Alexander J Freund, Timothy A. Romer, Dat Luong, Jaxson Wirth

Miami University

Department of Computer Science and Software Engineering

Oxford, OH, USA

{freundaj,romerta,luongdq,wirthj}@miamioh.edu

1. **Introduction:**

What should the CIA do to identify the leader of a terrorist group if the resources spent on investigating a person are costly? Where should a series of hospitals be constructed to maximize the number of people that are located near one? By measuring the centralities of a network, we can tell how important/ central a node is with respect to the rest of the network. Centrality measurement methods have been used for biological networks, applied to gene regulatory networks (Koshützki and Schreiber), and also used to assess the prospects for criminal intelligence (Sparrow) or even solving a financial crisis (Kuzubas et al.).

Calculating these centrality measures can be quite expensive in a computational sense which means that it takes a lot of time to calculate. To extend this even further, ranks are taken based on the calculation so nodes in the network each have a rank to compare rather than comparing the centrality values. The goal of this research is to identify whether certain centrality ranks can be accurately predicted using ranks of other centrality measures. To form predictions, we will generate synthetic networks, measure the centrality of their nodes, and use these measures to build machine learning models. Forming such accuracy predictions has two advantages: we can use computationally efficient measures, and run them through a fast machine learning model, to predict the outcome of a more computationally demand measure. In addition to saving computation time, researchers can also spend less time generating repetitive information: if one measure can be directly derived from another that has already been analyzed, then there is no need to perform or analyze yet another measurement.

In this paper, we will be identifying various centrality measures and identifying whether a calculated metric can be predicted by other centrality metrics, with regards many types of network. Specifically, our main contributions are as follows as follows:

* Examines 7 different metrics for measuring centrality
* Assesses the correlation between metrics
* Computes the prediction of a metric based on other metrics
* Computes the accuracy of such prediction to identify whether that metrics is truly worth doing and interpreting

The organization of this paper is as follows. In the next section, we will provide a background on the metrics we will be using and the mathematical definitions behind them. Section two will also contain examples of a randomly generated graph and the centrality calculations/ranks will be discussed. Following this, the last section will provide detailed methods on how we generate different types of networks, compute centralities, compute the correlations, and prediction of the metrics.

1. **Background:**

Centrality is the measure of how central a particular node in a graph or network is and the general assumption is that it is typically located in the center of a network or graph (Freeman). There are many forms of centrality measure including simplistic measure such as degree, closeness, betweenness, and load. There are also more complicated measures that have been developed throughout history for other applications such as Page Rank.

In order to know if certain centrality measures are able to predict other centrality measures using different machine learning models, the centrality measures will have to be calculated on graphs that are generated a specified amount of times. These graphs feature scale-free, small-world, random, and scale-free small-world networks. Scale-free networks have degree distributions that follow the power law which means that few nodes have many edges while many nodes have few edges. The small-world network type has a high clustering coefficient which means there are a lot of cliques of nodes and there are few connections to other cliques. Random networks are networks that have a specified number of nodes and are randomly connected with a specified number of edges. The scale-free small-world networks have the characteristics of both the scale-free and small-world networks. Below will be a discussion on the different centrality measures that are used in this research.

**IIA. Degree Centrality:**

In network science, degree centrality has traditionally been considered to be the simplest measure of centrality and first item to look at when examining centrality (Opsahl et al.). Degree centrality can be defined as the ability for a node to receive information that is flowing through a network. This is measured by the number of links that node has to other nodes (Opsahl et al.).

Consider a given graph , defined as for vertices and edges. The total number of vertices is denoted by N. The degree centrality for a vertex is defined as where is the degree of vertex (Opsahl et al.).

**IIB. Closeness Centrality:**

Within graphs, a node is considered to have a high value of ‘closeness’ if it has a relatively low average of shortest path distance to all other nodes (Rochat). The average shortest path distance is finding each of the shortest paths to all of the other nodes in the network then will find the average of those distances. This is insinuating that a node with a high value is generally closer to all other nodes in the graph. To calculate this value, take one less than the number of nodes in a graph and divide it by the sum of the shortest path between a node and all other nodes in a graph.

The closeness centrality can be defined mathematically as follows. Closeness centrality is defined as where is the length of the shortest path between vertex y and x (Rochat).

**IIC. Betweenness Centrality:**

Betweenness centrality measures how many times a particular node is situated on the shortest path between two other nodes. This centrality measure is similar to the closeness centrality because both of them involve the calculation of the shortest path between nodes. To calculate this value for a given node v, count how many shortest paths between all pairs of nodes traverse v, and divide by the total number of shortest paths. (Brandes, "Maintaining the duality of closeness and betweenness centrality.").

The betweenness centrality can be defined mathematically as follows. The closeness centrality for a vertex is defined as where is the number of shortest paths between s and t given they contain vertex v and is the number of shortest paths between s and t (Brandes, "Maintaining the duality of closeness and betweenness centrality.").

**IID. Load Centrality:**

The load centrality measure is similar to the betweenness centrality in that it measures the amount of flow that goes through a particular node; however, the load centrality measures the unit amount of information that get split between other nodes. Information is continually split between adjacent nodes until the target is reached. The total amount of information that passes through the node is defined as its load (Brandes, "On variants of shortest-path betweenness centrality and their generic computation.").

The load centrality can be defined mathematically as follows. The load centrality for a vertex is defined as where is the quantity of information that is passed through vertex and are the set of vertices (Maccari et al.).

**IIE. Local Reaching Centrality:**

The local reaching centrality is the measure for a node and its proportion of all other nodes that are reachable in a graph of that particular node. This gives a fundamental assumption that all nodes that are reachable for a node are located in some finite distance away (Mones et al.).

The local reaching centrality can be defined mathematically as follows. The local reaching centrality for a vertex , for a given graph , where the graph can be defined as for vertices and edges is defined as where is the distance formula and is the number of vertices in a graph (Mones et al.).

**IIF. Harmonic Centrality:**

The harmonic centrality measure is similar to the closeness centrality however it addresses the issues of unreachable nodes. The harmonic difference will correct the issues with the average shortest path measure because disconnected nodes can have a potentially misleading value because the average distance could be low if the graph is almost entirely disconnected (Boldi and Vigna).

The harmonic centrality can be defined mathematically as follows. The closeness centrality for a vertex , for a given graph , where the graph can be defined as for vertices and edges is defined as where is the shortest average distance function (Boldi and Vigna).

**IIG. Page Rank:**

The page rank centrality measure has many applications including being the basis for how Google designed its search function (Page et al.). Page rank will rank websites based on the quality of websites that reference that particular website. In terms of networks, page rank will work in a very similar sense as it works in populating a search result. Page rank can be calculated iteratively and will return a probability that a node is accessed via another link. The page rank value at any particular time can be shown as where is the node that is accessed. The page rank for a vertex , for a given graph , where the graph can be defined as for vertices and edges is defined as where is the number of links from a node and c is a factor for normalization. page). Although this is a simplified definition of the page rank calculation, it should suffice for our simplified network application rather than ranking web pages for a search engine (Page et al.).

**IIH. Centrality Ranks:**

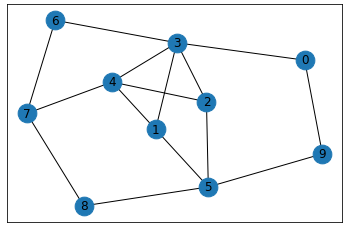


Figure 1. A randomly generated graph with ten nodes and fifteen edges.

Table 1. The centrality ranks based on the graph depicted in Figure 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Vertex | Degree Rank | Closeness Rank | Betweeness Rank | Load Rank | Reach Rank | Harmonic Rank | Page Rank |
| 0 | 7 | 7 | 5 | 5 | 7 | 7 | 8 |
| 1 | 4 | 2 | 8 | 8 | 4 | 4 | 5 |
| 2 | 4 | 2 | 8 | 8 | 4 | 4 | 5 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 2 | 2 | 3 | 3 | 2 | 2 | 3 |
| 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 6 | 7 | 7 | 10 | 10 | 7 | 7 | 10 |
| 7 | 4 | 6 | 4 | 4 | 6 | 6 | 4 |
| 8 | 7 | 7 | 5 | 6 | 7 | 7 | 9 |
| 9 | 7 | 10 | 7 | 7 | 10 | 10 | 7 |

Using the centrality measures that were highlighted in this section. A random graph was generated and centrality measures were calculated using Networkx 2.3 python library that specializes in networks. Based on the calculations, ranks were obtained after calculating the centrality values then comparing the values with the other vertices in the graph. Figure 1 shows the random graph that was created by the library and Table 1 shows the centrality ranks from the graph.

**IIJ. Literature Review:**

In regards to *“Characterization of topological keystone species Local, global and ‘meso-scale’ centralities in food webs“* by Ernesto Estrada; this paper adds conclusions that can be considered in this research. The researcher wished to study the impact of removing a keystone species (a species whose impact on the community are large) and used centrality measures to determine the impact. The researcher included centralities that fit into three different categories. Local centrality measures included degree and betweenness centrality. Global centrality measures included closeness, eigenvector, and information centrality. Finally, meso-scale measures included closed walk and subgraph centrality. The results of this research had insights into the research highlighted in this research. An important note is that all of the centrality measures generally have high correlation values with each other however, based on the centralities used, the rankings produced by the different centralities were quite different. The author specifically noted the highest correlations between degree centrality and information centrality, closeness centrality and information centrality, and subgraph centrality and eigenvector centrality. Removing different nodes in a network can have a variety of effects and differ for each centrality measure.

The paper, “*Robustness envelopes of networks”* by Stojan Trajanovski, Javier Martίn-Hernández, Wynand Winterbach, and Piet Van Mieghem discusses how targeted attacks on networks should highlight the worst-case attacks on networks rather than the average/random attacks. The researchers utilized the computational technique called envelopes which calculates that change in energy before and after a node removal. For targeted attacks, the researchers determined that removing nodes based on centrality values is sufficient for maximizing damage to a network. The researchers developed a technique to compare centrality metrics by coming up with a metric where a and b are different centrality measures and k is the percent of nodes to include. The more overlap of ranks there are for the two centralities after k ranks, the more similar the centrality measures are. For example, if it can be expected that a high level of overlap between ranks for these two centrality measures for the top 50% of ranks. The results found that using degree and eigenvector centrality were the most efficient for simulating the worst-case scenario for targeted attacks which indicated that the had little overlap.

In the paper *Topological Measures in Real-World Networks*, the authors establish they will be arguing there is an over abundance of network measurements and that this unnecessarily complicates attempts to determine a definite measure set. First, the authors describe how many network properties are “posed within a particular research interest, resulting in a characterization of real-world networks from a specific domain” (Jamakovic and Uhlig 346). In other words, many of the existing network properties have been developed specific to a domain, and can therefore have intrinsic overlap with measurements from other domains. Following this, many of the existing network properties are discussed and analyzed in the paper. These various methods are defined and described in detail, including Degree Centrality, Distance Distribution, and Clustering Coefficient. The authors then begin analyzing real-world networks to find similarities between different measurements. They find that “networks with smaller distance are much more likely to have high-degree nodes that form tight and well-interconnected subgraphs” (Jamakovic and Uhlig 351). In other words, average distance is inversely related to degree. Additionally, they find that the average node betweenness has no apparent affect on the overall connectivity of a graph. Overall, this paper shows that some topological measures are more correlated than others, which implies that there is intrinsic overlap between them and they may not all be necessary.

The paper *Correlation Coefficient Analysis of Centrality Metrics for Complex Network Graphs* addresses a similar but broader topic with a much more numerical approach. Rather, this paper seeks to find the general correlation coefficients between popular network centrality measures. For example, rather early on the author states that “as the variation in the degree distribution of vertices increases, the correlation coefficient between the two classes of centrality metrics increases” (Meghanathan 11). The author then goes on to define and describe several of the most popular centrality metrics, such as Degree, Eigenvector, Betweenness, Closeness, Farness, and Eccentricity. After sufficient analysis of real-world networks, the author uses the Pearson Correlation Coefficient formulation to evaluate correlations between centrality metrics, which is very similar to what we are doing in our Network Centrality project. In conclusion, the author finds that when the variation of node degrees is low, betweenness and closeness are ranked equivalently. Consequently, when the network becomes scale-free, these two shortest-path metrics become less correlated. Overall, the two degree-based metrics (degree and eigenvector) are highly correlated and the four shortest-path metrics are moderately correlated. Additionally, the degree-based and shortest-path based centralities are not correlated in random networks, but are moderately correlated in networks that exhibit scale-free nature. The author concludes by stating that “the level of correlation between a degree-based centrality metric and a shortest path-based centrality metric increases with increase in variation of node degree” (Meghanathan 19).

The first paper to be looked at is titled ‘Correlation between centrality metrics and their application to the opinion model’ studied by Cong Li, Qian Li, Piet Van Mieghem, H Eugene Stanley, and Huijuan Wang; of the Delft University of Technology and Boston University. This paper is based around the idea of studying correlation of centrality metrics in terms of their Pearson correlation metrics - and from there introducing a new centrality measure deemed the degree mass. Degree mass is defined as: di(m)=k=1m+1(Aku)i=j=1N(k=0mAk)ijdj. This paper is used as a proof of this new centrality metric. Through some simulations it is found that the lowest-order degree masses are strongly correlated with the betweenness, closeness, and the components of the principal eigenvector - all of which prove difficult to compute. It is found that the 0th-order degree mass is the degree and the high order degree mass is proportional to the principal eigenvector x1.

The next paper is ‘Consistency and differences between centrality measures across distinct classes of networks’ studied by Stuart Oldham, Ben Fulcher, Linden Parkes, Aurina Arnatkevic̆iūtė, Chao Sao, and Alex Fornito. This study is focused on finding the correlation between different centrality measures, and in turn finding that a comparative approach can inform in regarding to nodal roles of complex networks. In this paper spearman’s correlation is used to find the correlation between separate centrality measures, which finds that they tend to be positively correlated. In fact, there tends to be a medium to high correlation across all networks - however there was rather high variability across different networks. It was found that the density, global efficiency, modularity, majorization gap, and spectral gap were correlated with the centrality measure correlation; with the majorization gap being the most so.

The paper *On the Structural Properties of Social Networks and their Measurement-calibrated Synthetic Counterparts* analyzes 120 real-world networks and 480 models-generated generated by the four network models: Barabasi–Albert, stochastic block, forest-fire, and 2K. For each of the graphs, they measure 17 different structure measurements that cumulatively represent the whole graph. Then for each of the two graphs from real-word and models-generated, a distance will be computed based on the structure measurement, however, there is too many structure measurement to choose from, in order to choose the best one, a grid search optimization is used to choose the best structure to minimize the distance. Moreover, when they try to measure the distance between two graphs, they used Spearman’s rank correlation because of its ability to measure non-linear relations as well as outliers-sensitive graphs. Finally, they test whether it is able to synthesize networks, which is to determine if it is possible to generate a real-world network or there are some measurements that the model-generated networks cant capture. To do that, they fit each generated model into real-world ones, then we compute the structure metrics and compare it the that of the real-world by calculating the mean distance between them. From the result, they have concluded that 2K and the degree corrected stochastic block models are the only two models that can generally synthesize the real networks.

Up to these days, there are many centrality indices that have been introduced but only some of them are actually correlated. Knowing the correlation between indices is crucial because we can omit the redundant one or compute the cheaper indices. However, many reports show inconsistency in results, that’s why the paper *Correlations among centrality indices and a class of uniquely ranked graphs* will show that such correlations are determined mainly by structural properties of a network. The research proved that by showing the research by Valente et al and show that the correlation computation result is incorrect because of the preorder induced by the neighborhood-inclusion, it increases the ranking of the centrality much more than it should be in uniquely ranked graph or more specifically, threshold graphs. Because correlation results from threshold graphs are incorrect, we need to quantify the difference between a random graph and threshold graphs, in which the paper presents out 7 different methods. The result shows that further structural properties other than the neighborhood-inclusion preorder will affect the correlation result, thus making the structural properties very importance to compute the correlation.

In the paper *Correlation Analysis of Complex Network Metrics on the Topology of the Internet*,

1. **Methods:**

**IIIA. Resources Used**  
   
In order to work with network based systems and analyzing them computationally, we worked with four important Python libraries; Pandas 0.23.4, Numpy 1.14.6, NetworkX 2.2, and scikit-learn 0.21.3. Pandas and Numpy primarily are our methods to work with data in general, NetworkX is used to generate network data with associated methods, and scikit-learn implements our methods of machine learning.

**IIIB. Overall Process**

A screenshot of a cell phone

Description automatically generated  
Figure 8, Overall process

Our data comes from generating data for each of the combinations of scale free and small world; being random, only scale free, only small world, and both scale free and small world. We generate 100 graphs for each type, further subdivided in that by taking 25 of each specified size within the array [100, 200, 400, 800]; thus creating a total of 400 different graph samples.

The specific data we are working with is found from a variety of NetworkX libraries that allow us to take the data given a graph. The specifics we are working with are as follows; both degree and closeness rank are found via NetworkX’s degree centrality method, betweenness ranking is taken from the betweenness centrality method, load rank is taken from the load centrality method, and reach rank is taken from the global reaching centrality.

Using this data we then calculated correlations, between each element. To calculate this we simply used Pandas’ correlation method. From there we specified to calculate for these three different correlations: pearson, spearman, and kendall.

**IIIC. Data Generation**  
   
Our data comes from generating data for each of the combinations of scale free and small world; being random, only scale free, only small world, and both scale free and small world. We generate 100 graphs for each type, further subdivided in that by taking 25 of each specified size within the array [100, 200, 400, 800]; thus creating a total of 400 different graph samples.

The specific data we are working with is found from a variety of NetworkX libraries that allow us to take the data given a graph. The specifics we are working with are as follows; both degree and closeness rank are taken from NetworkX’s degree centrality method, betweenness ranking is taken from the betweenness centrality method, load rank is taken from the load centrality method, and reach rank is taken from the global reaching centrality.

Using this data we then calculated correlations, between each element. To calculate this we simply used Pandas’ correlation method. From there we specified to calculate for these three different correlations: pearson, spearman, and kendall.

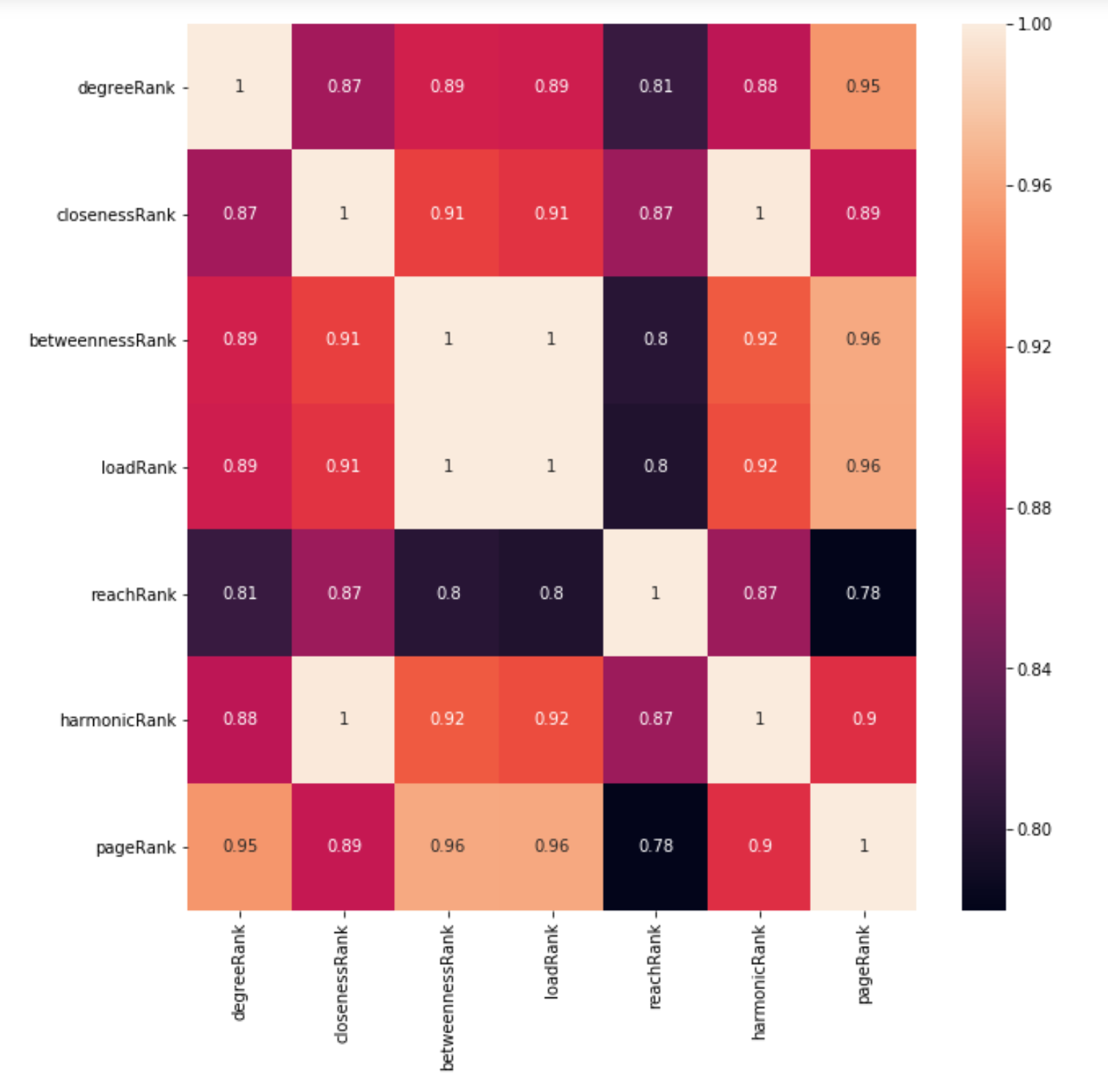


Fig. 8. Correlation Heatmap

**IIID. Data Preparation**  
  
As with all forms of machine learning, in order to produce accurate results it is important that we thoroughly clean the data. The process is rather consistent across each separate machine learning process we use, and will be detailed in this section.

The first part we must go about doing is creating the class outcome columns. The first step of this is to create a dictionary with the max ranking node for each network and centrality type. From there we simply create a binary list for each network type, where the list is comprised of 1 if a network is within the top 25%, and a 0 otherwise.

Following from that, we must simply split the data and balance it. We split the data with an 80:20 training:testing ratio. The balancing is a process of removing the majority until it has the same population as the minority in this situation. Finally, we must split the training data into 10 training and validation folds.

**IIIE. Machine Learning**  
  
Our machine learning process is completed via scikit-learn, where we use three different classifiers. The classifiers used are the random forest classifier, decision tree classifier, and the support vector classifier. These classifiers are all supplied via scikit-learn. Due to the computing power necessary to calculate what is needed, we ran the software on Miami University’s Redhawk supercluster.

1. **References**

[1] Koschützki, Dirk, and Falk Schreiber. “Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks.” Gene Regulation and Systems Biology, vol. 2, 2008, doi:10.4137/grsb.s702.

[2] Sparrow, Malcolm K. “The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects.” Social Networks, North-Holland, 3 July 2002, https://www.sciencedirect.com/science/article/pii/037887339190008H.

[3] Kuzubaş, Tolga Umut, et al. “Network Centrality Measures and Systemic Risk: An Application to the Turkish Financial Crisis.” Physica A: Statistical Mechanics and Its Applications, North-Holland, 12 Mar. 2014, https://www.sciencedirect.com/science/article/pii/S0378437114002003.

[4] Boldi, Paolo, and Sebastiano Vigna. "Axioms for centrality." *Internet Mathematics* 10.3-4 (2014): 222-262.

[5] Brandes, Ulrik. "On variants of shortest-path betweenness centrality and their generic computation." *Social Networks* 30.2 (2008): 136-145.

[6] Brandes, Ulrik, Stephen P. Borgatti, and Linton C. Freeman. "Maintaining the duality of closeness and betweenness centrality." *Social Networks* 44 (2016): 153-159.

[7] Freeman, Linton C. "Centrality in social networks conceptual clarification." *Social networks* 1.3 (1978): 215-239.

[8] Goh, K-I., Byungnam Kahng, and Doochul Kim. "Universal behavior of load distribution in scale-free networks." *Physical Review Letters* 87.27 (2001): 278701.

[9] Langville, Amy N., and Carl D. Meyer. "A survey of eigenvector methods for web information retrieval." *SIAM review* 47.1 (2005): 135-161.

[10] Mones, Enys, Lilla Vicsek, and Tamás Vicsek. "Hierarchy measure for complex networks." *PloS one* 7.3 (2012): e33799.

[11] Opsahl, Tore, Filip Agneessens, and John Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths." *Social networks* 32.3 (2010): 245-251.

[12] Page, Lawrence, et al. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab, 1999.

[13] Rochat, Yannick. *Closeness centrality extended to unconnected graphs: The harmonic centrality index*. No. CONF. 2009.

[14] Estrada, Ernesto. "Characterization of topological keystone species: local, global and “meso-scale” centralities in food webs." *Ecological Complexity* 4.1-2 (2007): 48-57.

[15] Trajanovski, Stojan, et al. "Robustness envelopes of networks." *Journal of Complex Networks* 1.1 (2013): 44-62.

[16] Meghanathan, Natarajan. “Correlation Coefficient Analysis of Centrality Metrics for Complex Network Graphs.” *SpringerLink*, Springer, Cham, 27 Apr. 2015, https://link.springer.com/chapter/10.1007/978-3-319-18503-3\_2.

[17] Jamakovic, Almerima, and Steve Uhlig. “On the Relationships Between Topological Measures in Real-World Networks.” *Networks and Heterogeneous Media*, June 2008, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.2216&rep=rep1&type=pdf.