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

## I. 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.).

Computing these centrality measures can be computationally intensive. To make matters worse, ranks of each node are taken based on the various centrality calculations for that node, so that nodes in the network each have a ranking to use for the training process. 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 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 demanding 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 to several types of networks. Specifically, our main contributions are as follows:

- Establishes a method for predicting one centrality metric based on several others
- Determines the correlation between different centrality measures, demonstrating that some possess more intrinsic overlap than others

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.

## II. 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 located in the center of a network or graph (Freeman). There are many forms of centrality measures including simplistic measures such as degree, closeness,

**Commented [r1]:** What should the CIA

**Commented [A2]:** Very nice opening

**Commented [r3]:** Syntax error

**Commented [A4]:** That may be a little rough to understand for a reader who doesn't already know about the 'ranking' thing. Maybe one more sentence?

**Commented [A5]:** Colloquial.

**Commented [A6]:** Betweenness centrality is a variant of stress centrality, because it's simply based on normalizing the number of shortest paths. But not all centrality measures are variants of each other: there is very little (computation-wise) in common between degree, closeness, and betwennness.

**Commented [r7]:** demanding

**Commented [A8]:** Good paragraph

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 machine learning classifiers, the centrality measures will have to be calculated on graphs that are randomly generated many times over. 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. Next, we will go into detail on the different centrality measures that are used in this research.

Commented [A9]: You may have to explain what they are, at least in just a sentence for each.

Commented [A10R9]: Ditto myself. Still needs to be done here.

## 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 $G$, defined as $G := (V, E)$ for $|V|$ vertices and $|E|$ edges. The total number of vertices is denoted by N. The degree centrality for a vertex $v$ is defined as $C_D = \frac{\deg(v)}{(N-1)}$ where $\deg(v)$ is the degree of vertex $v$ (Opsahl et al.).

Commented [A11]: Deg(v) has not been introduced. Please describe, otherwise the definition is not finished.

Commented [A12R11]: This problem still needs to be addressed.

## 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.

Commented [A13]: May need a brief definition

Commented [A14R13]: Still needs it.

The closeness centrality can be defined mathematically as follows. Closeness centrality is defined as $C_C(x) = \frac{(N-1)}{\sum_y d(y,x)}$ where $d(y,x)$ 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 $v$ is defined as $C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$ where $\sigma(s,t|v)$ is the number of shortest paths between s and t given they contain vertex v and $\sigma(s,t)$ 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 $v$ is defined as $C_l(v) = \sum_{s,d \in V} \theta_{s,d}(v)$ where $\theta$ is the quantity of information that is passed through vertex $v$ and $s$ & $d$ 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 $v$, for a given graph $G$, where the graph can be defined as $G \coloneqq (V, E)$ for $|V|$ vertices and $|E|$ edges is defined as $C_R(i) = \frac{1}{N-1} \sum_{j=0<d(i,j)<\infty} \frac{1}{d(i,j)}$ where $d(i,j)$ is the distance formula and $N$ 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 $v$, for a given graph $G$, where the graph can be defined as $G \coloneqq (V, E)$ for $|V|$ vertices and $|E|$ edges is defined as $C_H(v) = \sum_{d(y,x)<\infty, y \neq x} \frac{1}{d(y,x)}$ where $d(y,x)$ 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 $R(u)$ where $u$ is the node that is accessed. The page rank for a vertex $v$, for a given graph $G$, where the graph can be defined as $G \coloneqq (V, E)$ for $|V|$ vertices and $|E|$ edges is defined as $R(u) = c \sum_{v \in V} \frac{R(v)}{N_u}$ where $N_u$ 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:

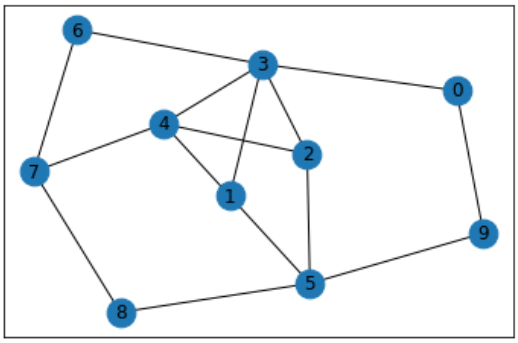


Figure 1. A randomly generated graph with ten nodes and fifteen edges.

| 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 |

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

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:

The following section is dedicated to related works in centrality analysis. The following research papers discuss elements of this research and provide a key background to the research.

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.

**Commented [A15]:** This is very important for your work. Hence it'd be worth listing which specific centralities, and which correlations. You also need to explain how he arrived at these conclusions. Otherwise, the summary is at the level of the abstract and passes over the relevant details from the paper itself.

**Commented [A16]:** Not sure about what you mean here.

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 Martín-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 $M_{a,b}(k)$ 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 $M_{closeness,eigenvector}(0.5) > 0.5$ 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 $M_{a,b}(k)$ had little overlap.

In the paper *Topological Measures in Real-World Networks*, the authors argue that there is an overabundance of network measurements and that this 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 effect 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 broader topic with a much more numerical approach. Instead of searching for overlap between centrality measures in different contexts, this paper seeks to find the general correlation coefficients between popular network centrality measures. For example, 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 analyzing data from several real-world

**Commented [A17]:** Odd grammar?

**Commented [A18]:** This is worth detailing, perhaps with an example. Do you see a room for something like this in your work?

**Commented [A19]:** Pick one

**Commented [RTA20]:** Over-abundance

**Commented [A21]:** Compared to what? Necessary complications? 😉

**Commented [RTA22]:** Not sure a comma is necessary here

**Commented [A23]:** Good observation. Clearly written.

**Commented [RTA24]:** Lower case

**Commented [A25]:** This is useful information for your work. It means that, for each network generated, you can also compute the average distance. It may help to get higher accuracy values.

**Commented [RTA26]:** effect

**Commented [RTA27]:** them,

**Commented [A28]:** "It's the same but not quite" is a common writing issue. Clearly state what they have in common and where they differ, if it helps make a smooth transition.

**Commented [RTA29]:** Not necessary to capitalize

**Commented [A30]:** Meaning?

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 paper titled C*orrelation between centrality metrics and their application to the opinion mode*' 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 analyzes the correlation of centrality metrics in terms of their Pearson correlation metrics - and from there introducing a new centrality measure deemed the degree mass. 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 are 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 $x_1$.

The next paper is *Consistency and differences between centrality measures across distinct classes of networks* studied by Stuart Oldham, Ben Fulcher, Linden Parkes, Aurina Arnatkevičiūtė, 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; unfortunately, the positively correlated measures are unspecified. 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, the researchers measured 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 was 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. The grid search optimization will optimize the best parameters to use for the model. Moreover, when they try to measure the distance between two graphs, the researchers used Spearman's rank correlation

**Commented [A31]:** Based on your readings, you may want to list which correlations have been done in previous research. You can then limit yourself to using these correlations, and you'll have references to say why.

**Commented [A32]:** Yet another useful feature to measure on your generated networks.

**Commented [RTA33]:** Networks but

**Commented [RTA34]:** In this paper,

**Commented [A35]:** May need to briefly state what "type" they are so we can see the connection with your work. For instance Barabasi-Albert is a scale-free network generator.

**Commented [A36]:** This will be very useful for your own work.

because of its ability to measure non-linear relations as well as outliers-sensitive graphs. Finally, testing on 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 can't capture. To do that, the researchers fit each generated model into real-world ones, then the structure metrics were computed and compared to the real-world networks 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.

There are many centrality indices that have been introduced but only some of them are correlated. Knowing the correlation between indices is crucial because some of the redundant or cheaper indices can be omitted. However, many reports show inconsistency in results, that is 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. This result is important to the research being demonstrated in this research.

In the paper *Correlation Analysis of Complex Network Metrics on the Topology of the Internet*, the researchers studied whether if there were correlations between different network metrics. These network metrics where specifically studied within the domain of the topology of the internet. The researchers examined many different network metrics including connectivity, clustering, distance, centrality, hierarchical, and scaling of network metrics. The researchers also examined the normalized values of the centrality metrics based on the number of vertices in graphs examined. The researchers found significant average correlation among the metrics studied. The researchers found that the centrality measures correlated highly with other network metrics such as betweenness is highly correlated with normalized average path length. However, it also exposed centralities the don't correlate well with anything such as eigenvector centrality did not show significant correlations with many of the metrics studied. The researchers exposed many other ways of examining networks other than centralities and it could expand the research being performed in this paper.

In the paper *The Correlation of Metrics in Complex Networks with Applications in Functional Brain Networks*, the researchers set out to analyze network metrics that applied to brain networks. They studied four different network models including Bárabasi–Albert, Erdös–Rényi, and functional brain models in order to examine if the network metrics are correlated. The purpose of this paper was very similar to the research done in the previous paper however the

**Commented [A37]:** This is technically advanced so it may need a bit more explaining/illustrating to get the point.

**Commented [A38]:** Look for words that are underlined before you submit.

**Commented [A39]:** You authored that paper too? 😉

**Commented [RTA40]:** Strange wording

**Commented [A41]:** Not sure what this means…

**Commented [A42]:** That's an important point for your work. Always try to emphasize the connection between "factoids" from previous readings and the work done here.

applications differ between the topology of the internet and brain networks. The researchers studied the correlations with each of the different degree distributions (different models) and wanted to examine if they were uniform or different. The research proved that the metrics are topology dependent and are highly correlated. The metrics that the researchers analyzed were: degree diversity, eigenvalues, assortativity, clustering coefficient, average hopcount, and global efficiency.

Two other papers that could be included in future literary reviews are *Axioms for Centrality* and *Centrality in Social Networks Conceptual Clarifications*. Both of these papers discuss the different centrality measures that were analyzed in this research thoroughly. The later paper is an old paper that discusses the basics of three centrality measures.

The information in table 2 discusses the differences in the related works and display the different things they are talking about. The different columns in the table represent the centrality measures examined, what did the researcher's study, what networks did they use, other network characteristics that were analyzed, and any conclusions.

| Paper | Centrality measures | Relationship between measures | Networks studied | Other network characteristics | Conclusions | Notes |
|---|---|---|---|---|---|---|
| **Characterization of topological keystone species Local, global and 'meso-scale' centralities in food webs** | Degree, Between-ness, Closeness, Eigenvector, Information, Subgraph, | Correlations and rankings of centrality measures | Ecosystems with keystone species | N/a | Picking the most central node using any measure is an effective strategy to identifying keystones in the ecosystem | Keystone species is a species that when removed, causes serious issues in the ecosystem. Biological application of networks. |
| **Robustness envelopes of networks** | Degree, Betweenane-ss, Closeness, Eigenvector | Studied targeted attacks on nodes with high centrality values | Erdos-Renyi, Watts-Strogatz, Barabasi-Albert, Real World Networks: Western US & Western EU power grids | Degree assortativity, Degree-preserving rewiring | Even if some networks have similar average case performance under attack, they may differ significantly in different attack sequences from each other. | Defined similarity metric. |
| **Topological Measures in Real-World Networks** | Degree Centrality | Correlation between network metrics to compare the | N/a | Distance Distribution and | Smaller distance is much more | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | rankings between different real-world networks | | Clustering Coefficient | likely to have high-degree nodes that form tight subgraphs | |
| **Correlation Coefficient Analysis of Centrality Metrics for Complex Network Graphs** | Degree, Eigenvector, Betweenness, Closeness, Farness, and Eccentricity | Correlations of popular centrality measures | Real World Networks | N/a | When variation in degree distribution increases, the correlation between centrality metrics increase. | |
| **Correlation between centrality metrics and their application to the opinion model** | Eigenvector, betweenness, closeness, degree | Correlation between the centrality measures and the degree mass metric | Real World Networks | Degree mass | 0th-order degree mass is the degree and the high order degree mass is proportional to the principal eigenvector | |
| **Consistency and differences between centrality measures across distinct classes of networks** | Degree, Eigenvector, Katz, PageRank, Leverage, H-Index, Laplacian, Shortest-Path, Closeness, Subgraph, | Correlation and consistency of network centralities across different networks | 107 Networks from Ghasemian and colleagues from Index of Complex Networks | density, global efficiency, modularity, majorization gap, and spectral gap | There is a medium to high correlation between centralities and these metrics | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Participation, Communicability, Information, Betweenness, Bridging | | | | | |
| **On the Structural Properties of Social Networks and their Measurement-calibrated Synthetic Counterparts** | N/a | Define correlation patterns and how they differ across domains | Online friendship networks, Reply networks, and co-authorship networks (scientific) | Assortativity, average clustering, average degree, pseudo diameter, interval degree probabilities, max degree divided by size | The fit depends heavily on domains | |
| **Correlations among centrality indices and a class of uniquely ranked graphs** | Indegree, outdegree, degree, Betweenness, s-Betweenness, Closeness-in, Closeness-out, S-Closeness, Integration, Radiality, S- | Study whether correlations are more indicative of network structure than relationships between indices | Used networks from previous studies highlighted in paper | N/a | Correlations among centralities are not indicative of formal and conceptual similarity | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | int/rad, Eigenvector | | | | | |
| **Correlation Analysis of Complex Network Metrics on the Topology of the Internet** | Closeness, Betweenness, Eigenvector | Correlations between the different measures and using machine learning to verify the non-redundant measures. | Autonomous System Networks (ASN) | Assortativity, Clustering, HClusteringD3, Degree Scaling, Nei Conn Scaling, Rich Club Scaling, H ClusteringD2, Clustering Scaling, Rich Club, Edge Betweenness, Diameter, Path Length, Dominance, Density, Max Degree, Nei Conn | There are clear relationship between subsets of metrics on ASN networks | ASN is a group of millions of host IP addresses |

| | | | | | | |
|---|---|---|---|---|---|---|
| **The Correlation of Metrics in Complex Networks with Applications in Functional Brain Networks** | N/a | Examine correlations between different metrics and different network types. Then examine correlations in functional brain networks | Barabasi-Albert, Erdos-Renyi, Watts-Strogatz, functional brain networks | Degree diversity, Assortativity, Clustering Coefficient, Average hopcount, Global efficiency, Spectral radius, Effective graph resistance, Algebraic Connectivity, Ratio | By examining the correlations between metrics on all network types, metrics | |

Table 2. Comparing and Contrasting the related works

## III. 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 the associated methods, and scikit-learn implements our methods of machine learning.

### IIIB. Overall Process

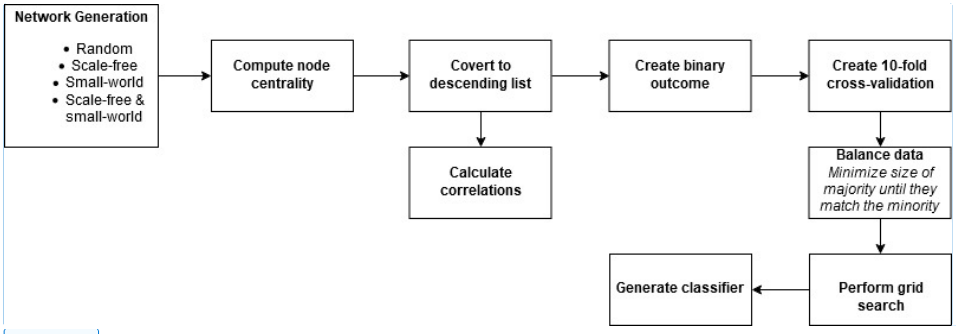


Figure 2, Overall process

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 200 graphs for each type, further subdivided in that by taking 50 of each specified size within the array [100, 200, 400, 800]; thus, creating a total of 800 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. PageRank and Harmonic

**Commented [r43]:** Network-based

**Commented [A44]:** Very decent process overview. A few tweaks to finalize this image for your last milestone:
1) You start with "network generation". So that box should be first. Scale-free, small-world, scale-free & small-world, random… can be written under the box as bullet points or equivalent. They're "cases of" or "details about" network generation.
2) Then you "compute node centrality". So that's the 2nd box. Again, the specific metric can be written underneath.
3) You turn centrality measures into a descending rank. That's another box.
4) A) You calculate correlation. That's a final box, nothing leaves this one.
B) You create a binary outcome. The details should state based on what.
5) Classification needs to show the stages. First there is a 10-fold cross-validation. Then within each fold, the "data to build" is balanced (how? Write details underneath). Then, within each balanced training subset, you optimize parameters (how? Write that it's a grid search). And so on.

**Commented [A45]:** Check the numbering before submission as part of proofreading.

**Commented [r46]:** Discuss the other network metrics that were generated for milestone05 here

centralities were also used and apart of the NetworkX library.

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.

Commented [r47]: This may also be a good place to discuss the need for multithreading and how that was implemented

### IIIC. 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 8:2 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 to find the optimal parameters via a grid search. Once we find the optimal parameters, we use the 8:2 training:testing ratio and train it using the results of the grid search.

### IIID. 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, naïve-bayes, gradient descent, and the support vector classifier. These classification algorithms 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.

Commented [r48]: We also used naïve-bayes and SGDClassifier

Commented [RTA49R48]:

Commented [A50]: The text will need to present and discuss the results.

## IV. Results

Our experiments result in strong findings, with high accuracy in our prediction columns. In the end, we are using: network type, size, degree rank, closeness rank, betweenness rank, load rank, reach rank, harmonic rank, page rank, average distance, average clustering, number of cliques, and strongly connected nodes as prediction metrics. And from that, we are predicting binary results as to whether it is within the top 25% or not, using degrees, closeness, betweenness, load, reach, harmonic, and page rank.

Below are the overall results for every classifier we tested. The result with both the decision tree and random forest proves extraordinary, with high metrics all around.

| | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| **Decision Tree** | 0.982 | 0.9383 | 0.9627 |
| **Random Forest** | 0.9831 | 0.9524 | 0.9604 |
| **Bernoulli Naïve Bayes** | 0.7927 | 0.0357 | 0.0224 |
| **Linear Support Vector** | 0.9022 | 0.7944 | 0.8387 |
| **Stochastic Gradient Descent** | 0.9045 | 0.7173 | 0.8140 |

Table 2. Overall results for each classifier

Displayed below are individual results for every type of graph and every metric.

| | **Overall** | **Scale-Free** | **Small-World** | **Random** | **Small-World/ Scale-Free** |
|---|---|---|---|---|---|
| **Decision Tree** | 0.9820 | 0.9582 | 0.9853 | 0.9933 | 0.9913 |
| **Random Forest** | 0.9831 | 0.9602 | 0.9870 | 0.9942 | 0.9912 |
| **Bernoulli Naïve Bayes** | 0.7927 | 0.8945 | 0.7723 | 0.7470 | 0.7569 |
| **Linear Support Vector** | 0.9022 | 0.9142 | 0.9185 | 0.9329 | 0.8434 |
| **Stochastic Gradient Descent** | 0.9045 | 0.9128 | 0.9240 | 0.8910 | 0.8903 |

Table 3. The accuracy of each classifier for specific graph types

| | **Overall** | **Scale-Free** | **Small-World** | **Random** | **Small-World/ Scale-Free** |
|---|---|---|---|---|---|
| **Decision Tree** | 0.9383 | 0.8190 | 0.9649 | 0.9981 | 0.9812 |
| **Random Forest** | 0.9524 | 0.8713 | 0.9695 | 0.9887 | 0.9801 |

| | | | | | |
|---|---|---|---|---|---|
| **Bernoulli Naïve Bayes** | 0.0357 | 0.1429 | 0.0 | 0.0 | 0.0 |
| **Linear Support Vector** | 0.7944 | 0.7384 | 0.9095 | 0.9811 | 0.7812 |
| **Stochastic Gradient Descent** | 0.7173 | 0.5815 | 0.7462 | 0.8982 | 0.6431 |

Table 4. The precision of each classifier for specific graph types

| | **Overall** | **Scale-Free** | **Small-World** | **Random** | **Small-World/ Scale-Free** |
|---|---|---|---|---|---|
| **Decision Tree** | 0.9627 | 0.9083 | 0.9739 | 0.9851 | 0.9833 |
| **Random Forest** | 0.9604 | 0.8937 | 0.9766 | 0.9881 | 0.9831 |
| **Bernoulli Naïve Bayes** | 0.0224 | 0.0896 | 0.0 | 0.0 | 0.0 |
| **Linear Support Vector** | 0.8387 | 0.7884 | 0.8252 | 0.9811 | 0.7601 |
| **Stochastic Gradient Descent** | 0.8140 | 0.6686 | 0.9048 | 0.7660 | 0.9167 |

Table 5. The recall of each classifier for specific graph types

Commented [r51]: You may also want to discuss how each of the centralities perform. This is important because it is one of the essential questions to our research.

Displayed in Table 6 is the predicted optimal centrality measure for each type of graph type and for each classifier.

| | **Scale-Free** | **Small-World** | **Random** | **Small-World/ Scale-Free** |
|---|---|---|---|---|
| **Decision Tree** | Load | Load | Harmonic/ Reach | Reach |
| **Random Forest** | Load | Betweenness | Harmonic/ Reach | Harmonic |
| **Bernoulli Naïve Bayes** | Betweenness/ Load | Betweenness | Harmonic | Degree |
| **Linear Support Vector** | Load/ Closeness | PageRank | Harmonic | Degree |

| | | | | |
|---|---|---|---|---|
| **Stochastic Gradient Descent** | Betweenness | Degree | Reach | Harmonic |

Table 6. The optimal centrality measure for each type and classifier

Commented [RTA52]: Also, what ever happened to a discussion for the correlations that were calculated? That should also be included here.

## V. References


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

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

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

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

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

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

Jamakovic, Almerima, and Steve Uhlig. "On the relationships between topological measures in real-world networks." NHM 3.2 (2008): 345-359.

Koschützki, Dirk, and Falk Schreiber. "Centrality analysis methods for biological networks and their application to gene regulatory networks." Gene regulation and systems biology 2 (2008): GRSB-S702.

Kuzubaş, Tolga Umut, Inci Ömercikoğlu, and Burak Saltoğlu. "Network centrality measures and systemic risk: An application to the Turkish financial crisis." Physica A: Statistical Mechanics and its Applications 405 (2014): 203-215.

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

Meghanathan, Natarajan. "Correlation coefficient analysis of centrality metrics for complex network graphs." Computer Science On-line Conference. Springer, Cham, 2015.

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

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

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

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

Sparrow, Malcolm K. "The application of network analysis to criminal intelligence: An assessment of the prospects." Social networks 13.3 (1991): 251-274.

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