MATH 167 PROJECT November 14th 2015

PageRank Computation in Google

When I query "Applied Linear Algebra" on the google search bar, the first two responses are:

Applied Linear Algebra: Lorenzo Sadun: 9780821844410 ...
www.amazon.com ... Algebra Linear Amazon.com, Inc.

and

Linear algebra - Wikipedia, the free encyclopedia
en.wikipedia.org/ wiki/ Linear_algebra Wikipedia

Google search deemed the listing on Amazon "best" and Wikipedia next. When I used the same keywords on the Yahoo search bar, the two responses are reversed in order of importance.

For this project, we will use a component of the PageRank algorithm developed by Page and Brin to rank the connectness and thus the importance of the webpages of persons 1 to n. Consider yourself to have the webpage denoted by 1 and let $n = 5, 10, 100, 1000, 10000, 100000$

Let us consider the directed graph where each number represents a webpage and the edge from i to j corresponds to the link from webpage i to webpage j. Let G be the (n x n) adjacency matrix of this graph. That is, if webpage i is connected to webpage j then $g_{ij} = 1$, otherwise $g_{ij} = 0$. If the matrix contains all 1's, then all of the webpages are connected to all the other webpages. If the matrix only has 1's along the diagonal and zeros everywhere else, then each webpage has only a link to itself. For n large, special numerical techniques may be needed.

Now, let c_j and r_i be the column and row sums of G; that is

$$c_j = \sum_{1 \le i \le n} g_{ij} \quad \text{and} \quad r_i = \sum_{1 \le j \le n} g_{ij}$$

Let p be the fraction of time that a websurfer at webpage i follows one of the links available on that page, and (1-p) is the fraction of time that a surfer jumps to a random webpage chosen uniformly. Choose $p = 0.85$ initially and change it to see how it's response are reflected in the eigenvalues and eigenvectors once your program is working.

The transition matrix $\mathbf{M}$ is thus

$$m_{ij} = p\left(\frac{g_{ij}}{r_i}\right) + \frac{(1-p)}{n}$$

where m_{ij} is the probability that we visit page j from your current location of page i.

A nonnegative normalized left eigenvector that satisfies $\vec{v}M = \vec{v}$ is called a steady-state vector. The theory of Markov chains shows that such a vector exists. At this point, you should show that you have a Markov process and using Perron's theorem show that such a solution exists starting from any state and this solution is unique. The limiting probaility of visiting page i at time t as $t \to \infty$ is v_i, the ith component of $\vec{v}$. So the largest element of $\vec{v}$ has the highest PageRank.

There are many ways to obtain $\vec{v}$ which is the eigenvector corresponding to the eigenvalue 1 (show). One way is to write $M^T\vec{w} = \vec{w}$, where $\vec{w} = \vec{v}^T$ is the right eigenvector of M^T.(show)

For n large, use the power method to obtain $\vec{w}$. To do this, first let the initial guess for $\vec{w}$ be a random vector denoted by $\vec{z}^{\,0}$ and let $\vec{w}^{\,0} = \frac{\vec{z}^{\,0}}{||\vec{z}^{\,0}||}$ be the corresponding normalized vector. Then compute $M^T\vec{w}^{\,0} = \vec{z}^{\,1}$, and let $\vec{w}^{\,1} = \frac{\vec{z}^{\,1}}{||\vec{z}^{\,1}||}$ and so on. In general $M^T\vec{w}^{\,k} = \vec{z}^{\,k+1}$, where $\vec{w}^{\,k+1} = \frac{\vec{z}^{\,k+1}}{||\vec{z}^{\,k+1}||}$ for $k = 1, 2, 3, \cdots$. This iteration should converge to the eigenvector corresponding to the largest eigenvalue of M^T. (Prove that the power method indeed converges to the eigenvector corresponding to the largest eigenvalue (1 in this case).

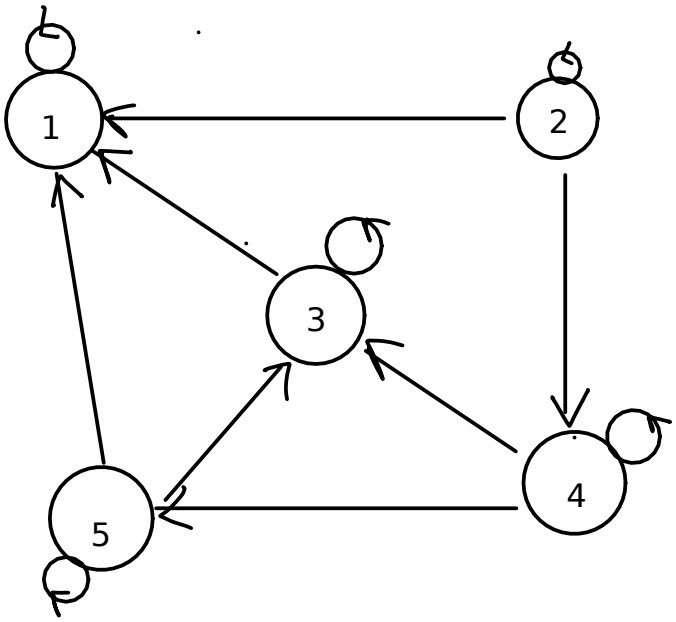
If you want to verify that the corresponding eigenvalue converges to 1, add the following to your program: $\lambda_1^{(m)} = \frac{\vec{z}_k^{(m)}}{\vec{w}_k^{(m-1)}}$ where the subscript k refers to the k^{th} component of the vector, and m refers to the m^{th} iteration of the power method.

This pagerank alone does not determine the final ranking, and has to be combined with other algorithms such as text-matching to determing the final ranking.

(A) test your program on this (see attached) small network of web pages.

(B) Analyze the solution, by changing the links, the indegree c_j and the outdegree r_i, the value of p, the size of n and so on.

Duedate *Friday, December 4th* 2015 *10am before class*



Each node is linked to itself.
Arrows indicate the direction. For example, node 5 has a link to node 1, but node 1 does not have a line to node 5.