Spectral Graph Theory Lecture 9

Trees and Distances

Daniel A. Spielman September 26, 2012

9.1 About these notes

These notes are not necessarily an accurate representation of what happened in class. The notes written before class say what I think I should say. The notes written after class way what I wish I said.

9.2 Overview

The goal of this lecture is to present an approach of Stone and Griffing [SG09] to reconstructing evolutionary trees. They take the distances between all pairs of leaves in a tree and use this to reconstruct the tree. There are many other algorithms for doing this. I will add some references when I revise the notes.

I present this approach both because it reveals the power of the second eigenvector of the Laplacian, and because I think it can be made to get around some of the deficits of some of the other algorithms.

9.3 Idealized Evolutionary Trees

I will now give an idealized view of how we could reconstruct the tree of life from the DNA of existing organisms. Evolutionary theory tells us that species split off from each other through mutation. Thus, we should be able to arrange the species that have existed into a very large tree with one vertex for each species. The root should be the first organism, and the children of each vertex should be those that have split off from the corresponding species. The leaves of the tree are those species that presently exist. OK, really any species that did not spawn off new species should be a leaf. But, it seems difficult to get DNA samples from any that don't exist, so we will ignore the extinct leaves.

Consider assigning each edge a length measuring the difference between the DNA of the organisms on either end of the edge. A naive measure would be the number of base-pairs in which the DNA sequences differ. We can see that this is too naive, as the number of base-pairs can change. A more sophisticated measure would be edit-distance. I doubt that it is good enough for practice, but that won't concern us here. Intuitively, the distance between the DNA of two leaf species should be close to the sum of the lengths of the edges on the path between those species. This won't be quite right either, as the same mutations can happen in different parts of the tree, some mutations

can be made and unmade, and stranger things can happen (like viruses transferring DNA between organisms). For this lecture, we will pretend that the distance between leaves is exactly the sum of the lengths of the edges between them in the tree.

The question is, under these assumptions can we recover the tree? It has long been known that we can. In this lecture, we will see a spectral approach that is implicit in the work of Stone and Griffing [SG09]. An interesting question is whether this method can be made to perform well under real situations. There is some reason for hope, as many spectral methods behave well in the presense of noise, as we will see later in the semester.

9.4 Recovering the Laplacian

In the last class, we proved that effective resistance is a distance between vertices, where I recall that the effective resistance is given by

$$\mathrm{R}_{\mathrm{eff}}(a,b) = (\boldsymbol{e}_a - \boldsymbol{e}_b)^T \boldsymbol{L}^+ (\boldsymbol{e}_a - \boldsymbol{e}_b).$$

We now consider the inverse problem: given a list of distances between vertices, can one recover the Laplacian matrix? That is, assume that we are given a matrix $\boldsymbol{R}$ such that

$$\boldsymbol{R}(a,b) = \mathrm{R}_{\mathrm{eff}}(a,b).$$

If we are told that $\boldsymbol{R}$ is the matrix of effective resistance distances of some weighted graph, we can recover that graph. We begin by recovering the pseudo-inverse of its Laplacian.

First, we observe that

$$(\boldsymbol{e}_a - \boldsymbol{e}_b)^T \boldsymbol{R} (\boldsymbol{e}_a - \boldsymbol{e}_b) = -2\boldsymbol{R}(a,b).$$

So, the matrix $-\boldsymbol{R}/2$ will have all the correct differences. The other thing we know about the matrix $\boldsymbol{L}^+$ is that the all-1s vector is in its nullspace. To create a matrix that satisfies this, let $\boldsymbol{\Pi}$ be the projection orthogonal to the all-1s vector. That is,

$$\boldsymbol{\Pi} = \boldsymbol{I} - \boldsymbol{J}/n = (1/n)\boldsymbol{L}_{K_n},$$

where we recall that $\boldsymbol{J}$ is the all-1s matrix. We then multiply by $\boldsymbol{\Pi}$ on either side to get

$$-\boldsymbol{\Pi}\boldsymbol{R}\boldsymbol{\Pi}/2.$$

Lemma 9.4.1.

$$-\boldsymbol{\Pi}\boldsymbol{R}\boldsymbol{\Pi}/2 = \boldsymbol{L}^+.$$

You will prove this in problem set 2.

9.5 The Distance on the Leaves

That was nice enough, but it doesn't solve our problem, we only know the distances between the leaves of the tree. The first thing we will do is verify that the distance between the leaves of a tree *is* the same as the effective resistance distance, at least if we set the resistance of each edge to be its length.

Claim 9.5.1. *Let T be a tree whose edges have lengths. Then, for every two vertices a and b, the effective resistance between the vertices equals the sum of the lengths of the edges between them.*

Proof. This will be on the problem set. The intuition is that the edges that are not on the path between a and b do not carry any flow. But, that is not a formal proof. □

This leads us to ask the following natural question: is there a graph $\widehat{G}$ whose vertices are just the leaves such that the effective resistance between vertices in $\widehat{G}$ is the same as in the tree? We will see that the answer is "yes", that it does not depend at all on the graph being a tree, and that it comes from Gaussian Elimination.

9.6 Elimination of one Vertex

Let's begin by eliminating just one vertex. We'll start with a graph $G = (V, E, w)$, and we will identify one vertex, say vertex 1, that we want to eliminate. I will think of this as an *interal* vertex and of all the others as the *boundary* vertices. Name the boundary vertices $B = \{2, \ldots, n\}$.

We now want to construct a graph $\widehat{G}$ on vertex set B so that for all $a, b \in B$,

$$\mathrm{R}_{\mathrm{eff}\,G}(a, b) = \mathrm{R}_{\mathrm{eff}\,\widehat{G}}(a, b).$$

When we are computing the effective resistance between a and b in G, we know what happens to the voltage of vertex 1: it is set to the weighted average of its neighbors. This fact will enable us to at least construct a matrix $\boldsymbol{M}$ that will have the desired property:

$$(\boldsymbol{e}_a - \boldsymbol{e}_b)^T \boldsymbol{M}^+ (\boldsymbol{e}_a - \boldsymbol{e}_b) = \mathrm{R}_{\mathrm{eff}\,G}(a, b).$$

Let $\boldsymbol{v}_B$ be the voltages that will be assigned to the vertices in B. We can then write $\boldsymbol{v}(1)$ as a function of $\boldsymbol{v}_B$ by

$$\boldsymbol{v}(1) = \frac{1}{d(1)} \sum_{(1,c) \in E} w_{1,c} \boldsymbol{v}(c) = -(1/\boldsymbol{L}(1,1)) \boldsymbol{L}(1, B) \boldsymbol{v}_B.$$

We thus want the matrix $\boldsymbol{M}$ for which

$$\boldsymbol{v}_B^T \boldsymbol{M} \boldsymbol{v}_B = \begin{pmatrix} -(1/\boldsymbol{L}(1,1)) \boldsymbol{L}(1, B) \boldsymbol{v}_B \\ \boldsymbol{v}_B \end{pmatrix}^T \boldsymbol{L} \begin{pmatrix} -(1/\boldsymbol{L}(1,1)) \boldsymbol{L}(1, B) \boldsymbol{v}_B \\ \boldsymbol{v}_B \end{pmatrix}.$$

We will see in a moment that the matrix $\boldsymbol{M}$ is exactly the matrix one obtains when using Gaussian elimination to eliminate vertex 1 in $\boldsymbol{L}$, and that this matrix is a Laplacian! In the following, I will write $\boldsymbol{L}(B,B)$ to indicate the submatrix of $\boldsymbol{L}$ indexed by rows and columns of B. We have

$$\begin{aligned}
&\begin{pmatrix} -(1/\boldsymbol{L}(1,1))\boldsymbol{L}(1,B)\boldsymbol{v}_B \\ \boldsymbol{v}_B \end{pmatrix}^T \boldsymbol{L} \begin{pmatrix} -(1/\boldsymbol{L}(1,1))\boldsymbol{L}(1,B)\boldsymbol{v}_B \\ \boldsymbol{v}_B \end{pmatrix} \\
&= \boldsymbol{v}_B^T \boldsymbol{L}(B,B)\boldsymbol{v}_B + \boldsymbol{L}(1,1)\left(-(1/\boldsymbol{L}(1,1))\boldsymbol{L}(1,B)\boldsymbol{v}_B\right)^2 + 2\boldsymbol{v}(1)\boldsymbol{L}(1,B)\left(-(1/\boldsymbol{L}(1,1))\boldsymbol{L}(1,B)\boldsymbol{v}_B\right) \\
&= \boldsymbol{v}_B^T \boldsymbol{L}(B,B)\boldsymbol{v}_B + \left(\boldsymbol{L}(1,B)\boldsymbol{v}_B\right)^2 / \boldsymbol{L}(1,1) - 2\left(\boldsymbol{L}(1,B)\boldsymbol{v}_B\right)^2 / \boldsymbol{L}(1,1) \\
&= \boldsymbol{v}_B^T \boldsymbol{L}(B,B)\boldsymbol{v}_B - \left(\boldsymbol{L}(1,B)\boldsymbol{v}_B\right)^2 / \boldsymbol{L}(1,1).
\end{aligned}$$

Let $\boldsymbol{M}$ be the matrix that realizes this quadratic form. We can see that

$$\boldsymbol{M} = \boldsymbol{L}(B,B) - \boldsymbol{L}(B,1)\boldsymbol{L}(1,B)/\boldsymbol{L}(1,1).$$

Claim 9.6.1. *The matrix $\boldsymbol{M}$ is a Laplacian matrix.*

Proof. We first observe that all of the off-diagonal entries on $\boldsymbol{M}$ are non-positive. This is clearly true of $\boldsymbol{L}(B,B)$. And, the matrix that we subtract is non-negative. To show that $\boldsymbol{M}$ is a Laplacian matrix, it now suffices to show that all of its row-sums are 0. As the row-sums in $\boldsymbol{L}$ are zero, we know that

$$\boldsymbol{L}(B,B)\mathbf{1}_B + \boldsymbol{L}(B,1) = \mathbf{0}.$$

So,

$$\left(\boldsymbol{L}(B,B) - \boldsymbol{L}(B,1)\boldsymbol{L}(1,B)/\boldsymbol{L}(1,1)\right)\mathbf{1}_B = -\boldsymbol{L}(B,1) - \boldsymbol{L}(B,1)(\boldsymbol{L}(1,B)\mathbf{1}_B)/L(1,1) = \mathbf{0}_B,$$

as $\boldsymbol{L}(1,B)\mathbf{1}_B = -d(1) = -L(1,1)$. □

We now observe that $\boldsymbol{M}$ is the matix we obtain when we use Gaussian elimination to eliminate the entries in the first column of $\boldsymbol{L}$. To eliminate a non-zero in the $(c,1)$ entry of $\boldsymbol{L}$, we subtract the first row of $\boldsymbol{L}$ times $\boldsymbol{L}(c,1)/\boldsymbol{L}(1,1)$. So, the rows in B after this elimination look like

$$\boldsymbol{L}(B,:) - (\boldsymbol{L}(B,1)/\boldsymbol{L}(1,1))\boldsymbol{L}(1,B).$$

The submatrix on the columns in B is exactly $\boldsymbol{M}$.

9.7 Eliminating Many Vertices

We can of course use the same procedure to eliminate many vertices. We begin by partitioning the vertex set into *boundary* vertices B and *internal* vertices I. We can then use Gaussian elimination to eliminate all of the internal vertices. This will result in a graph $\widehat{G}$ on the boundary vertices such that the effective resistances between all pairs of vertices in B are the same in G as in $\widehat{G}$.

I'd like to take a moment to see exactly what this produces. Given a setting of voltages $\boldsymbol{v}_B$ on vertices in B, the voltages in I should be assigned so that the voltage at each node is the appropriate

average of the voltages as its neighbors. That is, when we take the entire vector of voltages and multiply it by the rows of $\boldsymbol{L}$ in I, we should get the zero vector. We want

$$\boldsymbol{L}(I,I)\boldsymbol{v}_I + \boldsymbol{L}(I,B)\boldsymbol{v}_B = \mathbf{0}_I.$$

Provided that the induced subgraph on the vertices in I is connected, the argument we stated at the end of last lecture we proves that $\boldsymbol{L}(I,I)$ is invertible. So, we can set

$$\boldsymbol{v}_I = -\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B.$$

If we re-arrange the vertices so that those in I come first, we can write the quadratic form we obtain in $\boldsymbol{v}_B$ as

$$\begin{aligned}
&\begin{pmatrix} -\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B \\ \boldsymbol{v}_B \end{pmatrix}^T \begin{pmatrix} \boldsymbol{L}(I,I) & \boldsymbol{L}(I,B) \\ \boldsymbol{L}(B,I) & \boldsymbol{L}(B,B) \end{pmatrix} \begin{pmatrix} -\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B \\ \boldsymbol{v}_B \end{pmatrix}^T \\
&= \boldsymbol{v}_B^T\boldsymbol{L}(B,B)\boldsymbol{v}_B + (\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B)^T\boldsymbol{L}(I,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B - 2\boldsymbol{v}_B^T\boldsymbol{L}(B,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B \\
&= \boldsymbol{v}_B^T\boldsymbol{L}(B,B)\boldsymbol{v}_B + \boldsymbol{v}_B^T\boldsymbol{L}(B,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B - 2\boldsymbol{v}_B^T\boldsymbol{L}(B,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B \\
&= \boldsymbol{v}_B^T\boldsymbol{L}(B,B)\boldsymbol{v}_B - \boldsymbol{v}_B^T\boldsymbol{L}(B,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{v}_B.
\end{aligned}$$

The matrix that realizes this quadratic form,

$$\boldsymbol{L}(B,B) - \boldsymbol{L}(B,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B),$$

is called the *Schur complement* of $\boldsymbol{L}$ with respect to the vertices in I. It is also the Laplacian of the graph we want on the vertices in B.

It is easy to show that this matrix is in fact a Laplacian. I claimed last lecture (and have put a proof in the lecture notes) that every entry of the matrix $\boldsymbol{L}(I,I)^{-1}$ is positive. So, every entry of $\boldsymbol{L}(B,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)$ is positive. This tells us that all of the off-diagonal entries of the Schur complement are non-positive. To prove that this matrix is a Laplacian, it just remains to show that its row-sums are 0.

To this end, I first compute

$$L(I,I)^{-1}\boldsymbol{L}(I,B)\mathbf{1}_B.$$

Well, computing this is a pain. Pure thinking works better. If we set the voltage of every vertex in B to 1, then the voltage of every vertex in I must also be 1. This is a consequence of every vertex in I having a voltage equal to the average of its neighbors. So,

$$\left(\boldsymbol{L}(B,B) - \boldsymbol{L}(B,I)\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\right)\mathbf{1}_B = \boldsymbol{L}(B,B)\mathbf{1}_B - \boldsymbol{L}(B,I)\mathbf{1}_I = \boldsymbol{L}(B,:)\mathbf{1}_V = \mathbf{0}_B.$$

9.8 Reconstructing Trees

Let T be a tree, let B be the set of leaf vertices, and let I be the internal vertices. The induced graph on the internal vertices is connected.

Given all the effective resistances between pairs of boundary vertex B, we now know how to construct a Laplacian matrix $\widehat{\boldsymbol{L}}$ with vertex set B that produces these effective resistances. From the discussion in the previous section, we know that this Laplacian is the Schur complement of the Laplacian of the tree with respect to the internal vertices. Our task is to now reconstruct this tree.

This is a type of problem that is usually very ill-conditioned: given the effective resistances between the points of a structure it is very difficult to reconstruct that structure. Small errors in your measurements can lead to very large reconstruction errors. I think the case is better for trees.

Let m be the number of leaves in B, let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $\widehat{\boldsymbol{L}}$, and let $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ be the corresponding eigenvectors[1]. I claim that if $\boldsymbol{\psi}_2$ is never zero, then we should consider dividing the leaves into two sets based on whether they are assigned positive or negative values by $\boldsymbol{\psi}_2$. This division corresponds to breaking exactly one edge of the tree. Once we have divided the tree into two parts by cutting an edge, we can reconstruct the subtrees recursively. We will prove most of these statements now.

To begin, again order the vertices in the tree so that those in I come first. Let's see what an eigenvector of $\widehat{\boldsymbol{L}}$ looks like in the whole tree. By setting $\boldsymbol{v}_I = -\boldsymbol{L}(I,I)^{-1}\boldsymbol{L}(I,B)\boldsymbol{\psi}_2$, we find a vector that satisfies the equation

$$\begin{pmatrix} \boldsymbol{L}(I,I) & \boldsymbol{L}(I,B) \\ \boldsymbol{L}(B,I) & \boldsymbol{L}(B,B) \end{pmatrix} = \lambda_2 \begin{pmatrix} \boldsymbol{0}_I \\ \boldsymbol{\psi}_2 \end{pmatrix}.$$

We will now analyze this by using the technique of Fiedler from Lecture 9. We begin by constructing the matrix

$$\boldsymbol{M} \stackrel{\text{def}}{=} \boldsymbol{V} \begin{pmatrix} \boldsymbol{L}(I,I) & \boldsymbol{L}(I,B) \\ \boldsymbol{L}(B,I) & \boldsymbol{L}(B,B) - \lambda_2 \boldsymbol{I} \end{pmatrix} \boldsymbol{V},$$

where $\boldsymbol{V}$ is the diagonal matrix with $\boldsymbol{v}_I, \boldsymbol{\psi}_2$ on the diagonal.

As in Lecture 7, we can show that the matrix $\boldsymbol{M}$ can be expressed as sum of elementary edge-Laplacians, and that the number of negative terms in this sum equals the number of sign changes across edges, which equals the number of negative eigenvalues of $\boldsymbol{M}$.

By Sylvester's Law of Intertia, we know that the number of negative eigenvalues of $\boldsymbol{M}$ equals the number of negative eigenvalues of

$$\begin{pmatrix} \boldsymbol{L}(I,I) & \boldsymbol{L}(I,B) \\ \boldsymbol{L}(B,I) & \boldsymbol{L}(B,B) - \lambda_2 \boldsymbol{I}_B \end{pmatrix}.$$

We will show that this matrix has exactly one negative eigenvalue.

To do this, consider the family of matrices parameterized by λ:

$$\begin{pmatrix} \boldsymbol{L}(I,I) & \boldsymbol{L}(I,B) \\ \boldsymbol{L}(B,I) & \boldsymbol{L}(B,B) - \lambda \boldsymbol{I}_B \end{pmatrix}.$$

We know that the eigenvalues of a matrix are a continuous function of its entries. When $\lambda = 0$, we have the matrix $\boldsymbol{L}$ which has one zero eigenvalue and $n-1$ positive eigenvalues. As we start to

[1]Note that $\boldsymbol{\psi}_2$ is actually the dominant eigenvector of $\widehat{\boldsymbol{L}}^+$, which is the matrix we actually construct from the distances.

subtract multiples of $\boldsymbol{I}_B$, the eigenvalues shift down. So, for small enough λ there will be exactly one negative eigenvalue. I mainting that this is true up until $\lambda = \lambda_2$.

To see this, note that the number of negative eigenvalues cannot increase until one of them crosses the origin. At this point, the matrix has determinant zero. To evaluate the determinant, recall that row-operations do not change the determinant. We perform the row operations used to obtain the Schur complement. These give the matrix

$$\begin{pmatrix} \boldsymbol{L}(I,I) & \boldsymbol{L}(I,B) \\ 0_{B,I} & \widehat{\boldsymbol{L}} - \lambda \boldsymbol{I}_B. \end{pmatrix}$$

The determinant of this matrix is the product of the determinant of $L(I,I)$ with $\widehat{\boldsymbol{L}} - \lambda \boldsymbol{I}_B$. We know that the matrix $\boldsymbol{L}(I,I)$ is positive definite, so its determinant is positive. The determinant of $\widehat{\boldsymbol{L}} - \lambda \boldsymbol{I}_B$ is zero when $\lambda = \lambda_2$, and non-zero for λ between 0 and λ_2. So, no eigenvalue crosses the origin until λ hits λ_2.

I have neglected on thing in the argument above. The argument that I just gave works provided that no entry of $\boldsymbol{\psi}_2$ or $\boldsymbol{v}_I$ is zero. Assuming that every internal node of our tree has degree at least 3, we could show that there is a zero entry in $\boldsymbol{v}_I$ if and only if there are zero entries in $\boldsymbol{\psi}_2$. We can even do more than that, and I might put these statements on a future problem set.

References

[SG09] Eric A Stone and Alexander R Griffing. On the fiedler vectors of graphs that arise from trees by schur complementation of the laplacian. *Linear Algebra and Its Applications*, 431(10):1869–1880, Jan 2009.