Graphs and Networks Lecture 5

PageRank

Lecturer: Daniel A. Spielman September 20, 2007

5.1 Intro to PageRank

PageRank, the algorithm reportedly used by Google, assigns a numerical rank to every web page. More important pages get higher rankings. The more in-links a page has, the higher its ranking should be. But, more importantly, a page has a higher rank if it is pointed to by high-rank pages. Low-rank pages influence the rank of a page less. If one page points to many others, it will have less influence on their rankings than if it just points to a few.

To algebraicize this intuitively appealing idea, PageRank treats the web as a directed graph, with web pages as vertices and links as directed edges. The rank of vertex v is denoted $r(v)$, and is supposed to satisfy the formula:

$$r(v) = \sum_{u:(u,v)\in E} r(u)/d_+(u), \tag{5.1}$$

where $d_+(u)$ is the number of edges going out of u. Note that this sum is over edges going in to v.

To express this in matrix form, we let $\boldsymbol{D}_+$ be the diagonal matrix whose uth diagonal is $d_+(u)$. We then $\boldsymbol{A}$ be the directed adjacency matrix of the graph, where $\boldsymbol{A}(v,u) = 1$ if there is an edge from u to v. Yes, I know that this looks backwards. But, it is what I have to do if I want to make $\boldsymbol{r}$ be a column vector.

We then find that $\boldsymbol{r}$ must satisfy the equation

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{D}_+^{-1}\boldsymbol{r}. \tag{5.2}$$

That is to say that $\boldsymbol{r}$ is an eigenvector of eigenvalue 1 of the matrix $\boldsymbol{A}\boldsymbol{D}_+^{-1}$.

However, $\boldsymbol{A}\boldsymbol{D}_+^{-1}$ is **not** a symmetric matrix, and is not in any way similiar to a symmetric matrix. So, some of the eigenvalues of this matrix can be complex, it might not have n eigenvectors, and the eigenvectors it does have can have complex entries. Nevertheless, in this lecture we will show that

1. If the graph has no vertices of out-degree 0, then 1 is an eigenvalue.
2. If the graph is strongly connected, then the eigenvalue 1 has multiplicity 1.
3. If the graph is strongly connected, then the unique solution (5.2) is strictly positive.

Before I go further, I would like to point out that this measure of importance was first suggested in the social network community in the paper by Phillip Bonacich, "Factoring and weighting approaches to status scores and clique identification", Journal of Mathematical Sociology, 1972.

I should also point out that $\boldsymbol{r}$ can be understood as the stable distribution of the directed random walk on the graph G. But, random walks on directed graphs are more complicated than on undirected graphs.

5.2 Eigenvalue 1

Set

$$\boldsymbol{M} \stackrel{\text{def}}{=} \boldsymbol{A}\boldsymbol{D}_+^{-1}$$

Lemma 5.2.1. *If G has no vertices of out-degree 0, then 1 is an eigenvalue of $\boldsymbol{M}$.*

Proof. If G has no vertices of out-degree 0, then every column of $\boldsymbol{A}$ has at least one non-zero entry. In fact, the uth column of $\boldsymbol{A}$ has $d_+(u)$ non-zero entries, so the uth column of $\boldsymbol{A}\boldsymbol{D}_+^{-1}$ has sum 1. This implies that

$$\mathbf{1}\boldsymbol{M} = \mathbf{1},$$

and so $\boldsymbol{M}$ has an eigenvector of eigenvalue 1. □

This is similar to the undirected case–in both cases the walk matrix has the vector $\mathbf{1}$ as a left-eigenvector. However, it differs in that we do not know any simple expression for the corresponding right-eigenvector, $\boldsymbol{r}$.

Lemma 5.2.2. *If G is strongly connected, then the eigenvalue 1 has multiplicity 1. In particular, if*

$$\boldsymbol{v}\boldsymbol{M} = \boldsymbol{v},$$

then we must have $\boldsymbol{v} = c\mathbf{1}$ for some constant c.

The proof of this is similar to the proof in the undirected case, so we will skip it.

5.3 $\boldsymbol{r}$ is positive

Lemma 5.3.1. *If G is strongly connected and if the solution of (5.2) is non-negative, then it is positive.*

Proof. First, note that the solution to (5.2) cannot be the all-zero vector. So, if it is non-negative, it must have at leave one positive coordinate. So, assume that $\boldsymbol{r}(z) > 0$. Now, let v be any node

that z points to. Equation (5.1) tells us that

$$\begin{aligned}\boldsymbol{r}(v) &= \sum_{u:(u,v)\in E} \boldsymbol{r}(u)/d_+(u) \\ &\geq \boldsymbol{r}(z)/d_+(z) \\ &> 0.\end{aligned}$$

In general, for every node z for which $\boldsymbol{r}(z) > 0$, every node v that z points to must have $\boldsymbol{r}(v) > 0$. Since the graph is strongly connected, we can apply induction to show that $\boldsymbol{r}(v) > 0$ for all $v \in V$. □

Now, we must show that $\boldsymbol{r}$ is non-negative. To do this, we will consider the matrix

$$\boldsymbol{M}^* \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=0}^{n-1} \boldsymbol{M}^i.$$

We need to establish a few properties of $\boldsymbol{M}^*$.

Claim 5.3.2. *If $\boldsymbol{M}\boldsymbol{r} = \boldsymbol{r}$, then $\boldsymbol{M}^*\boldsymbol{r} = \boldsymbol{r}$. Similarly, $\mathbf{1}\boldsymbol{M}^* = \mathbf{1}$.*

Claim 5.3.3. *The matrix $\boldsymbol{M}^*$ has no negative or zero entries.*

Proof. As $\boldsymbol{M}$ is non-negative, it follows immediately that $\boldsymbol{M}^*$ is non-negative. To show that $\boldsymbol{M}^*$ has no zero entries, note that $\boldsymbol{M}^t(b,a)$ is equal to the probability that a random walk starting at a hits b in exactly t time steps. As the graph is strongly connected, for every pair of vertices a and b, there is some t less than n for which this probability is non-zero (you may prove this in the same way you proved 5.3.1). As $\boldsymbol{M}^*(b,a)$ is the average of these probabilities for t between 0 and n, it is non-zero as well. □

Theorem 5.3.4. *The equation $\boldsymbol{M}\boldsymbol{r} = \boldsymbol{r}$ has a non-negative solution.*

Proof. We will show that it has a solution in which all the signs are the same, which implies that it has a non-negative solution (flip all signs if necessary). But, we will work with the matrix $\boldsymbol{M}^*$, which we proved also satisfies

$$\boldsymbol{M}^*\boldsymbol{r} = \boldsymbol{r}. \tag{5.3}$$

Assume by way of contradiction that $\boldsymbol{r}$ is not sign-uniform. That is, that $\boldsymbol{r}$ has both positive and negative entries. We will use the fact that if $\boldsymbol{x}$ is some vector with both positive and negative entries, then

$$\left|\sum_u \boldsymbol{x}(u)\right| < \sum_u |\boldsymbol{x}(u)|.$$

From equation (5.3), we have that for all u,

$$\boldsymbol{r}(u) = \sum_v \boldsymbol{M}^*(v,u)\boldsymbol{r}(v),$$

and so

$$|\boldsymbol{r}(u)| = \left| \sum_v \boldsymbol{M}^*(v,u)\boldsymbol{r}(v) \right|.$$

As we have assumed that $\boldsymbol{r}$ is not sign-uniform, and $\boldsymbol{M}^*(v,u)$ is always positive, we have the inequality

$$\left| \sum_v \boldsymbol{M}^*(v,u)\boldsymbol{r}(v) \right| < \sum_v \boldsymbol{M}^*(v,u) \left|\boldsymbol{r}(v)\right|,$$

which implies

$$|\boldsymbol{r}(u)| < \sum_v \boldsymbol{M}^*(v,u) \left|\boldsymbol{r}(v)\right|.$$

If we now sum over all u, we get

$$\begin{aligned}
\sum_u |\boldsymbol{r}(u)| &< \sum_u \sum_v \boldsymbol{M}^*(v,u) \left|\boldsymbol{r}(v)\right| \\
&= \sum_v \sum_u \boldsymbol{M}^*(v,u) \left|\boldsymbol{r}(v)\right| \\
&= \sum_v \left|\boldsymbol{r}(v)\right| \sum_u \boldsymbol{M}^*(v,u) \\
&= \sum_v \left|\boldsymbol{r}(v)\right|,
\end{aligned}$$

as $\mathbf{1}\boldsymbol{M}^*(v,u) = \mathbf{1}$ is equivalent to

$$\sum_u \boldsymbol{M}^*(v,u) = 1.$$

But, we have derived a contradiction. □

5.4 Closer to PageRank

Brin and Page tell us that they don't actually take A to be the original web graph. Rather, they consider a "random surfer" who actually jumps to a random web page with some fixed probability at each time step. We can model this by including an edge between all pairs of vertices, giving that edge low weight. Since we haven't discussed weighting edges yet, let me instead say that this is equivalent to forcing $\boldsymbol{r}$ to satisfy the equation

$$((1-\alpha)\boldsymbol{M} + \alpha\boldsymbol{J}/n)\boldsymbol{r} = \boldsymbol{r}, \tag{5.4}$$

where α is the probability of jumping to a random web page at any moment, and $\boldsymbol{J}$ is the all-1s matrix.

This equation is actually much nicer than the original. First of all, it gives us an all-positive matrix. So, we know that the solution will be all positive. It also eliminates the issue of nodes with no out-edges.

If we decide that we are going to normalize $\boldsymbol{r}$ so that $\mathbf{1}\boldsymbol{r} = 1$, then we have that

$$\boldsymbol{J}\boldsymbol{r} = \mathbf{1},$$

so equation (5.4) becomes

$$(1-\alpha)\boldsymbol{M}\boldsymbol{r} + (\alpha/n)\mathbf{1} = \boldsymbol{r},$$

which is equivalent to

$$(\alpha/n)\mathbf{1} = (\boldsymbol{I} - (1-\alpha)\boldsymbol{M})\boldsymbol{r},$$

and

$$\left((\boldsymbol{I} - (1-\alpha)\boldsymbol{M})\right)^{-1}(\alpha/n)\mathbf{1} = \boldsymbol{r}.$$

That is, $\boldsymbol{r}$ is now given by the solution to a system of linear equations.

Even better, we can solve these equations quickly. We have that

$$\left(\boldsymbol{I} - (1-\alpha)\boldsymbol{M})\right)^{-1} = \sum_{t=0}^{\infty}\left((1-\alpha)\boldsymbol{M}\right)^t.$$

(This is just like the formula you learned for $1/(1-x)$, but for matrices. It is true as long as the sum converges). Moreover, this sum converges very quickly. Brin and Page suggest using $\alpha = .15$. We know that every entry of $\boldsymbol{M}^t$ is at most 1, so every entry of $((1-\alpha)\boldsymbol{M})^t$ is at most 0.85^t, which becomes small very quickly as we increase t. So, we can quickly approximate $\boldsymbol{r}$ by using the first few terms from this series.