3-22-2012

# Analysis of Social Network Measures with Respect to Structural Properties of Networks

Joshua D. Guzman


# ANALYSIS OF SOCIAL NETWORK MEASURES WITH RESPECT TO STRUCTURAL PROPERTIES OF NETWORKS

THESIS

Joshua D Guzman, Second Lieutenant, USAF

AFIT/OR/MS/ENS/12-12

**DEPARTMENT OF THE AIR FORCE**
**AIR UNIVERSITY**

***AIR FORCE INSTITUTE OF TECHNOLOGY***

**Wright-Patterson Air Force Base, Ohio**

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/OR/MS/ENS/12-12

# ANALYSIS OF SOCIAL NETWORK MEASURES WITH RESPECT TO STRUCTURAL PROPERTIES OF NETWORKS

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Joshua D Guzman, BA

Second Lieutenant, USAF

March 2012

AFIT/OR/MS/ENS/12-12

# ANALYSIS OF SOCIAL NETWORK MEASURES WITH RESPECT TO STRUCTURAL PROPERTIES OF NETWORKS

Joshua D Guzman, BA
Second Lieutenant, USAF

Approved:

| | |
|---|---|
| //SIGNED// | 12 March 2010 |
| Richard F. Deckro, DBA, (Advisor)<br>Professor of Operations Research<br>Department of Operational Sciences | date |
| //SIGNED// | 12 March 2010 |
| Matthew J. Robbins, Maj, USAF, PhD (Reader)<br>Assistant Professor of Operations Research<br>Department of Operational Sciences | date |

## Abstract

Social Network Analysis (SNA), the study of social interactions within a group, spans many different fields of study, ranging from psychology to biology to information sciences. Over the past half century, many analysts outside of the social science field have taken the SNA concepts and theories and have applied them to an array of networks in hopes to formulate mathematical descriptions of the relations within the network of interest. More than 50 descriptive measures of networks have been identified across these fields; however, little research has examined the findings of these measures for possible relationships. This thesis tests a set of widely accepted SNA measures for correlation and redundancies with respect to the most accepted network structural properties; size, clustering coefficient, and scale-free parameter. The goal of this thesis is to investigate the SNA measures' ability to discriminate and identify different actors in a network. As a result this study not only identifies high correlation amongst many of the tested measures, it also aids analysts in identifying which measure best suits a network with specific structural properties and its efficiency for a given analysis goal.

*To my Family and Friends*
*Thank you for all your love and support*

## Acknowledgements

I would like to express my heartfelt appreciation to my faculty advisor, Dr. Richard Deckro, for his guidance and genuine support throughout the course of this thesis effort. His insight and experience in the field, and as a mentor, was most certainly appreciated. To my reader and class advisor, Maj Robbins, I thank you for to your mentorship and support throughout my time at AFIT. I would, also, like to thank Mr. James Morris and Mr. Nicholas Ballester, who spent their valuable time explaining the processes and procedures that they used in the fields of Social Network Analysis and computer programming.

I am also indebted to the many professionals of the Operational Science Department that helped keep me progressing thorough out the past 18 months. Lastly, I would like to thank my fellow classmates for helping me push through over the past year and a half.

Joshua D Guzman

## Table of Contents

## List of Figures

## List of Tables

# ANALYSIS OF SOCIAL NETWORK MEASURES WITH RESPECT TO STRUCTURAL PROPERTIES OF NETWORKS

## 1. Introduction

*"After a half century of focusing on a Major Theater War with a near-peer competitor, the nation awoke on Sept. 11, 2001 to find out that a new principal threat to the U.S. is terrorism."* - *C. Clark, Captain, USAF, 2005*

### 1.1. Background

As Clark notes in the quote above, new light has been shed on the study of the unconventional organizational structures that have come to be known as terrorist networks. The United States and the world have been thrown into a race to accurately describe and model these networks and the behaviors of the players within them. Although the study of social based relationships and social interactions in groups have been undertaken for decades before September 11, 2001 by social scientists, the world found itself initially lacking accurate ways to measure and model these terrorist networks. Within the last decade, many theories and models have been developed to assist decision makers with the analysis of different types of networks.

As of 2012, there are over 50 Social Network Analysis (SNA) descriptive measures that have been published across the various fields of studies that analyze relations and connections within a given social network (Hagberg, Schult, & Swart, 2008; Borgatti, Everett, & Freeman, 2002). For this study, these SNA measures are divided into four distinct types of output, and identify everything from the individuals whom are most well connected to an individual through which most information is expected to flow. The four groups of network measures are those that describe the overall graph or network, those that describe each node or actor, those that describe each relationship or tie and those that describe the subgroups or clusters of the network.

Since SNA has been used across many fields of study, there is a need for research that compares the performance of these measures, and their computational times. It is necessary to investigate both the efficacy and efficiency of each measure with respect to known network characteristics. This research helps to guide analysts, who wish to study a specific network structure, in selecting the correct measure, saving valuable time and resources.

## 1.2. Problem Definition

The over 50 SNA measures span fields of studies, from sociology to mathematics, but have not been collectively examined from an efficiency and efficacy viewpoint. SNA analysts often utilize the measures that they are most comfortable with in order to draw conclusions about a given network. This technique can lead to a shallow interpretation of the network; or worse, an analysis that improperly describes the network relationships. Moreover, this practice can lead to an inefficient analysis that could be more readily resolved by using a similar measure. These inefficiencies are the result of the wide ranging types of networks and the lack of understanding of the array of network measures and their interactions with the network topology. This thesis investigates each measure's ability to discriminate among actors, identify different actors with respect to network structural properties and the computational times at which it accomplishes these tasks. Specifically, correlation and redundancy of each measure is addressed, as well as which measure best suits a network's structural properties.

## 1.3 General Assumptions and Scope

As with every experiment and test, it is important to state the assumptions applicable to the experiment and results. The first assumption is that the analysts, using the results herein, possess the means to collect complete and accurate data on the social network of interest. Second, all social network connections, or relationships, between members are undirected; flow

between actors may traverse an edge in either direction. This assumption allows for the investigation of a greater set of SNA measures that apply to both directed and undirected networks. Third, though social networks can be viewed dynamically, the networks examined throughout this thesis are measured at an instant in time. This allows for the analysis to be focused on the measure and not the stochastic nature of network actors. Fourth, in order to conduct a proper Design of Experiment (DOE) the randomization in which data is collected must be addressed in order to not introduce error due to outside factors. This randomization in networks is appropriately accounted for within the computer generated networks algorithm used in this experiment (Morris, O'Neal, & Deckro, 2011). Fifth, all networks edges are generated according to the power law with an estimated scale free exponent α (Morris, O'Neal, & Deckro, 2011). Finally, nodal rankings are nonparametric and cannot be described by a general distribution. Therefore, Spearman's Rank Correlation Coefficient (ρ), Kendall's Tau (τ) tests for correlation, and plots and overlays are utilized to indicate further investigations. These assumptions are addressed in further detail in Chapters 2 and 3.

This research's primary focus is the efficiency and efficacy of a group of the most widely accepted SNA measures that describe and rank each node in a graph. With the help of Subject Matter Experts (SMEs) and review of SNA literature, 29 node descriptive measures were chosen for testing, outlined in Table 1. These metrics include measures of centrality, betweenness and clusterability. Testing the correlation of more descriptive measures increases by $\binom{n}{2}$ for n measures; therefore this thesis limits the number of measures to 29. Even with this limitation a great deal of insight is found to aid analysts in choosing the appropriate choice in particular network settings.

**Table 1: List of SNA Measures Analyzed**

| | | | |
|---|---|---|---|
| Degree Centrality | Betweenness Centrality | Closeness Centrality | Eigenvector Centrality |
| Diversity | Clustering | Soffer's Clustering | Flow Betweenness |
| Length Betweenness | Endpoint Betweenness | Communicability Centrality | General Diversity |
| Newman's Betweenness | Linear Betweenness | Communicability Betweenness | k-Betweenness |
| PageRank | Closeness Vitality | Proximal Betweenness (S) | Stress Centrality |
| Load Centrality | Squares Clustering | Proximal Betweenness (T) | Neighbor |
| Hubs | Current Flow | Approx. Current Flow | Core Number |
| Authorities | | | |

## 1.4 Research Objectives

The objectives of this research are as follows:

- Provide well tested results of preference and correlation between well known SNA measures.
- Determine the efficiency and efficacy of descriptive network measures with respect to each other.
- Provide guidance for SNA analysts to choose the appropriate descriptive measure with respect to network structural properties.

## 1.5 Thesis Overview

The organization of this thesis is as follows. Chapter 2 presents a review of pertinent literature in Design of Experiments (DOE), network structural properties and Social Network Analysis to support this research. The use of a network generator, the Spearman's Rank Correlation Coefficient and Kendall's Tau tests for correlation are also addressed within chapter 2. Chapter 3 provides an overview of the complete methodology used; its general assumptions and the hypotheses tested. Chapter 4 describes the results and analysis of the research. Chapter 5 reviews overall, general conclusions as well as recommendations for future research.

# 2. Literature Review

## 2.1. Introduction

This chapter provides an overview of the basic foundation that supports this research. Design of Experiments, network structural properties and Social Network Analysis as well as the use of a network generator and statistical tests preformed are reviewed. While this section does not provide full detail of past research, it gives a general base of knowledge. Cited references can be used to further research a topic outside the scope of this thesis.

## 2.2. Social Network Analysis

In this day and age, networks encompass a great deal of our everyday life. Probably the most prominent example is the World Wide Web as well as the multitude of social networking websites, such as Facebook©, Twitter© and Google+©. Measuring the connections and players in these networks bring substantial insight into the network, especially relationships that may not be noticed while studying a conventional organizational chart. This insight can be enhanced through Social Network Analysis.

The thought that comes to nearly every mind when the words 'social network' are heard is that of modern day, online social networking websites such as Facebook. Although these can be modeled using SNA, the study of social networks came long before the creation of the computer and the World Wide Web. Sociologists have long since studied groups of people, organizations and systems. In fact, "the true foundation of the field is usually attributed to psychiatrist Jacob Moreno, a Romanian immigrant to America who in the 1930s became interested in the dynamics of social interactions within groups of people" (Newman, 2010, p. 36). Over the next few decades the basis for SNA was laid by many researches who used SNA to study everything from friendships to the spread of diseases.

In the 1960s, psychologist, Stanley Milgram performed his now famous Small World Experiment which later came to be associated with the name "Six Degrees of Separation". He used SNA to determine the average number of people it took to get from any one person to another (Newman, 2010, p. 55). In his experiment, Milgram mailed out 96 packages to recipients in Omaha, Nebraska; detailed instructions asked the recipient to attempt to get an official looking passport to a specified person living in Boston, Massachusetts. The catch was that the only information included was the target's name, address, occupation and that they must only try to get the passport to the target by giving it to someone they knew on a first name basis that they thought may have a better connection. At the conclusion of the experiment, Milgram observed (out of the 18 passports that made it to their target) that on average there were about six people needed to get from the recipient to the target (Newman, 2010, p. 55).

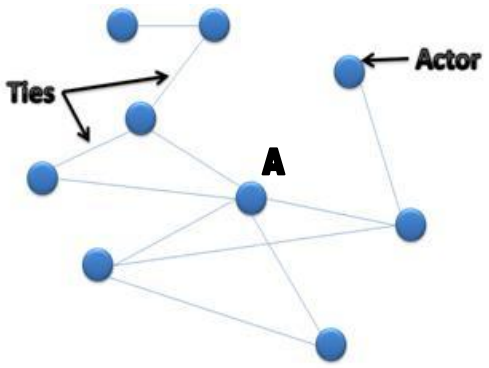


**Figure 1: Simple Network**

Although there were many critics to Milgram's experiment, his research sparked others to look in to patterns such as this and in turn create metrics to measure different aspects of social networks. In the more recent decades many social scientists, as well as experts in a variety of other fields, have taken an interest in SNA. Current and widely accepted sources on the subject can be found in texts such as those written by Wasserman & Faust and Newman (1994; 2010).

The universally accepted definition of a social network is "the set of actors and the ties among them" (Wasserman & Faust, 1994, p. 9). In this definition we find that actors and ties can represent a wide variety of people, places and things. For example, an actor could represent a

single person and the ties between two people could represent whether or not those two are considered friends. Other possible representations for an actor are organizations, or groups of people, objects, such as a computer or router in the World Wide Web, or places like distribution centers for a business. Ties also can portray many different bonds between actors, such as relationships, information flow, influence, physical connection (wire between routers), and so forth.

As shown in Figure 1, networks are often represented by graphs where the vertices are refered to as actors and the edges connecting the actors are called ties. This convention comes from graph theorists notation describing a graph, G, with *n* nodes and *m* arcs, as well as using adjacency matrices, n x n, to show which actors are connected. Examples of adjacency matrices are shown in Figure 2 (Weisstein, 2005). One reason that networks are often represented by graphs is because of the relative ease that these graphs can be described by mathematical means. Another reason to use graphs similar to the one in Figure 1 is because it is often easy to visually see patterns in the network that may not be evident in the classic organization chart or a matrix representation for those not versed in the mathematics of network models and graph theory. Answers to critical questions, such as "Who are the "go to" persons before making a decision?" may be very visible in a graph

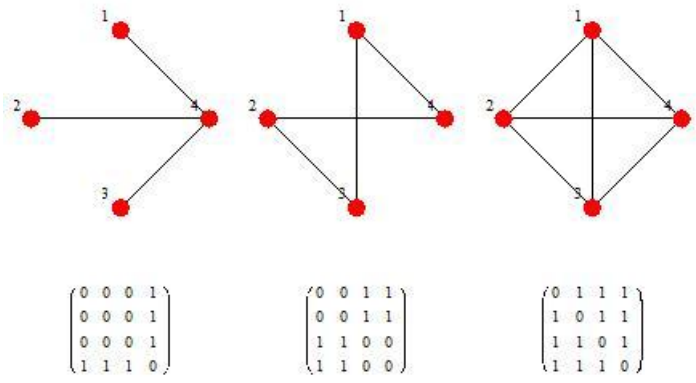

**Figure 2: Example of Adjacency Matrices**
**(Weisstein, 2005)**

that models employees as actors and who they would consult as the relationship ties. Actor A in Figure 1 may be such a person. This actor is well connected to the rest of the actors and may be a person who is highly influential to the decisions made in this network.

## 2.3. Types of Networks

This thesis focuses on four distinct types of networks that are found in network analysis. Specification of these network types limit the scope of the research problem and allows for proper assumptions to be made.

### 2.3.1 Weighted vs. Binary-Weighted Networks

One aspect that has proven to be a challenge in the modeling of social networks is representing the characteristics of a relationship between two actors. It may be convenient to assume that if actor A is connected to actor B then both A and B influence or flow information at an equal rate between each other, but the fact is that sometimes ties flow with unequal weights. As with a hierarchical business, a connection is most certainly present between an employee and their supervisor, but the weight of the influence that one has over the other is certainly not symmetric. Once properly defined, these weights can be used to mathematically solve different measures of the network, but may make the process more complex. For the research at hand unweighted, or binary-weighted, networks are used to allow computations of larger sized graphs to solve in a practical amount of time. In addition, not all measures were derived for weighted graphs, therefore in order to use all of the

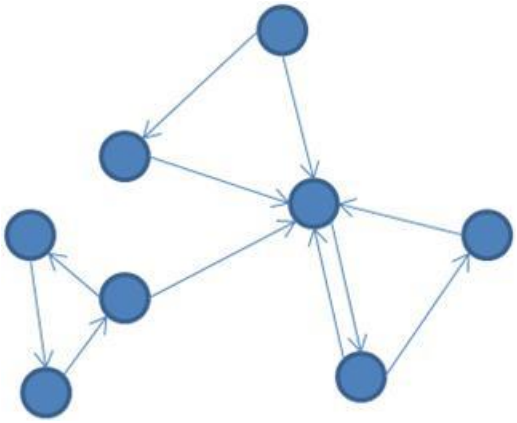

**Figure 3: Simple Directed Network**

mentioned measures the weight of an arc is limited to 1 if there is a tie between two actors or 0 if there is not (Hagberg, Schult, & Swart, 2008).

### 2.3.2 Directed vs. Undirected Networks

Likewise, most researchers assume that if actor A is connected to actor B then actor B is likely to be connected to actor A, but, again, sometimes ties flow with direction. Figure 3 shows an example of a directed network. As with weighted networks, there are measures that do not allow for measurement of directed networks, this also reduces the number of measures that can be used in this study. Another drawback to directed networks is that they have the tendency to increase the computation time of the SNA measure. Measures for undirected networks are more readily available and used throughout the field of SNA. These have been widely studied since the 1930's (Newman, 2010; Hagberg, Schult, & Swart, 2008; Borgatti, Everett, & Freeman, 2002). For these reasons, this research also restricts the networks tested to undirected networks. A few betweenness measures are also used in this study that were derived for directed networks but allow for undirected networks by assuming that a tie could be seen as two separate and opposite ties.

### 2.3.3 Flat vs. Layered Networks

In addition to whether a network is weighted/directed or not, they may also be classified into several other categories. The break out of these categories over a single network highlights what is known as a layered network and is illustrated in Figure 4 (Hamill, 2006, p. 6). The layers in this network highlight the different affiliations that

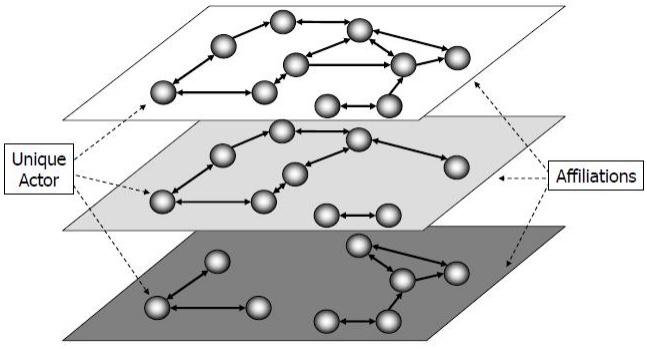


**Figure 4: Simple Layered Network**

**(Hamill, 2006)**

the actors in that network may have. For example, the top layer of Figure 4 may represent the acquaintances of the unique actor in a day, while the middle layer may show those who are considered friends, and lastly the bottom layer may show those who are related to the actor. Layers can represent a plethora of distinctions within a network. Layers themselves can also be weighted when determining which is more important for the question at hand. More on weighting layered networks can be found in the studies by Clark, Hamill and Geffre (2005; 2006; 2007). Very few measure have been defined for layered networks; therefore this research restricts networks to a single layer.

### 2.3.4 Dark vs. Bright Networks

Another distinguishing attribute that seperates out a particular type of network is whether the organization is "trying to hide their structures or are unwilling to provide information regarding their operations; examples include criminal networks, secret societies, and, most importantly, clandestine terrorist organizations" (Hamill, 2006, p. 3). These networks are often referred to as dark (a term coined by Raab &Milward in 2003), secret, covert, non-cooperative or clandestine networks (Raab & Milward, 2003). In the past decade a great deal of research has been carried out on clandestine networks in attempt to influence, disrupt, disband or destabilize these types of networks. The greatest motivation for the surge in this area comes from the increased threat from terrorist groups. The challenge for analysts when trying to model a clandestine network is that these networks depend on the secrecy for their continued existence and therefore data collection on the interactions of the network is difficult, if not nearly impossible (Geffre, 2007). A good description as to why dark networks need to be distinguished from its counterpart, bright or cooperative networks, is best described by Clark (2005, pp. 2-1):

> Many of the assumptions made by social scientist about individual importance to a network are based on the members' connections; more connections imply

> greater importance. Leaders of clandestine networks practicing good OPSEC [operational security], however, by design will likely have very few connections, and *may* not be uncovered through classic SNA techniques.

This concept can be visualized by the recent US mission that ended in the death of Osama Bin Laden. Intelligence of where Bin Laden was located is said to be from tracking his personal courier along with other personnel. Intelligence analysts were able to detect that the courier was acting as a gatekeeper of most information passed to the leader of Al-Qaeda (Labott & Lister, 2011). However, if only a rudimentary SNA had been conducted, Bin Laden might have seemed to be an isolated node.

Counter to clandestine networks are bright or cooperative networks. Most of SNA theories and measures are geared for these 'open source' networks. Cooperative networks freely give information to a SNA analyst. This is not to say that all information is disclosed or correct, as some participants may forget about a relationship with another individual or lie on a survey question (Newman, 2010, pp. 39-49). Some examples of cooperative networks are businesses, open source information given on internet social networking sites or even organizations within failing states seeking help to strengthen or restructure. Other bright networks can include physical structures such as a power grid or internet routing. For these reasons, and because of the difficulty of generating a truly dark network, this thesis assumes that all data is accurate and complete.

## 2.4. Structural Properties of Networks

This section gives a description of a few of the network structural properties that explain the network as a whole. The following network structural properties are looked at in the generated networks to find significance, if any, in choosing a descriptive measure. Throughout this thesis the following notation will be used to describe a network. A graph $G = (V, E)$

containing of the set of $V = \{1, 2, \ldots, n\}$ nodes (actors) and the set of $E$ edges (ties) within the network.

### 2.4.1 Size

This study defines the size of a network as the number of nodes in the given network. This attribute is often the first and most simple way to convey the scope of a social network. In SNA, nodes are most often people; therefore this measure is easiest to describe as the number of people in the network. Equation (1) gives the mathematical representation of the size of a graph.

$$n = |V| \tag{1}$$

where $|V|$ is the number of nodes in the set $V$.

### 2.4.2 Density

The density of a graph is defined as the number of edges or lines in a graph ($l$), expressed as a proportion of the maximum possible number of lines (Scott, 1987, p. 18). The formula for the density is as follows:

$$Density = \frac{l}{n(n-1)/2} \tag{2}$$

### 2.4.3 Diameter and Radius

The diameter and radius of a network draw upon the measure of eccentricity of all nodes in the graph; that is, the maximum distance from one node to any other node. The diameter is defined as the maximum of the eccentricities of the graph, whereas the radius is defined as the minimum (Wasserman & Faust, 1994, p. 111).

$$Eccentricity_i = e_i = Max(d(i, j)) \quad \forall j \tag{3}$$

where $d(i, j)$ is the minimum distance from node $i$ to node $j$

$$Diameter = Max(e_i) \quad \forall i \tag{4}$$

$$Radius = Min(e_i) \quad \forall i \tag{5}$$

### 2.4.4 Clusterability, $c_i$

The clusterability of a graph is a measure of the groupings of the nodes within the network. Clustering is defined as the formation of triangles with edges and nodes (Watts & Strogatz, 1998). The clusterability, also known as the clustering coefficient, is said to be the average clustering of all nodes; where the clustering coefficient of a node is the number of triangles that include the node over the number of possible triangles containing the node. Equation **(**6**)** describes this measure (Soffer & Vazquez, 2005, pp. 1-2).

$$c_i = \frac{2t_i}{\deg(i)[\deg(i)-1]} \tag{6}$$

where $t_i$ is the number of triangles containing node $i$ and deg($i$) is the degree or number of edges that include node $i$.

## 2.5. Design of Experiments

Proper design of an experiment is very important in order to ensure that the results are nonbiased and allow analysis with minimum nuisance factors. In a DOE the following four steps are used to make sure that the results are useful in the experiment. First, the analyst must recognize the problem. This may seem like a simple step, but it is often overlooked or misinterpreted. Knowledge of the problem will allow for the experiment to flow smoothly into the second step; choosing factors, levels and ranges. The correct choice of factors, levels and ranges depend on the system being analyzed and what influences the response. Potential design factors are factors that the experimenter wishes to vary over the levels chosen. The third step, choose a response variable, may be performed before, concurrent or after step two. Choosing a

response variable is vital to what the results tell the analyst. The response should provide useful information about the system and be geared toward the objective of the study. Lastly, the analyst must choose the type of experimental design. Different designs have their benefits in different situations (Montgomery, 2009, pp. 11-19). When choosing a design, several aspects of the system should be looked at:

- Is there a time suspense for the experiment?
- Is there a cost to the experiment?
- How many responses are there?
- How many factors, levels and ranges are there?
- How many (if any) replications will be performed?

Experimental designs include general factorial, multiple factorial or randomized block design along with many others. For further information on design and analysis of experiments see the texts of author Douglas Montgomery (2009, pp. 11-19).

### 2.6. Network Generators

There are several network generators that have become well known over the years. These generators have gone through the growth of SNA starting at small world to random graphs to scale-free networks. In addition, there has been increased interest in degree based generators. The three generators that this thesis is interested in are the Erdos-Renyi (ER) random graph generator, the Barabasi-Albert (BA) scale-free graph generator, and the Prescribed Node Degree, Connected Graph (PNDCG) generator. Each of these generators has their pros and cons. The ER graph generator was developed in 1960 by Erdos and Renyi in hopes of producing networks that describe a 'real situation' (1960, pp. 17-18). Unfortunately, after years of use this generator was suspended as Barabasi and Albert preformed a study that resulted with the realization that

most real world networks were not random, they followed, what they called, a scale-free power law distribution (1999, pp. 1-2). PNDCG was developed to give more accurate generations and allow for more parameter control.

## 2.7. Node Descriptive Measures

This section gives a description of each measure tested within this experiment. Each graph generated ran through all 29 measures and their nodal ranks and computational times were recorded, as is discussed in later chapters.

### 2.4.1 Degree Centrality, $C_D^{'}(p_k)$

Many of the 29 measures used in SNA are aimed at identifying the important or most visible person within the network. Degree centrality falls into this category. The idea of centrality has been discussed at length as early as 1950. The measure of degree centrality is relatively intuitive, as noted by the well known SNA analyst Linton Freeman (1979, pp. 215-239). Degree centrality takes the notion that an important person is one who is well connected. Equation **(**7**)** gives the mathematical formulation to degree centrality as used in this study. Each node receives a degree centrality score that corresponds to the number of edges connected to the node. For this study degree centrality score is normalized by dividing $n - 1$, since at most each node is connected to that many nodes.

$$C_D^{'}(p_k) = \frac{\deg(p_k)}{n-1} \qquad (7)$$

### 2.4.2 Betweenness Centrality, $C_B^{'}(p_k)$

Freeman also discusses the concept of betweenness. He states that an important person can also be a person in the network who controls the flow of information, communication or

other flows throughout the network edges. This is the idea of betweenness centrality. Freeman goes on to derive Equation **(**8**)** in order to capture this concept (Freeman, 1979, pp. 222-223).

His equation makes use of the notion of geodesic paths. A geodesic path is the shortest path from one node to another (Newman, 2010, p. 139). The likelihood, $b_{ij,}$ that a point $p_k$ falls on a randomly selected geodesic linking $p_i$ with $p_j$ is given by the number of geodesic paths from $p_i$ to $p_j$ containing $p_k$ divided by the total number of geodesic paths from $p_i$ to $p_j$. The betweenness centrality score can be calculated for a node by summing these probabilities for every pair of nodes. Again, this study normalizes the betweenness centrality score by dividing it by the maximum possible score as derived by Freeman and seen in Equation **(**8**)** (Freeman, 1979, pp. 222-223).

$g_{ij}(p_k)$ = the number of geodesics linking $p_i$ and $p_j$ that contain $p_k$ where $k \neq i \neq j$

$$b_{ij} = \frac{g_{ij}(p_k)}{g_{ij}}$$

$$C_B(p_k) = \sum_{i<j}^{n}\sum_{j}^{n} b_{ij}(p_k)$$

$$C_B^{'}(p_k) = \frac{2C_B(p_k)}{n^2 - 3n + 2} \qquad (8)$$

### 2.4.3 Closeness Centrality, $C_C^{'}(p_k)$

Another measure that Freeman discusses is closeness centrality. This measure's intent is to determine the interdependence of a node by calculating how far a node is from every other node. Equation **(**9**)** was derived in 1965 by the behavioral scientist Beauchamp (Freeman, 1979, p. 226). It determines the distances from one node to another by summing the number of edges (assuming a length of 1), and normalizes them by dividing by the maximum number of edges that are traversed from a node to directly reach another node. This measure when first derived

was a measure of decentrality because the larger the score meant that a node was very far from the other nodes. This was corrected by Beauchamp by using the reciprocal.

$$C_C^{'}(p_k) = \frac{n-1}{\sum_{i=1}^{n} g_{ik}} \quad (9)$$

### 2.4.4 Eigenvector Centrality, $C_E(p_i)$

An extension to degree centrality is eigenvector centrality. Degree centrality looks at how many neighboring nodes are connected to another, but not all nodes are equivalent. Eigenvector centrality takes this into account by looking at that neighboring node's degree centrality (Newman, 2010, pp. 169-172). The concept of eigenvector centrality is that of being connected to a few well connected people is better than being connected to number of unconnected people. First proposed by Bonacich eigenvector centrality uses a network's adjacency matrix (Bonacich, 1987, pp. 1172-1173). As the name reveals, this measure uses the eigenvectors and eigenvalues to calculate the degree of a node weighted by its neighbors' degrees. Equation **(**10**)** shows the mathematical formulation used in this study.

$$C_E(p_i) = \kappa_1^{-1} \sum_j A_{ij} C_E(p_j) \quad (10)$$

where $\kappa_1$ = the largest value of matrix **A**'s eigenvalues and $A_{ij}$ = the adjacency matrix of the network.

### 2.4.5 Stress Centrality, $C_S(p_k)$

Stress centrality is a close relative of betweenness centrality. It was created to measure how much 'work' is done by each node within a network. Similar to betweenness centrality, the creator of this measure defines work as the flow controlled by the node. This flow is measured by the number of geodesic paths that contain the node. Work, or stress on a node, is calculated, as shown in Equation (11) (Koschützki, Lehmann, Peeters, & Richter, 2005, pp. 28-29).

$$C_S(p_k) = \sum_{i \neq j} \sum_{j} g_{ij}(p_k) \quad (11)$$

### 2.4.6 Load Centrality, $C_L(p_k)$

Like stress centrality, load centrally is closely related to betweenness centrality. The load centrality of a node measures a fraction of the geodesic paths that contain the node (Hagberg, Schult, & Swart, 2008). Newman noted that in many social networks flow does not use the geodesic path 100 percent of the time (Newman, 2001, p. 3). Equation **(**12**)** uses this concept to iteratively calculate the load that each node bares.

$$\delta(p_i) = 1 \quad \forall p_i \in V \qquad \text{initialization}$$

$$\delta(p_j) = \delta(p_j) + \frac{\delta(p_k)}{|\Pr ed(p_k)|} \qquad \forall p_j \in \Pr ed(p_k)$$

$$C_L(p_k) = C_L(p_k) + \delta(p_k) \quad (12)$$

where $\delta(p_i)$ is called the dependence of the source on node $p_i$ and Pred($p_k$) is the set of nodes that precede $p_k$ from the source.

### 2.4.7 Communicability Centrality, $C_{COM_C}(p_k, p_j)$

Communicability Centrality is another measure concerned with flow between two nodes. This measure uses the adjacency matrix to calculate eigenvalues and eigenvectors and in turn sum the number of closed walks of different lengths; where a walk is a sequence of adjacent nodes (Estrada & Hatano, 2008, pp. 2-3). Estrada & Hatano derived communicability centrality as a measure of how well nodes are able to communicate with others (Estrada & Hatano, 2008, pp. 2-3). The more ability a node has to pass flow to another node, the higher the score it recieves. It also uses the fact that flow does not travel on the geodesic path all of the time. Equation (13) shows the mathematical formulation of communicability centrality.

$$C_{COM_C}(p_k, p_j) = \sum_{j=1}^{n} \phi_j(p_k)\phi_j(p_j)e^{\lambda_j} \quad (13)$$

where $\phi_j(p_k)$ is the *u*th element of the *j*th orthonormal eigenvector of the adjacency matrix associated with the eigenvalue $\lambda_j$.

### 2.4.8 Simple Diversity, $D(p_k)$

Diversity is a measure that considers nodes in the neighborhoods within the network. The neighborhood of a node is the set of adjacent nodes. For this measure, a node is said to be more diverse if it has few neighbors in common with its neighbors. The concept of this measure is that an important node is diverse in the neighbors that it has. For example, someone who is friends with many tire repairmen is less diverse, and thus less valuable, than someone who knows one tire repairmen, a radiator mechanic, a stereo installer and an engine mechanic. The formulation derived by Lui *et al.* is seen in Equation (14) (2010, pp. 4-6).

$$D(p_k) = \sum_{p_i \in N(p_k)} \left(1 - \frac{|N(p_k) \bigcap N(p_i)|}{|N(p_i)|}\right) \quad (14)$$

where *N(p)* denotes the set of *p*'s neighbors.

### 2.4.9 General Diversity, $D_G(p_k)$

Lui *et al.* also derived an equation for general diversity, which takes into account the size of the neighborhood and the weight of its neighbors (2010, pp. 4-6). In Equation **(**14**)** a node could receive a high score simply by not having any neighbors. This node should not be considered diverse in the network. In addition, general diversity, like eigenvector centrality, takes into account if a neighbor is itself very diverse. The mathematical formulation of general diversity is seen in Equation (15).

$$D_G(p_k) = \sum_{p \in N(p_k)} w_k(p_i) * F(p_i, p_k) \quad (15)$$

$$F(p_i, p_k) = 1 - \alpha * S(p_i, N_{-p_i}(p_k))$$

$$S(p_i, p_x) = \begin{cases} \mu^{(d(p_i, p_x)-1)}, \ 0 < \mu < 1 \ if \ d(p_i, p_x) \leq r \\ 0 \ otherwise \end{cases}$$

$$S(p_i, N_{-p_i}(p_k)) = \frac{\sum_{p_x \in N_{-p_i}(p_k) \cap N_{-p_k}(p_i)} (w_k(p_x) * S(p_i, p_x))}{\sum_{p_x \in N_{-p_k}(p_i)} S(p_i, p_x)}$$

where "$F(p_i, p_k)$ is a function evaluating the dissimilarity between $p_i$ and other neighbors of $p_k$ in the set radius $r$, i.e., the set $N_{-p_i}(p_k)$" (Liu, *et al.*, 2010, p. 5). $N_{-p_i}(p_k)$ denotes the set of *v's* neighbors which excludes the nodes that become *v's* neighbors through *u*. $w_k(p_i)$ is the weight of $p_i$ in $p_k$'s neighborhood, "we define $S(p_i, N_{-p_i}(p_k))$ as the average similarity between $p_i$ and each node $p_x$ of $N_{-p_i}(p_k)$" (Liu, *et al.*, 2010, p. 5) where α is its weight and $\mu$ "is the damping factor to reflect the notion that nodes farther apart share less similarity" (Liu, *et al.*, 2010, p. 6).

### 2.4.10 Flow Betweenness, $C_F(p_k)$

Flow betweenness, as the name implies, is a measure that is also concerned with the control of flow. Proposed by Freeman in 1991, flow betweenness looks at betweenness in respect to the maximum amount of flow through a node (Freeman, Borgatti, & White, 1991, pp. 147-148). This measure, like others, does not just examine the shortest paths through a node but notes that flow can take alternative paths to get from one node to another. As Equation (16) shows, flow betweenness sums that flow through a node. For this study every edge has a maximum flow of one therefore once a path is used it cannot be used again for that calculation.

$$C_F(p_k) = \sum_{i<j} \sum_{j} M_{ij}(p_k) \quad (16)$$

where $M_{ij}(p_k)$ is the maximum amount of flow from node *i* to node *j* that passes

through node *k*.

### 2.4.11 Endpoint Betweenness, $C_{EB}(p_k)$

Endpoint betweenness makes a slight adjustment to betweenness centrality by allowing the source or target node to also be the node being measured. Brandes proposes this measure, remarking that "it may be inappropriate to have pairs of vertices depend on intermediaries, but not on themselves" as is the case with betweenness centrality (2008, pp. 6-7). In his proposal of this measure each node goes up in importance with the number of neighbors that it has. Equation **(**17**)** shows this in mathematical form where $b_{ij}$ is defined in Equation **(**8**)**.

$$C_{EB}(p_k) = \sum_{i<j}^{n} \sum_{j}^{n} b_{ij}(p_k) \qquad (17)$$

where *k* may equal *i* or *j*.

### 2.4.12 Length-Scaled Betweenness, $C_{B_{dist}}(p_k)$

Another alternative to betweenness centrality is length-scaled betweenness proposed by Borgatti and Everett (Brandes, 2008, p. 9; Borgatti, 2003, pp. 245-247). This measure allows for flow to traverse paths of all lengths but weights that path by the inverse of its distance. Therefore, as the path lengths get longer they have less weight. The concept is that the control of a longer path is less valuable than a shorter path. This is calculated by Equation **(**18**)**.

$$C_{B_{dist}}(p_k) = \sum_{i \neq j \in V} \frac{1}{d(p_i, p_j)} * \frac{g_{ij}(p_k)}{g_{ij}} \qquad (18)$$

### 2.4.13 Linearly-Scaled Betweenness, $C_{B_{Lin}}(p_k)$

Linearly-scaled betweenness is yet another variant to betweenness centrality. This measure not only takes in to account the distance from the source to the target but also the distance from the source to the node being measured. The idea here is that the farther away from the source node (thus closer to the target) the more control a node has over the flow to the target.

This is similar to proximal betweenness but broader in that the node being measured varies in its distance. Equation 19 shows the mathematical equation used in this study (Brandes, 2008, p. 10).

$$C_{B_{Lin}}(p_k) = \sum_{i \neq j \in V} \frac{d(p_i, p_k)}{d(p_i, p_j)} * \frac{g_{ij}(p_k)}{g_{ij}} \quad (19)$$

### 2.4.14 Communicability Betweenness, $C_{COM_B}(p_k)$

Introduced by Estrada and Hatano, communicability betweenness is concerned with the ability of a node to communicate with other nodes within the network (Estrada & Hatano, 2008, pp. 2-3). This measure works by computing the betweenness of a node with respect to removing edges attached to the node. The concept here, again, is the idea of control of flow as importance. If a node is contained in every path from one node to another then it has control of flow between the two nodes. Conversely, if it is not in any of the paths it has little or no importance in controlling flow between the two nodes. Equation (20) the walks, or paths, from node *i* to node *j* is subtracted by the walks that contain node *k*, giving node *k* a communicability betweenness score.

$$C_{COM_B}(p_k) = \frac{\sum_i \sum_j \frac{G_{ij}(p_k)}{G_{ij}}}{(n-1)^2 - (n-1)} \quad i \neq k \neq j \quad (20)$$

where $G_{ij}(p_k) = (e_{ij}^A - (e^{A+E(k)}))_{ij}$ is the number of walks from node *i* to node *j* that include node *k*, and $G_{ij} = (e^A)_{ij}$ is the number of walks from node *i* to node *j*.

### 2.4.15 *k*-Betweenness, $C_{B_{(k)}}(p_v)$

As noted by Borgotti and Everett, sometimes the longer paths measured in betweenness centrality are not always realistic in a true network. This is taken in to account in their measure *k*-betweenness (Brandes, 2008, p. 9; Borgatti & Everett, 2006, pp. 475-476). This measure

restricts the path length to only length *k*. Equation (21) shows this small variant to betweenness centrality that better describes some networks.

$$C_{B_{(k)}}(p_v) = \sum_{i,j \in V: dist(i,j) \le k} \frac{g_{ij}(p_v)}{g_{ij}} \quad (21)$$

### 2.4.16 Newman's Betweenness, $C_{B_{NEW}}(p_k)$

This measure, derived by Newman, uses the idea of random walks through a network. A random walk is defined as the flow from node *i* to node *j* that traverses the edges with the probability that it will travel a given edge given by the uniform distribution; that is that once the information flow reaches a node it chooses an edge to take at random until it reaches the target (Newman, 2005, pp. 5-9). As Equation **(**22**)** shows, the betweenness centrality measure is used after factoring in the probabilities for each edge at a node. Newman gives the following steps in order to account for the flow traveling back and forth through the measured node.

1. Construct the matrix $\boldsymbol{D} - \boldsymbol{A}$, where $\boldsymbol{D}$is the diagonal matrix of vertex degrees and $\boldsymbol{A}$is the adjacency matrix.
2. Remove any single row, and the corresponding column. For example, one could remove the last row and column.
3. Invert the resulting matrix and then add back in a new row and column consisting of all zeros in the position from which the row and column were previously removed (e.g., the last row and column). Call the resulting matrix $\boldsymbol{T}$, with elements $T_{ij}$.
4. Calculate the betweenness from Equation 22 (Newman, 2005, p. 9).

$$C_{B_{NEW}}(p_k) = \sum_{i<j}^{n} \sum_{j}^{n} \frac{g_{ij}(p_k)}{g_{ij}} \; i \neq j \neq k \quad (22)$$

### 2.4.17 Source Proximal Betweenness, $C_{PS}(p_k)$

As previously stated, proximal betweenness is similar in definition and purpose to length-scaled betweenness. This measure uses the concept of proxies by giving weight to the node that

is one step away from the target, thus giving it more influence of the flow (Brandes, 2008, pp. 7-8). A proxy of the target essentially controls the flow into the target. This is called source proximal betweenness and is shown in Equation (23).

$$C_{PS}(p_k) = \sum_{\substack{s \in V \\ t:(k,t) \in E}} \frac{g_{ij}(p_k)}{g_{ij}} \tag{23}$$

### 2.4.18 Target Proximal Betweenness, $C_{PT}(p_k)$

Like source proximal betweenness, target proximal betweenness looks at the proxies, but this time they are proxies to the source node. These nodes control the flow out of the source and to the target. Equation (24) shows this in mathematical form (Brandes, 2008, pp. 7-8).

$$C_{PT}(p_k) = \sum_{\substack{t \in V \\ s:(s,k) \in E}} \frac{g_{ij}(p_k)}{g_{ij}} \tag{24}$$

### 2.4.19 Clustering Coefficient, $C_{Clust}(p_k)$

The next three measures examine the clustering ability of a node. Clustering examines how tightly connected a node is from its neighbors, and therefore more important within the clique of nodes. As stated in section 2.4.4 this clustering coefficient looks at triangles and their proportion to the maximum number of triangles. A triangle connecting three nodes together is said to be a tight connection between the three nodes (Soffer & Vazquez, 2005, pp. 1-2). This measure is shown in Equation (25).

$$C_{Clust}(p_k) = \frac{2t_i}{\deg(p_k)[\deg(p_k)-1]} \tag{25}$$

### 2.4.20 Soffer's Clustering Coefficient, $C_{SC}(p_k)$

Soffer and Vazquez discuss the fact that the definition of clustering is erroneous in those nodes with higher degree scores (Soffer & Vazquez, 2005, pp. 2-3). Therefore, they propose a new measure of calculating clustering that is not biased in this manner. Equation **(**26**)** shows their formulation of this coefficient that calculates "the maximum possible number of edges between the neighbors of a vertex, given their degrees" (Soffer & Vazquez, 2005, pp. 2-3).

$$C_{SC}(p_k) = \frac{t_k}{\omega_k} \quad (26)$$

where $\omega_k$ is the maximum number of edges that can be drawn among the neighbors of a vertex *k*, given the degrees of its neighbors.

### 2.4.21 Squares Clustering Coefficient, $C_4(v)$

This measure, introduced by Lind *et al.*, is the same concept as the normal clustering coefficient but instead of using triangle, squares are used. "While [the clustering coefficient] gives the probability that two neighbors of node $p_i$ are connected with each other, [square clustering] is the probability that two neighbors of node $p_i$ share a common neighbor (different from $p_i$)" (Lind, González, & Herrmann, 2005, p. 2). They claim that the measure is equivalent to the previously discussed clustering coefficient over the total network average (Lind, González, & Herrmann, 2005, pp. 1-2). Equation **(**27**)** shows their formulation that measures the proportion of squares to the possible number of squares of a node.

$$C_4(v) = \frac{\sum_u \sum_{w=u+1} q_v(u,w)}{\sum_u \sum_{w=u+1} [a_v(u,w) + q_v(u,w)]} \quad (27)$$

where $q_v(u, w)$ are the number of common neighbors of $u$ and $w$ other than $v$,

$$a_v(u,w) = (k_u - (1 + q_v(u,w) + \theta_{uv}))(k_w - (1 + q_v(u,w) + \theta_{uw})),$$

and $\theta_{uw} = 1$ if $u$ and $w$ are connected and 0 otherwise.

### 2.4.22 Current Flow Betweenness, $C_{CF}(p_k)$

Social networks are not limited to people and their acquaintances, in fact this measure is based on the flow of electronic information. Communication via electronic current is plentiful this day and age. Brandes and Fleischer propose this electronic based version of betweenness centrality (Brandes & Fleischer, 2005, pp. 536-540). Current flow betweenness is similar to the random walk of Newman's betweenness in that current flows randomly from one node to the next. This measure is shown in Equation (28).

$$C_{CF}(p_k) = \frac{\sum_{i,j \in V} \tau_{ij}(p_k)}{(n-1)(n-2)} \qquad (28)$$

$$\tau(p_k) = \frac{1}{2}\left(-|b(p_k)| + \sum_{e:k \in e} |x(\vec{e})|\right)$$

> where "$-|b(v)|$ accounts for the fact that only inner vertices are considered in the definition of shortest-path betweenness centrality. To include start and end vertex, it should be replaced by $+|b(v)|$. Accordingly, the throughput of an edge $e \in E$ is defined as $\tau(e) = |x(-e)|$ "
> (Brandes & Fleischer, 2005, pp. 536-540).

### 2.4.23 Approximate Current Flow Betweenness, $C'_{APROX}(p_i)$

Brandes and Fleischer note that with larger networks the current flow betweenness becomes difficult to calculate and more time consuming. Therefore they introduce the approximation of the measure. They state that "the basic idea is that the betweenness of a vertex, i.e. the throughput over all *st*-currents, can be approximated using a small fraction of all pairs $s \neq t \in V$" (Brandes & Fleischer, 2005, p. 542). This approximation is given in Equation **(**29**)**.

$$C'_{APROX}(p_i) = 0 \quad \forall p_i \in V \quad \text{initialization}$$

$$k = \left\lceil \left( \frac{n(n-1)}{\frac{(n-1)(n-2)}{\varepsilon}} \right)^2 \log n \right\rceil$$

select $p_i \neq p_j \in V$ uniformly at random

for every $p_k \in V \setminus \{p_i, p_j\}$ and for every edge out of $p_k$

$$C'_{APROX}(p_k) = C'_{APROX}(p_k) + |\tilde{p}_k - \tilde{N}(p_k)| * \frac{n(n-1)}{2k(n-1)(n-2)} \quad (29)$$

where $\tilde{p}_k - \tilde{N}(p_k)$ corresponds to the effective resistance, or distance of an edge.

### 2.4.24 Closeness Vitality, $C_{CV}(p_k)$

Closeness vitality is another measure concerned with the communication of one node to another. Koschützki *et al.* state that "the closeness vitality denotes how much the transport costs in an all-to-all communication will increase if the corresponding element *x* is removed from the graph" (Koschützki, Lehmann, Peeters, & Richter, 2005, pp. 36-38). This measure uses the Wiener index; that is, the sum of all the distances from all node pairs. Equation **(**30**)** shows the formulation of this measure, taking the network's Wiener index subtracted by the Weiner index once removing the node being measured.

$$C_{CV}(p_k) = I_w(G) - I_w(G \setminus \{p_k\}) \quad (30)$$

where $I_{p_j}(G) = \sum_{p_i \in V} \sum_{p_j \in V} d(p_i, p_j)$.

### 2.4.25 PageRank, $C_{PR}(p_k)$

This measure was created to, as the name implies, rank web pages by the number of incoming links (Langville & Meyer, 2004, pp. 2-4). The concept of this measure is that the more links to a page the more important that page will be. Each incoming link is viewed as a

recommendation for that page. Equation (31) shows how each node is ranked as the proportion of incoming edges to the number of outgoing links from its neighbors.

$$C_{PR}(p_k) = \sum_{p_i \in B_k} \frac{C_{PR}(p_i)}{\deg[p_i]} \quad (31)$$

### 2.4.26 Hits (Hubs and Authorities), $x^{(k)}$, $y^{(k)}$

This measure was developed for similar reasons as PageRank. Hits of a web page are actually an iterative pair of measures. Hits refer to both the authority ($x$) and hub ($y$) scores of a web page. Authorities estimate a node's score based on the incoming links. Hubs estimate the node's score based on outgoing links. It is said that a good (important) authority points to good hubs and vice versa (Langville & Meyer, 2004, pp. 4-8). Equations **(**32**)** and (33) show the iterative formulas for authorities and hubs.

1. Initialize: $y^{(0)} = e$, where e is a column vector of all ones. Other positive starting vectors may be used.
2. Until convergence, do
   $x^{(k)} = A^T y^{(k-1)}$ (32)
   $y^{(k)} = A x^{(k)}$ (33)
   $k = k + 1$

### 2.4.27 Average Neighbor Degree, $C_{AND}(p_k)$

Average neighbor degree is another measure that uses the idea that important people are those who know other well connected people. Just as the name implies, this measure takes the average of degrees for all the neighbors of a node. Equation **(**34**)** is the mathematical formula for this measure used in this study (Hagberg, Schult, & Swart, 2008).

$$C_{AND}(p_k) = \frac{\sum_{j \in N(k)} \deg[p_j]}{|N(k)|} \quad (34)$$

### 2.4.28 Core Number, *core[v]*

The core number of a node is a score that demonstrates a node's reach. The importance of a node is measured by how many other nodes it can communicate with. Batagelj & Zaversnik give the following definition when deriving the formula for core number. "If from a given graph $G = (V, L)$ we recursively delete all vertices, and lines incident with them, of degree less than k, the remaining graph is the k-core" (Batagelj & Zaversnik, 2003, pp. 1-3). From this definition they define the core number of a node as the maximum k that a node can obtain. This is also shown in the algorithm used in this study, as shown below.

```
Compute the degrees of vertices;
Order the set of vertices V in increasing order of their degrees;
        for each v ∈ V in the order do begin
                core[v] := degree[v];
        for each u ∈ Neighbors(v) do
                 if degree[u] > degree[v] then begin
                        degree[u] := degree[u] − 1;
                        reorder V accordingly
            end
    end;
```

(35)

(Batagelj & Zaversnik, 2003, p. 3)

## 2.6 Rank Correlations

Spearman's Rank Correlation Coefficient, or Spearman's Rho (ρ), is named after Charles Spearman, an English psychologist known for his work in statistics (*Spearman*, 2012). Spearman's Rho is a non-parametric measure of statistical dependence between two variables. Spearman's Rank Correlation Coefficient is shown in Equation (36) (Conover, 1980, p. 252).

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\left(\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)\left(\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)}} \quad (36)$$

Where $X$ = random variables of sample size $n$ and $R(X_i)$ = are the ranks of $X_i$ compared to the other values of $X$.

Spearman's Rank Correlation Coefficient was chosen for testing this data set because of its ability to take ranks, which we were given from each measure, and output a value that can test the hypothesis that $X$ and $Y$ are mutually independent. This measure of correlation can be used on data that is ordinal without regarding the scale of the measurement or the type of distribution (Conover, 1980, p. 251).

In addition to measuring the ranks of the measures for correlation another correlation test was calculated on the raw output of the measures. Kendall's Tau ($\tau$) similarly measures the correlation between two variables and also does not depend on distribution of the two variables, $X$ and $Y$. As seen in Equation **(**37**)** instead of measuring the differences in the ranks, Kendall's Tau measures the probabilities of observing concordant and discordant pairs (Conover, 1980, p. 256). "Two observations, for example, (1.3, 2.2) and (1.6, 2.7) are called concordant if both members of one observation are larger than their respective members of the other observation" (Conover, 1980, pg 256). "A pair of observations, such as (1.3, 2.2) and (1.6, 1.1), are called discordant if the two numbers in one observation differ in opposite directions from their respective members of the other observation" (Conover, 1980, p. 256).

$$\tau = \frac{N_c - N_d}{n(n-1)/2} \qquad \textbf{(37)}$$

where $n =$ *the number of observations*
$N_c =$ *the number of concordant pairs*
$N_d =$ *the number of discordant pairs*

As observed by Conover, a professor of statistics, Kendall's Tau and Spearman's Rho produce equivalent results in testing the hypothesis that $X$ and $Y$ are mutually independent so did the results of this thesis (Conover, 1980, p. 258).

## 2.7 Similarity Tests

Given Equation **(**36**)** for Spearman's rank correlation, the occurrence of one measure's ranks being exactly identical to another would happen if their differences equaled zero. Therefore to confirm the Spearman's rank correlations density plots and paired measure scatter plots could be utilized. The overlay of two density plots shows the analyst if the raw data follows the same ranking pattern. In order to correct the difference in scale of the raw scores one could standardize all of the raw scores by Equation **(**38**)**. This standardization shifts the mean of the raw scores to zero and scales them to the standard deviation. Plotting these standardized raw scores against one another allows for a visual of the correlation between the measures.

$$x_{i_{stz}} = \frac{x_i - \overline{x}}{\sigma} \qquad \textbf{(38)}$$

where $x_i$ is the $i$th observation, $\sigma$ is the standard deviation of $X$

and $\overline{x}$ is the mean of $X$

The last test for similarity was to test the nodes in the top and bottom 10 ranks. An analyst may only be interested in certain subsets of rankings. This test examines the subsets of the top 10 rankings as well as the bottom 10 in order to see if the measures identify the same nodes. The process of this measure of similarity is concerned with counting the number of unique nodes in the subsets for each pair of measures. For example, if Measure 1 identified, in rank order {5, 9, 2, 4, 8} and Measure 2 identified {5, 2, 9, 3, 8}, there would be six unique numbers. As seen in the example, order is not taken into account within this test, but it is not an

unreasonable assumption that an analyst looking into a subset will examine all nodes identified. Therefore, this test was coded up in Visual Basic and a scale from 0 to 1 was formed, as seen in Equation **(**39**)**, to identify those measures that are substitutable.

$$\beta = \frac{20 - U}{10} \qquad (39)$$

where $U$ is the number of unique nodes in the top or bottom 10 ranks.

## 2.8 Summary

This chapter presented the foundation of this thesis by providing an overview of Social Network Analysis, network structural properties and Design of Experiments. In addition, the use of a network generator and statistical tests preformed were discussed. The next chapter explains the methods used to conduct this study.

# 3. Methodology

## 3.1. Introduction

This chapter outlines the design of the experiment, as well as the methods and procedures used in this thesis. This methodology was created following the pertinent literature as discussed in Chapter 2. Figure 7 displays the methods use throughout this study.

## 3.2. Design of Experiment

With the help of SMEs and review of SNA literature, several network structural properties were laid out to be considered as the factors and their levels of this experiment. This thesis was also limited to the input possibilities of the network generator used. After some discussion, size, scale-free parameter (∝) and encouraged clustering were chosen for their role in defining and differentiating networks. Once these factors were chosen the next step was to choose the levels of these factors to obtain a broad range of networks to test the measures. For the size factor the levels that were chosen were discussed with SMEs. One restriction that limited the size levels was the potential computational times from previous uses of the generator as the number of nodes grew. Therefore, 50 node, 100 node, 500 node and 1000 node graphs were utilize as the four levels of size.

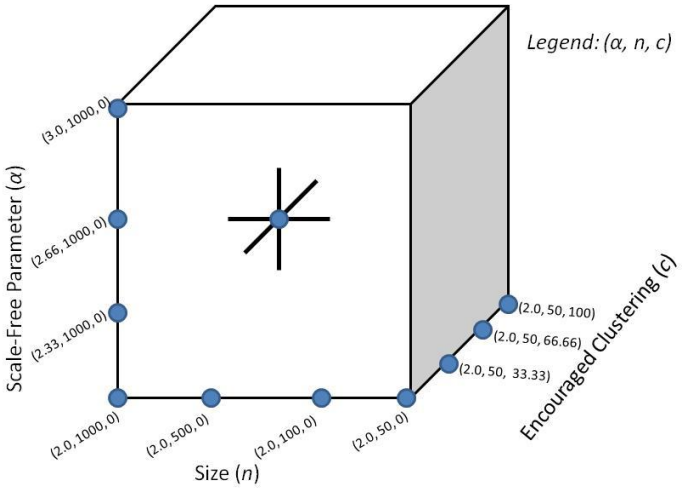


**Figure 5: Experimental Design Space**

As discussed in section 2.6 the scale-free parameter aids in determining the edge distribution of the network. In reading the literature of scale-free networks by Barabasi and

Bonabeau (2003) and Morris, O'Neal, & Deckro (2011) it becomes clear with testing that the parameter α typically falls between two and three. For this reason 2.0, 2.33, 2.66 and 3.0 are chosen as the levels to diversify the types of networks.

Lastly, the encouraged clustering found in the PNDCG algorithm ranges from 0%, or no encouraged clustering, to 100% encouraged clustering. In Figure 6 Morris, O'Neal, & Deckro demonstrate the PNDCG's range of the clustering coefficient, using encouraged clustering, against the other well known graph generators (2011, p. 21). This lead this study to choose 0%, 33.33%, 66.66% and 100% as the appropriate levels for encouraged clustering to span the full range of network clustering. This factor is called encouraged clustering because the generator does not guarantee the exact average clustering but promotes the probablity of the level set. Figure 6 shows the actual clustering coefficient with respect to the precentage input encouraged clustering.

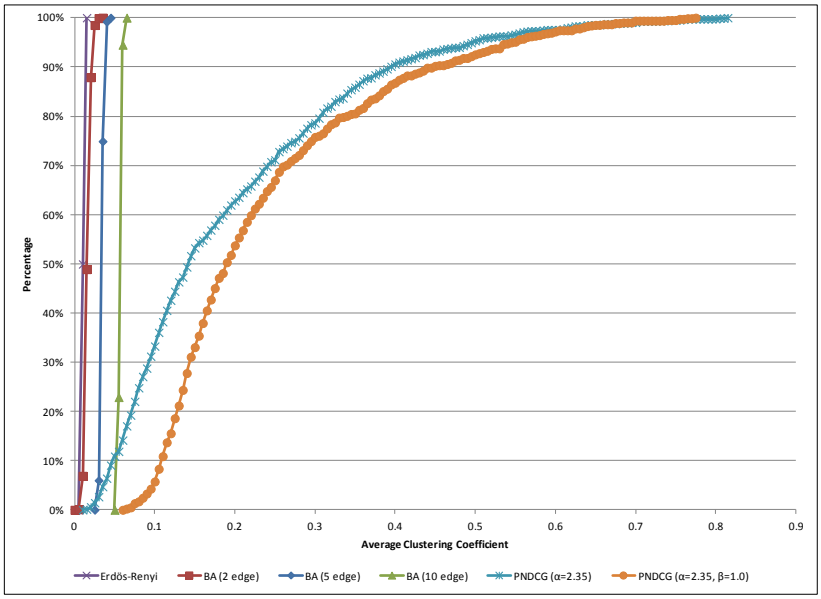


**Figure 6: Clustering Coefficient CDFs by Algorithm**

**(Morris, O'Neal, & Deckro, 2011)**

Once the factors and levels were chosen, the next step was to choose a design for the experiment. The PNDCG algorithm was coded up in C++ using Microsoft® Visual Studio software; therefore the addition of a small amount of Visual Basic code was then used to repeatedly generate the networks needed (Morris, O'Neal, & Deckro, 2011, p. 16). Since the only direct cost being expended for the generation of these networks was time, a full $4^3$ factorial

with five replications was decided upon. With this experimental design the total networks generated came to 320. Design-Expert® was utilized in this study to layout the number of networks to generate and at which factor levels. Figure 5 is an example of the experimental space tested over all the levels of factors. Each corresponding point within the space matches up to a network with the given parameters.

### 3.4 SNA Measures and Data

Once all the networks were generated they were then run through Python© code that implemented NetworkX's built in measures as well as written code (Hagberg, Schult, & Swart, 2008; van Rossum & Drake, 2001). The measures were computed in the cloud with an Amazon 32 Core Processor Remote Desktop Instance®. Each network ran through all 29 measures and recorded the data in an output file. Four output files were generated that include overall network structural properties, the raw scores for each measure, the node ranks for each measure and the computational times for each measure. This data was then read into the statistical computing software R© to compute the following correlation and similarity tests (2008).

### 3.5 Rank Correlation and Similarity Tests

As discussed in section 2.6, four statistical tests were executed on the data to narrow the measures down to those that showed significant correlations with one another. Simultaneously both the Spearman's Rank Correlation Coefficient and the Kendall's tau test for correlation were measured using R© over all pairs of measures and all networks. This program used the definitions stated above to calculate these correlation values and gave great insight into which measures have potential to be interchangeable.

These correlations exhibited a clear grouping over the full set of data as well as a smaller subset. Four well defined clusters of highly correlated measures provided this study with a manageable subset of the 29 measures. This can be seen in Table 2 and Table 3. The next step was to take these measures and test for similarity and confirmation. As noted, density plots, overlays and paired measure scatter plots allowed for a visual aspect of the correlation test. The last tests performed on the data were similarity tests that looked at the top 10 and bottom 10 ranked nodes over the subsets of measures. This test counted the number of unique nodes for each pair of measures. The best possible outcome being that there were 10 unique nodes out of the 20 and therefore every node was in each measure's top or bottom 10. The worst outcome would be that of the 20 there are 20 unique nodes and therefore no matching pairs in each measure's top or bottom 10. This measure was then scaled between zero and one by Equation (39).

### 3.6 Computational Time Testing

Once each measure was analyzed from all four of these tests there was a suitable understanding of each measure's similarity. Finally, the computational times for these measures were averaged over each size 50, 100, 500 and 1000 node graph and then statistically compared to one another, via the Z two sample for comparing means. The Z-test is a statistical test with a null hypothesis that the means are equal to one another, whereas the alternate hypothesis is that they are significantly different, either faster or slower. These levels showed a statistically significant difference over a few of the measures.

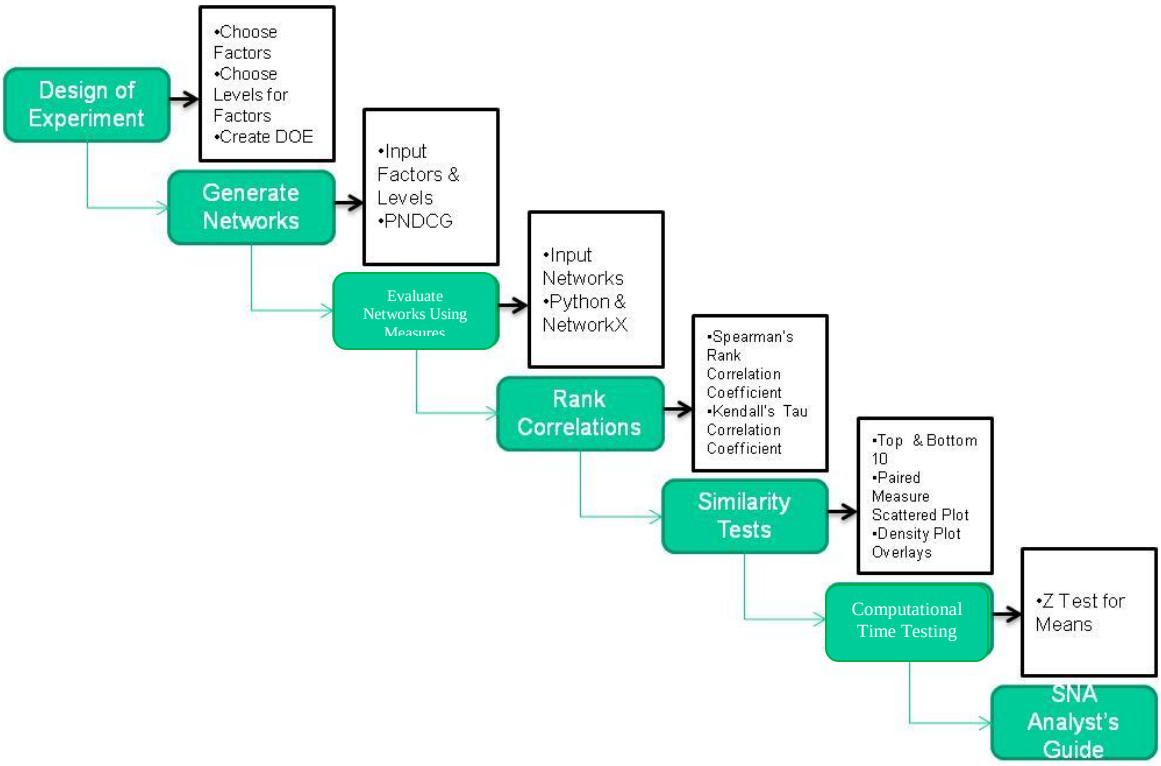


**Figure 7: Methodology Flow Chart**

## 3.7 Summary

In this chapter the techniques and tests used were outlined in order to allow for replication of this study. The complete design of experiment, correlation and similarity tests were summarized, along with the comparison of computational times. As shown in Figure 7, the culmination of these statistical techniques was used to develop a SNA Analyst's Guide to assist analysts in choosing an appropriate measure. The next chapter will explore the results of these methods and the analysis of those results.

# 4. Results and Analysis

## 4.1. Introduction

This chapter describes the results gathered throughout this study and provide guidance for SNA analysts on selecting the most effective and efficient measure. A description of the data, correlation results and similarity tests will lead to the SNA Analyst Guide created from this study.

## 4.2. Data Description

Once the networks were generated using the PNDCG, they were run through the 29 measures chosen for analysis. It rapidly became apparent that some measures preformed slower than others. The first measure to be noticed to cause problems as the size of the graph grew, was flow betweenness. Looking back at the definition of flow betweenness we see that for every node, all paths to every other pair of nodes is computed and recomputed for every node scored. This was not as severe a problem with the 50 node graphs but as the size grew so did the computational time at what appeared to be an exponential rate. Figure 8 shows flow betweenness as the size increases.

Newman notes the flow betweenness can always be calculated in time $O(n^2m)$ (2005). In addition to this measurement the minimum number of edges needed to create a connected graph, $n-1$, gives a suitable lower bound for the number of calculations used to compute flow betweenness. As seen in Figure 8, as expected, there is an exponential increase with an increase of size. This characteristic of flow betweenness caused the total computational times to compute all scores to take approximately five day to run a single 1000 node graph. Therefore, it was decided to only compute flow betweenness for size 50 node and 100 node graph in order to give a general guide for the other graphs. Other measures that slightly increased the computational

time are Newman's betweenness, communicability, hits and current flow. These were kept in the computations for all size networks because their times did not increase the time as significantly as flow betweenness.

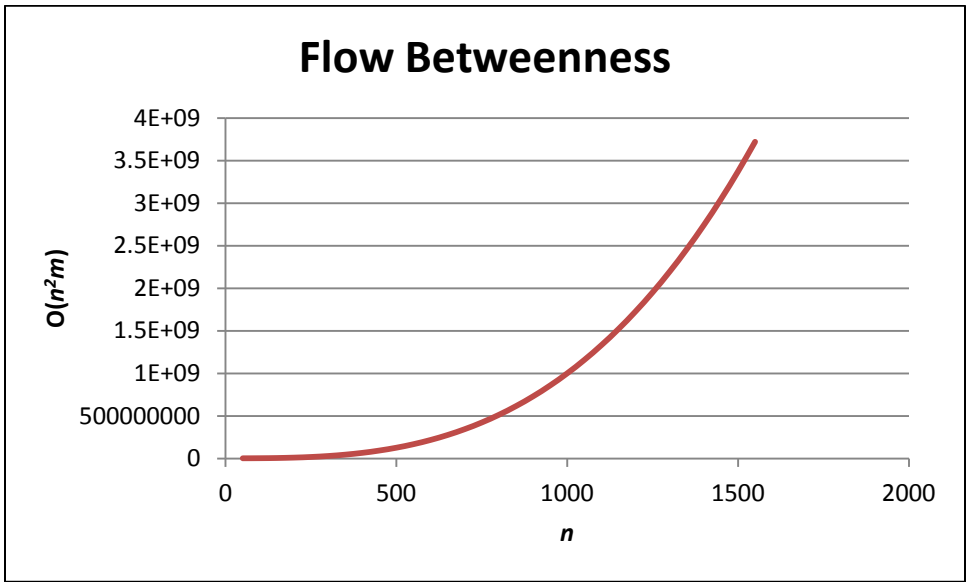


**Figure 8: Big O Notation for Flow Betweenness**

The Python© code for the nodal scoring was computed on an Amazon EC2 32 Core Processor Instance® allowing for up to 32 graphs to run through the measures at one time. Even with this major computing power the calculation of all the 320 graphs over the 29 measures took 48 hours. Next, the resulting outputs of raw data were read into R© to compute the correlations of each measure pair.

### 4.3. Correlation Analysis

After calculating the Spearman's Rank Correlation Coefficient, we see that the measures chosen for this study are in fact highly correlated amongst each other. Table 2 and Table 3 show the resulting correlation matrix of the paired measures' Spearman's Rank Correlation. The matrix shown in Table 2 takes the correlations of all the measures except for flow betweenness for reasons previously discussed. Table 3 displays the same but it is for the subset of only 50

node and 100 node graphs. These correlation groups remain constant over the different size groups. Therefore it can be deduced that these measures' correlations are not affected by the size of the graph. The four groups identified in both of these tables determined which measures were to be tested further to determine if they give statistically equivalent rankings. Using these results, 16 measures were found to possess high correlations with one another within their group.

This inter-connectedness suggests the need for further statistical testing to see if these measures are in fact calculating the same ranks for the nodes in a network with several structural properties. In addition, this offers insight of how social network analyst can use a measure that is more time efficient. Another interesting finding is that, not surprisingly, all of the clustering measures were very highly correlated with one another. The correlation values for this group are all over 0.85 on the test set suggesting that these measures could be used. The same can be said about the other three groups.

**Table 2: Cluster Groups of Spearman's Rho (Flow Betweenness Excluded)**

| | Closeness | Eigen | Clustering | Soffers | Square | Betweenness | Length | Linear | Newman | Prox1 | Stress | KBetweenness | Load | Endpoint | Prox2 | Comm2 | Aprox | Pagerank | Degree | Current | Vitality | AND | Diversity | GenDiv | Comm1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Closeness | 1.00 | 0.03 | 0.55 | 0.56 | 0.55 | 0.13 | 0.12 | 0.08 | 0.13 | 0.08 | 0.10 | 0.08 | -0.08 | -0.09 | -0.07 | 0.06 | 0.31 | 0.04 | 0.36 | 0.22 | -0.56 | 0.66 | 0.30 | 0.05 | 0.52 |
| Eigen | 0.03 | 1.00 | 0.23 | 0.23 | 0.22 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.02 | 0.00 | 0.00 | 0.18 | 0.24 | 0.35 | 0.28 | 0.20 | 0.02 | 0.15 | 0.25 | 0.22 | -0.07 |
| Clustering | 0.55 | 0.23 | 1.00 | 1.00 | 0.85 | 0.30 | 0.30 | 0.28 | 0.30 | 0.27 | 0.30 | 0.28 | 0.17 | 0.16 | 0.18 | 0.19 | 0.62 | 0.25 | 0.45 | 0.43 | -0.15 | 0.35 | 0.52 | -0.02 | 0.23 |
| Soffers | 0.56 | 0.23 | 1.00 | 1.00 | 0.85 | 0.31 | 0.31 | 0.28 | 0.31 | 0.28 | 0.30 | 0.28 | 0.17 | 0.16 | 0.18 | 0.20 | 0.63 | 0.26 | 0.45 | 0.43 | -0.15 | 0.35 | 0.52 | -0.01 | 0.23 |
| Square | 0.55 | 0.22 | 0.85 | 0.85 | 1.00 | 0.45 | 0.43 | 0.40 | 0.45 | 0.38 | 0.42 | 0.40 | 0.22 | 0.20 | 0.22 | 0.17 | 0.68 | 0.30 | 0.49 | 0.48 | -0.14 | 0.30 | 0.57 | 0.15 | 0.28 |
| Betweenness | 0.13 | 0.18 | 0.30 | 0.31 | 0.45 | 1.00 | 0.96 | 0.96 | 1.00 | 0.94 | 0.96 | 0.96 | 0.64 | 0.55 | 0.52 | 0.20 | 0.74 | 0.57 | 0.59 | 0.71 | 0.25 | -0.12 | 0.76 | 0.69 | 0.13 |
| Length | 0.12 | 0.18 | 0.30 | 0.31 | 0.43 | 0.96 | 1.00 | 1.00 | 0.96 | 0.99 | 0.99 | 0.99 | 0.71 | 0.62 | 0.60 | 0.19 | 0.71 | 0.58 | 0.57 | 0.70 | 0.34 | -0.11 | 0.78 | 0.72 | 0.14 |
| Linear | 0.08 | 0.18 | 0.28 | 0.28 | 0.40 | 0.96 | 1.00 | 1.00 | 0.96 | 0.99 | 0.99 | 1.00 | 0.71 | 0.63 | 0.61 | 0.18 | 0.70 | 0.58 | 0.56 | 0.70 | 0.36 | -0.14 | 0.77 | 0.72 | 0.12 |
| Newman | 0.13 | 0.18 | 0.30 | 0.31 | 0.45 | 1.00 | 0.96 | 0.96 | 1.00 | 0.94 | 0.96 | 0.96 | 0.64 | 0.55 | 0.52 | 0.20 | 0.74 | 0.57 | 0.59 | 0.71 | 0.25 | -0.12 | 0.76 | 0.69 | 0.12 |
| Prox1 | 0.08 | 0.18 | 0.27 | 0.28 | 0.38 | 0.94 | 0.99 | 0.99 | 0.94 | 1.00 | 0.98 | 0.99 | 0.71 | 0.62 | 0.61 | 0.19 | 0.69 | 0.58 | 0.56 | 0.69 | 0.37 | -0.14 | 0.76 | 0.72 | 0.12 |
| Stress | 0.10 | 0.18 | 0.30 | 0.30 | 0.42 | 0.96 | 0.99 | 0.99 | 0.96 | 0.98 | 1.00 | 1.00 | 0.72 | 0.63 | 0.60 | 0.19 | 0.71 | 0.57 | 0.57 | 0.71 | 0.33 | -0.12 | 0.78 | 0.71 | 0.13 |
| KBetweenness | 0.08 | 0.18 | 0.28 | 0.28 | 0.40 | 0.96 | 0.99 | 1.00 | 0.96 | 0.99 | 1.00 | 1.00 | 0.72 | 0.63 | 0.60 | 0.19 | 0.71 | 0.57 | 0.56 | 0.71 | 0.34 | -0.13 | 0.77 | 0.71 | 0.13 |
| Load | -0.08 | 0.02 | 0.17 | 0.17 | 0.22 | 0.64 | 0.71 | 0.71 | 0.64 | 0.71 | 0.72 | 0.72 | 1.00 | 0.96 | 0.94 | 0.14 | 0.44 | -0.04 | -0.06 | 0.43 | 0.71 | -0.03 | 0.38 | 0.29 | 0.10 |
| Endpoint | -0.09 | 0.00 | 0.16 | 0.16 | 0.20 | 0.55 | 0.62 | 0.63 | 0.55 | 0.62 | 0.63 | 0.63 | 0.96 | 1.00 | 0.98 | 0.13 | 0.40 | -0.13 | -0.15 | 0.35 | 0.71 | -0.01 | 0.31 | 0.21 | 0.10 |
| Prox2 | -0.07 | 0.00 | 0.18 | 0.18 | 0.22 | 0.52 | 0.60 | 0.61 | 0.52 | 0.61 | 0.60 | 0.60 | 0.94 | 0.98 | 1.00 | 0.12 | 0.37 | -0.15 | -0.17 | 0.33 | 0.72 | 0.01 | 0.30 | 0.19 | 0.12 |
| Comm2 | 0.06 | 0.18 | 0.19 | 0.20 | 0.17 | 0.20 | 0.19 | 0.18 | 0.20 | 0.19 | 0.19 | 0.19 | 0.14 | 0.13 | 0.12 | 1.00 | 0.25 | 0.22 | 0.16 | 0.19 | 0.08 | -0.19 | 0.18 | 0.16 | 0.03 |
| Aprox | 0.31 | 0.24 | 0.62 | 0.63 | 0.68 | 0.74 | 0.71 | 0.70 | 0.74 | 0.69 | 0.71 | 0.71 | 0.44 | 0.40 | 0.37 | 0.25 | 1.00 | 0.51 | 0.62 | 0.71 | 0.03 | 0.05 | 0.74 | 0.44 | 0.18 |
| Pagerank | 0.04 | 0.35 | 0.25 | 0.26 | 0.30 | 0.57 | 0.58 | 0.58 | 0.57 | 0.58 | 0.57 | 0.57 | -0.04 | -0.13 | -0.15 | 0.22 | 0.51 | 1.00 | 0.91 | 0.50 | -0.14 | -0.29 | 0.72 | 0.71 | -0.12 |
| Degree | 0.36 | 0.28 | 0.45 | 0.45 | 0.49 | 0.59 | 0.57 | 0.56 | 0.59 | 0.56 | 0.57 | 0.56 | -0.06 | -0.15 | -0.17 | 0.16 | 0.62 | 0.91 | 1.00 | 0.55 | -0.34 | -0.04 | 0.80 | 0.64 | 0.07 |
| Current | 0.22 | 0.20 | 0.43 | 0.43 | 0.48 | 0.71 | 0.70 | 0.70 | 0.71 | 0.69 | 0.71 | 0.71 | 0.43 | 0.35 | 0.33 | 0.19 | 0.71 | 0.50 | 0.55 | 1.00 | 0.08 | -0.02 | 0.67 | 0.49 | 0.15 |
| Vitality | -0.56 | 0.02 | -0.15 | -0.15 | -0.14 | 0.25 | 0.34 | 0.36 | 0.25 | 0.37 | 0.33 | 0.34 | 0.71 | 0.71 | 0.72 | 0.08 | 0.03 | -0.14 | -0.34 | 0.08 | 1.00 | -0.32 | -0.01 | 0.08 | -0.18 |
| AND | 0.66 | 0.15 | 0.35 | 0.35 | 0.30 | -0.12 | -0.11 | -0.14 | -0.12 | -0.14 | -0.12 | -0.13 | -0.03 | -0.01 | 0.01 | -0.19 | 0.05 | -0.29 | -0.04 | -0.02 | -0.32 | 1.00 | -0.03 | -0.27 | 0.47 |
| Diversity | 0.30 | 0.25 | 0.52 | 0.52 | 0.57 | 0.76 | 0.78 | 0.77 | 0.76 | 0.76 | 0.78 | 0.77 | 0.38 | 0.31 | 0.30 | 0.18 | 0.74 | 0.72 | 0.80 | 0.67 | -0.01 | -0.03 | 1.00 | 0.72 | 0.06 |
| GenDiv | 0.05 | 0.22 | -0.02 | -0.01 | 0.15 | 0.69 | 0.72 | 0.72 | 0.69 | 0.72 | 0.71 | 0.71 | 0.29 | 0.21 | 0.19 | 0.16 | 0.44 | 0.71 | 0.64 | 0.49 | 0.08 | -0.27 | 0.72 | 1.00 | -0.05 |
| Comm1 | 0.52 | -0.07 | 0.23 | 0.23 | 0.28 | 0.13 | 0.14 | 0.12 | 0.12 | 0.12 | 0.13 | 0.13 | 0.10 | 0.10 | 0.12 | 0.03 | 0.18 | -0.12 | 0.07 | 0.15 | -0.18 | 0.47 | 0.06 | -0.05 | 1.00 |

High Positive Correlation

High Negative Correlation

**Table 3: Cluster Groups of Spearman's Rho (Size 50&100) (All Measures)**

| | Closeness | Eigen | Clustering | Soffers | Square | Flow | Betweenness | Length | Linear | Newman | Prox1 | Stress | KBetweenness | Load | Endpoint | Prox2 | Comm2 | Aprox | Pagerank | Degree | Current | Vitality | AND | Diversity | GenDiv | Comm1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Closeness | 1.00 | 0.60 | 0.55 | 0.55 | 0.52 | 0.30 | 0.23 | 0.26 | 0.22 | 0.23 | 0.22 | 0.23 | 0.22 | 0.09 | 0.06 | 0.09 | 0.61 | 0.29 | 0.19 | 0.42 | 0.26 | -0.37 | 0.61 | -0.61 | 0.61 | 0.82 |
| Eigen | 0.60 | 1.00 | 0.49 | 0.50 | 0.47 | 0.41 | 0.37 | 0.38 | 0.36 | 0.37 | 0.35 | 0.36 | 0.37 | 0.18 | 0.14 | 0.15 | 0.74 | 0.39 | 0.44 | 0.52 | 0.31 | -0.12 | 0.45 | -0.70 | 0.43 | 0.61 |
| Clustering | 0.55 | 0.49 | 1.00 | 1.00 | 0.79 | 0.54 | 0.34 | 0.37 | 0.34 | 0.34 | 0.34 | 0.35 | 0.34 | 0.25 | 0.24 | 0.27 | 0.53 | 0.57 | 0.43 | 0.53 | 0.37 | -0.04 | 0.15 | -0.40 | 0.38 | 0.58 |
| Soffers | 0.55 | 0.50 | 1.00 | 1.00 | 0.80 | 0.55 | 0.35 | 0.37 | 0.35 | 0.35 | 0.34 | 0.36 | 0.35 | 0.26 | 0.24 | 0.27 | 0.54 | 0.57 | 0.44 | 0.53 | 0.37 | -0.03 | 0.15 | -0.41 | 0.38 | 0.59 |
| Square | 0.52 | 0.47 | 0.79 | 0.80 | 1.00 | 0.60 | 0.47 | 0.49 | 0.47 | 0.47 | 0.45 | 0.48 | 0.47 | 0.33 | 0.32 | 0.34 | 0.53 | 0.60 | 0.46 | 0.56 | 0.41 | 0.01 | 0.10 | -0.37 | 0.38 | 0.59 |
| Flow | 0.30 | 0.41 | 0.54 | 0.55 | 0.60 | 1.00 | 0.94 | 0.94 | 0.94 | 0.94 | 0.93 | 0.95 | 0.94 | 0.78 | 0.71 | 0.69 | 0.79 | 0.87 | 0.72 | 0.73 | 0.64 | 0.42 | -0.21 | -0.34 | 0.37 | 0.51 |
| Betweenness | 0.23 | 0.37 | 0.34 | 0.35 | 0.47 | 0.94 | 1.00 | 0.99 | 0.99 | 1.00 | 0.98 | 0.99 | 0.99 | 0.82 | 0.72 | 0.70 | 0.74 | 0.81 | 0.72 | 0.71 | 0.61 | 0.48 | -0.28 | -0.30 | 0.30 | 0.42 |
| Length | 0.26 | 0.38 | 0.37 | 0.37 | 0.49 | 0.94 | 0.99 | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.83 | 0.74 | 0.74 | 0.73 | 0.79 | 0.71 | 0.70 | 0.60 | 0.50 | -0.26 | -0.31 | 0.35 | 0.44 |
| Linear | 0.22 | 0.36 | 0.34 | 0.35 | 0.47 | 0.94 | 0.99 | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.83 | 0.74 | 0.74 | 0.73 | 0.79 | 0.71 | 0.69 | 0.60 | 0.51 | -0.28 | -0.29 | 0.33 | 0.43 |
| Newman | 0.23 | 0.37 | 0.34 | 0.35 | 0.47 | 0.94 | 1.00 | 0.99 | 0.99 | 1.00 | 0.98 | 0.99 | 0.99 | 0.82 | 0.72 | 0.70 | 0.74 | 0.81 | 0.72 | 0.71 | 0.62 | 0.48 | -0.28 | -0.30 | 0.30 | 0.42 |
| Prox1 | 0.22 | 0.35 | 0.34 | 0.34 | 0.45 | 0.93 | 0.98 | 0.99 | 0.99 | 0.98 | 1.00 | 0.98 | 0.98 | 0.83 | 0.74 | 0.74 | 0.72 | 0.78 | 0.70 | 0.69 | 0.59 | 0.53 | -0.28 | -0.29 | 0.33 | 0.42 |
| Stress | 0.23 | 0.36 | 0.35 | 0.36 | 0.48 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 1.00 | 1.00 | 0.83 | 0.75 | 0.73 | 0.73 | 0.80 | 0.70 | 0.69 | 0.61 | 0.49 | -0.27 | -0.30 | 0.34 | 0.43 |
| KBetweenness | 0.22 | 0.37 | 0.34 | 0.35 | 0.47 | 0.94 | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 1.00 | 1.00 | 0.83 | 0.75 | 0.73 | 0.73 | 0.80 | 0.70 | 0.69 | 0.61 | 0.50 | -0.27 | -0.32 | 0.34 | 0.43 |
| Load | 0.09 | 0.18 | 0.25 | 0.26 | 0.33 | 0.78 | 0.82 | 0.83 | 0.83 | 0.82 | 0.83 | 0.83 | 0.83 | 1.00 | 0.95 | 0.94 | 0.53 | 0.64 | 0.29 | 0.26 | 0.49 | 0.74 | -0.20 | -0.17 | 0.52 | 0.41 |
| Endpoint | 0.06 | 0.14 | 0.24 | 0.24 | 0.32 | 0.71 | 0.72 | 0.74 | 0.74 | 0.72 | 0.74 | 0.75 | 0.75 | 0.95 | 1.00 | 0.98 | 0.44 | 0.58 | 0.18 | 0.14 | 0.44 | 0.75 | -0.17 | -0.15 | 0.57 | 0.40 |
| Prox2 | 0.09 | 0.15 | 0.27 | 0.27 | 0.34 | 0.69 | 0.70 | 0.74 | 0.74 | 0.70 | 0.74 | 0.73 | 0.73 | 0.94 | 0.98 | 1.00 | 0.43 | 0.56 | 0.17 | 0.14 | 0.42 | 0.74 | -0.15 | -0.15 | 0.59 | 0.42 |
| Comm2 | 0.61 | 0.74 | 0.53 | 0.54 | 0.53 | 0.79 | 0.74 | 0.73 | 0.73 | 0.74 | 0.72 | 0.73 | 0.73 | 0.53 | 0.44 | 0.43 | 1.00 | 0.69 | 0.65 | 0.73 | 0.53 | 0.08 | 0.29 | -0.66 | 0.50 | 0.73 |
| Aprox | 0.29 | 0.39 | 0.57 | 0.57 | 0.60 | 0.87 | 0.81 | 0.79 | 0.79 | 0.81 | 0.78 | 0.80 | 0.80 | 0.64 | 0.58 | 0.56 | 0.69 | 1.00 | 0.69 | 0.71 | 0.66 | 0.30 | -0.21 | -0.31 | 0.27 | 0.45 |
| Pagerank | 0.19 | 0.44 | 0.43 | 0.44 | 0.46 | 0.72 | 0.72 | 0.71 | 0.71 | 0.72 | 0.70 | 0.70 | 0.70 | 0.29 | 0.18 | 0.17 | 0.65 | 0.69 | 1.00 | 0.93 | 0.48 | 0.11 | -0.35 | -0.30 | -0.13 | 0.15 |
| Degree | 0.42 | 0.52 | 0.53 | 0.53 | 0.56 | 0.73 | 0.71 | 0.70 | 0.69 | 0.71 | 0.69 | 0.69 | 0.69 | 0.26 | 0.14 | 0.14 | 0.73 | 0.71 | 0.93 | 1.00 | 0.51 | -0.04 | -0.18 | -0.41 | 0.03 | 0.38 |
| Current | 0.26 | 0.31 | 0.37 | 0.37 | 0.41 | 0.64 | 0.61 | 0.60 | 0.60 | 0.62 | 0.59 | 0.61 | 0.61 | 0.49 | 0.44 | 0.42 | 0.53 | 0.66 | 0.48 | 0.51 | 1.00 | 0.22 | -0.12 | -0.25 | 0.23 | 0.37 |
| Vitality | -0.37 | -0.12 | -0.04 | -0.03 | 0.01 | 0.42 | 0.48 | 0.50 | 0.51 | 0.48 | 0.53 | 0.49 | 0.50 | 0.74 | 0.75 | 0.74 | 0.08 | 0.30 | 0.11 | -0.04 | 0.22 | 1.00 | -0.45 | 0.15 | 0.16 | -0.07 |
| AND | 0.61 | 0.45 | 0.15 | 0.15 | 0.10 | -0.21 | -0.28 | -0.26 | -0.28 | -0.28 | -0.28 | -0.27 | -0.27 | -0.20 | -0.17 | -0.15 | 0.29 | -0.21 | -0.35 | -0.18 | -0.12 | -0.45 | 1.00 | -0.47 | 0.54 | 0.61 |
| Diversity | -0.61 | -0.70 | -0.40 | -0.41 | -0.37 | -0.34 | -0.30 | -0.31 | -0.29 | -0.30 | -0.29 | -0.30 | -0.32 | -0.17 | -0.15 | -0.15 | -0.66 | -0.31 | -0.30 | -0.41 | -0.25 | 0.15 | -0.47 | 1.00 | -0.54 | -0.58 |
| GenDiv | 0.61 | 0.43 | 0.38 | 0.38 | 0.38 | 0.37 | 0.30 | 0.35 | 0.33 | 0.30 | 0.33 | 0.34 | 0.34 | 0.52 | 0.57 | 0.59 | 0.50 | 0.27 | -0.13 | 0.03 | 0.23 | 0.16 | 0.54 | -0.54 | 1.00 | 0.80 |
| Comm1 | 0.82 | 0.61 | 0.58 | 0.59 | 0.59 | 0.51 | 0.42 | 0.44 | 0.43 | 0.42 | 0.42 | 0.43 | 0.43 | 0.41 | 0.40 | 0.42 | 0.73 | 0.45 | 0.15 | 0.38 | 0.37 | -0.07 | 0.61 | -0.58 | 0.80 | 1.00 |

High Positive Correlation

High Negative Correlation

This study found that the Kendall's Tau Correlation Coefficient identified the same 16 measures. As noted in the texts of Conover, the Kendall's Tau Correlation Coefficients produced nearly identical results for all measures (1980, p. 258). In fact, every absolute value of Kendall's Tau Correlation Coefficients was consistently 0.05 to 0.20 units lower than Spearman's Rank Correlation Coefficients, as was expected. Table 4 and Table 5 show these results that compare to those of the Spearman's Rank Correlation Coefficient. Again, Kendall's Tau was computed on the full set of graph sizes, without flow betweenness, as well as the subset.

### 4.4. Similarity Testing

In order to confirm the results of the correlation coefficients both the density plots and pair-wise scatter plots of each group were reviewed. These plots are shown in Figures 9-16. Group 1, consisting of the three clustering measures, confirms the results shown in the correlation matrices. The density distribution plot in Figure 9 shows the cumulative distribution function (CDF) for the three measures. The CDF is equal to the area under the curve and shows the probability of being several standard deviations away from the mean. In group 1's density plot it is clearly visible that both clustering and Soffer's clustering adhere to a very similar distribution. In addition, looking at the pair-wise scatter plot, shown in Figure 10, one can observe the same correlation between clustering and Soffer's clustering. This scatter plot does not show an exact correlation but there is a definite relation between the pair-wise raw scores and thus high association between the two. In addition, square clustering is shown to have discrepancies to these other measures. These plots both follow the same results as both Spearman's and Kendall's correlation coefficients. That is, clustering and Soffer's clustering are very similar in its description of the networks, and square clustering, in fact, is not an equivalent stand alone measure in a scale-free network. The authors of the square correlation coefficient

state that this measure could be used to strengthen the other two correlation measures as well as provide a way to measure bipartite graph clustering, in which triangles are absent.

## Table 4: Cluster Groups of Kendall's Tau (Flow Betweenness Excluded)

| | Closeness | Eigen | Clustering | Soffers | Square | Betweenness | Length | Linear | Newman | Prox1 | Stress | KBetweenness | Load | Endpoint | Prox2 | Comm2 | Aprox | Pagerank | Degree | Current | Vitality | AND | Diversity | GenDiv | Comm1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Closeness | 1.00 | 0.07 | 0.43 | 0.57 | 0.45 | 0.10 | 0.12 | 0.11 | 0.11 | 0.10 | 0.08 | 0.08 | 0.00 | -0.03 | 0.01 | 0.05 | 0.24 | 0.07 | 0.25 | 0.21 | -0.25 | 0.48 | 0.29 | 0.04 | 0.21 |
| Eigen | 0.07 | 1.00 | 0.17 | -0.06 | 0.17 | 0.09 | 0.13 | 0.13 | 0.13 | 0.13 | 0.14 | 0.15 | 0.03 | 0.02 | 0.01 | 0.22 | 0.20 | 0.24 | 0.24 | 0.10 | 0.20 | 0.11 | -0.02 | 0.18 | -0.21 |
| Clustering | 0.43 | 0.17 | 1.00 | 0.99 | 0.49 | 0.15 | 0.20 | 0.20 | 0.17 | 0.20 | 0.18 | 0.20 | 0.21 | 0.29 | 0.24 | 0.10 | 0.29 | 0.12 | 0.32 | 0.27 | -0.08 | 0.15 | 0.44 | -0.11 | 0.11 |
| Soffers | 0.57 | -0.06 | 0.99 | 1.00 | 0.49 | 0.22 | 0.21 | 0.21 | 0.17 | 0.21 | 0.18 | 0.20 | 0.22 | 0.30 | 0.24 | 0.10 | 0.29 | 0.13 | 0.31 | 0.28 | -0.07 | 0.14 | 0.44 | -0.11 | 0.11 |
| Square | 0.45 | 0.17 | 0.49 | 0.49 | 1.00 | 0.27 | 0.21 | 0.19 | 0.26 | 0.18 | 0.26 | 0.29 | 0.13 | 0.16 | 0.19 | 0.10 | 0.37 | 0.18 | 0.40 | 0.36 | -0.11 | 0.12 | 0.39 | 0.07 | 0.15 |
| Betweenness | 0.10 | 0.09 | 0.15 | 0.22 | 0.27 | 1.00 | 0.73 | 0.83 | 0.76 | 0.83 | 0.60 | 0.73 | 0.56 | 0.51 | 0.44 | 0.14 | 0.43 | 0.46 | 0.50 | 0.62 | 0.16 | -0.08 | 0.51 | 0.42 | 0.09 |
| Length | 0.12 | 0.13 | 0.20 | 0.21 | 0.21 | 0.73 | 1.00 | 0.95 | 0.63 | 0.96 | 0.68 | 0.81 | 0.63 | 0.63 | 0.53 | 0.13 | 0.42 | 0.45 | 0.50 | 0.60 | 0.36 | -0.08 | 0.67 | 0.64 | 0.11 |
| Linear | 0.11 | 0.13 | 0.20 | 0.21 | 0.19 | 0.83 | 0.95 | 1.00 | 0.63 | 0.97 | 0.68 | 0.82 | 0.64 | 0.73 | 0.54 | 0.12 | 0.41 | 0.45 | 0.49 | 0.59 | 0.42 | -0.09 | 0.61 | 0.62 | 0.09 |
| Newman | 0.11 | 0.13 | 0.17 | 0.17 | 0.26 | 0.76 | 0.63 | 0.63 | 1.00 | 0.83 | 0.61 | 0.73 | 0.56 | 0.52 | 0.44 | 0.14 | 0.43 | 0.46 | 0.49 | 0.63 | 0.16 | -0.08 | 0.50 | 0.47 | 0.09 |
| Prox1 | 0.10 | 0.13 | 0.20 | 0.21 | 0.18 | 0.83 | 0.96 | 0.97 | 0.83 | 1.00 | 0.66 | 0.79 | 0.82 | 0.73 | 0.55 | 0.12 | 0.40 | 0.44 | 0.49 | 0.59 | 0.42 | -0.09 | 0.62 | 0.63 | 0.09 |
| Stress | 0.08 | 0.14 | 0.18 | 0.18 | 0.26 | 0.60 | 0.68 | 0.68 | 0.61 | 0.66 | 1.00 | 0.82 | 0.56 | 0.49 | 0.46 | 0.13 | 0.42 | 0.44 | 0.42 | 0.61 | 0.25 | -0.08 | 0.56 | 0.51 | 0.10 |
| KBetweenness | 0.08 | 0.15 | 0.20 | 0.20 | 0.29 | 0.73 | 0.81 | 0.82 | 0.73 | 0.79 | 0.82 | 1.00 | 0.71 | 0.61 | 0.57 | 0.09 | 0.55 | 0.44 | 0.44 | 0.61 | 0.33 | -0.09 | 0.65 | 0.62 | 0.05 |
| Load | 0.00 | 0.03 | 0.21 | 0.22 | 0.13 | 0.56 | 0.63 | 0.64 | 0.56 | 0.82 | 0.56 | 0.71 | 1.00 | 0.71 | 0.64 | 0.10 | 0.30 | 0.34 | 0.10 | 0.51 | 0.39 | -0.07 | 0.32 | 0.28 | 0.06 |
| Endpoint | -0.03 | 0.02 | 0.29 | 0.30 | 0.16 | 0.51 | 0.63 | 0.73 | 0.52 | 0.73 | 0.49 | 0.61 | 0.71 | 1.00 | 0.68 | 0.09 | 0.28 | 0.28 | -0.08 | 0.45 | 0.38 | -0.05 | 0.26 | 0.26 | 0.04 |
| Prox2 | 0.01 | 0.01 | 0.24 | 0.24 | 0.19 | 0.44 | 0.53 | 0.54 | 0.44 | 0.55 | 0.46 | 0.57 | 0.64 | 0.68 | 1.00 | 0.08 | 0.25 | 0.27 | -0.01 | 0.41 | 0.37 | -0.04 | 0.30 | 0.23 | 0.05 |
| Comm2 | 0.05 | 0.22 | 0.10 | 0.10 | 0.10 | 0.14 | 0.13 | 0.12 | 0.14 | 0.12 | 0.13 | 0.09 | 0.10 | 0.09 | 0.08 | 1.00 | 0.21 | 0.06 | 0.14 | 0.04 | 0.04 | -0.11 | 0.15 | 0.12 | 0.03 |
| Aprox | 0.24 | 0.20 | 0.29 | 0.29 | 0.37 | 0.43 | 0.42 | 0.41 | 0.43 | 0.40 | 0.42 | 0.55 | 0.30 | 0.28 | 0.25 | 0.21 | 1.00 | 0.40 | 0.47 | 0.65 | 0.02 | 0.00 | 0.52 | 0.29 | 0.13 |
| Pagerank | 0.07 | 0.24 | 0.12 | 0.13 | 0.18 | 0.46 | 0.45 | 0.45 | 0.46 | 0.44 | 0.44 | 0.44 | 0.34 | 0.28 | 0.27 | 0.06 | 0.40 | 1.00 | 0.66 | 0.42 | 0.15 | -0.20 | 0.52 | 0.58 | -0.13 |
| Degree | 0.25 | 0.24 | 0.32 | 0.31 | 0.40 | 0.50 | 0.50 | 0.49 | 0.49 | 0.49 | 0.42 | 0.44 | 0.10 | -0.08 | -0.01 | 0.14 | 0.47 | 0.66 | 1.00 | 0.49 | -0.18 | -0.06 | 0.57 | 0.54 | 0.03 |
| Current | 0.21 | 0.10 | 0.27 | 0.28 | 0.36 | 0.62 | 0.60 | 0.59 | 0.63 | 0.59 | 0.61 | 0.61 | 0.51 | 0.45 | 0.41 | 0.04 | 0.65 | 0.42 | 0.49 | 1.00 | 0.18 | -0.02 | 0.48 | 0.48 | 0.09 |
| Vitality | -0.25 | 0.20 | -0.08 | -0.07 | -0.11 | 0.16 | 0.36 | 0.42 | 0.16 | 0.42 | 0.25 | 0.33 | 0.39 | 0.38 | 0.37 | 0.04 | 0.02 | 0.15 | -0.18 | 0.18 | 1.00 | -0.22 | 0.02 | -0.04 | -0.17 |
| AND | 0.48 | 0.11 | 0.15 | 0.14 | 0.12 | -0.08 | -0.08 | -0.09 | -0.08 | -0.09 | -0.08 | -0.09 | -0.07 | -0.05 | -0.04 | -0.11 | 0.00 | -0.20 | -0.06 | -0.02 | -0.22 | 1.00 | -0.04 | -0.18 | 0.34 |
| Diversity | 0.29 | -0.02 | 0.44 | 0.44 | 0.39 | 0.51 | 0.67 | 0.61 | 0.50 | 0.62 | 0.56 | 0.65 | 0.32 | 0.26 | 0.30 | 0.15 | 0.52 | 0.52 | 0.57 | 0.48 | 0.02 | -0.04 | 1.00 | 0.54 | 0.02 |
| GenDiv | 0.04 | 0.18 | -0.11 | -0.11 | 0.07 | 0.42 | 0.64 | 0.62 | 0.47 | 0.63 | 0.51 | 0.62 | 0.28 | 0.26 | 0.23 | 0.12 | 0.29 | 0.58 | 0.54 | 0.48 | -0.04 | -0.18 | 0.54 | 1.00 | -0.03 |
| Comm1 | 0.21 | -0.21 | 0.11 | 0.11 | 0.15 | 0.09 | 0.11 | 0.09 | 0.09 | 0.09 | 0.10 | 0.05 | 0.06 | 0.04 | 0.05 | 0.03 | 0.13 | -0.13 | 0.03 | 0.09 | -0.17 | 0.34 | 0.02 | -0.03 | 1.00 |

High Positive Correlation

High Negative Correlation

**Table 5: Cluster Groups of Kendall's Tau (Size 50&100) (All Measures)**

| | Closeness | Eigen | Clustering | Soffers | Square | Flow | Betweenness | Length | Linear | Newman | Prox1 | Stress | KBetweenness | Load | Endpoint | Prox2 | Comm2 | Aprox | Pagerank | Degree | Current | Vitality | AND | Diversity | GenDiv | Comm1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Closeness | 1.00 | 0.46 | 0.42 | 0.54 | 0.44 | 0.26 | 0.18 | 0.20 | 0.23 | 0.20 | 0.23 | 0.18 | 0.19 | 0.12 | 0.05 | 0.12 | 0.46 | 0.23 | 0.23 | 0.30 | 0.24 | -0.12 | 0.43 | -0.36 | 0.40 | 0.55 |
| Eigen | 0.46 | 1.00 | 0.37 | 0.44 | 0.38 | 0.35 | 0.18 | 0.29 | 0.34 | 0.33 | 0.34 | 0.28 | 0.31 | 0.20 | 0.10 | 0.18 | 0.60 | 0.32 | 0.41 | 0.39 | 0.28 | 0.10 | 0.30 | -0.29 | 0.22 | 0.33 |
| Clustering | 0.42 | 0.37 | 1.00 | 0.98 | 0.49 | 0.31 | 0.18 | 0.20 | 0.23 | 0.19 | 0.24 | 0.19 | 0.21 | 0.25 | 0.39 | 0.27 | 0.20 | 0.27 | 0.24 | 0.38 | 0.22 | 0.00 | 0.10 | -0.31 | 0.28 | 0.36 |
| Soffers | 0.54 | 0.44 | 0.98 | 1.00 | 0.49 | 0.32 | 0.25 | 0.21 | 0.24 | 0.20 | 0.25 | 0.20 | 0.22 | 0.25 | 0.40 | 0.28 | 0.21 | 0.28 | 0.25 | 0.48 | 0.23 | 0.00 | 0.10 | -0.32 | 0.29 | 0.37 |
| Square | 0.44 | 0.38 | 0.49 | 0.49 | 1.00 | 0.37 | 0.26 | 0.28 | 0.27 | 0.26 | 0.25 | 0.28 | 0.30 | 0.23 | 0.25 | 0.29 | 0.22 | 0.31 | 0.29 | 0.49 | 0.28 | 0.01 | 0.07 | -0.28 | 0.34 | 0.39 |
| Flow | 0.26 | 0.35 | 0.31 | 0.32 | 0.37 | 1.00 | 0.67 | 0.69 | 0.92 | 0.76 | 0.92 | 0.71 | 0.79 | 0.78 | 0.64 | 0.67 | 0.47 | 0.61 | 0.55 | 0.66 | 0.57 | 0.43 | -0.11 | -0.26 | 0.29 | 0.34 |
| Betweenness | 0.18 | 0.18 | 0.18 | 0.25 | 0.26 | 0.67 | 1.00 | 0.67 | 0.92 | 0.82 | 0.91 | 0.68 | 0.77 | 0.75 | 0.72 | 0.62 | 0.40 | 0.52 | 0.59 | 0.62 | 0.55 | 0.37 | -0.14 | -0.22 | 0.22 | 0.32 |
| Length | 0.20 | 0.29 | 0.20 | 0.21 | 0.28 | 0.69 | 0.67 | 1.00 | 0.96 | 0.76 | 0.96 | 0.69 | 0.78 | 0.75 | 0.57 | 0.68 | 0.39 | 0.50 | 0.57 | 0.54 | 0.51 | 0.45 | -0.13 | -0.24 | 0.27 | 0.35 |
| Linear | 0.23 | 0.34 | 0.23 | 0.24 | 0.27 | 0.92 | 0.92 | 0.96 | 1.00 | 0.76 | 0.97 | 0.69 | 0.78 | 0.76 | 0.82 | 0.68 | 0.38 | 0.49 | 0.56 | 0.64 | 0.50 | 0.49 | -0.14 | -0.17 | 0.25 | 0.32 |
| Newman | 0.20 | 0.33 | 0.19 | 0.20 | 0.26 | 0.76 | 0.82 | 0.76 | 0.76 | 1.00 | 0.91 | 0.68 | 0.77 | 0.75 | 0.70 | 0.62 | 0.40 | 0.52 | 0.59 | 0.64 | 0.55 | 0.37 | -0.14 | -0.23 | 0.24 | 0.32 |
| Prox1 | 0.23 | 0.34 | 0.24 | 0.25 | 0.25 | 0.92 | 0.91 | 0.96 | 0.97 | 0.91 | 1.00 | 0.67 | 0.76 | 0.90 | 0.81 | 0.68 | 0.37 | 0.47 | 0.55 | 0.64 | 0.49 | 0.48 | -0.14 | -0.17 | 0.26 | 0.32 |
| Stress | 0.18 | 0.28 | 0.19 | 0.20 | 0.28 | 0.71 | 0.68 | 0.69 | 0.69 | 0.68 | 0.67 | 1.00 | 0.79 | 0.67 | 0.59 | 0.56 | 0.39 | 0.50 | 0.55 | 0.53 | 0.52 | 0.39 | -0.13 | -0.23 | 0.26 | 0.33 |
| KBetweenness | 0.19 | 0.31 | 0.21 | 0.22 | 0.30 | 0.79 | 0.77 | 0.78 | 0.78 | 0.77 | 0.76 | 0.79 | 1.00 | 0.76 | 0.67 | 0.65 | 0.65 | 0.62 | 0.55 | 0.50 | 0.54 | 0.43 | -0.13 | -0.23 | 0.27 | 0.31 |
| Load | 0.12 | 0.20 | 0.25 | 0.25 | 0.23 | 0.78 | 0.75 | 0.75 | 0.76 | 0.75 | 0.90 | 0.67 | 0.76 | 1.00 | 0.77 | 0.70 | 0.36 | 0.46 | 0.51 | 0.38 | 0.49 | 0.48 | -0.13 | -0.04 | 0.36 | 0.30 |
| Endpoint | 0.05 | 0.10 | 0.39 | 0.40 | 0.25 | 0.64 | 0.72 | 0.57 | 0.82 | 0.70 | 0.81 | 0.59 | 0.67 | 0.77 | 1.00 | 0.74 | 0.30 | 0.41 | 0.44 | 0.11 | 0.44 | 0.48 | -0.11 | -0.10 | 0.28 | 0.29 |
| Prox2 | 0.12 | 0.18 | 0.27 | 0.28 | 0.29 | 0.67 | 0.62 | 0.68 | 0.68 | 0.62 | 0.68 | 0.56 | 0.65 | 0.70 | 0.74 | 1.00 | 0.29 | 0.39 | 0.44 | 0.31 | 0.42 | 0.47 | -0.10 | -0.01 | 0.41 | 0.31 |
| Comm2 | 0.46 | 0.60 | 0.20 | 0.21 | 0.22 | 0.47 | 0.40 | 0.39 | 0.38 | 0.40 | 0.37 | 0.39 | 0.65 | 0.36 | 0.30 | 0.29 | 1.00 | 0.56 | 0.60 | 0.55 | 0.52 | 0.01 | 0.22 | -0.47 | 0.36 | 0.54 |
| Aprox | 0.23 | 0.32 | 0.27 | 0.28 | 0.31 | 0.61 | 0.52 | 0.50 | 0.49 | 0.52 | 0.47 | 0.50 | 0.62 | 0.46 | 0.41 | 0.39 | 0.56 | 1.00 | 0.55 | 0.54 | 0.61 | 0.22 | -0.11 | -0.23 | 0.21 | 0.34 |
| Pagerank | 0.23 | 0.41 | 0.24 | 0.25 | 0.29 | 0.55 | 0.59 | 0.57 | 0.56 | 0.59 | 0.55 | 0.55 | 0.55 | 0.51 | 0.44 | 0.44 | 0.60 | 0.55 | 1.00 | 0.67 | 0.32 | 0.33 | -0.24 | -0.19 | -0.03 | 0.09 |
| Degree | 0.30 | 0.39 | 0.38 | 0.48 | 0.49 | 0.66 | 0.62 | 0.54 | 0.64 | 0.64 | 0.64 | 0.53 | 0.50 | 0.38 | 0.11 | 0.31 | 0.55 | 0.54 | 0.67 | 1.00 | 0.43 | 0.16 | -0.12 | -0.27 | 0.09 | 0.34 |
| Current | 0.24 | 0.28 | 0.22 | 0.23 | 0.28 | 0.57 | 0.55 | 0.51 | 0.50 | 0.55 | 0.49 | 0.52 | 0.54 | 0.49 | 0.44 | 0.42 | 0.52 | 0.61 | 0.32 | 0.43 | 1.00 | 0.26 | -0.07 | -0.15 | 0.20 | 0.25 |
| Vitality | -0.12 | 0.10 | 0.00 | 0.00 | 0.01 | 0.43 | 0.37 | 0.45 | 0.49 | 0.37 | 0.48 | 0.39 | 0.43 | 0.48 | 0.48 | 0.47 | 0.01 | 0.22 | 0.33 | 0.16 | 0.26 | 1.00 | -0.32 | 0.17 | 0.05 | 0.02 |
| AND | 0.43 | 0.30 | 0.10 | 0.10 | 0.07 | -0.11 | -0.14 | -0.13 | -0.14 | -0.14 | -0.14 | -0.13 | -0.13 | -0.13 | -0.11 | -0.10 | 0.22 | -0.11 | -0.24 | -0.12 | -0.07 | -0.32 | 1.00 | -0.32 | 0.39 | 0.45 |
| Diversity | -0.36 | -0.29 | -0.31 | -0.32 | -0.28 | -0.26 | -0.22 | -0.24 | -0.17 | -0.23 | -0.17 | -0.23 | -0.23 | -0.04 | -0.10 | -0.01 | -0.47 | -0.23 | -0.19 | -0.27 | -0.15 | 0.17 | -0.32 | 1.00 | -0.33 | -0.36 |
| GenDiv | 0.40 | 0.22 | 0.28 | 0.29 | 0.34 | 0.29 | 0.22 | 0.27 | 0.25 | 0.24 | 0.26 | 0.26 | 0.27 | 0.36 | 0.28 | 0.41 | 0.36 | 0.21 | -0.03 | 0.09 | 0.20 | 0.05 | 0.39 | -0.33 | 1.00 | 0.58 |
| Comm1 | 0.55 | 0.33 | 0.36 | 0.37 | 0.39 | 0.34 | 0.32 | 0.35 | 0.32 | 0.32 | 0.32 | 0.33 | 0.31 | 0.30 | 0.29 | 0.31 | 0.54 | 0.34 | 0.09 | 0.34 | 0.25 | 0.02 | 0.45 | -0.36 | 0.58 | 1.00 |

High Positive Correlation

High Negative Correlation

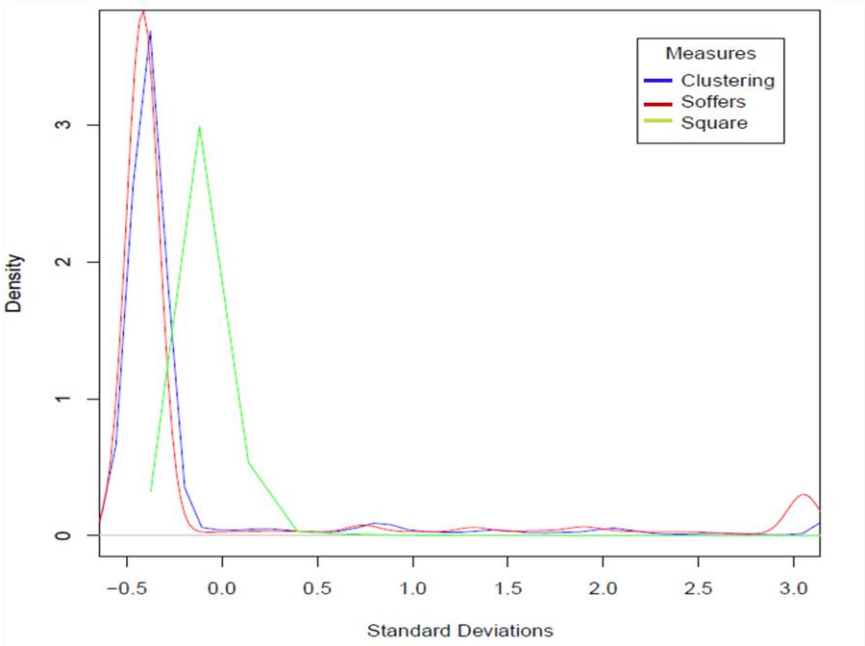


**Figure 9: Group 1 Density Distributions**

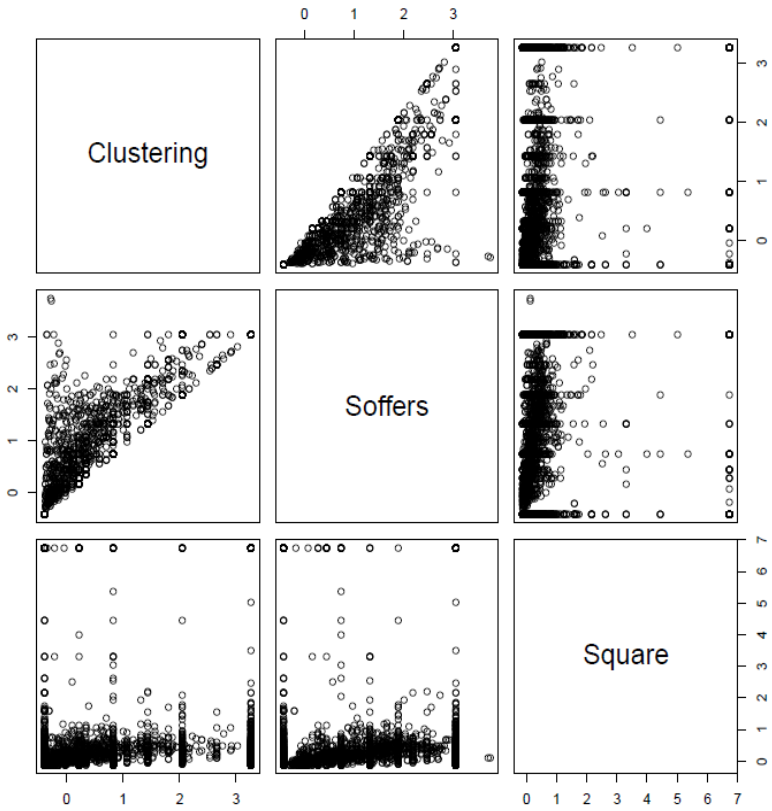


**Figure 10: Group 1 Pair-wise Scatter Plots**

Next, the measures of group 2, flow betweenness, betweenness centrality, length-scaled betweenness, linear-scaled betweenness, Newman's betweenness, source proximal betweenness, stress centrality, and k-betweenness, are compared using the same techniques as group 1. Shown in Figure 11, there are two significantly different sets of distributions. In observation of this plot, it is evident that flow betweenness, betweenness centrality and Newman's betweenness share a very similar distribution, where as length-scaled betweenness, linear-scaled betweenness, source proximal betweenness, stress centrality, and k-betweenness are following another. This result does not reject the hypothesis that this group of measures is similar in scoring networks; it just explains possible stronger relationships within the overall group. In addition, the pair-wise scatter plots show a definite link between the scoring of these measures. The more a linear relationship can be defined within the scatter plots, the more correlated the two measures are over the set of networks. Figure 12 confirms the relationships over the whole group. Flow betweenness is shown to have a clear relationship with all the measures in group two, but most defined is its relationship with stress centrality and k-betweenness. In addition, betweenness centrality and Newman's betweenness are shown to have a near 1.0 correlation. This is not surprising because of the fact they both use the same equation with minor alterations. Another subgroup that shows a substantial relationship is length-scaled betweenness, linear-scaled betweenness and source proximal betweenness. These relationships confirm that these subgroups could be interchangeable with little loss of accuracy.

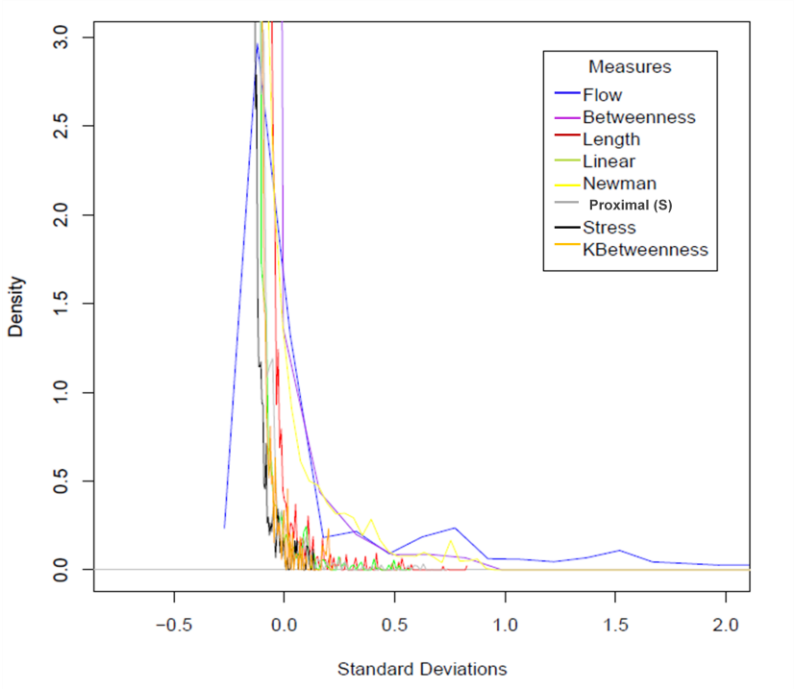


**Figure 11: Group 2 Density Distributions**

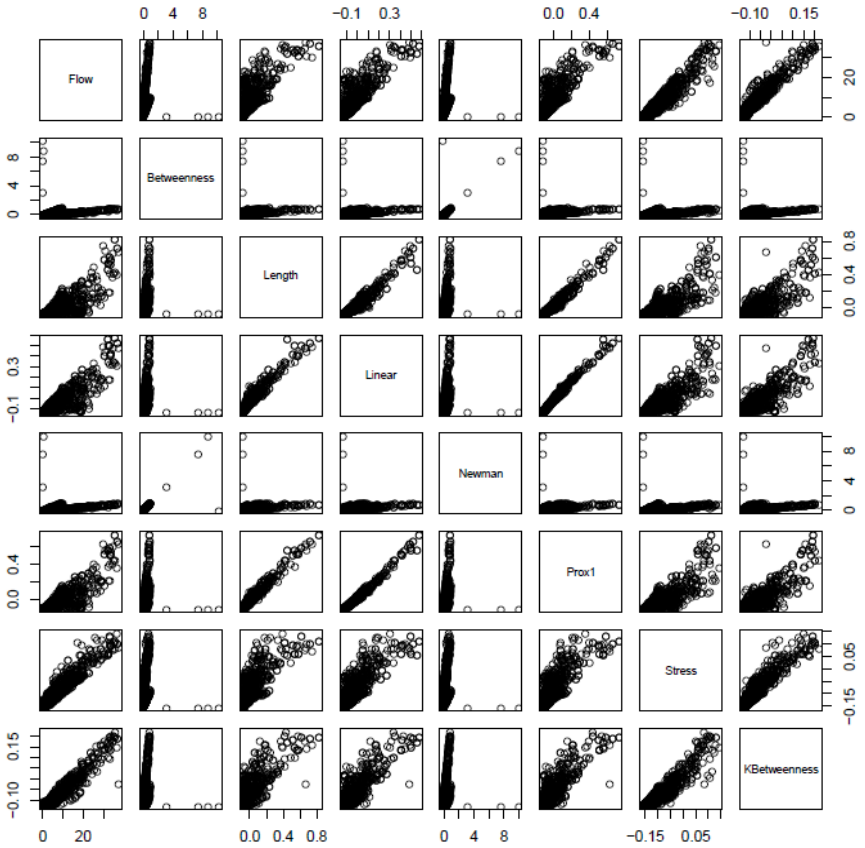


**Figure 12: Group 2 Pair-wise Scatter Plots**

Group 3 consists of the measures target proximal betweenness, load centrality and endpoint betweenness. Again, it is visibly clear that these three measures follow very similar distributions. When looking at Figure 13 there is little deviation from each other. This is also seen in the pair-wise scatter plots in Figure 14. All three measures have a well defined linear relationships, also endpoint and load centrality have what looks like a perfect 1-1 relationship, which corresponds with the 0.98 correlation shown in Table 2.

The last group of measures that express a high correlation with one another is the measures PageRank and degree centrality. Figure 15 displays the density plots of these two measures. Within this plot there are visible discrepancies between the measures, this suggest that these measures, while quite similar, do not identify the same nodes in certain points of their scoring. Figure 16 also shows this with the linear nature that begins in the lower left corner and slowly diverges from each other. This does not mean that these measures should not be inter-changeable, just that caution should be taken and the knowledge that they will not give the same scores to the nodes should be known. These measures do have a 0.91 correlation and show a better relationship then some of the above measures. The same knowledge pertains to those measures as well.

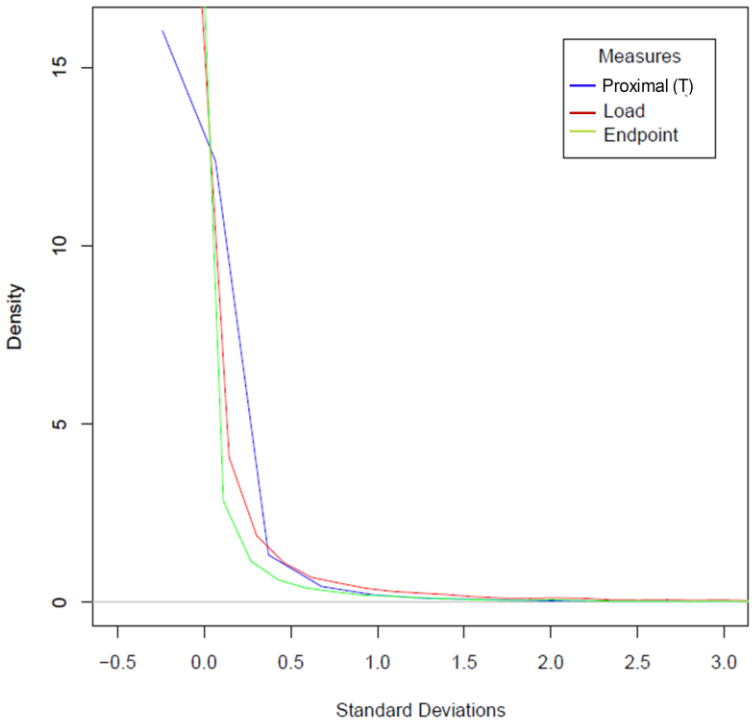


**Figure 13: Group 3 Density Distributions**

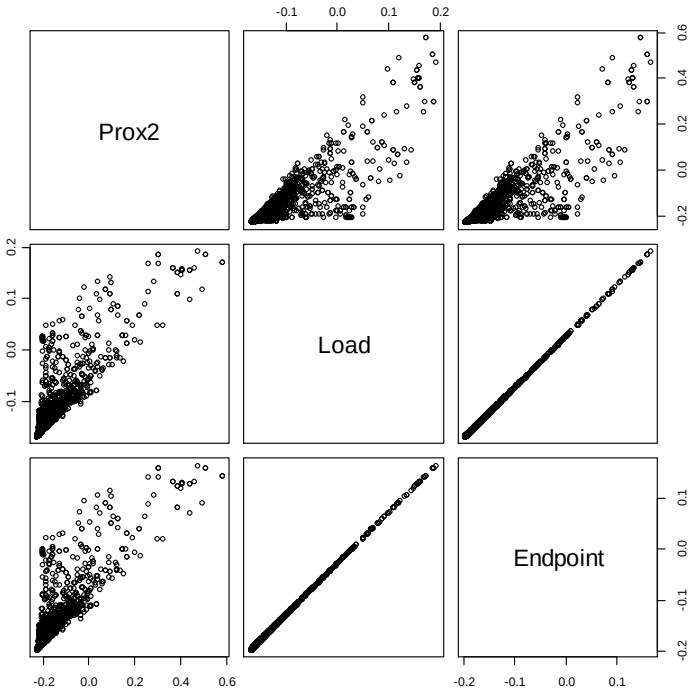


**Figure 14: Group 3 Pair-wise Scatter Plots**

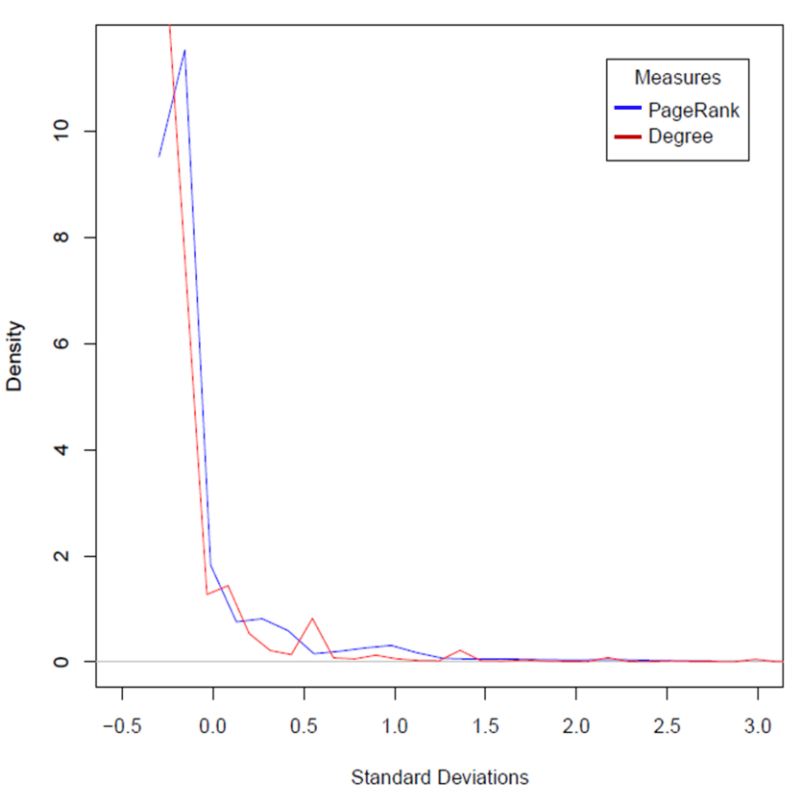


**Figure 15: Group 4 Density Distributions**

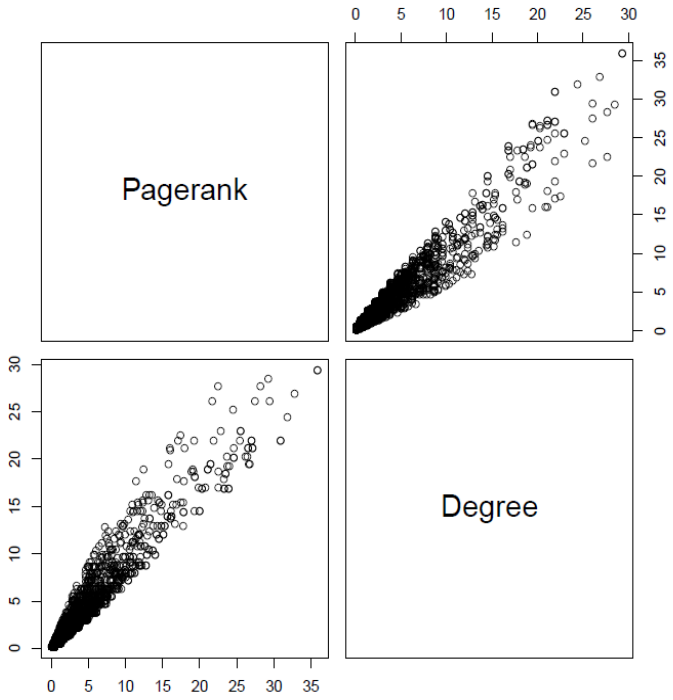


**Figure 16: Group 4 Pair-wise Scatter Plots**

Figures 9-16 help to confirm the results and insights gained from the correlation analysis. Another test of similarity utilized in this study involves only a certain group of rankings. In particular circumstances an analyst may only be interested in that top ranking individuals. These nodes may represent high ranking officials in an organization along with many other representations. This study tested each measure in the groups of interest to see how they perform in identifying the top 10 nodes, averaged over all 320 networks. Equation **(**39**)** allows the analyst to scale the number of unique nodes identified in a pair of nodes. This number permits the study to average over the entire set of networks, and gain a unique node identification score. This score tells at what accuracy the two measure identify the same top 10 (i.e. 0.95 of identification score means that on average one would expect the two measures to identify 9.5 of the same nodes in their top ten). Table 6 shows that there are a sizeable number of measures that identify on average the same 9 or more nodes as in the top 10. It should be noted that the average run time for degree centrality was 0.001 seconds whereas the average run time for flow betweenness was 472.497 seconds, and yet they identify the same top 10 every time.

**Table 6: Top 10 Identification Similarity Score**

| | | | | |
|---|---|---|---|---|
| Degree-Length | Degree-Linear | Between-Newman | Endpoint-Length | Endpoint-Linear |
| 0.911 | 0.912 | 0.945 | 0.942 | 0.937 |
| Degree-Prox2 | Clust-Soffers | Endpoint-Flow | Endpoint-Prox2 | Endpoint-Stress |
| 0.916 | 0.992 | 0.920 | 0.939 | 0.983 |
| Endpoint-Load | Flow-Length | Endpoint-Prox1 | Degree-Flow | Between-Flow |
| 1.00 | 0.920 | 0.919 | 1.00 | 0.920 |
| Endpoint-Kbet | Flow-Newman | Flow-Linear | Flow-Prox2 | Flow-Stress |
| 0.984 | 0.920 | 0.920 | 0.920 | 0.920 |
| Flow-Load | Length-Linear | Length-Load | Linear-Load | Length-Prox1 |
| 0.920 | 0.983 | 0.942 | 0.937 | 0.966 |
| Flow-Kbet | Length-Stress | Length-Endpoint | Linear-Kbet | Prox1-Stress |
| 0.920 | 0.937 | 0.948 | 0.942 | 0.913 |
| Length-Prox2 | Linear-Prox2 | Linear-Stress | Load-Kbet | Prox1-Kbet |
| 0.988 | 0.994 | 0.932 | 0.984 | 0.923 |
| Linear-Prox1 | Load-Prox2 | Load-Stress | Prox1-Prox2 | Stress-Kbet |
| 0.974 | 0.939 | 0.983 | 0.978 | 0.988 |
| Load-Prox1 | Prox2-Stress | Prox2-Kbet | | |
| 0.919 | 0.934 | 0.944 | | |

Similarly, the bottom 10 of each measure was compared over all of the networks as well. The unique node identification score for the bottom 10 nodes could aid analysts in multiple objectives, such as finding a member of a dark network that may be the member most likely to dissent. As shown in Table 7, there are fewer pairs of measures that identify the same nodes for the bottom ranks.

**Table 7: Bottom 10 Identification Similarity Score**

| | | | |
|---|---|---|---|
| Between-Flow | Clust-Soffers | Between-Newman | Endpoint-Stress |
| 0.938 | 0.952 | 0.929 | 0.933 |
| Endpoint-Load | Flow-Newman | Linear-Prox1 | Flow-Stress |
| 0.999 | 0.939 | 0.946 | 0.921 |
| Endpoint-Kbet | Length-Linear | Linear-Prox2 | Length-Prox1 |
| 0.920 | 0.923 | 0.957 | 0.934 |
| Flow-Load | Length-Prox2 | Stress-Kbet | Prox1-Prox2 |
| 0.939 | 0.951 | 1.00 | 0.979 |

## 4.5 Computational Times

The computational times, or the amount of time on average it took to run a measure, for most measures took approximately 10 seconds or less at the largest size networks. As discussed in section 4.2, there were a few exceptions to this, which grew large as the size of the network increased. One component of the SNA analyst's guide is giving alternative measures to reduce the amount of resources used in describing a network, i.e. time. Table 8 shows the computational times experienced in this study over the sizes of networks. In this table of the measures found in the four groups it is clear to see those who increase significantly with the size of the nodes. Other measures highlighted in Table 8 are those that are over two minutes up to half an hour.

**Table 8: Computational Times**

| Time In Seconds | Total Avg | n=50 Avg | n=100 Avg | n=500 Avg | n=1000 Avg |
|---|---|---|---|---|---|
| Degree Time: | 0.001 | 0.000 | 0.000 | 0.001 | 0.001 |
| Betweenness Time: | 2.692 | 0.015 | 0.077 | 1.961 | 8.760 |
| Clustering Time: | 0.019 | 0.001 | 0.002 | 0.0189604 | 0.05318331 |
| Soffer's Time: | 0.061 | 0.003 | 0.008 | 0.061 | 0.175 |
| Endpoint Time: | 94.426 | 0.093 | 0.875 | 42.405 | 336.618 |
| Flow Time: | 472.497 | 128.736 | 1755.346 | NA | NA |
| Length Time: | 168.284 | 0.112 | 0.928 | 42.606 | 634.577 |
| Linear Time: | 94.870 | 0.110 | 0.930 | 42.317 | 338.433 |
| Newman's Time: | 328.692 | 0.197 | 2.834 | 131.820 | 1188.933 |
| Load Time: | 2.469 | 0.013 | 0.082 | 1.825 | 8.002 |
| Prox1 Time: | 95.277 | 0.116 | 1.143 | 44.455 | 337.853 |
| Prox2 Time: | 95.200 | 0.115 | 1.155 | 44.426 | 337.562 |
| Stress Time: | 102.1153 | 0.107984 | 0.6177193 | 40.465497 | 365.994928 |
| *k*-Betweenness Time: | 93.53987 | 0.095886 | 1.0113288 | 43.088458 | 332.956507 |
| Square Time: | 0.389658 | 0.002783 | 0.0107055 | 0.3128932 | 1.22741483 |
| PageRank Time: | 19.47025 | 0.019464 | 0.6687019 | 26.963857 | 49.9858583 |

In order to ensure that these measures are in fact significantly different a Z two sample test for means were used. This statistical test examines the average computational time of a pair

of measures to reject or not reject the null hypothesis that they are in fact statistically equal. As seen in Table 9, this test uses the sample variance of these large samples of computational times along with the number of observations for each measure to calculate a Z statistic. This statistic is then compared to the Z distribution to determine if the two computational times are significantly different. In the case of Table 9, the absolute value of this statistic is greater than the Z critical value, therefore the null hypothesis is rejected and we can say that with 95% confidence flow betweenness is statistically slower in run time than betweenness centrality. Similarly, the absolute value of Table 10's Z statistic is less than the critical value. Therefore, it cannot be said that, statistically, simple clustering and Soffer's clustering are different in run times.

Table 9 and Table 10 show a sample of these tests that were computed for each group of measures. The results from these tests showed that, indeed, flow betweenness is significantly higher in computational time compared to ever measure in group 2. These tests culminate the computations of this thesis and lead way for an SNA analyst's guide. As seen in Table 9, this test uses the sample variance of these large samples of computational times along with the number of observations for each measure to calculate a Z statistic. This statistic is then compared to the Z distribution to determine if the two computational times are significantly different. This is given by the null hypothesis that the two measures' computational times are statically equivalent. In the case of Table 9, the absolute value of this statistic is greater than the Z critical value, therefore the null hypothesis is rejected and we can say that with 95% confidence flow betweenness is statistically slower in run time than betweenness centrality. Similarly, the absolute value of Table 10's Z statistic is less than the critical value. Therefore, it cannot be said that, statistically, simple clustering and Soffer's clustering are different in run times.

**Table 9: Z Two Sample Test for Means (Betweenness & Flow)**

| | *BetTime:* | *FlowTime:* |
|---|---|---|
| Mean | 0.015 | 128.736 |
| Known Variance | 0.008 | 195.806 |
| Observations | 80 | 80 |
| Hypothesized Mean Diff | 0 | |
| z | -82.276 | |
| Significance Level | $\alpha = 0.05$ | |
| P(Z<=z) two-tail | <0.001 | Reject $H_0$ |
| z Critical two-tail | 1.960 | |

**Table 10: Z Two Sample Test for Means (Clustering & Soffer's)**

| | *ClustTime:* | *SofferTime:* |
|---|---|---|
| Mean | 0.0006 | 0.0025 |
| Known Variance | 0.00023 | 0.00083 |
| Observations | 80 | 80 |
| Hypothesized Mean Diff | 0 | |
| z | -0.512 | |
| Significance Level | $\alpha = 0.05$ | |
| P(Z<=z) two-tail | 0.608 | Fail to Reject $H_0$ |
| z Critical two-tail | 1.960 | |

## 4.6 SNA Analyst's Guide

This section of this thesis provides a detail guide to aid the SNA analysts in selecting an appropriate descriptive measure with respect to network structural properties. This guide takes in to account two tests for correlation; overlaid CDFs for similarity, pair-wise scatter plots, top and bottom identification similarity and computational times. Table 11 gives analysts six types of measures to choose from with a description of the intended outcome. After the type of measure is chosen, the rows display the measures within the group and which pairs are best to substitute for one another as well as the estimated computation time. Being tested over several levels of structural properties of networks, and the fact that the patterns of correlation remains unchanged, lead this study to believe that size, scale-free parameter and clusterability do not play

an important role in the relationships of this data set. Further research should be done to test this hypothesis.

The SNA analyst's guide directs the analysis to choosing the proper measure by giving all of the above information as well as which measure is on average preferred within its group. This preference takes in to account computational time and the closeness to real world network attributes, as written in the measures definition. For example, Soffer's clustering is preferred to simple clustering (Soffer's clustering $\succ$ simple clustering), not because these two measures are statistically different in computational time, but because Soffer's clustering takes into account the errors that occur in simple clustering for nodes with higher degree (Soffer & Vazquez, 2005). For the betweenness-based centrality #1 measures, there are three sub groups observed and thus three groups of preferences. Preference for the top and bottom 10 ranked nodes is denoted by †. This also takes into account computational times and real world accuracy based on the definition of the measure. Those pairs who do not have a preferred measure are defined as indifferent to substituting one for the other. When using this guide, the analysis should carefully choose which type of measure to select by defining their objective, reading each description of measure type and reading the group of measures definitions as found in Chapter 2.

**Table 11: SNA Analyst's Guide**

| Types of Measure | Description* | Measures | Substitutable Measures** | Estimated Computational Times (1000 Node Network) in Seconds |
|---|---|---|---|---|
| Degree-Based Centrality | These measures evaluate each node's incoming and outgoing edges in an attempt to find the most 'central' or well connected in the network. The concept of these measures is that an important person is a well connected person. | Degree Centrality<br>PageRank | Degree Centrality ⟷ PageRank<br><br>**Preference‡**<br>Degree Centrality ≻ PageRank | Degree Centrality- 0.00125<br>PageRank- 49.986+ |
| Clustering Measures | These measures evaluate each nodes ability to interact with its neighbors within the network. Triangles an squares made with edges between three or four node are considered a cluster and strongly connected amongst each other. | Simple Clustering<br>Soffer's Clustering<br>Square Clustering | Simple Clustering ⟷ Soffer's<br><br>**Preference**<br>Simple Clustering ≻ Soffer's | Simple Clustering- 0.0187<br>Soffer's Clustering- 0.0610<br>Square Clustering- 1.227+ |
| Betweenness-Based Centrality #1 (Flow) | These measures evaluate each nodes ability to control the flow throughout the network. The control of flow equates to control of the network , thus importance within the network. The measures in this group measure the control of flow by how dependent a node is to the measured node. | Flow Betweenness<br>Betweenness Centrality<br>Length-Scaled Betweenness<br>Linear-Scaled Betweenness<br>Newman's Betweenness<br>Source Proximal Betweenness<br>Stress Centrality<br>*k*-Betweenness | Flow Betweenness ⟷ Stress Centrality<br>Flow Betweenness ⟷ K-Betweenness<br>Betweenness Centrality ⟷ Newman's Betweenness<br>Length Betweenness ⟷ Linear Betweenness<br>Length Betweenness ⟷ Proximal (S) Betweenness<br>Stress Centrality ⟷ *k*-Betweenness<br><br>**Preference**<br>Linear ≻ Proximal (S) ≻ Length Betweenness ≻ Newman's<br>*k*-Betweenness ≻ Stress ≻ Flow Betweenness | Flow Betweenness- 1287361929.52+<br>Betweenness Centrality- 8.760<br>Length-Scaled Betweenness- 634.577++<br>Linear-Scaled Betweenness- 338.433<br>Newman's Betweenness- 1188.933+<br>Proximal (S) Betweenness- 337.853<br>Stress Centrality- 365.995<br>*k*-Betweenness- 332.957 |
| Betweenness-Based Centrality #2 (Flow) | These measures evaluate each nodes ability to control the flow throughout the network. The control of flow equates to control of the network , thus importance within the network. These measures evaluate the dependence of the source or target node on the measured node creating a sore for each node. | Load Centrality<br>Endpoint Betweenness<br>Target Proximal Betweenness | Load Centrality ⟷ Endpoint Betweenness<br><br>Load Centrality ~ Proximal (T) Betweenness***<br><br>Proximal (T) Betweenness ~ Endpoint Betweenness***<br><br>**Preference**<br>Load Centrality ≻ Proximal (T) ≻ Endpoint Betweenness | Load Centrality- 8.002<br>Endpoint Betweenness- 336.618+<br>Proximal (T) Betweenness- 337.562+ |

* Equations and descriptions of each measure can be found in Chapter 2.

** Substitutable measures do not guarantee an exact 1-1 correlation. In addition, these correlations have not been tested on graphs over 1000 node. For average correlation see section 4.3

*** ~ denotes measures are similar but caution should be used if accuracy is a concern. Alternate correlated measures can be found in section 4.3

\+ Measure is significantly slower in computational time.

++ Measure is significantly slower in computational time for graphs of size greater than 500 nodes.

‡ For description of preferences see section 4.6

**Table 11: SNA Analyst's Guide Continued**

| | | | | Identification Similarity Accuracy |
|---|---|---|---|---|
| **Top 10 Ranked Nodes** | These measures are those who identify nodes of high importance with high accuracy of substitution. | Degree Centrality<br>Simple Clustering<br>Soffer's Clustering<br>Flow Betweenness<br>Betweenness Centrality<br>Length-Scaled Betweenness<br>Linear-Scaled Betweenness<br>Newman's Betweenness<br>Source Proximal Betweenness<br>Stress Centrality<br>K-Betweenness<br>Load Centrality<br>Endpoint Betweenness<br>Target Proximal Betweenness | Degree Centrality† ↔ Length Betweenness | 0.911 |
| | | | Degree Centrality† ↔ Linear Betweenness | 0.912 |
| | | | Degree Centrality † ↔ Proximal (T) Betweenness | 0.916 |
| | | | Simple Clustering ↔ Soffer's Clustering | 0.992 |
| | | | Flow Betweenness ↔ Degree Centrality† | 1.00 |
| | | | Flow Betweenness ↔ Betweenness Centrality† | 0.920 |
| | | | Flow Betweenness ↔ Length Betweenness† | 0.920 |
| | | | Flow Betweenness ↔ Linear Betweenness† | 0.920 |
| | | | Flow Betweenness ↔ Newman's Betweenness† | 0.920 |
| | | | Flow Betweenness ↔ Proximal (T) Betweenness† | 0.920 |
| | | | Flow Betweenness ↔ Stress Centrality† | 0.920 |
| | | | Flow Betweenness ↔ K-Betweenness† | 0.920 |
| | | | Flow Betweenness ↔ Load Centrality† | 0.920 |
| | | | Flow Betweenness ↔ Endpoint Betweenness† | 0.920 |
| | | | Betweenness Centrality† ↔ Newman's Betweenness | 0.945 |
| | | | Length Betweenness ↔ Linear Betweenness† | 0.983 |
| | | | Length Betweenness ↔ Proximal (S) Betweenness† | 0.966 |
| | | | Length Betweenness ↔ Proximal (T) Betweenness | 0.988 |
| | | | Length Betweenness ↔ Stress Centrality† | 0.937 |
| | | | Length Betweenness ↔ Load Centrality† | 0.942 |
| | | | Length Betweenness ↔ Endpoint Betweenness† | 0.942 |
| | | | Length Betweenness ↔ K-Betweenness† | 0.948 |
| | | | Linear Betweenness † ↔ Proximal (S) Betweenness | 0.974 |
| | | | Linear Betweenness ↔ Proximal (T) Betweenness | 0.994 |
| | | | Linear Betweenness ↔ Stress Centrality | 0.932 |
| | | | Linear Betweenness ↔ K-Betweenness | 0.942 |
| | | | Linear Betweenness ↔ Load Centrality† | 0.937 |
| | | | Linear Betweenness ↔ Endpoint Betweenness | 0.937 |
| | | | Proximal (S) Betweenness ↔ Proximal (T) Betweenness | 0.978 |
| | | | Proximal (S) Betweenness† ↔ Stress Centrality | 0.913 |
| | | | Proximal (S) Betweenness ↔ Load Centrality† | 0.919 |
| | | | Proximal (S) Betweenness ↔ Endpoint Betweenness | 0.919 |
| | | | Proximal (S) Betweenness ↔ K-Betweenness | 0.923 |
| | | | Proximal (T) Betweenness ↔ Stress Centrality | 0.934 |
| | | | Proximal (T) Betweenness ↔ Load Centrality† | 0.939 |
| | | | Proximal (T) Betweenness ↔ Endpoint Betweenness | 0.939 |
| | | | Proximal (T) Betweenness ↔ K-Betweenness | 0.944 |
| | | | Stress Centrality ↔ Load Centrality† | 0.983 |
| | | | Stress Centrality ↔ Endpoint Betweenness | 0.983 |
| | | | Stress Centrality ↔ K-Betweenness | 0.988 |
| | | | Load Centrality† ↔ Endpoint Betweenness | 1.00 |
| | | | Load Centrality† ↔ K-Betweenness | 0.984 |
| | | | Endpoint Betweenness ↔ K-Betweenness | 0.984 |
| **Bottom 10 Ranked Nodes** | These measures are those who identify nodes of low importance with high accuracy of substitution. | Simple Clustering<br>Soffer's Clustering<br>Flow Betweenness<br>Betweenness Centrality<br>Length-Scaled Betweenness<br>Linear-Scaled Betweenness<br>Newman's Betweenness<br>Source Proximal Betweenness<br>Stress Centrality<br>Load Centrality<br>Endpoint Betweenness<br>Target Proximal Betweenness<br>K-Betweenness | Simple Clustering ↔ Soffer's Clustering | 0.952 |
| | | | Betweenness Centrality† ↔ Flow Betweenness | 0.938 |
| | | | Betweenness Centrality† ↔ Newman's Betweenness | 0.929 |
| | | | Flow Betweenness ↔ Newman's Betweenness† | 0.939 |
| | | | Flow Betweenness ↔ Stress Centrality† | 0.921 |
| | | | Flow Betweenness ↔ Load Centrality† | 0.939 |
| | | | Length Betweenness ↔ Linear Betweenness† | 0.923 |
| | | | Length Betweenness ↔ Proximal (S) Betweenness† | 0.934 |
| | | | Length Betweenness ↔ Proximal (T) Betweenness | 0.951 |
| | | | Linear Betweenness† ↔ Proximal (S) Betweenness | 0.946 |
| | | | Linear Betweenness ↔ Proximal (T) Betweenness | 0.957 |
| | | | Proximal (S) Betweenness ↔ Proximal (T) Betweenness | 0.979 |
| | | | Endpoint Betweenness ↔ Stress Centrality | 0.933 |
| | | | Endpoint Betweenness ↔ Load Centrality† | 0.999 |
| | | | Endpoint Betweenness ↔ K-Betweenness | 0.920 |
| | | | Stress Centrality ↔ K-Betweenness | 1.00 |

† Preferred measure. For description of preferences see section 4.6

## 4.6 Summary

This chapter reviewed the findings of this thesis, to include the describing the data and problems that were found. In addition, the results of Spearman's Rank Correlation Coefficients Kendall's Tau Correlation Coefficients and similarity tests were discussed in detail. Lastly, this chapter introduced an SNA analyst's guide for efficient and effective measures. The next Chapter will summarize the findings of this theses and offer recommendations for future research.

# 5. Conclusions

## 5.1. Overview

This chapter provides a summary of the results and analysis given in this thesis. In addition, the methods used are summarized and recommendations for future research will be offered.

## 5.2 Thesis Contribution

This thesis provides insight into to the study of social network measures. Prior to this research there had been little exploration of the measures that define the study of SNA. Through the use of statistical tests like Spearman's Rank Correlation and Kendall's Tau Correlation Coefficient it was possible to identify a group of measures that were highly correlated with one another. The density plots, overlays and paired measure scatter plots for visual conformation and the top 10 unique ranks test, allowed the ultimate goal of creating a SNA analyst's guide. The computational times also aided in the creation of this guide by determining which factors were significant within this study. The three objectives discussed in section 1.4, also shown below were met, achieving a great step forward in the SNA community.

- Provide well tested results of preference and correlation between well known SNA measures.
- Determine the efficiency and efficacy of descriptive network measures with respect to each other.
- Provide guidance for SNA analysts to choose the appropriate descriptive measure with respect to network structural properties.

## 5.3 Recommendations for Future Research

This research provides a step forward for the SNA community, but it also raised several questions about the structural properties of networks and more. In analyzing the correlations and density plots there was an obvious relationship between measures. A future study may be interested in looking into theses plots and seeing if they are consistent for a single measure over many networks, and if so can the plots indicate were an analyst should focus on. Another study related to this thesis would be to fit regression models for each pair of measures so that a formula could be used to speed up calculations. Yet another study related to this thesis would be to test the SNA analyst's guide for accuracy on known real life networks. Lastly, this study can be extended for directed or weighted networks since these types of measures tend to take more computational time.

# Bibliography


Barabasi, A., & Bonabeau, E. (2003). Scale-Free Networks. *Scientific America* , 50-59.

Barabasi, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science , 286* (509).

Batagelj, V., & Zaversnik, M. (2003). An O(m) Algorithm for Cores Decomposition of Networks.

Bonacich, P. F. (1987). Power and Centrality: A Family of Measures. *The American Journal of Sociology , 92* (5), 1170-1182.

Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks , 28*, 466-484.

Borgatti, S. (2003). The Key Player Problem. *National Academy of Sciences Press* , 241-252.

Borgatti, S., Everett, M., & Freeman, L. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology , 25* (2), 163-177.

Brandes, U. (2008). On Variants of Shortest-Path Betweenness Centrality and Their Generic Computation.

Brandes, U., & Fleischer, D. (2005). Centrality Measures Based on Current Flow. *Symp. Theoretical Aspects of Computer Science (STACS)* , pp. 533-544.

Brass, D. J. (1995). A SOCIAL NETWORK PERSPECTIVE ON HUMAN RESOURCES MANAGEMENT. *Personnel and Human Resource Management* , 39-79.

Clark, C. R. (2005). *MODELING AND ANALYSIS OF CLANDESTINE NETWORKS.* Masters Thesis, Air Force Institute of Technology.

Conover, W. J. (1980). Measures of Rank Correlation. In *Practical Nonparametric Statistics.* John Wiley & Sons, Inc.

Erdos, P., & Renyi, A. (1960). On the Evolution of Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci & Academic Press, London , 5* (17).

Estrada, E., & Hatano, N. (2008). Communicability in complex networks. *Phys. Rev. E , 77* (036111).

Freeman, L. C., Borgatti, S. P., & White, D. R. (1991). Centrality in valued graphs:A measure of betweenness based on network flow. *Social Networks , 13*, 141–154.

Freeman, L. (1979). *Centrality in social networks: conceptual clarification.* Lausanne: Elsevier Sequoia S.A.

Geffre, J. L. (2007). *A LAYERED SOCIAL AND OPERATIONAL NETWORK ANALYSIS.* Masters Thesis, Air Force Institute of Technology.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). "Exploring network structure, dynamics, and function using NetworkX". *Proceedings of the 7th Python in Science Conference (SciPy2008)*, (pp. 11-15). Pasadena.

Hamill, J. T. (2006). *ANALYSIS OF LAYERED SOCIAL NETWORKS.* Air Force Institute of Technology.

Koschützki, D., Lehmann, K. A., Peeters, L., & Richter, S. (2005). Centrality Indices. *Lecture Notes in Computer Science , 3418*, 16-61.

Labott, E., & Lister, T. (2011, May 3). *Courier Who Led U.S to Osama bin Laden's Hideout Identified.* Retrieved Feburary 14, 2012, from CNN World: http://articles.cnn.com/2011-05-03/world/bin.laden.courier_1_al-qaida-members-al-qaeda-leader-tora-bora?_s=PM:WORLD

Langville, A. N., & Meyer, C. D. (2004). A Survey of Eigenvector Methods for Web Information Retrieval.

Lind, P. G., González, M. C., & Herrmann, H. J. (2005). Cycles and clustering in bipartite networks. *Physical Review E , 72* (056127).

Liu, L., Zhu, F., Chen, C., Yan, X., Han, J., Yu, P., et al. (2010). Mining Diversity on Networks. *Database Systems for Advanced Applications* , 384-398.

Montgomery, D. C. (2009). *Design and Analysis of Experiments.* Hoboken: John Wiley & Sons, Inc.

Morris, J. F., O'Neal, J. W., & Deckro, R. F. (2011). *A Random Graph Generation Algorithm for the Analysis of Social Networks.* Air Force Institute of Technology, Future Operations Investigation Laboratory, Wright-Patterson Air Force Base.

Newman, M. E. (2005). A Measure of Betweenness Centrality Based on Random Walks. *Social Networks* .

Newman, M. E. (2010). *Networks: An Introduction.* Oxford University Press Inc.

Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *PHYSICAL REVIEW E , VOLUME 64* (016132).

Raab, J., & Milward, H. B. (2003). Dark Networks as Problems. *Journal of Public Administration Research and Theory , 13*, 413-439.

Renfro, R. S. (2001). *MODELING AND ANALYSIS OF SOCIAL NETWORKS.* Air Force Institute of Technology.

Scott, J. (1987). *Social Network Analysis a Handbook.* London: Sage Publications.

Soffer, S. N., & Vazquez, A. (2005). Network Clustering Coefficient Without Degree-Correlation Biases. *Physics Review E , 71*, 1-4.

Spearman. (2012, January 26). *Human Intelligence: Charles Spearman*. Retrieved February 17, 2012, from Indiana University: http://www.indiana.edu/~intell/spearman.shtml

Team, R. D. (2008). *R: A language and environment for statistical computing*. Retrieved Feb 2012, from R Foundation for Statistical Computing: http://www.R-project.org

van Rossum, G., & Drake, F. (2001). *Python Reference Manual.* Retrieved Feb 2012, from Python Programming Language: http://www.python.org

Wackerly, D., Mendenhall, W., & Scheaffer, R. (2008). *Mathematical Statistics with Applications.* Belmont: Thomson Learning, Inc.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge University Press.

Watts, D., & Strogatz, S. (1998). Collective Dynamics of 'Small-World' Networks. *Nature , 393*, 440-442.

Weisstein, E. W. (2005). Adjacency Matrix. Retrieved Feburary 13, 2012

## Vita

Second Lieutenant Joshua D. Guzman graduated from Floyd B. Buchanan High School in Clovis, California in June 2005. He entered undergraduate studies at California State University, Fresno, California where he graduated with a Bachelor of Arts degree in Mathematics in June 2010. He received his commission through Detachment 035 AFROTC and was accepted in the Graduate School of Engineering and Management, Air Force Institute of Technology, Ohio for his first assignment. Upon graduation, he will be assigned to Kirtland AFB, New Mexico where he will work for the Air Force Test and Evaluation Center.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 074-0188*



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From – To)* |
|---|---|---|
| 22-03-2012 | **Master's Thesis** | Sep 2010 - Mar 2012 |

**4. TITLE AND SUBTITLE**
**Analysis of Social Network Measures With Respect to Sstructural Properties of Networks**

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Guzman, Joshua D, Second Lieutenant, USAF

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)**
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 642
WPAFB OH 45433-7765

**8. PERFORMING ORGANIZATION REPORT NUMBER**

AFIT/OR/MS/ENS/12-12

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
NASIC/GTRB
Attn: Mr. Robert K. Mussen
4180 Watson Way DSN: 787-2895
WPAFB OH 45433-7765 e-mail: Robert.Mussen@wpafb.af.mil

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
DISTRIBUTION STATEMENT A: APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Social Network Analysis (SNA), the study of social interactions within a group, spans many different fields of study, ranging from psychology to biology to information sciences. Over the past half century, many analysts outside of the social science field have taken the SNA concepts and theories and have applied them to an array of networks in hopes to formulate mathematical descriptions of the relations within the network of interest. More than 50 measures of networks have been identified across these fields; however, little research has examined the findings of these measures for possible relationships. This thesis tests widely accepted SNA measures for correlation and redundancies with respect to the most excepted network structural properties; size, clustering coefficient and scale-free parameter. The goal of this thesis is to investigate the SNA measures' ability to discriminate and identify different actors in a network. As a result this study not only identifies high correlation amongst many of the measures, it also aids analysts in identifying which measure best suits a network with specific structural properties and its efficiency for a given analysis goal.

**15. SUBJECT TERMS**
Social Network Analysis, Measures, Analyst, Guidance, Guide, Centrality, Betweenness, Clustering

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Dr. Richard F. Deckro, AFIT/ENS |
| U | U | U | UU | 79 | **19b. TELEPHONE NUMBER** *(Include area code)* (937) 785-3636, ext 4325; e-mail: Richard.Deckro@afit.edu |

**Standard Form 298 (Rev. 8-98)**
Prescribed by ANSI Std. Z39-18