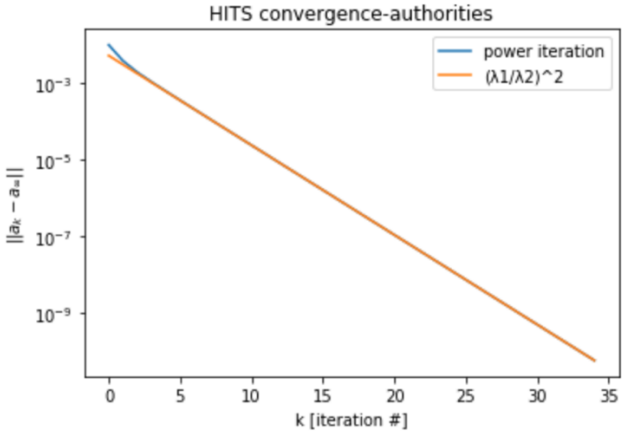
Orkhan Bayramli
Network Science
Homework 2

HITS (Hyperlink-Induced Topic Search)

HITS algorithm works in a way that a good hub represents a page that pointed to many other pages, while a good authority represents a page that is linked by many different hubs. In other words, authorities are interpreted as content providers, while hubs are experts. That's why, each node has authority and hub scores. The algorithm performs a series of iterations, each consisting of two basic steps:

- **Authority update**: Update each node's *authority score* to be equal to the sum of the *hub scores* of each node that points to it. That is, a node is given a high authority score by being linked from pages that are recognized as Hubs for information.
- **Hub update**: Update each node's *hub score* to be equal to the sum of the *authority scores* of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

To find a vector of authority scores, we can use the power iteration method in which we will be finding the eigenvector and eigenvalue of the $M = A^T A$ matrix by iterating and finding the converged eigenvector. At each step, the vector a is multiplied by the matrix M and gets normalized. When the vector a converges, we find the eigenvector of M which contains the authority scores.



In this graph, we are getting the errors which are the differences between a_t and a_{t+1} in each iteration. According to the formula, we are reaching the final authority score vector in the case of:

$$\left|\left|a_t - a_{t+1}\right|\right|_2 \leq 2\sqrt[2]{N}\left(\frac{\lambda_2}{\lambda_1}\right)^t$$

The blue line represents the error in the form of:

$$\frac{\left|\left|a_t - a_{t+1}\right|\right|_2}{2\sqrt[2]{N}}$$

As we can see, from the, approximately, 3rd iteration we have got our final authority vector compared to the second eigenvalue of matrix.

PageRank

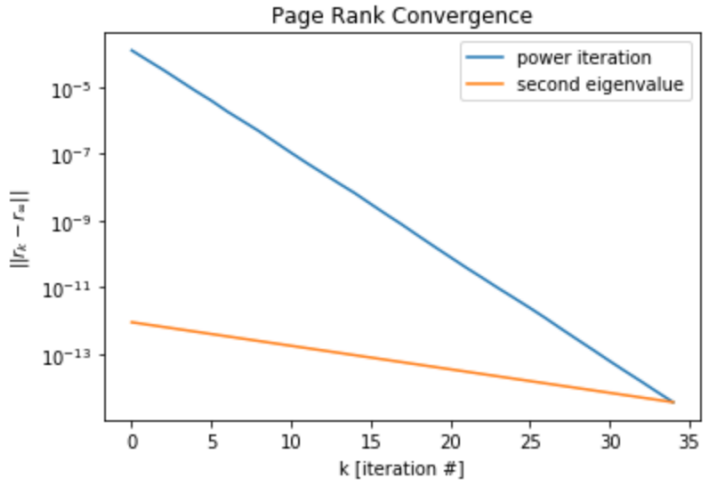
PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. We first beginning with creating a sparse matrix which contains PageRank scores according to their out-degrees. Then according to the following formulas, we find the PageRank scores vector through the damping factor and vector q with all probabilities equal to 1/N.

$$M_1 = cM + (1 - c)q$$
$$p_{t+1} = M_1 p_t$$

Another method is using the power iteration method which requires updating vector p at each iteration till it converges as in the case of HITS but with the damping factor and teleportation vector.

$$p_{t+1} = cMp_t + (1 - c)q$$

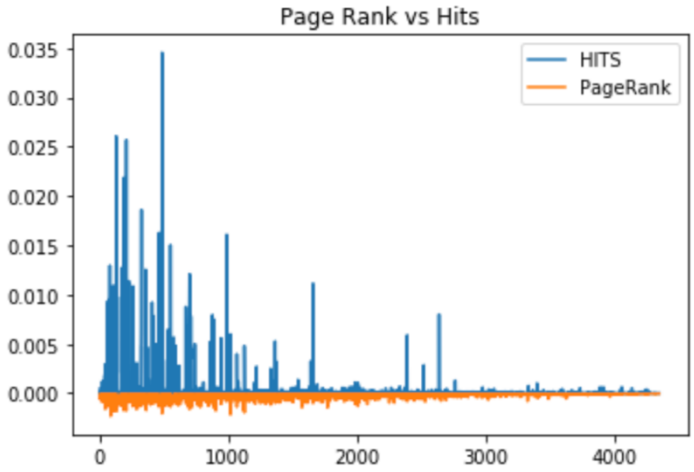
In this formula, we are taking into account that the user will land to the nodes with the probability of c but also on the random nodes with the probability of 1 – c.



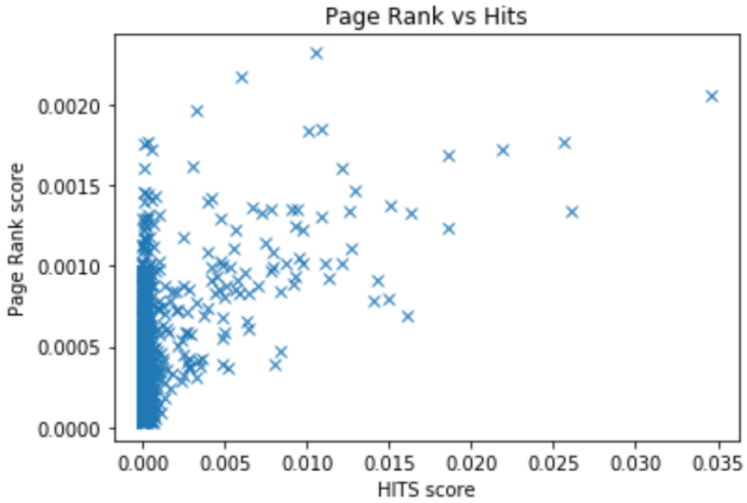
As we can see from the graph, we can state that the PageRank score vector will converge in, approximately, 35 iterations.

PageRank vs HITS

After removing the dead ends and cleaning the network, we are left with the 4352 nodes which are shown along the x axis of graph. On the other hand, we can see the HITS scores in the positive and PageRank scores in the negative interval. The PageRank score can be lower compared to the HITS ones. However, it is possible to observe the slightly matching, more visually like power-law distribution, shape of two different algorithms.



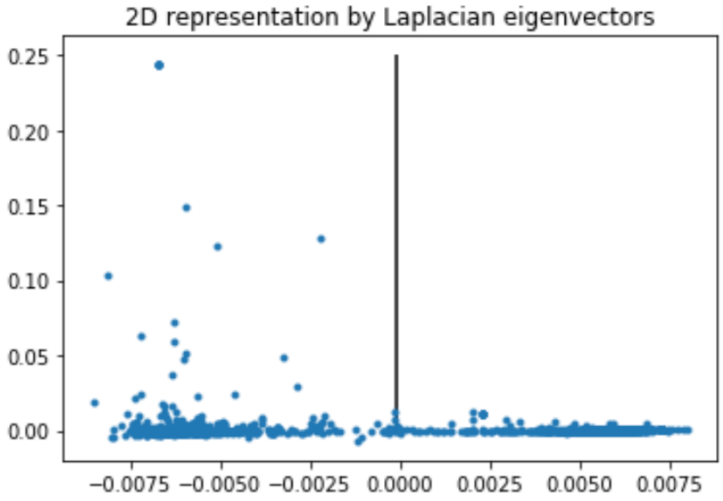
Another graph also represents the comparison of PageRank and HITS. It also proves that the majority of ranking scores obtained through the HITS for the same nodes are higher.



Spectral Clustering

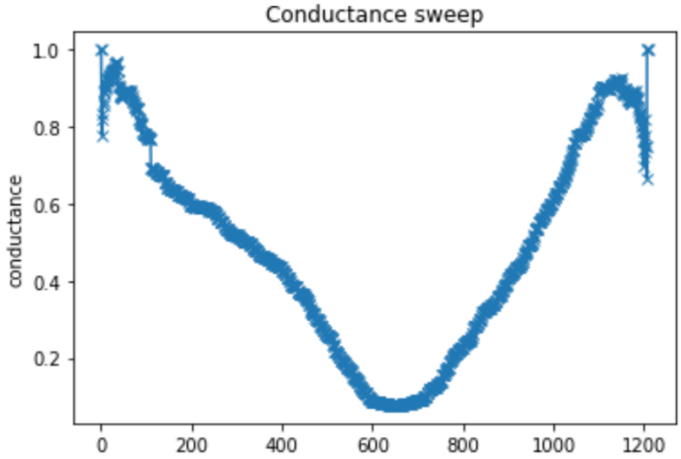
Spectral clustering is the technique for partitioning the data using the eigenvector and eigenvalue of the Laplacian matrix. The main idea is to maximize the number of within-cluster connections and minimize the number of between-cluster connections. In order to do this, the conductance score which determines how well is the partitioning is taken into account. To begin, Laplacian

matrix should be formulated. Laplacian matrix is a matrix whose diagonal consists of the degrees of nodes and the rest shows the connection between nodes being marked as {-1, 0}. Then, the second smallest eigenvalue and corresponding eigenvector of this Laplacian matrix is chosen.



$$\lambda_2 = min \frac{\sum_{(i,j)\in E}(x_i - x_j)^2}{\sum_i x_i^2}$$

According to the above formula, eigenvectors are balanced to 0. Moreover, we are looking for the partition that has lowest distance between the nodes within either right or left side of graph. The following graph shows the conductance of each node in the network.



$$\phi(A) = \frac{|\{(i,j) \in E; i \in A, j \notin A\}|}{\min(vol(A), 2m - vol(A))} = \frac{cut(A)}{vol(A)}$$

We can derive the result that somewhere between the nodes 600 and 700 the conductance is the lowest and it is our partitioning point.