the transition kernel for that vanilla random walk.

$$\mathbb{P}\left[x_{t+1} = j | x_t = i\right] = \begin{cases} \frac{1}{d} & \text{if } i \sim j \\ 0 & \text{otherwise.} \end{cases}$$

If we write this as a transition matrix operating on a (row) vector, then we have that

$$p(t) = sW^t,$$

where W is the transition matrix, and where $s = p(0)$ is the initial distribution, with $\|s\|_1 = 1$. Then, $p = \vec{1}^T D / (\vec{1}^T D \vec{1})$ is the stationary distribution, i.e.,

$$\lim_{t \to \infty} \mathbb{P}\left[x_t = i\right] = \frac{d_i}{\text{Vol}(G)},$$

independent of $s = p(0)$, as long as G is connected and not bipartite. (If it is bipartite, then let $W \to W_{LAZY} = \frac{1}{2}\left(I + D^{-1}A\right)$, and the same results holds.)

There are two common interpretations of this asymptotic random walk.

- Interpretation 1: the limit of a random walk.
- Interpretation 2: a measure of the importance of a node.

With respect to the latter interpretation, think of an edge as denoting importance, and then what we want to find is the important nodes (often for directed graphs, but we aren't considering that here). Indeed, one of the simplest *centrality measures* in social graph analysis is the degree of a node. For a range of reasons, e.g., since that is easy to spam, a refinement of that is to say that important nodes are those nodes that have links to important nodes.

This leads to a large area known as *spectral ranking* methods. This area applies the theory of matrices or linear maps—basically eigenvectors and eigenvalues, but also related things like random walks—to matrices that represent some sort of relationship between entities. This has a long history, most recently made well-known by the PageRank procedure (which is one version of it). Here, we will follow Vigna's outline and description in his "Spectral Ranking" notes—his description is very nice since it provides the general picture in a general context, and then he shows that with several seemingly-minor tweaks, one can obtain a range of related spectral graph methods.

18.2 Basics of spectral ranking

To start, take a step back, and let $M \in \mathbb{R}^{n \times n}$, where each column/row of M represents some sort of entity, and M_{ij} represents some sort of *endorsement or approval* of entity j from entity i. (So far, one could potentially have negative entries, with the obvious interpretation, but this will often be removed later, basically since one has more structure if entries must be nonnegative.) As Vigna

describes, Seeley (in 1949) observed that one should define importance/approval recursively, since that will capture the idea that an entity is important/approved if other important entities think is is important/approved. In this case, recursive could mean that

$$r = rM,$$

i.e., that the index of the i^{th} node equals the weighted sum of the indices of the entities that endorse the i^{th} node. This isn't always possible, and indeed Seeley considers nonnegative matrices that don't have any all-zeros rows, in which case uniqueness, etc., follow from the Perron-Frobenius ideas we discussed before. This involves the left eigenvectors, as we have discussed; one could also look at the right eigenvectors, but the endorsement/approval interpretation fails to hold.

Later, Wei (in 1952) and Kendall (in 1955) were interested in *ranking* sports teams, and they said essentially that better teams are those teams that beat better teams. This involves looking at the rank induced by

$$\lim_{k\to\infty} M^k \vec{1}^T$$

and then appealing to Perron-Frobenius theory. The significant point here is three-fold.

- The motivation is very different than Seeley's endorsement motivation.
- Using dominant eigenvectors on one side or the other dates back to mid 20th century, i.e., well before recent interest in the topic.
- The relevant notion of convergence in both of these motivating applications is that of *convergence in rank* (where rank means the rank order of values of nodes in the leading eigenvector, and not anything to do with the linear-algebraic rank of the underlying matrix). In particular, the actual values of the entires of the vector are not important. This is very different than other notions of convergence when considering leading eigenvectors of Laplacians, e.g., the value of the Rayleigh quotient.

Here is the generalization. Consider matrices M with real and positive dominant eigenvalue λ and its eigenvector, i.e., vector r such that $\lambda r = rM$, where let's say that the dimension of the eigenspace is one.

Definition 1. *The* left spectral ranking *associated with M is, or is given by, the dominant left eigenvector.*

If the eigenspace does not have dimension one, then there is the usual ambiguity problem (which is sometimes simply assumed away, but which can be enforced by a reasonable rule—we will see a common way to do the latter in a few minutes), but if the eigenspace has dimension one, then we can talk of *the* spectral ranking. Note that it is defined only up to a constant: this is not a problem if all the coordinates have the same sign, but it introduces an ambiguity otherwise, and how this ambiguity is resolved can lead to different outcomes in "boundary cases" where it matters; see the Gleich article for examples and details of this.

Of course one could apply the same thing to M^T. The mathematics is similar, but the motivation is different. In particular, Vigna argues that the endorsement motivation leads to the left eigenvectors, while the influence/better-than motivation leads to the right eigenvectors.

The next idea to introduce is that of "damping." (As we will see, this will have a reasonable generative "story" associated with it, and it will have a reasonable statistical interpretation, but it is also important for a technical reason having to do with ensuring that the dimension of the eigenspace is one.) Let M be a zero-one matrix; then $\left(M^k\right)_{ij}$ is the number of directed paths from $i \to j$ in a directed graph defined by M. In this case, an obvious idea of measuring the importance of i, i.e., to measure the number of paths going into j, since they represent recursive endorsements, which is given by

$$\vec{1}\left(I + M + M^2 + \cdots\right) = I + \sum_{k=0}^{\infty} M^k,$$

does *not* work, since the convergence of the equation is not guaranteed, and it does not happen in general.

If, instead, one can guarantee that the spectral radius of M is less than one, i.e., that $\lambda_0 < 1$, then this infinite sum does converge. One way to do this is to introduce a damping factor α to obtain

$$\vec{1}\left(I + \alpha M + \alpha^2 M^2 + \cdots\right) = I + \sum_{k=0}^{\infty} (\alpha M)^k .$$

This infinite sum does converge, as the spectral radius of αM is strictly less than 1, if $\alpha < \frac{1}{\lambda_0}$. Katz (in 1953) proposed this. Note that

$$\vec{1}\sum_{k=0}^{\infty} (\alpha M)^k = \vec{1}\,(I - \alpha M)^{-1} .$$

So, in particular, we can compute this index by solving the *linear system* as follows.

$$x\,(I - \alpha M) = \vec{1}.$$

This is a particularly well-structured system of linear equations, basically a system of equations where the constraint matrix can be written in terms of a Laplacian. There has been a lot of work recently on developing fast solvers for systems of this form, and we will get back to this topic in a few weeks.

A generalization of this was given by Hubbel (in 1965), who said that one could define a status index r by using a recursive equation $r = v + rM$, where v is a "boundary condition" or "exogenous contribution." This gives

$$r = v\,(I - M)^{-1} = v\sum_{k=0}^{\infty} M^k.$$

So, we are generalizing from the case where the exogenous contribution is $\vec{1}$ to an arbitrary vector v. This converges if $\lambda_0 < 1$; otherwise, one could introduce an damping factor (as we will do below).

To see precisely how these are all related, let's consider the basic spectral ranking equation

$$\lambda_0 r = rM.$$

If the eigenspace of λ_0 has dimension greater than 1, then there is no clear choice for the ranking r. One idea in this case is to perturb M to satisfy this property, but we want to apply a "structured perturbation" in such a way that many of the other spectral properties are not damaged. Here is the relevant theorem, which is due to Brauer (in 1952), and which we won't prove.

Theorem 1. *Let $A \in \mathbb{C}^{n\times n}$, let*

$$\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$$

be the eigenvalues of A, and let $x \in \mathbb{C}^n$ be nonzero vector such that $Ax^T = \lambda_0 x^T$. Then, for all vectors $v \in \mathbb{C}^n$, the eigenvalues of $A + x^T v$ are given by

$$\lambda_0 + vx^T, \lambda_1, \ldots, \lambda_{n-1}.$$

That is, if we perturb the original matrix by an rank-one update, where the rank-one update is of the form of the outer product of an eigenvector of the matrix and an arbitrary vector, then one eigenvalue changes, while all the others stay the same.

In particular, this result can be used to split degenerate eigenvalues and to introduce a *gap* into the spectrum of M. To see this, let's consider a rank-one convex perturbation of our matrix M by using a vector v such that $vx^T = \lambda_0$ and by applying the theorem to αM and $(1-\alpha)\, x^T v$. If we do this then we get

$$\lambda_0 r = r\left(\alpha M + (1-\alpha)\, x^T v\right).$$

Next, note that $\alpha M + (1-\alpha)\, x^T v$ has the same dominant eigenvalues as M, but with algebraic multiplicity 1, and all the other eigenvalues are multiplied by $\alpha \in (0,1)$.

This ensures a unique r. The cost is that it introduces extra parameters (α is we set v to be an all-ones vector, but the vector v if we choose it more generally). These parameters can be interpreted in various ways, as we will see.

An important consequence of this approach is that r is defined only up to a constant, and so we can impose the constraint that $rx^T = \lambda_0$. (Note that if $x = \vec{1}$, then this says that the sum of r's coordinates is λ_0, which if all the coordinates have the same sign means that $\|r\|_1 = \lambda_0$.) Then we get

$$\lambda_0 r = \alpha r M + (1-\alpha)\, \lambda_0 v.$$

Thus,

$$r\left(\lambda_0 - \alpha M\right) = (1-\alpha)\, \lambda_0 v.$$

From this it follows that

$$\begin{aligned} r &= (1-\alpha)\, \lambda_0 v \left(\lambda_0 - \alpha M\right)^{-1} \\ &= (1-\alpha)\, v \left(1 - \frac{\alpha}{\lambda_0} M\right)^{-1} \\ &= (1-\alpha)\, v \sum_{k=0}^{\infty} \left(\frac{\alpha}{\lambda_0} M\right)^k \\ &= (1-\lambda_0\beta)\, v \sum_{k=0}^{\infty} (\beta M)^k, \end{aligned}$$

which converges if $\alpha < 1$, i.e., if $\beta < \frac{1}{\lambda_0}$.

That is, this approach shows that the Katz-Hubbel index can be obtained as a rank-one perturbation of the original matrix. In a little bit, we will get to what this rank-one perturbation "means" in different situations.

To review, we started with a matrix M with possibly many eigenvectors associated with the dominant eigenvalue, and we introduced a structured perturbation get a specific eigenvector associated with λ_0, given the boundary condition v.

The standard story is that if we start from a generic nonnegative matrix and normalize its rows to get a stochastic matrix, then we get a Markovian spectral ranking, which is the limit distribution of the random walk. Here, we are slightly more general, as is captured by the following definition.

Definition 2. *Given the matrix M, the sampled spectral ranking of M with boundary condition v and dumping factor α is*

$$r_0 = (1-\alpha)\, v \sum_{k=0}^{\infty} \left(\frac{\alpha}{\lambda_0} M \right)^k,$$

for $|\alpha| < 1$.

The interpretation of the boundary condition (from the sampled to the standard case) is the following.

- In the damped case, the Markov chain is restarted to a fixed distribution v, and there is a single stationary distribution which is the limit of every starting distribution.
- In the standard case, v is the starting distribution from which we capture the limit using an eigenprojection.

While in some sense equivalent, these two interpretations suggest different questions and interpretations, and we will consider both of these over the next few classes.

18.3 A brief aside

Here is an aside we will get back to over the next few classes.

Consider a vanilla random walk, where at each time step, we follow the graph with some probability β, and we randomly jump to any uniformly-chosen node in the graph with probability $1-\beta$. This Markov chain has stationary distribution

$$p_\beta = \frac{1-\beta}{n}\vec{1} + \beta p_\beta W.$$

This is often known as PageRank, which has received a lot of attention in web ranking, but from the above discussion it should be clear that it is one form of the general case of spectral ranking methods. We can also ask for a "personalized" version of this, by which we informally mean a ranking of the nodes that in some sense is conditioned on or personalized for a given node or a given seed set of nodes. We can get this by using a personalized PageRank, by randomly jumping (not to any node chosen uniformly at random but) to a "seed set" s of nodes. This PPR is the unique solution to

$$p_\beta(s) = (1-\beta)\, s + \beta p_\beta(s) W,$$

i.e., it is of the same form as the expression above, except that an all-ones vector has been replaced by a seed set or boundary condition vector V. This latter expression solves the linear equation

$$(I - \beta W)\, p_\beta(s) = (1-\beta)\, s.$$

We can write this expression as an infinite sum as

$$p_\beta(s) = (1-\beta)\, s + \beta \sum_{t=0}^{\infty} \beta^t \, (sW)^t$$

Thus, note that the following formulations of PageRank (as well as spectral ranking more generally) are equivalent.

- $(I - \beta W)\, x = (1-\beta)\, s$
- $(\gamma D + L)\, z = \gamma s$, where $\beta = \frac{1}{1+\gamma}$ and $x = Dz$.

As noted above, this latter expression is of the form of Laplacian-based linear equations. It is of the same form that arises in those semi-supervised learning examples that we discusses. We will talk toward the end of the term about how to solve equations of this form more generally.