AuthorEchelon Algorithm for Erdos1 Authors

March 23, 2014

ABSTRACT

We analyze coauthor and article networks populated by authors with an Erdos number of 1, and articles linked by citation respectively. We develop an algorithm to rank the authors in terms of relative influence, taking into account the quality of published works. When the only relationship between authors is coauthorship, only general centrality measures can be readily applied to attempt to distinguish influential network members. Simple models which only consider node degree measures fail to provide a purposeful way to weight links, limiting the detail of the analysis. We present a method to rank coauthors by first constructing an article network composed of articles written by members of the coauthor network. We require that the articles cite other papers in the network so that relative article rankings can be calculated.These article rankings are used to determine link weights in a symmetric coauthor(coendorsement) network, which can be thought of as the work quality of an author.Weighting coauthor links in this manner treats endorsement by a prominent author as more valuable than endorsement by a lesser one. With this furniture in place, a variation of a pagerank algorithm is applied to the coauthor network. We find the Perron-Frobenious eigenvector of the matrix composed of authorrank coefficients (quality weights), whose elements when normalized to one are the author ranks of the coauthor network. We discuss conclusions concerning influential members of the Erdos1 subnetwork and general applicability of this ranking method to other types of networks. We conclude with suggestions for further modifications to this approach, such as treating Erdos like a random web surfer in the Erdos1 network, where the pagerank damping coefficient is the probability that Erdos will coauthor with you given that he pays you a visit.

Keywords: Network, Node, Target, Centrality, Heimdall Node
Software Used: MATLAB, R

Contents

1 Introduction

In the digital world we live in today, with the constant flow of information between people, businesses, countries, etc. made possible by the leaps and bounds we have made in technology, analyzing and visualizing these webs of interactions has recently become a hot topic in mathematics. The ability to study and quantify relationships, whether it be in biological processes, social media, or the publishing network of mathematicians can reveal many exciting things, and the advancement in technology has also aided us in this area. Combining mathematics with this new technology, our aim is to construct an algorithm to determine the most influential mathematician in what we have labeled as the ”Erdos1” network.

There are three main components of this project: the construction and analysis of the Erdos1 network, the construction of a test set of published mathematical articles and the development of ArticleImport, a variation of a page-rank algorithm to rank each of these according to the number of citations and structure of the network, and the construction of a sub-network of the Erdos1 network and completion of our AuthorEchelon algorithm, which builds upon the algorith constructed in the previous component. AuthorEchelon takes the ranks created in component two and factors these into its ranking of the authors in the Erdos1 network, accounting for both importance of research and connectedness within the network. Before continuing futher, our vocabulary must be established, as it will be seen throughout the rest of the text.

Node, Edge: Each element of a network will be referred to as a node, with the connections between nodes being the edges.
Centrality: The centrality of a vertex measures its relative importance within a graph. A centrality index is a real-valued function on the nodes of a graph.
Betweenness Centrality: The measure of a node’s cenrality in a network, which equals the number of shortest paths from all vertices to all others that pass through that particular node. A more rigorous mathematical definition is that the betweenness centrality is g(v) = $\sum_{S\neq V\neq t}\frac{r_{St}(V)}{r_{St}}$ where r_{St} is the total number of shortest paths from node s to node t, and $r_{St}(V)$ is the number of those paths which pass through node V.
Closeness Centrality: The average shortest distance from node i to every other node. This value is lower for vertices which are more central in the graph.
Stress Centrality: The stress of a node is the number of shortest paths that travel through that node.
Heimdall (Gatekeeper) Node: We say a node V is a Heimdall node if for some other two nodes X and Y, every path from X to Y passes through X. The stress and betweenness centrality measures are good indicators of which nodes are Heimdall nodes.

2 Building the Coauthor Network of Erdos1 Authors

Supplied with the list of Erdos authors and being prompted to create a network of said authors, part one of the project was divided into three sections: data extraction, data formatting and generation of network visualization. For data extraction and formatting, RStudio was used while Cytoscape was utilized for creating a graphical user interface and performing networking analysis.

2.1 Tools Used for Data Extraction and Network Visualization

2.1.1 RStudio

RStudio is a free and open source software used mainly for statistical/data analysis. It is useful for sifting through large data sets and creating usable data structures for modeling. This software was chosen because of familiarity and suitability for the task at hand.

2.1.2 Cytoscape

Cytoscape is also an open sourceware for visualizing data networks. While originally intended more for biological systems, it can be used for any general network ($www.cytoscape.org/what_iscytoscape.html$). In addition to the base Cytoscape application, an application named Centiscape was downloaded. This application calculates various influence measures based on the structure of a network (defined below). Cytoscape appears to be a well-maintained site, reliable in its operation. The standards of centrality measures are well-defined through it, and the visualization options aid in understanding the details of a given network.

2.2 Extracting the Erdos1 Subnetwork

Data was supplied via a website or included by the COMAP competition and consisted of over 18,000 lines of data, including names of authors and coauthor relationships. An example of a few lines of the raw data format is:

ABBOTT, HARVEY LESLIE 1974
Aull, Charles E.
Brown, Ezra A.
Dierker, Paul F.
Exoo, Geoffrey

To create a network of authors only with Erdos numbers equal to one, this data set had to be modified with those authors of Erdos number equal to two removed. The function erdosnetwork.R was composed to handle the data extraction and formatting. After various attempts at constructing an appropriate patter using regular expressions in R, it was noticed that all Erdos 1 author names were spelled in all uppercase letters, leading to a relatively easy extraction of the relevant data while maintaining the coauthor relationships. The function takes the data saved as a .txt file and writes the relevant data into a new .txt file formatted into two columns, where horizontal rows represent a coauthor relationship. A sample of the returned data is:

```
> #newdata <- erdosnetwork("Erdos.txt")
> #head(newdata)
```

This data was then loaded into Cytoscape to construct the network and perform analyses via .

2.3 Graphing Erdos1 Subnetwork

Cytoscape accepts Microsoft Excel files with the data formatted in columns, one being the column of source nodes, and the other the corresponding target nodes. Microsoft Excel in turn can accept formatted text files. In this way the data assembled in R is turned into the input for the software to build the network. Once the network has been assembled, Cytoscape offers various options for controlling the directedness of the graph, removing duplicate nodes and edges, and with Centiscape performing network analysis caluclations. Specifically, this software was used to calculate closeness, betweenness, stress, mean average path, and other structural-related influence measures.

2.4 Limitations of this Model

While this network contains all researchers in the Erdos1 set, the only determining factors are those from the inherent structure of the network. Therefore, we concluded that the node most central and well connected in the network was the most influential, with connectedness being determined by summing up the values of each node's closeness, betweenness, and stress. While this model served as a good beginning point, there are many factors to consider when considering the influence of a researcher. The influence was thought of as influence due to connectedness and influence due to importance of research. For example, if a researcher is connected to many others but their research isn't of much worth when thinking about advancing that field, that researcher's influence wouldn't be as high as somebody in a comparable situation with more significant research. A more thorough understanding of all of the structural influence measures, such as eccentricity, clustering, and others could give a more accurate depiction of which influence is most influencial. This graph is also lacking direction, meaning that there is no apparent way to weight how each node will affect the relationship they share. From a practical standpoint, this model is somewhat limited when considering general application to other networks in the fact that the data extraction process requires a precise, network specific algorithm. If the data is presented in any other way than the Erdos1 set was presented, the model would require some alterations to be able to reach the analysis stage. These limitations lead us to consider creating a more thorough model, one which will also take into account research importance.

3 Construction of Article Network 1

4 Pagerank Algorithm for Directed Graphs

4.1 Applicability

The pagerank algorithm is the reason why google was an important player among search engines early on. Instead of simply considering the web as an undirected graph and considering only total node (page) degree, the pagerank takes into consideration the direction of links between web pages and treats links conceptually as votes for the webpages they are directed to. Backlinks to a page are links to that page from other pages. Simply counting backlinks and using those values to rank web pages fails to account for the idea that important pages should contribute more than unimportant pages when when votes are being tallied for a page which they both link to. Pagerank does consider such weighted votes, and is therefore far superior to simpler ranking algorithms. In the context of an article network, one must declare what variable is providing the link between articles. In the case where the goal is to apply knowledge of article rankings to weight links in a coauthorship network, it makes sense to treat citations as backlinks. As with ranking pages, one should avoid simply taking the sum of indegree (Being cited) and outdegree (Citing someone) and treating that as the rank. Doing so would not consider who is citing who, which wouldn't provide very good information about which articles are influential. Thus, it is a better model to divide an articles' rank among the articles it cites, distributing the vote equally among them. This will have the same effect as a basic pagerank algorithm, giving more weight to votes from more influential articles. Of course, this model is still simple in that it assumes equal distribution of vote, which implies that an author relied upon each source equally which is an unlikely event. Setting up the article network in this fashion allows a system of equations to be established which expresses the relationships between article ranks given the outdegree for each network member. Equal distribution does have a nice effect, namely that in the absence of dangling nodes(nodes with zero outdegree) and nodes with zero indegree, the matrix produced is non-negative and column stochastic. That is, each entry is greater than or equal to 0, and the sum of the values in any given column is equal to 1. In fact, this matrix will be guaranteed to have 1 as an eigenvalue. Taking the eigenvector corresponding to eigenvalue 1 and normalizing its' components to 1 gives article rankings which take into account indegree, outdegree, and weighted importance of vote (citation). Difficulties arise when dangling nodes (Articles which do not cite) or nodes with zero indegree (articles which are not cited by anyone in the network) appear. When such things occur, patches can sometimes be made, but one may find an alternative approach is preferable.

4.2 Implementation

Let r_n be the rank of a node n in a directed network. Define $r_n = \sum_{j \epsilon L_n} \frac{r_j}{k_j}$ where $L_n = \{$ nodes with links to node n $\}$ and $k_j =$ number of out citations. . The system of equations produced from this can be written as follows:

$$\begin{gathered}
r_1 = c_{12}r_2 + c_{13}r_3 + ... + c_{1n}r_n \\
r_2 = c_{21}r_1 + c_{23}r_3 + ... + c_{2n}r_n \\
r_3 = c_{31}r_1 + c_{32}r_3 + ... + c_{3n}r_n \\
. \\
. \\
. \\
r_n = c_{n1}r_1 + c_{n2}r_2 + ... + c_{nn}r_n
\end{gathered}$$

This corresponds to a the matrix product Cr where r is the rank column vector and C is the matrix of coefficients for this system. This matrix is column stochastic (the column values sum to one in each time). A column stochastic matrix is guaranteed to have an eigen value of 1. Calculate the eigenvector of C corresponding to eigenvalue 1, normalize the eigenvector to 1, and those are the ranks of the nodes.

4.2.1 Applying Pagerank to Article Network 1

As described above, a small database of articles was compiled, each citing one of the others in the network. There were a total of nine nodes, and eighteen directed edges, and this data was represented by an excel data table which was exported to MATLAB for analysis. Many attempts were made trying to implement the pagerank algorithm to this network, but it was realized that the presence of dangling nodes and those with indegree 0 were throwing a wrench into the algorithm.

4.3 Thorny Issues With Pageranking Articles

As mentioned earlier, dangling nodes and those with zero indegree create problem for this article pagerank algorithm. In the case of dangling nodes, the matrix is no longer column stochastic. Instead, it is column substochastic, meaning the sum of the values in any given column is less than or equal to 1. This removes the guarantee that the matrix will have 1 as an eigenvalue. One may definitely cry alas! However, hope is not yet lost thanks to a striking theorem from linear algebra independently discovered by Oskar Perron and Georg Frobenius in 1907 and 1912 respectively. See below for explanation. Nodes with zero indegree lead to a more troubling problem. In this model, such nodes receive zero votes due to their not being cited. This causes their corresponding rows and columns in the rank coefficient matrix to zero out, making it futile to perform the eigenvector analysis as before. When this situation is encountered, it is probably best to find a way to modify the algorithm, or to develop a scheme to artificially assign equal ranks to such nodes from the onset. This allows for a similar algorithm to be carried out which preserves informative relative rankings of the remaining nodes in the graph. After the ranks are computed, one can then reassign the rank of the zero indegree nodes using an appropriate scheme giving the other rankings. Admittedly, this is not an ideal patch, but it does provide meaningful rankings. Furthermore, it makes sense that articles with zero indegree will have equal or similar rankings given that neither is receiving a citation. They could be interpreted as new editions to the network and therefore have not been around long enough to be cited, or as articles with less than significant value. Thus, their true rank is indeterminate, but should be considered non-zero because they did manage to get published.

4.4 Attempt to Patch Article Pagerank Method

4.4.1 Perron-Frobenius Theorem

5 ArticleImport Algorithm for Ranking Articles

This algorithm was developed when nodes with zero indegree were encountered. It is meant to be implemented instead of the other pagerank variations.

5.1 Implementation

5.2 Circumventing Dangling Nodes and Those with Zero Indegree

This algorithm deals with the presence of nodes(articles) with zero indegree and dangling nodes by assigning the former a rank of 1. Therefore, zero indegree nodes have a unit rank to distribute evenly among the articles they cite.This action uniquely defines the ranks for the rest of the network, thereby giving ranks to the dangling nodes, and allowing the matrix corresponding to the system of equations expressing node relationships to be placed in reduced-row echelon form. As before, this vector when normalized is the set of rankings for the articles, in this case before rank reassignment of zero indegree nodes. The rankings of nodes which were not artificially set to 1 allow determination of which articles are influential in the network. MATLAB was used to create ArticleImport. To create the matrix, MATLAB had to loop through the data in order to create the ordered pairs representing edges between two nodes, then use this to create a matrix representing the system of equations based on the number of edges for each node. Arbitrary values were then assigned to nodes with indegree of zero, the system solved, and the appropriate eigenvector reported as set of ranks of corresponding to each article.

5.3 Conclusions Regarding Influential Nodes in Article Network 1

6 Construction of Article Network 2

In much the same way, a database of articles was created for a subnetwork of Erdos1. The search was laborious, seeking articles citing other articles authored by those within Erdos1 and also having authors connected by coauthorship. The search engine http://citeseerx.ist.psu.edu was utilized in the sesarch, and the resulting article network was one of ten nodes, which then determined our subgroup of Erdos1 as a network of 11 nodes. An attempt was made to build upon an authorrank algorithm in order to better analyze these networks.

7 Using Modified Authorrank Algorithm on Erdos1 Subnetwork

8 Modified Authorrank Algorithm

In *Co-Authorship Networks in the Digital Library Research Community* by Lui, Bollen et al., an algorithm is developed to rank authors within a coauthorship network. In this *authorrank* algorithm, a generalization of google's pagerank algorithm, strength of collaboration between authors is quantified and used an link weights in a symmetric (bidirectional) coauthorship network. Weighting the links in this way aims to weight the vote (endorsement) contributed by an author to a coauthor. Thus, an endorsement by a prominent author is worth more than an endorsement by a less prominent author. This approach is very promising, but there are other options for what can be used to set the link weight in the coauthorship network. If a network of articles is constructed which cite each other, and each article is authored by a member of the coauthorship network, a measure of work quality can be used to weight links between coauthors. Using the ArticleImport algorithm, weights (magnitude of importance) are assigned to all articles in a relative fashion. Then, for each author in the coauthorship network, the weight of its' outward links can be set as the average of the article weights for the articles that author helped write. Thus, an endorsement (coauthoring) from (with) an author who has produced many important articles will contribute more to the "endorsement receiving" author's rank than an endorsement from an author who has produced lower quality articles, or fewer high quality articles. Deciding when to give an author more endorsement to transfer to others based on the quality of his work seems to be preferable to deciding the level of endorsement based on the strength of the collaboration between two authors. MATLAB was used to create ArticleImport. To create the matrix, MATLAB had to loop through the data in order to create the ordered pairs representing edges between two nodes, then use this to create a matrix representing the system of equations based on the number of edges for each node. Arbitrary values were then assigned to nodes with indegree of zero, the system solved, and the appropriate eigenvector reported as set of ranks of corresponding to each article.

8.1 Implementation

Let W_i be the rank of an article i as computed with the ArticleImport algorithm. Let Z_n be the average of the ranks of the W_i's associated with author n. Then, the weight given to the link from author n to author m is Z_n. Take the matrix composed of these new weights which are normalized to one, and find perron frobenius eigenvector. These arew author ranks after normalization to 1.

8.2 Dangling Nodes, Zero Indegree, and Perron-Frobenius Eigenvector Revisited

8.3 Applying Modified Authorrank Algorithm to Article Network 2 and Erdos1 Subnetwork

9 Utility of Modeling Influence and Impact Within Networks

10 Conclusion

Works Cited

Citeseerx. Web. 10 Feb 2014. <http://citeseerx.ist.psu.edu>.
Cytoscape Network Software
Matlab 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.
"THE PERRON-FROBENIUS THEOREM." Prentice-Hall Inc.. Prentice-Hall Inc., n.d. Web. 09 Feb 2014. <http://www.prenhall.com/divisions/esm/app/ph-linear/leon/html/perron.html>.
RStudio
Xiaoming, Liu. "In much the same way, a database of articles was created for a subnetwork of Erdos1. The search was laborious, seeking articles citing other articles authored by those within Erdos1 and also having authors connected by coauthorship. The search engine http://citeseerx.ist.psu.edu was utilized in the sesarch, and the resulting article network was one of ten nodes, which then determined our subgroup of Erdos1 as a network of 11 nodes.." Elsevier Science. (2008): n. page. Print.