

Using Mathematical Models to Rank the Members of Criminal Networks

Lucas Chirino
High Point University

August 26, 2019

1 Abstract

Different mathematical approaches to ranking were used to determine the level of importance of every member in a criminal network. Two different data sets consisting of phone records provided by the FBI were analyzed. The first data set consisted of call logs over a three year span of members in a drug ring. The second data set consisted of the call logs of members in a gang. After analyzing the results of the rankings, properties of the two networks are discussed to provide further insight into why certain ranking algorithms performed well and others did not.

2 Background

Rankings are a familiar part of life found in sports, movies, and even dining locations. Despite the importance of knowing which is best, it is challenging to rank a set of data objectively and fairly. Numbers and statistics are impartial to sentiment, so a mathematical approach to ranking seems appropriate. However, there are many ways to mathematically rank a set of data, each of which has a unique approach and emphasis which in turn produces a variety of results. In this paper, we explore several ranking methods including the PageRank algorithm, the Massey and Colley method, Keener, and Elo, which are used in various applications across disciplines. PageRank is used in ranking the importance of webpages and is a component of Google's search engine. The Massey and Colley methods are used throughout the NCAA football season when producing rankings. One method used to produce rankings in the NFL is the Keener method, while the Elo method is used in the ranking of chess competitors. We set out to determine whether these methods could effectively rank members of a drug ring and then a gang, using only their phone records.

The first data set consists of call logs of eight people in a drug ring investigated by the FBI. Over the span of approximately three years, the eight people placed or received a total of 58,513 calls between 129 people. The goal was to

apply the various ranking methods to the call logs to ascertain which mathematical method produced the most accurate results. Accuracy is defined by matching the FBI's ranked list of the criminal network as closely as possible.

To further compare effectiveness, each method was applied to a similar data set of phone records from members of a gang. This data set was considerably smaller and consisted of the phone records of 23 individuals; the call log was approximately half as plentiful and spanned a year and a half. Again, the goal was to find the best method to match the FBI's ranked list as closely as possible.

3 The Markov Method and Google's PageRank

Prior to the appearance of Google in 1998, internet search engines were not very useful because they could be easily gamed. When Google arrived with its own search engine, users found it gave good results [3]. These results came in part from their innovative ranking approach called PageRank ([3], [4]). In constructing their search algorithm, the founders of Google changed the way search engines displayed results. Instead of fully relying on text relevancy (which was easily abused), they used stochastic matrices of relevant web pages that represented random web traffic ([3], [4]). The stochastic matrix is created by counting the links between pages and dividing each entry in a column by the column's sum.

The following example illustrates how PageRank works. Suppose we start with a simple web consisting of four web pages as follows, where the direction of the arrow represents the direction of the link on the internet.

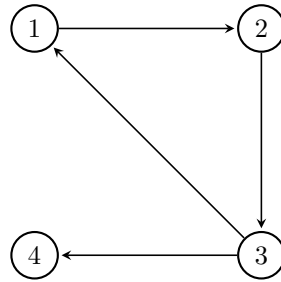


Figure 1: Directed Web Graph of four Nodes

Then, an associated matrix A is created where

$$A_{i,j} = \begin{cases} \frac{1}{n} & \text{if webpage } j \text{ links to webpage } i, n = \text{number of directed edges out of } j \\ 0 & \text{if } j \text{ does not link to } i. \end{cases}$$

The associated matrix for this small example would be:

$$\begin{bmatrix} 0 & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

First, we apply the Markov Chain process, which can be summarized as finding the eigenvector corresponding to the dominant eigenvalue which gives the long term behavior of the system [6]. Matrix A has an eigenvalue less than 1 by the Perron-Frobenius Theorem because the values are all non-negative and the columns sum to at most 1 ([5], [6]). This matrix is also column sub-stochastic. To make it fully stochastic, we must fix the column of zeros to sum to 1 as well.

When a node doesn't link to another node, as is the case with node 4, it is called a *dangling node*. This disrupts the method because dangling nodes create a column of zeros, therefore making the matrix sub-stochastic. We can proceed to find the Perron eigenvalue, or we may choose to make that node link to every other node equally. Using the example from above, we could imagine placing an arrow from node 4 to every other node including itself. Then, the associated matrix can be fixed by placing $\frac{1}{4}$ in every element of the fourth column. Although this approach is changing elements of the matrix, it is not changing the solution because it is sending links to every node equally. If the dangling node had many links, then every web page will receive an equal share of these, thus not changing the state probabilities nor the long term probabilities. This produces the following revised matrix:

$$\begin{bmatrix} 0 & 0 & 1/2 & 1/4 \\ 1 & 0 & 0 & 1/4 \\ 0 & 1 & 0 & 1/4 \\ 0 & 0 & 1/2 & 1/4 \end{bmatrix}.$$

Non-negative matrices are known to have a dominant positive eigenvalue and corresponding dominant eigenvector ([5], [6]). The Markov method states that if our matrix is also stochastic, the dominant eigenvalue will be 1 and the long term behavior of the matrix will converge to the dominant eigenvector. To find the long term result for a stochastic matrix, we find the eigenvector corresponding to the eigenvalue of 1. In the above example, the resulting vector is

$$\begin{bmatrix} 1 \\ 5/4 \\ 3/2 \\ 1 \end{bmatrix}.$$

This means the first and fourth nodes are equally important, and the third node is most important. The third node is the most important because the probability of being on that page after convergence is higher than the probability of being on any other page. Typically the resulting eigenvector is scaled to sum to one so that the probabilities can be read as percentages.

The founders of Google expanded on this method. The current model assumes 100% of movement is continuous. If you imagine surfing this small web, moving around the websites, you can think of the associated matrix as the probabilities of where you click next. But what if the user decides to start over, to start at a new web page by typing in a new URL? To model this, Google founders Larry Page and Sergey Brin proposed the following equation to model web behavior [3]

$$G = (1 - \alpha)A + \alpha R. \tag{1}$$

Here A is the same matrix as before, matrix R contains all entries of $\frac{1}{n}$ where n is the number of nodes, and $0 < \alpha < 1$ is a constant chosen to characterize the network. A higher α value corresponds to more random networks. While, today, Google keeps much of their information on their programming private, they did reveal that they often chose $\alpha = 0.15$ for many simulations when first developing their search engine [4]. This means that the majority of the movement was through the directed web graph; roughly 85% of the time on the Internet you are moving around through the web, but 15% of the time you decide to start over. From here, we can proceed to find the dominant eigenvector as before because G will still be stochastic with dominant eigenvalue 1. In our example, using $\alpha = 0.15$ produced

$$G = \begin{bmatrix} 3/80 & 3/80 & 37/80 & 1/4 \\ 71/80 & 3/80 & 3/80 & 1/4 \\ 3/80 & 71/80 & 37/80 & 1/4 \\ 3/80 & 3/80 & 3/80 & 1/4 \end{bmatrix},$$

with corresponding dominant eigenvector after normalization

$$\begin{bmatrix} 0.242 \\ 0.253 \\ 0.457 \\ 0.048 \end{bmatrix}.$$

Now we see that the eigenvector no longer has two equal rankings providing a hierarchy for all nodes, and the overall top rank is still the same. According to the results, the most important node is 3 followed by 2 and 1 while 4 is the least important.

The Google PageRank method can be applied to any data set when looking for a ranking of nodes, and members of a criminal network are an interesting and relevant set to further examine. The first data set given consisted of 129 people in a drug ring, labeled Person A through Person F, Foreign 1, Foreign 2, and Person 1 through 121 (inputted in that order). Persons A-F were the central figures of the gang, while persons 1-121 were possible members or contacts of the gang. The first step is to input the call log data of all 129 people into a matrix. Using the method outlined above, a person calling another person was viewed as that person “linking” to the other “page”. If person A called person

C 376 times, then $A_{3,1}=376$, similar to techniques used in [7]. Since many of the people only called a handful of other people, a fairly sparse 129×129 matrix was created. Each column of matrix A was then scaled by the reciprocal of its sum to make the matrix column stochastic. Next the dangling nodes were fixed using the aforementioned method, to eliminate columns of zeros. In this context, a dangling node was someone who never made an outgoing call. Our data consisted of five of these people. After this was done, both the Markov method and Google's PageRank method were applied and produced the following top 10 results respectively:

$$\begin{aligned} & [A \ E \ F \ C \ G \ B \ D \ 1 \ 2 \ 71] \text{ and} \\ & [A \ E \ F \ C \ G \ 1 \ D \ 71 \ 74 \ 2]. \end{aligned}$$

The results of each method will be compared to each other and the FBI's ranked list.

4 Massey and Colley

The Massey method takes a different approach that can easily be applied to sports rankings. It uses point differentials and number of games won against each opponent to create a ranking. The idea behind this method is that if you know the rankings of two teams, not only can you predict the winner, but you can also predict the margin of victory. This idea can be expressed in a system of equations given by

$$r_i - r_j = p_k, \tag{2}$$

where p_k is the point differential between teams i and j while r_i is the rating of team i and r_j is the rating of team j . The point differential is calculated as points i scored minus points j scored. After a season of play, we know the point differentials, but we don't know the rankings. For each game played, we know the comparative difference in ranking between the two teams. Over a season there are many of these. For each team at the end of a season of m games, the equation for team i is

$$(r_i - r_1) + (r_i - r_2) + (r_i - r_3) + \dots + (r_i - r_m) = p_k,$$

which simplifies to

$$mr_i - r_1 - r_2 - r_3 - \dots - r_m = p_k. \tag{3}$$

If this is done for every team, then a matrix can be made from these equations. The numerical value placed on non-diagonal entries is the negative of how many times teams i and j played each other. Naturally, if two teams did not play one another, then the corresponding entries are zeros. This matrix M is of size $n \times n$ and gives rise to the equation $M\mathbf{r} = \mathbf{p}$ where \mathbf{r} is the ratings vector and

\mathbf{p} is the vector of cumulative point differentials for each team. This is a system of equations which has infinitely many solutions. To adjust this matrix to have a unique solution, the last row is replaced with ones. No information is lost with this change because all of the other teams carry a piece of the last team's information on their own rows.

For our data, the Massey matrix was a very sparse 129×129 matrix. The point differential for our data was how many calls a person made to another person minus how many calls that person made to them. After finding the ratings, the numbers were sorted to produce the following ranking:

$$[F \ A \ G \ 47 \ E \ 46 \ 62 \ 26 \ 27 \ 73] .$$

The Colley method, summarized by the equation $C\mathbf{r} = \mathbf{b}$ is equal to $(2I + M)\mathbf{r} = \mathbf{p}$ where M is the adjusted Massey matrix. However, to find Colley without first finding the Massey matrix, C and \mathbf{b} are the results of, $b_i = 1 + \frac{1}{2}(w_i - l_i)$ where w_i and l_i are the wins and losses of i respectively, and

$$C_{i,j} = \begin{cases} 2 + t_i & i = j \\ -n_{i,j} & i \neq j \end{cases},$$

where t_i is number of games played by team i and $n_{i,j}$ is the number of games team i played team j . The ranking produced by the Colley method turned out to be the exact same ranking as that of Massey for the drug ring data.

5 Keener

The Keener method, primarily used in the NFL, creates ratings for teams based on a single metric at the user's discretion. After choosing your metric, the first step is to create the matrix

$$S_{i,j} = \text{value of metric team } i \text{ got against team } j.$$

The next step is to adjust these values using Laplace's rule of succession. Doing this normalizes the data so a person/team can not skew the results by having an outlier performance in terms of the chosen metric. Laplace's Rule of Succession is expressed as

$$\frac{S + 1}{S + F + 2},$$

where S and F represent the number of successes and failures respectively. After doing this, we have the following new matrix A

$$A_{i,j} = \frac{S_{i,j} + 1}{S_{i,j} + S_{j,i} + 2}.$$

Lastly, a skew function may be used to separate people of equal ratings. We used the standard Keener skew function [5]

$$h(x) = \frac{1}{2} + \frac{\text{sgn}\{x - \frac{1}{2}\}\sqrt{|2x - 1|}}{2}.$$

To provide an example of this method, here is a single entry from the matrix we used. Let $S_{3,1} = 376$ and $S_{1,3} = 676$. To find $A_{i,j}$, we use the above formulas which yields

$$A_{3,1} = \frac{376 + 1}{376 + 676 + 2} = 0.3577 \quad \text{and} \quad A_{1,3} = \frac{676 + 1}{376 + 676 + 2} = 0.6423.$$

Notice that $A_{3,1} + A_{1,3} = 1$.

For the criminal network data, the number of calls was the chosen statistic, which was the same initial matrix from the Markov method and yielded the following top 10 rankings:

$$[B \quad C \quad 63 \quad 5 \quad 10 \quad 116 \quad 15 \quad \text{Foreign} \quad 2 \quad 92 \quad 105].$$

6 Elo

The Elo method of ranking has wide usage ranging from chess to E-sports such as League of Legends. Named after Arpad Elo, who created the system to rank chess players, the Elo system changes subjects' rating based on how far the subjects' performance is away from its current rating. The basis for this is the equation

$$r_{(new)} = r_{(old)} + K(S - \mu)$$

where K is a value (sometimes adjusted based on importance of match-up), and S is a constant that can be defined as

$$S_{i,j} = \begin{cases} 1 & \text{if } i \text{ beats } j \\ 0 & \text{if } i \text{ loses to } j \\ 1/2 & \text{if } i \text{ and } j \text{ tie} \end{cases}$$

and μ is a value that can be defined as

$$\mu_{i,j} = \frac{1}{1 + 10^{-d_{i,j}/400}}.$$

Here $d_{i,j}$ is calculated based on the difference between the ratings of two subjects to find the expected margin for who is to win

$$d_{i,j} = \text{rank}_{i(ol)} - \text{rank}_{j(ol)}.$$

To provide an example of this, let's consider two chess players. One player is rated 1300 while another is rated at 1500, using a K of 25. First calculate the new rating of the 1500 player if they lose to the 1300 player:

$$d = 1500 - 1300 = 200$$

$$r_{(new)} = 1500 + 25 * (0 - 1/(1 + 10^{-d/400})) = 1481.$$

Now for the 1300 player who beats the 1500 player:

$$d = 1300 - 1500 = -200$$

$$r_{(new)} = 1300 + 25 * (1 - 1/(1 + 10^{-d/400})) = 1319.$$

Notice that the winner gained points and the loser lost points by an equal magnitude.

By construction, the Elo method works over time and not all at once. The data that we used had the entire log of calls spanning three years arranged chronologically. It also featured FROM and TO columns. Each call was considered to be a “match”. In total there were 58,513 calls. To quickly run a rating update after every call to the correct callers, a simple script in Matlab was created. In this method, everyone starts out at an arbitrary, but equal initial rating. The K value was kept constant to make sure every call is equally valued.

When a call is made, there are two different interpretations of who the winner is, both results from each interpretation are included below. If the person making the call is the “loser”, we are saying that the most important people are the ones receiving the calls and this interpretation produced the following top ten:

$$[3 \ 107 \ 71 \ 26 \ 52 \ 1 \ 2 \ 50 \ 8 \ 53].$$

The other interpretation is that the person making the call is the “winner”. A reason this might make sense that that in a criminal network, it is possible the people at the head of the organization make the most calls. If choosing the caller to be the winner, then the result becomes:

$$[F \ C \ A \ G \ E \ 7 \ D \ B \ 104 \ 42].$$

7 Analysis of Results

After running the data through the various ranking methods, we needed to compare them to the actual rankings, as assigned by the FBI. An intelligence analyst at the FBI provided additional information about the actual rankings of the criminals in the drug ring. Note that while there is a hierarchy, there is also lateral structure, meaning that there are several instances where more than one person would practically be of equal importance and rank. There were two outside foreign contacts that were of significance in the larger structure of the network. In this data set, Person E and F are actually the same person who simply used two phones. A further exploration could be to combine both E and F’s call logs and look at the results. However, the rankings were never combined into one in order to better extrapolate the findings of these methods on networks of which one has less information.

Looking collectively at all of the ranking methods together with the actual assigned ranking by the FBI in Table 1 below, we see that the method that best approximated this ranking was Google’s PageRank method. The two

forms of Elo are to provide both possible interpretations of who the winner is when a phone call is made. The Elo(From=Winner) outperformed the Elo (From=Loser), providing that people at the top of the network are the ones making more calls. Other points of interest include that Foreign 1 and 2 were contacts from Mexico that supplied the operations. The only methods that had the foreign contacts appear even remotely important were the Massey, Colley, and Keener methods, and Keener was the only one that had Foreign 2 in its top ten.

FBI	Google	Massey	Colley	Keener	Elo(F=W)	Elo(F=L)
A	A	F	F	B	F	3
E/F	E	A	A	C	C	107
E/F	F	G	G	63	A	71
C	C	47	47	5	G	26
G	G	E	E	10	E	52
D	B	46	46	116	7	1
1	D	62	62	15	D	2
	1	26	26	For. 2	B	50
	2	27	27	92	104	8
	71	73	73	105	42	53

Table 1: Drug Ring Final Rankings

To statistically analyze and compare algorithm performance, we used Kendall’s tau-b, a non-parametric test of rank order correlation adjusted to handle ties [1]. Here the Google algorithm performed best ($\tau_B = 0.324, p = 0.000$). This indicates a statistically significant moderate relationship between the algorithm’s prediction and the FBI rankings. The best of the Elo algorithms ($\tau_B = 0.235, p = .001$), and the Keener algorithm ($\tau_B = .148, p = .039$) also revealed statistically significant but weaker relationships between their predictions and the FBI rankings. In all cases, it is important to note that the large number of cases unranked by the FBI artificially lowers the correlation coefficients.¹

8 Second Data Set

After analyzing the drug ring data set, we obtained a second set of data from the FBI to consider. This data set was considerably smaller, consisting of 27,510 calls over a period of approximately a year and a half between 23 individuals in a gang. All the previous methods discussed were used on the entirety of the data set and the following rankings were obtained.

These rankings compared to the rankings given by the FBI yield less than favorable results. There were no instances in this data set of a single person using multiple phones but this data set had some aspects of a lateral structure. The FBI correspondent informed us that multiple people were considered to be at equal rank. In the results below, names on the same line indicate that people are equal in rank in reference to lateral structure.

FBI	Google	Massey(FL)	Massey(FW)	Elo(FW)	Elo(FL)	Keener
3	17	23	23	1	9	17
12	11	22	8	14	22	5
9	2	4	16	8	15	9
4	20	8	7	16	10	18
8/1/22/17/23	18	16	3	7	18	13
5	15	7	13	3	19	15
14/2	3	17	11	11	17	10
7	16	19	18	4	6	2
13	8	12	15	23	20	22
16	12	14	12	7	13	19

Table 2: Gang Member Rankings

FBI
3
12
9
4
8/1/22/17/23
5
14/2
7
13
16
18
20
15

9 Time Frames

For this second set of data, a specific time period, noted by the FBI, was more significant than the rest of the data. To see how this time period would affect the rest of the data, the same methods were used to rank the data split into two categories. The first subset was from the beginning of the call logs to the start of the time period, the second subset was from the start of the time period to the end of it.

The results again were mediocre at best. For any given method, the most important people, according to the FBI Intelligence Analyst, were either entirely absent within the top ten or not close to the top. When comparing an important person's ranking between the various time frames, it is equally inconsistent. Some important people rise while others fall.

Statistically with the same Kendall's tau-b comparison across the board for all time frames and scopes, the algorithms performed much worse in predicting the rank order of the gang data set. None of the algorithms were statistically

significantly related to the FBI's overall rankings at the $p < .05$ level. The top performers, however, were the Massey (FL) model ($\tau_B = 0.206, p = 0.190$) and the Massey (FW) and Elo models (all $\tau_B = 0.157, p = 0.323$). It is worth noting that the small sample size ($n = 23$) in these analyses is potentially a key reason why the p -values are comparatively large.

10 Network Properties

To understand why the ranking methods performed so much better on the first network than they did on the second, we analyzed various properties of the two networks. Several properties were computed and compared, including average degree, degree distribution, clustering coefficients, and average clustering coefficient [2]. The *degree* of a node is how many links are associated with the node. *Average degree* is the sum of the degrees for each node divided by total number of nodes (people)

$$\frac{1}{n} \sum_{i=1}^n deg_i,$$

where n is the number of nodes. The *density* of a network is how many connections there are divided by the number of potential connections, given by

$$\text{density} = \frac{L_i}{n(n-1)}.$$

The *clustering coefficient* of a node is defined as

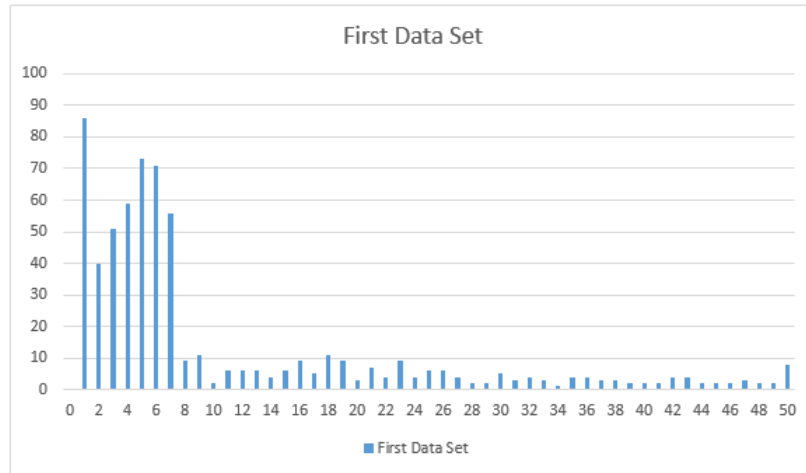
$$C_i = \frac{2L_i}{k_i(k_i-1)},$$

where L_i represents the number of links between the k_i neighbors of node i . The *Average Clustering Coefficient* is the sum of all clustering coefficients divided by the number of nodes.

The following are the values for the first data set consisting of members in a drug ring.

Density = 0.0268
 Average Clustering Coefficient 0.00085352
 Average Degree = 3.4264

Degree Distribution:



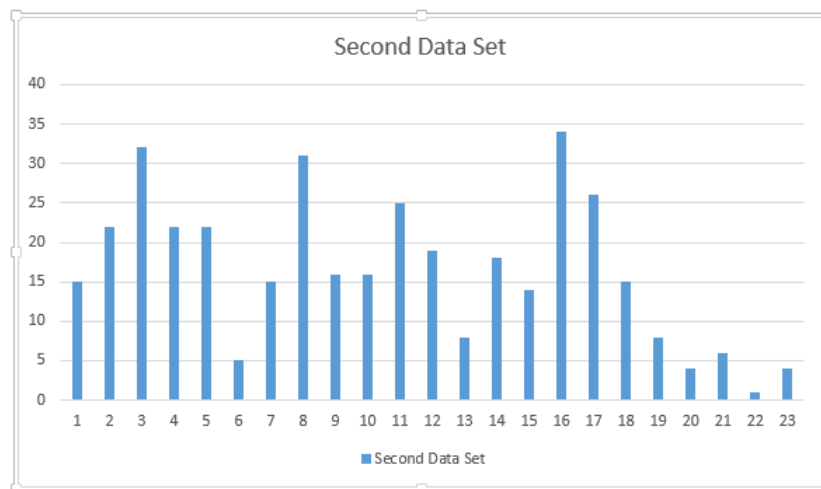
The following are the values for the second data set consisting of members in a gang.

Density = 0.3735

Average Clustering Coefficient 0.0125

Average Degree = 8.2174

Degree Distribution:



These key network properties tell us that the first data set is much more sparse. This more closely resembles the network structure of the Internet for

which the PageRank method is used, which may be why it performed so well. Nodes are less clustered, have a lower density, and have a lower average degree. There are simply fewer connections per node in the first data set than there are in the second. The high ranking individuals in the first set have much higher degrees than the rest of the members. This can be seen from the degree distribution charts above. Meanwhile, in the second data set, the difference in degrees from node to node is not nearly as large. This helps explain why the rankings in the first data set more closely resemble the true rankings than in the second set.

11 Conclusion

Despite the promising results of the rankings for the first data set, no conclusion can be drawn in favor of any particular method for all types of networks. For sparse networks, the PageRank method is likely to be more effective and the best for determining a ranking. Future analysis could be done on the network(s) such as k -means analysis to look at the clusters more closely and examine properties of those clusters in the network and their relative importance. These ranking methods can be applied to many more data sets to get a more complete comparison of their effectiveness.

References

- [1] A. Agresti, B. Finlay, *Statistical Methods for the Social Sciences* 3rd ed. New York: Prentice Hall, 1999, p. 277-81.
- [2] A-L. Barabasi *Network Science* Creative Commons License. Web September 2016
- [3] S. Brin, L. Page, *The anatomy of a largescale hypertextual Web search engine*, In Proceedings of the 7th World Wide Web Conference, 1998.
- [4] K. Bryan and T. Leise, *The \$25,000,000 Eigenvector: The Linear Algebra behind Google*, SIAM Review, Vol. 48 No. 3, September 2006, pp. 569-581
- [5] A.N. Langville, and C.D. Meyer. *Who's #1? : The Science Of Rating And Ranking*. Princeton: Princeton University Press, 2012. eBook Academic Collection (EBSCOhost). Web. 26 July 2016.
- [6] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.
- [7] L. Zack, R. Lamb and S. Ball, *An application of Google's PageRank to NFL rankings*, Involve A Journal of Mathematics, Vol. 5, no. 4, 2012.

Notes

¹Statistical analysis was provided by Dr. Alixandra Yanus from High Point University