UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

FELIPE GRANDO

On the Analysis of Centrality Measures for Complex and Social Networks

Dissertation submitted as a requirement for the degree of Master of Computer Science.

Advisor: Prof. Luís da Cunha Lamb

Porto Alegre
2015

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Grando, Felipe

On the Analysis of Centrality Measures for Complex and Social Networks / Felipe Grando. – 2015.

66 f. il.

Orientador: Prof. Dr. Luís da Cunha Lamb.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2015.

1. Introduction. 2. Complex Networks. 3. Centrality Measures. 4. Experimental Methodology. 5. Experimental Result and Discussion. 6. Conclusions and Further Work.

ABSTRACT

Over the last years, centrality measures have gained importance within complex and social networks research, e.g., as predictors of behavior, identification of powerful and influential elements, detection of critical spots in communication networks and in transmission of diseases. New measures have been created and old ones reinvented, but few have been proposed to understand the relation among measures as well as between measures and other structural properties of the networks. Our research analyzes and studies these relations with the objective of providing a guide to the application of existing centrality measures for new environments and new purposes. We shall also present evidence that the measures known as *Walk Betweenness*, *Information*, *Eigenvector* and *Betweenness* are substantially better than other metrics in distinguishing vertices in a network by their structural properties. Furthermore, we provide evidence that each metric performs better with respect to distinct kinds of networks. In addition, we show that most metrics present a high level of redundancy (over 0.8 correlation) and its simultaneous use, in most cases, is fruitless. The results achieved in our research reinforce the idea that to use centrality measures properly, knowledge about their underlying properties and behavior is valuable, as we show in this dissertation.

Keywords: Centrality Measures. Complex Networks. Social Networks.

RESUMO

Recentemente, as medidas de centralidade ganharam relevância nas pesquisas com redes complexas e redes sociais, atuando como preditores comportamentais, na identificação de elementos de poder e influência, na detecção de pontos estratégicos para a comunicação e para a transmissão de doenças. Novas métricas foram criadas e outras reformuladas, mas pouco tem sido feito para que se entenda a relação existente entre as diferentes medidas de centralidades, assim como sua relação com outras propriedades estruturais das redes em que elas são frequentemente aplicadas. Nossa pesquisa visa analisar e estudar essas relações para que sirvam de guia na aplicação das medidas de centralidade existentes em novos contextos e aplicações. Nós apresentamos também evidencias que indicam um desempenho superior das medidas conhecidas como *Walk Betweenness*, *Information*, *Eigenvector* and *Betweenness* na distinção de vértices das redes somente pelas suas características estruturais. Ainda, nós propiciamos detalhes sobre o desempenho distinto de cada métrica de acordo com o tipo de rede em que se trabalha. Adicionalmente, mostramos que várias das medidas de centralidade apresentam um alto nível de redundância e concordância entre si (com correlação superior a 0,8). Um forte indício que o uso simultâneo de várias métricas é improdutivo ou pouco eficaz. Os resultados da nossa pesquisa reforçam a ideia de que para usar apropriadamente as medidades de centralidade é de extrema importância que se saiba mais sobre o comportamento e propriedades das mesmas, fato que salientamos nessa dissertação.

Palavras-chave: Medidas de Centralidade. Redes Complexas. Redes Sociais.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVIATIONS AND ACRONYMS

C_b	Betweenness Centrality
C_c	Closeness Centrality
C_d	Degree Centrality
C_e	Eigenvector Centrality
C_i	Information Centrality
C_s	Subgraph Centrality
C_w	Walk-based Betweenness Centrality
C_x	Eccentricity as Centrality
M_{cs}	Community Structure Model
M_{er}	Erdős and Rényi Model
M_{gr}	Geographical Model
M_{sf}	Scale-free Model
M_{sw}	Small-world Model
N_{ni}	Non-isomorphic Networks

SUMARY

1 INTRODUCTION

Human beings naturally present social behavior as an important aspect that increases their adaptation to the environment. The association of any kind and purpose between several individuals form and characterize a social network. The construction of an optimal social network is relevant and pursued by every social being to accomplish any complex task, for instance, food manufacture (chain of production), reproduction (relationship networks) and knowledge production (scientific contributions).

The study of these networks are subject of complex networks analysis, which also includes networks composed by machine and any other kinds of elements. Their analysis are of fundamental importance not only to understand human behavior, but also to improve their effectiveness in several tasks and to apply it to artificial intelligence domains where individuals are simulated by agents and machines (BARABÁSI, 2002).

The fast development of the internet, the web and of powerful computers makes it possible to retrieve and analyze large amounts of data about social behavior for many application areas. This creates several new fields of study of complex and social networks for many objectives and opens new possibility of analysis in large scale with more complex algorithms and techniques (EASLEY and KLEINBERG, 2010).

The measurement of numerous characteristics and properties of networks is one of the most important tools for their analysis and understanding. However, they are most of the time related to graph theory and linear algebra. There are plenty of metrics for distinct and similar purposes that are applied in many contexts and studies of the area, such as, measures related with distance, clustering and cycles, degree distribution, vertex types, entropy, spectral and hierarchical measures, fractal dimensionality, correlations and centrality (COSTA et al., 2008).

Every one of them have their applications and importance for complex and social networks analysis but centrality measures are one of the most widely used metrics nowadays as a good alternative to uncover intrinsic behavior contained in networks structural properties. They are also especially useful in the context of social behavior and simulation, and they will be the focus of our work.

The concept of central position is relative to a particular context. For example, it can mean the center of a circle (perfectly defined by a formula) or the most important person within an organization or group (most of the time it is hard to identify with precision). As one can imagine, its meaning goes from a physical position to an abstract idea, such as, power,

influence, or responsibility. Nevertheless, it is important to identify the most central element for many applications and areas with different objectives in mind.

Centrality measures look to quantify how much central is an element, considering a defined environment and objective. Diverse environments and objectives have specific centrality measures, but whenever the central concept is abstract, it is generally undefined which centrality measure is more adequate to use (THILAGAM, 2010). Furthermore, many metrics can be used for the same application, which leads to distinct results that are difficult to compare.

Complex networks are an example where centrality plays an important role to identify powerful, critical or relevant elements, but it is difficult to detect which measure is adequate, given the large number of objectives and structural properties found in complex networks studies (EVERETT and BORGATTI, 2005). That are the several reasons for the intense expansion in studies about centrality measures in the last years.

Back in the 1940s, Bavelas (1948) introduced the centrality idea within the social context as an important element. He studied communication among groups and related centrality with influential individuals. Later, the idea of centrality was used by many authors in several studies:

- to identify prominence differences of individuals in distinct social networks determined by their position in their local network or community network (ADAH et al., 2013);
- in the analysis of individuals vulnerability to disease and infectivity in disease transmission networks (BELL et al., 1999);
- in the evaluation of vulnerable spots in communication networks (BORBA, 2013);
- in the identification of key players sets in social networks (BORGATTI, 2006; ORTIZ-ARROYO, 2010);
- in the reinforcement of communication networks to prevent attacks (CUNNINGHAM, 1985);
- in the mapping of actors social networks (DANOWSKI and CEPELA, 2010);
- in the visual analysis of networks (BRANDES et al., 2003; CORREA, 2011; DWYER et al., 2006);
- to identify the relationship between academic employment and departments prestige (HEVENSTONE, 2008);
- in network analysis of USA air transportation network (HUA et al., 2010);

- in the identification of high-status elements on knowledge and scientific networks (KAZA and CHEN, 2010);
- in the selection of influencers for optimal spread of viral marketing (KISS and BICHLER, 2008);
- in biological network analysis (KOSCHUTZKI and SCHREIBER, 2004);
- in the analysis of scientific collaboration networks (NEWMAN, 2001);
- to find community structures in networks (NEWMAN and GIRVAN, 2004);
- to identify the impact of structural location to individual and collective performance in social problem solving systems (NOBLE et al., 2015);
- to rank pages in the web (PAGE et al., 1999);
- in targeting optimization of intrusion detection systems in social networks (PUZIS et al., 2010);
- in the measurement of individual's connectedness and reachability in a network (VALENTE and FOREMAN, 1998);
- in journal impact analysis in coauthorship networks (YAN and DING, 2009).

Network measurements are therefore essential as a direct or subsidiary resource in many network investigations, including representation, characterization, classification and modeling (COSTA et al., 2008).

This expansion caused the creation of many new centrality measures and improved the algorithms used in its computation, but few were done to understand their actual relations between the network structure and among themselves. Some measures appear to lack theoretical focus because they may be applied in diverse contexts. For example, although networks terms such as centrality, prestige and power have coherent measurement definitions, theoretical definitions of these terms tend to be vague.

Some efforts have been done to classify the centrality measures and verify their actual usefulness or precision. This line of research started with Freeman (1978/79) and Freeman et al. (1979/80), who organized centrality measures in three main groups: *degree* (representing vertex visibility), *betweenness* (representing vertex control of communication), and *closeness* (representing vertex independency). In addition, they separated the concept of point centrality, which focuses on each element of a network, and network or graph centrality, which aims to measure how central is an entire network. He also made experiments to verify how close the metrics values are to real case environments where the correct values or at least a rank is known

earlier. Their experiments achieved many promising results by indicating that the centrality measures are highly correlated with real environments variables.

More recently, Borgatti (2006) highlighted the importance of traffic flow to determine the significance of a vertex. For him, the selection of centrality measures must take into consideration the kind of network flow to guarantee accuracy. Furthermore, he reinforces that most centrality measures do not specify characteristics about network structure and network flow, both relevant for its use. In addition, Borgatti and Everett (2006) proposed a new classification for centrality measures based on the concepts of nodal involvement (radial or medial), property of the walk assessed (volume or length) and type of walk considered (random or guided).

Our objectives do not involve creating a new centrality measure or discovering which measure is more accurate in a determined environment. These elements have already received a lot of attention from many researchers. Moreover, only point centrality measures for social networks are considered thus present lack of theoretical foundation and are gaining importance in many applications.

Our main objective is to improve the selection of new centrality measures for a problem, making use of network structure properties and/or the information about previously applied centrality measures to that problem.

We will accomplish that through the following tasks:

- selecting the main and most known centrality measures applied to social networks;
- comparing experimentally different centrality measures, and classifing them using a similarity criterion;
- relating network properties such as density (number of edges, diameter, maximum distance mean, clustering coefficient), connectivity (minimum degree, maximum degree, mean degree) and size (number of vertices), with each centrality measure;
- simplifying the variety of centrality measures by using similar experimental results to show which measures are significantly divergent from the others and which ones are best applied in different kinds of networks.

Both the characterization and classification of natural and human-made structures using complex networks imply the same important question of *how to choose the most appropriate metrics and evaluations of structural properties*. While such a choice should reflect the specific interests and application, it is unfortunate that there are no general model or formal procedure for identifying the best measurements. In addition, there is an unlimited set of metrics

variations, and they are often correlated, implying redundancy in many situations. Ultimately, one has to rely on his knowledge of the problem and available measurements in order to select a suitable set of features to be considered. For such reasons, it is of paramount importance to have a good knowledge not only of the most representative measurements, but also of their respective properties and interpretation (CORREA and MA, 2011; BRANDES et al., 2003).

Currently, the fast pace of developments of more sophisticated measures and new results reported in this very dynamic area makes it particularly difficult to follow and to organize the existing measurements (COSTA et al. 2008).

Bolland (1988), Nakao (1990), Friedkin (1991), Costenbader and Valente (2003), Goh et al. (2003), Koschützki and Schreiber (2004), Zemljič and Hlebec (2005), Borgatti et al. (2006), Butts (2006) are some of the most relevant work about centrality measures relation's with structural properties of the network and among themselves. Their work differs from ours because one or more of the following reasons: they considered only graph centrality (our work focus on vertex centrality), their experiments were restricted to few real networks (our experiments use thousands of random networks with distinct properties), they studied few centrality measures (we use not only the classical measures but also many of the newer measures), their analysis or objectives are distinct from ours. Further comparisons with their work will be presented in the coming Sections.

The importance of our work comes from the fact that it is useful to predict the behavior of a measure or its performance before applying it onto a network, using for that, information about both measure and network properties. This helps to select among several measures the one that will be most relevant for the application analyzed, preventing the need to try over all the measures, an increasing number nowadays, and yet achieve the desired objective.

Second, knowing in what way centrality measures relate to each other provides an important idea of how similar they are, creating an opportunity to use other close measures when it is known in advance at least one that is adapted for a determined problem. This is especially relevant if one wants more precision or accuracy and to achieve that, wants to select other measures like to the one already used. Applying similar measures on the same network can produce a small, but relevant, increase on precision, desired when dealing with hard objectives and especially whenever you already have reasonable but still not good enough results with a known measure. Another advantage can be achieved when it is possible to change a complex measure for a simpler one, the simpler are always more easily interpreted and faster to calculate.

The first part of this work (Section 2) explains the characteristics of general complex models and important studies done in the area. It also presents five complex network models that will be later used to generate synthetic networks for our experimental analysis.

Section 3 contextualizes and defines centrality measures, presenting eight of the most relevant metrics applied in many research applications. These metrics will be the focus of our research and analysis.

Section 4 explains the procedures of our experimental methodology in details, justifying our choices of steps and parameters with arguments retrieved from important references and from trial experimental tests.

The results of our experiments and their associated discussion and analysis are in Section 5, where we present all the important information retrieved from our experiments and discuss their implications in real world domains, exemplifying their importance and contribution for the area of complex and social networks.

The last part of this work (Section 6) concludes with the main results and contributions of our work, and we also discuss possible further investigations and the next steps of complementary research.

2 COMPLEX NETWORKS

A complex network is characterized for not being regular i.e., it does not present a clear and homogenous pattern of connections and its elements are usually unique in its environment. Complex networks appear frequently in various technological (internet, world wide web), social (scientific collaboration networks, lexicon or semantic networks, friendship networks, business relationships) and biological (neural networks, food webs, metabolic networks, interaction between proteins) domains. They are usually mapped (modeled, represented) by graphs where each vertex represents a node of the network (person, animal, company, object, element) and each edge represents any arbitrary relationship between entities.

Although graph theory is a well-established and developed area in mathematics and theoretical computer science, many of the recent developments in complex networks have taken place in areas such as sociology, biology and physics. The recent studies of complex networks where mainly supported by the availability of high performance computers and large data collections, providing important results and increasing the interest in this area. Current interest has focused not only on applying the developed concepts to many real data and situations, but also on studying the dynamical evolution of network topology (COSTA et al., 2008).

Stephenson and Zelen (1989) and Hevenstone (2008) reinforce the idea saying that networks are implicit in a wide range of social phenomena and that the overall structure of a network has consequences not only for individual members, but also for the entire group, extending well beyond individual behaviors and social roles. Assessing the quality of relations between entities and understanding connections patterns has generated much interest and research in various disciplines (KEARNS, 2012).

In the early history of the area, technical ideas often raced ahead of applications leading to a criticism that network analysis provided nothing more than a superfluous language that served to recognize what was patently obvious. Such criticism failed to recognize that these techniques yielded results that added substantially to our understanding of social and cultural processes and could not have been obtained by simple common sense notions for large and complex networks (STEPHENSON and ZELEN, 1989). Today, complex network analysis is viewed as fundamental to the understanding of human, social and economic activities and relationships (EASLEY and KLEINBERG, 2010; KEARNS, 2012).

Before we advance in the subject and explain the details of the complex networks models we will revise briefly some important mathematical definitions of graph theory, used also in the definitions of centrality measures and in the experimental analysis in later sections.

2.1 Mathematical Definitions

A graph consists of a set of vertices (points, vertices) and a set of edges (lines, links) connecting pairs of vertices. The edges can be valued (weighted) or unitary (unweighted), directed or undirected, but only unitary and undirected edges will be considered in the definitions below because our experiments use only undirected and unweighted graphs. The number of vertices of a graph will be represented by n and the number of edges by m.

Whenever two vertices are connected by an edge, they are adjacent, and the number of vertices that one is adjacent to is called the degree of that vertex.

A path between two vertices is a sequence of edges starting from one vertex to the other (possibly with many other vertices amid the way). If there exists a path between all pairs of vertices in the graph, the graph is considered connected. The number of edges in a path is a definition of distance between two vertices and the smallest path between two vertices is called geodesic.

There are a great number of useful structural properties of graphs, the ones used in this work are: minimum degree (minimum degree value of the graph), maximum degree (maximum degree value of the graph), mean degree (sum of all vertices degree values divided by the number of vertices), diameter (maximum geodesic distance between all pairs of vertices of the graph), maximum distance mean (sum of geodesics distances between all pairs of vertices divided by the number of vertices), clustering coefficient (three times the number of closed triangles divided by the number of connected triples) and density (number of edges divided by the maximum number of edges possible).

Recent studies of complex networks have shown that these networks have some interesting properties, such as, high clustering coefficients, “small-world” effect, scale-free effect and community organization. Such studies proposed distinct models of randomly generated networks mapping each of these characteristics with the objective to understand how real networks are organized (NEWMAN, 2003). These models are presented and briefly explained below where sample pictures generated by each model algorithm are also presented. The sample pictures of the network graphs were generated with a free software called Gephi where force atlas 2 (a simpler and faster version of force atlas) was used to organize the layout of the vertices. Force atlas layout algorithm simulates a gravitational force using vertices degree

value and their connections (attracting the ones which a vertex is connected to and, at the same time, repelling those which it is not). The resulting layout gives a perceptible idea of community structural organization and visually identifies their presence in a network.

2.2 Complex Networks Models

Each complex network (or class of networks) presents specific topological features, which characterizes their connectivity and highly influences the dynamics of processes executed on the network. The analysis, discrimination, and synthesis of complex networks therefore rely on the use of measurements capable of expressing the most topological features (COSTA et al., 2008).

Complex networks research can be conceptualized as lying at the intersection between graph theory and statistical mechanics, which endows it with a truly multidisciplinary nature. The main reason that complex networks became a focus of attention recently is the discovery that real data networks involve community structure, power-law degree distributions and hubs, among other structural features not explained by uniformly random connectivity (COSTA et al., 2008).

Some relevant works in the area include: Erdős and Rényi (1959), Watts and Strogatz (1998), Barabási and Albert (1999), Newman and Park (2003). Their main purpose was to create models that are capable of generating synthetic networks with pre-defined sizes and properties, which simulate the main characteristics presented by real social networks structures while keeping random elements that allows the creation of unique networks.

Their ideas will be explained in detail in the following Sections and will be used in our further experiments. The algorithms/procedures proposed by each author for the generation of networks with determined properties were implemented and tested by us in Java programming language. Later, these algorithms are used in our research to generate thousands of sample synthetic networks, useful to simulate numerous real networks samples and increasing the analysis statistical relevance. We show also that these models are substantial attributes for the appropriate selection and prediction of centrality measures behavior before the need of their actual computation.

We start presenting the more ancient models and finalize with the newer ones. This is important because many of the later models use ideas already discussed by the other models and some of them can be viewed as extensions to the previous model properties.

2.2.1 Random Graphs of Erdős and Rényi (M_{er})

That is considered the simplest model for complex networks, introduced to generate random graphs (ERDŐS AND RÉNYI, 1959). This model defines a fixed number of vertices *n* and a probability *p* of connection between each pair of vertices. The probability *p* is the same for all pairs of vertices.

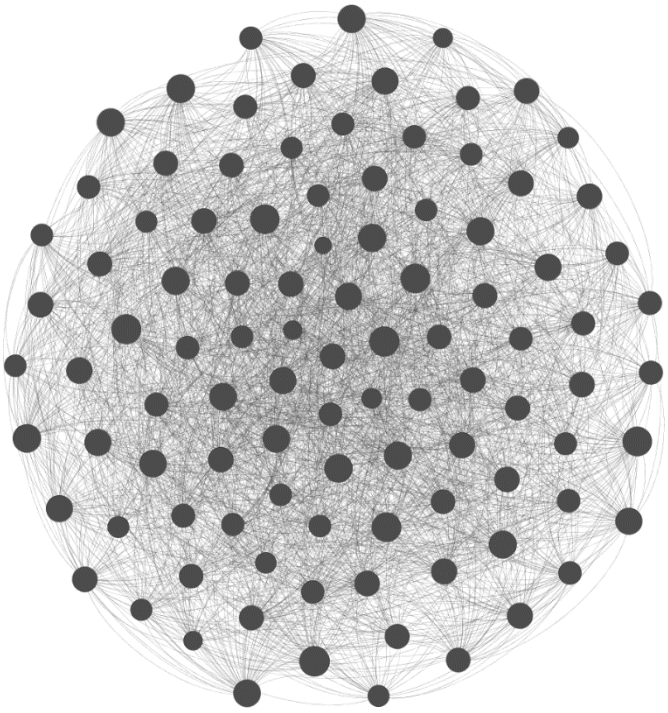
Their model does not fit well social networks structural properties, but some artificial agent networks are assembled in a very similar way.

The networks generated by the model are mainly characterized by an expected average degree of *p(n-1)* and by a degree distribution similar to a Poisson distribution.

For our experiments we used $p = \{0.1, 0.3, 0.5\}$, the higher is *p*, higher is the mean degree, clustering coefficient and density, and lower is the diameter. Higher values of *p* were not sampled because networks with such density are rare in real applications.

Figure 1 presents an example of network generated by this model with our algorithm (vertices size represents their degree value). We can see that this model generates a homogenous distribution of connections and presents no formation of distinguishable community structures or any other regular pattern structure format.

Figure 1 – M_{er} sample with 100 vertices and $p = 0.5$



Source: author (picture generated with Gephi)

2.2.2 Small-World Model of Watts and Stogatz (M_{sw})

Complex real networks usually present the Small-World properties, also known as the so-called “six degrees of separation”, where most vertices can be reached with short paths. In addition, these networks show a large number of small cycles, especially the ones of size three which represents the idea of the closely related communities (a close group of friends in a social network, for example).

Watts and Strogatz (1998) proposed a model to generate random networks with both properties, starting with a ring of connected vertices, each one adjacent with the k nearest neighbors. Then with a probability p, each edge can be random (uniformly) reassigned for any available position. This relinking method, with an intermediate or small p, will create paths among distant vertices while it keeps a high clustering coefficient among close neighbors.

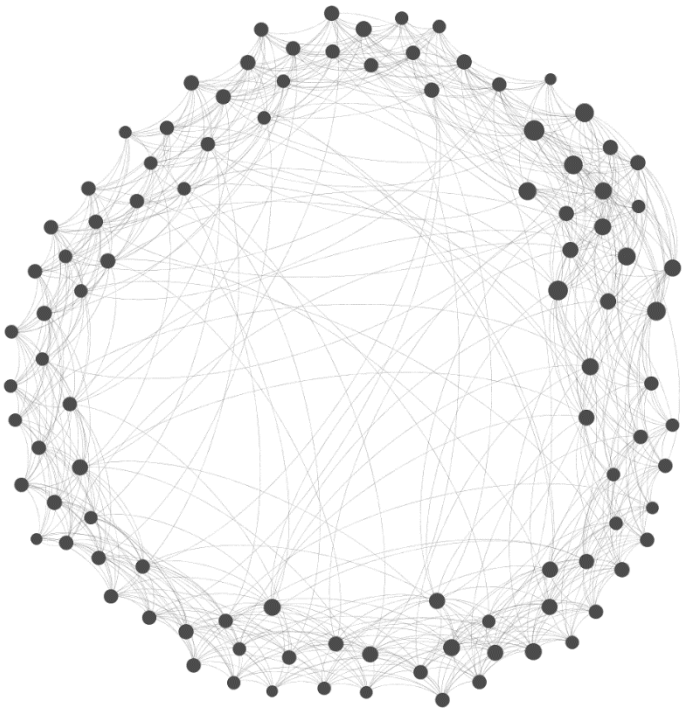
We tried every combination of $k = \{4, 8, 16\}$ with $p = \{0.1, 0.3, 0.5\}$, resulting in nine distinct configurations for the experiments. The higher is k, higher is the mean degree, clustering and density, although diameter decreases. In addition, the higher is p, lower is the clustering coefficient and diameter.

Higher values of p distort the properties of the model, turning it very similar to the random graphs of Erdős and Rényi (1959) and that is why we do not sample p values above 0.5. At the same time, k values could not be larger, relative to the size of the networks, to keep a faithful representation of real application networks behavior and properties where a large number of a closely fully connected community is pretty rare.

Many actual real networks present Watts and Strogatz model’s properties but they are generally not as homogenous and simple as the proposed generative model suggests. Still, it can represent very well many structural properties of many real social networks and it is very distinguishable from the other complex network models.

Figure 2 presents an example of network generated by this model with our algorithm (vertices size represents their degree value). It presents the ring layout of vertices forming distance related communities, and also the long range edges relinked in the generation process, which are responsible to lower the diameter of the network giving the small-world effect of the model while maintaining the high clustered groups of vertices. It is also visible that the degree distribution among vertices are closely homogenous due the fact that the random and uniform probability of edge relinking retains mostly the initial fully homogenous degree’s distribution.

Figure 2 – M_{sw} sample with 100 vertices, p = 0.1 and k = 16



Source: author (picture generated with Gephi)

2.2.3 Scale-free Networks of Barabási and Albert (M_{sf})

The analysis of large networks data show that degree follows a scale-free power-law distribution. Barabási and Albert (1999) explained this fact by the facts that networks expand continuously by adding new vertices and, that these new vertices attach preferentially to vertices already well-connected (with higher degree).

The model proposed by them with both characteristics starts with a k number of fully connected vertices and keeps adding new vertices with k connections, defined by a preferential attachment formula. The probability of a vertex p_i to receive a new connection takes into consideration the degree d of the vertex divided by the sum over the degree of all vertices, this formula (presented below) gives higher chance for a vertex with higher degree to receive new connections than for a vertex with lower degre.

$$p_i = \frac{d_i}{\sum_j^n d_j}$$

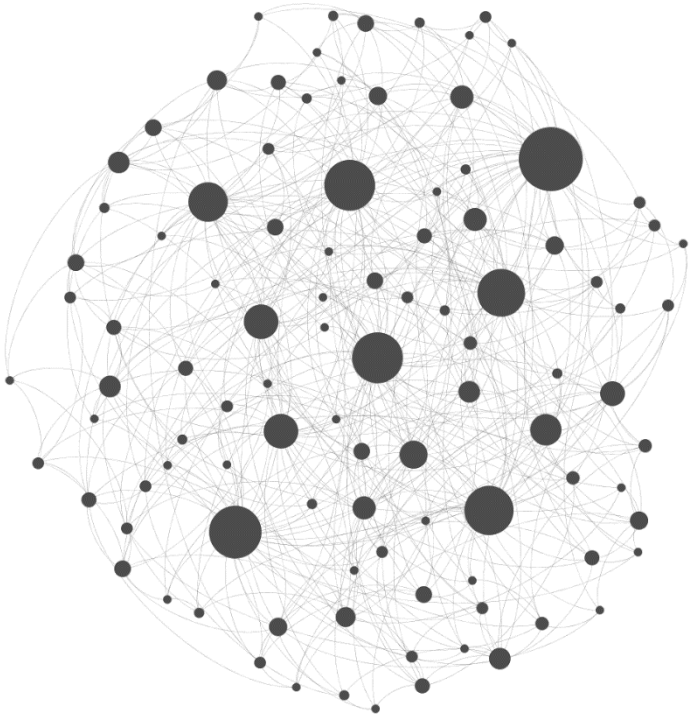
The parameter k = {2, 3, 5} was applied to the experiments, the range was chosen to keep the properties of the model which are altered with increasing values of k. With higher k, the mean degree, clustering coefficient and density is higher, while the diameter sinks.

Barabási (2002) presents evidence that the web and the internet are examples of networks represented by the model with high reliability. He also explains that this fact is related to the similarity of the generative model process with the network evolving reality, i.e., both model and real networks, such as the web, involve an analogous growing methodology of preferential attachment.

Figure 3 presents an example of network generated by this model with our algorithm (vertices size represents their degree value). It clearly presents the main characteristic of this model: a strong difference in vertices degree with few high degree vertices, also called hubs, and a large amount of vertices with low degree which represents the power-law degree distribution created by the preferential attachment formula during the generation of the network. This model also does not show a clear division into communities or groups of elements, being very similar in network structure (organization) of the networks generated by the model of Erdős and Rényi (1959) despite, of course, the noticeable difference in the degree's distribution.

This model is also the one presented in this work that shows the largest degree value for a given vertex and the largest difference between the lowest degree vertex and the highest degree vertex.

Figure 3 – M_{sf} sample with 100 vertices and $k = 5$



Source: author (picture generated with Gephi)

2.2.4 Networks with Community Structure of Newman and Park (M_{cs})

More recently, Newman and Park (2003) analyzed several social networks and discovered that these networks are formed by communities where each vertex has many connections with vertices inside the same community and less with vertices of others communities. In addition, they discovered that in these networks, vertices with high degree tend to be connected with vertices also with high degree and vertices with small degree connected with vertices with small degree (disassortativity behavior).

They proposed a model to generate random networks with these properties, their model starts defining *c* communities and an uneven distribution of vertices for each community that need to represent distinct sizes of groups to create a disassortativity behavior, also, each vertex can be assigned to more than one community. Then, each vertex has a fixed high probability *p* of being connected to each element of its communities and zero probability to be connected with vertices that do not share a community.

The number of communities was based on the size of network *n*, so *c* = {n/10, n/20, n/50} was used for experiments. Each vertex has 10% chance to belong to each community. The probability *p* = {0.5, 0.7} is used to define connections between members of the same community. Community sizes followed an arithmetic progression. The first community has $a_1 = \frac{100}{2c}$ percent of elements and with ratio *r*, defined by the equation below. The other communities' percentiles are calculated (the sum of all percentiles is a hundred percent) by:

$$r = \frac{200}{c^2 - c} - \frac{2a_1}{c - 1}$$

All configurations of *c* and *p* are combined for the experiments, resulting in six total combinations. The higher is *p* and the lower is *c*, higher is the mean degree, clustering and density, although diameter decreases. All parameters range took into consideration sample trials where more combinations of parameters were tested. To select the range of the parameters used in our experiments we analyzed the consistency of model's properties with real networks. That process was necessary because no formal guidance was provided by Newman (proposer of the generative model) to select the best-suited parameters values in such a way that the models characteristics were preserved.

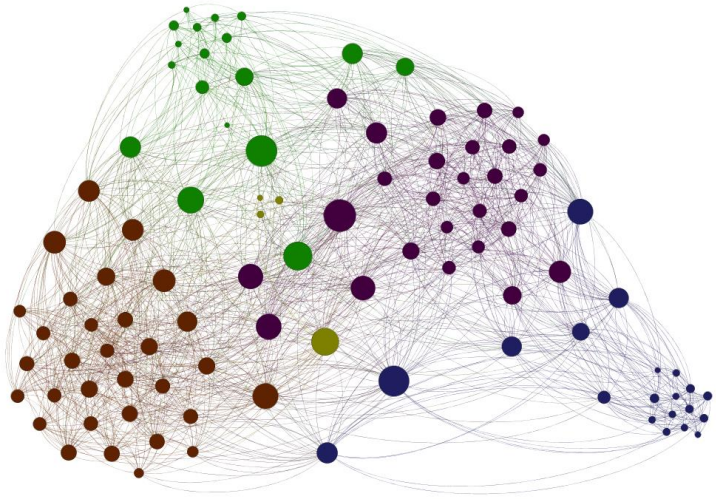
The community structure model also presents many of the characteristics of the small-world model proposed by Watts and Strogatz (1998), such as low diameter and a high number of small cycles. Therefore, it can be viewed as an extension of the small-world generative model.

Figure 4 presents an example of network generated by this model with our algorithm (vertices size represents their degree value, and colors represents the dominant community of

each vertex). Each vertex can belong to multiple communities, but the dominant community is the one assigned for that vertex on the first step of the algorithm. Later, some vertices will be connected to members of multiple communities representing the hubs (larger degree) of the network.

The main characteristics of this model are visible in the Figure 4: a clearer separation between communities, communities of distinct sizes and disassortativity behavior (vertices belonging to the same dominant community have similar degree). The vertices that are connected to multiple communities and play a role of hub are also the ones that deviates most of the disassortativity behavior. Another noticeable point is the fact that the difference between vertices degree in different communities becomes increasingly larger with network size and, at the same time, smaller with an increasing number of communities.

Figure 4 – M_{cs} sample with 100 vertices, 5 communities and $p = 0.7$



Source: author (picture generated with Gephi)

2.2.5 Geographical Models (M_{gr})

Complex networks are generally considered as lying in an abstract space, where the position of vertices has no particular meaning, but several kinds of networks model's physical interactions and so, in these latter networks, the positions of vertices characterize a higher probability for interaction with close neighbors than with distant ones.

A way to model this behavior is presented by Costa et al. (2008) where distances s_{ij} between vertices *i* and *j* are defined. Then the probability of vertices *i* and *j* are connected can be defined to decay with distance as shown in the formula below.

$$p_{ij} = e^{-s_{ij}}$$

Exploratory results showed that using the equation above the chance of creating a connected network is small, so instead of *e* we used a value $k = \{1.2, 1.5, 2\}$ in our experiments. That is because when *k* increases, the diameter grows and the mean degree, clustering coefficient and density are lowered.

The distance s_{ij} between vertices *i* and *j* was calculated with the formula below (representing a square grid of vertices). The formula used by us is based on the Manhattan distance metric.

$$s_{ij} = \left|\left\lfloor \frac{i}{\sqrt{n}} \right\rfloor - \left\lfloor \frac{j}{\sqrt{n}} \right\rfloor\right| + \left|(i \bmod \sqrt{n}) - (j \bmod \sqrt{n})\right|$$

The choice of parameters for this model is quite arbitrary, because each environment will have a specific configuration of parameters that suits better the reality. The choice of parameters can also produce networks with similar properties of the ones generated by the scale-free, small-world and community structure models.

Our choice of parameters for our experiments was biased by roads configurations inside a city (represented fairly well by a square grid of crossroads and a planar Euclidian distance formula). We also expect that our choice of parameters models equally well other bidimensional environments where the Euclidian distance makes sense. It does so because parameter k can simulate distinct arrangements of vertices in a way that the initial square grid organization are substantially and randomly modified to more complex configurations.

Figure 5 presents an example of network generated by this model with our algorithm (vertices size represents their degree value).

Differently from the other complex network models, as shown in Figure 5, the geographic model is sparser (restriction posed by the two dimensional space used in our experiments) and presents a more positional community structure with peripheral and central groups of vertices. Also, it is noticeable that the initial square grid formation of vertices (formation used in our implementation) is hugely modified in the process.

Figure 5 – M_{gr} sample with 100 vertices and k = 2

Source: author (picture generated with Gephi)

Apart from these characteristics, clearly visible in Figures 1 to 5, another important properties are representative for each model and were calculated during the generation of the networks used in our experimental analysis.

Table 1 and Table 2 summarize the means and standard deviations of the main properties of a hundred samples of each combination of parameters for each complex network model. Table 1 presents the data about the networks generated with 100 vertices while Table 2 presents the data of networks with 500 vertices. The standard deviation is not present in the cells where the mean value is equal to the absolute value of all networks generated, standard deviation in these cases are obviously zero. For example: the small-world model fixes the degree of the vertices on the first step of the generation method with parameter k, the relinking step does not change the number of edges present in the network, so the mean degree and the mean density remain unaltered while diameter and clustering coefficient varies.

There are cells that present zero standard deviation, but this happen when we rounded off the value, reducing the precision to a couple decimal cases to better fit the table length.

Table 1 – Mean and Standard Deviation of Networks with 100 vertices

Model	Degree	Diameter	Clustering	Density
M_{cs} (p = 0.5; c = n/10)	18.62±1.62	3.85±0.48	0.37±0.02	0.19±0.02
M_{cs} (p = 0.5; c = n/20)	21.04±1.85	3.55±0.52	0.41±0.02	0.21±0.02
M_{cs} (p = 0.5; c = n/50)	34.42±2.20	3.05±0.22	0.48±0.01	0.35±0.02
M_{cs} (p = 0.7; c = n/10)	25.52±2.36	3.12±0.33	0.49±0.02	0.26±0.02
M_{cs} (p = 0.7; c = n/20)	29.10±2.46	3.02±0.14	0.56±0.02	0.29±0.02
M_{cs} (p = 0.7; c = n/50)	48.79±2.80	2.58±0.52	0.66±0.01	0.49±0.03
M_{er} (p = 0.1)	9.91±0.43	4.00±0.00	0.10±0.01	0.10±0.00
M_{er} (p = 0.3)	29.73±0.64	2.24±0.43	0.30±0.01	0.30±0.01
M_{er} (p = 0.5)	49.50±0.69	2.00±0.00	0.50±0.01	0.50±0.01
M_{gr} (k = 1.2)	34.43±0.57	2.43±0.50	0.38±0.01	0.35±0.01
M_{gr} (k = 1.5)	13.50±0.41	4.00±0.00	0.22±0.01	0.14±0.00
M_{gr} (k = 2)	5.76±0.29	6.64±0.70	0.15±0.02	0.06±0.00
M_{sf} (k = 2)	3.94	5.40±0.49	0.06±0.01	0.04
M_{sf} (k = 3)	5.88	4.40±0.49	0.11±0.01	0.06
M_{sf} (k = 5)	9.70	3.86±0.35	0.17±0.01	0.10
M_{sw} (p = 0.1; k = 4)	4.00	10.58±1.29	0.36±0.03	0.04
M_{sw} (p = 0.1; k = 8)	8.00	5.41±0.51	0.48±0.02	0.08
M_{sw} (p = 0.1; k = 16)	16.00	3.78±0.42	0.54±0.02	0.17
M_{sw} (p = 0.3; k = 4)	4.00	7.35±0.52	0.17±0.03	0.04
M_{sw} (p = 0.3; k = 8)	8.00	4.26±0.44	0.25±0.02	0.08
M_{sw} (p = 0.3; k = 16)	16.00	3.00±0.00	0.32±0.01	0.17
M_{sw} (p = 0.5; k = 4)	4.00	6.63±0.51	0.08±0.02	0.04
M_{sw} (p = 0.5; k = 8)	8.00	4.01±0.10	0.13±0.01	0.08
M_{sw} (p = 0.5; k = 16)	16.00	3.00±0.00	0.21±0.01	0.17

Noticeably, all the complex network models, despite the parameters values, presented a low relative diameter to the size of the networks. All other properties are affected widely by the choice of the parameters values, but these variations were expected and predicted by each models' characteristics.

It is also important to highlight that the random seed used in our algorithm is very stable. This can be easily verified by looking at the density standard deviations of Erdős and Rényi model which are close to zero with just a hundred samples. It also visible that the mean degree values were also really close to the ones expected for the model given a parameter value.

Table 2 – Mean and Standard Deviation of Networks with 500 vertices

Model	Degree	Diameter	Clustering	Density
M_{cs} (p = 0.5; c = n/10)	158.60±4.73	2.78±0.48	0.41±0.01	0.32±0.01
M_{cs} (p = 0.5; c = n/20)	113.12±3.66	3.00±0.00	0.36±0.01	0.23±0.01
M_{cs} (p = 0.5; c = n/50)	92.92±3.71	3.00±0.00	0.37±0.01	0.19±0.01
M_{cs} (p = 0.7; c = n/10)	207.07±4.79	2.29±0.46	0.51±0.01	0.41±0.01
M_{cs} (p = 0.7; c = n/20)	151.90±5.27	3.00±0.00	0.46±0.01	0.30±0.01
M_{cs} (p = 0.7; c = n/50)	128.44±5.19	3.00±0.00	0.49±0.01	0.26±0.01
M_{er} (p = 0.1)	49.93±0.41	3.00±0.00	0.10±0.00	0.10±0.00
M_{er} (p = 0.3)	149.67±0.59	2.00±0.00	0.30±0.00	0.30±0.00
M_{er} (p = 0.5)	249.57±0.71	2.00±0.00	0.50±0.00	0.50±0.00
M_{gr} (k = 1.2)	69.16±0.40	3.00±0.00	0.23±0.00	0.14±0.00
M_{gr} (k = 1.5)	18.92±0.25	5.88±0.36	0.16±0.00	0.04±0.00
M_{gr} (k = 2)	6.97±0.14	10.88±0.67	0.13±0.01	0.01±0.00
M_{sf} (k = 2)	3.99	6.95±0.22	0.02±0.00	0.01
M_{sf} (k = 3)	5.98	5.54±0.50	0.03±0.00	0.01
M_{sf} (k = 5)	9.94	4.18±0.39	0.05±0.00	0.02
M_{sw} (p = 0.1; k = 4)	4.00	15.61±1.20	0.36±0.01	0.01
M_{sw} (p = 0.1; k = 8)	8.00	7.87±0.42	0.46±0.01	0.02
M_{sw} (p = 0.1; k = 16)	16.00	5.00±0.00	0.51±0.01	0.03
M_{sw} (p = 0.3; k = 4)	4.00	10.16±0.55	0.16±0.01	0.01
M_{sw} (p = 0.3; k = 8)	8.00	6.00±0.00	0.22±0.01	0.02
M_{sw} (p = 0.3; k = 16)	16.00	4.00±0.00	0.25±0.01	0.03
M_{sw} (p = 0.5; k = 4)	4.00	8.94±0.40	0.06±0.01	0.01
M_{sw} (p = 0.5; k = 8)	8.00	5.16±0.37	0.09±0.01	0.02
M_{sw} (p = 0.5; k = 16)	16.00	4.00±0.00	0.11±0.00	0.03

Further comparison between complex networks properties will be presented in the further sections, together with the centrality measures analysis, which have the potential to discover existing informal network patterns and behavior that are not noticed before and helps to understand networks and their members by its location and structure (ABBASI and HOSSAIN, 2013).

As a structural property, vertex's centrality was considered to be related to visibility, importance, involvement, control, independence or activity. The next Section will present and explain the centrality measures used in this work.

3 CENTRALITY MEASURES

Centrality measures can be viewed as a mathematical heuristic applied in social network analysis to identify important elements of the network from its structural properties. The heuristic nature of centrality measures is due to the inexistence of a formal definition of what it should measure.

Freeman (1978/79, p. 217) emphasizes that

> [...] centrality is an important structural attribute of social networks. All concede that it is related to a high degree to other important group properties and processes. But there consensus ends. There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is very little agreement on the proper procedure for its measurement.

Bell et al. (1999) reinforces the idea that there is no necessary relationship between a given problem and a centrality measure and many can be applied to the same problem and with the same objective in mind.

Over the years, many centrality measures have been proposed and applied to several contexts within social networks. Everett and Borgatti (2010) presented a method to construct centrality measures based on virtually any graph invariant, by simply calculating the invariant value difference between the complete network and the network without a given vertex.

The development of measures should help to clarify a concept by specifying its components and relationships. However, several measures are vaguely related to intuitive ideas they purport to index, and many are so complex (especially in number of parameters and algorithm complexity) that it is hard to discover what, if anything, they are measuring. Besides, complex measures are difficult or impossible to calibrate and use in huge networks due to time constraints.

This is why the proper selection of centrality measures is fundamental to a successful application in a problem. In this Section, we present the selected centrality measures for the experiments. They were chosen taking into consideration their use in the literature, simplicity (no parametric measure was chosen) and applicability inside social network environment.

Moreover, as our experiments (explained in details on Section 4) focus only on undirected and unweighted networks, no centrality measure exclusive for directed and weighted networks were selected. Many of the selected measures can be applied to any kind of network with the appropriate modifications, but no detailed explanation about their use on other kinds of networks besides the one used in our experiments will be presented in this Section.

As regards the use of centrality measures, it is not relevant which kind of relationship exists between the vertices and what does each vertex represents. Only the structural pattern of the graph/network is important.

We start by presenting the three main and simplest centrality measures: *closeness*, *betweenness* and *degree*. Then, we will explain the newer and more complex measures: *eigenvector*, *information*, *eccentricity*, *subgraph* and *walk-based betweenness*.

Table 3 classifies each one of the eight metrics used in our work according to the definitions of Freeman (1978/79). Each group reflects the underlying idea of centrality that each measure is trying to identify. It is important to notice that the presence of centrality measures belonging to a common group does not mean that they will rank vertices at the same order or evaluate them at the same magnitude in a network.

Table 3 – Centrality Measures Grouped by their underlying Theoretical Foundations

The central element is located between many paths	The central element is close to all other elements	The central element is the one that interacts with many others
Betweenness (C_B)	Closeness (C_C)	Degree (C_D)
Walk Betweenness (C_W)	Information (C_I)	Eigenvector (C_E)
	Eccentricity (C_X)	Subgraph (C_S)

3.1 Closeness Centrality

This centrality measure was the first used in a study of social networks. It was specially proposed for this purpose even though the term centrality measure was not widespread at that time.

The idea of closeness centrality measure (C_C) was presented first by Bavelas (1948) and rigorously defined by Sabidussi (1966) as being the sum of the geodesics inverse distances from the vertex analyzed to all other vertices.

$$C_C\,(p_k) = \frac{1}{\sum_{i=1}^{n} d\,(p_i, p_k)}$$

$$d\,(p_i, p_k) = geodesic\ distance\ from\ p_i\ to\ p_k$$

Freeman (1978/79) affirms that this measure is related to the concept of independency and efficiency of the vertex. In addition, Freeman (1980) highlights that both betweenness and

closeness are determined by the same structural elements of the network: betweenness a measure of control and closeness of independency from that control. For Borgatti (2005), closeness refers to time-until-arrival of "something" flowing in the network.

This measure has time complexity O(mn) for all vertices using the algorithm of Brandes (2001, 2008). The algorithm performs a simple breadth-first search, in which distance and shortest-path counts are determined from each vertex.

Closeness centrality measure is undefined for disconnected graphs. A common approach for unweighted disconnected graphs is to define $d\ (p_i, p_k) = n$ whenever there is no path from p_i to p_k because it is larger than the maximum distance possible between two vertices $(n - 1)$ in such graphs. This algorithm is unsuitable for directed networks because it considers that $d\ (p_i, p_k) = d\ (p_k, p_i)$, restriction that is not fulfilled by most directed graphs.

3.2 Betweenness Centrality

A vertex that is part of many communication paths between other points exhibits a potential communication control, thus it influences the group by withholding, distorting or facilitating the information being communicated during its transmission. Betweenness centrality (C_B) tries to measure this structural aspect of a vertex and was first presented by Shaw (1954), but only strictly defined later by Freeman (1977, 1978/79).

$$C_B\ (p_k) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{g_{ij}(p_k)}{g_{ij}}$$

$$g_{ij}\ (p_k) = number\ of\ geodesics\ between\ p_i\ and\ p_j\ that\ contains\ p_k$$

Borgatti (2005) sees betweenness as a heuristic method to predict frequency of arrival of "something" flowing through the network.

This measure also has time complexity O(mn) for all vertices (BRANDES, 2001, 2008). Similar to the closeness centrality algorithm, it starts performing a simple breadth-first search, in which distance and shortest-path counts are determined from each vertex. An additional step to calculate betweenness is executed by the algorithm. It visits all vertices in reverse order of their discovery, i.e. those farthest from the analyzed source first, to accumulate pair-wise dependencies using a recursive relation that allows the computation of a cubic number of dependencies without computing them all explicitly.

Betweenness centrality measure is defined for disconnected graphs, but its interpretation turns out to be confusing because each component of the graph needs to be calculated separately. It is also inapplicable for directed networks because it considers that $d\ (p_i, p_k) = d\ (p_k, p_i)$ and it has little use on weighted graphs because its algorithm does not consider valued edges distinct from unitary ones. There is a variation of this algorithm proposed by Brandes (2001, 2008) which considers weighted edges, but its time complexity increases to $O(nm + n^2 \log n)$.

3.3 Degree Centrality

Degree centrality was first introduced by Shaw (1954) as an index of vertex's importance and formally defined by Nieminen (1974). It is important to remind the reader that the vertices degree, degree distribution of a network and mean, minimum and maximum degree are also important network properties studied and applied by many other applications outside the centrality context.

The degree centrality (C_D) is simply calculated by the number of adjacencies of a vertex.

$$C_D\ (p_k) = \sum_{i=1}^{n} a\ (p_i, p_k)$$

$$a\ (p_i, p_k) = \begin{cases} 1\ if\ p_i\ and\ p_k\ are\ adjacent \\ 0\ otherwise \end{cases}$$

Freeman (1978/79) says that a vertex with a high degree value is considered to have high visibility and potential for communication activity. In addition, a vertex with high degree is defined as a hub.

This measure has time complexity O(m) for all vertices by counting each edge for both vertices connected by it. It is clearly the simplest and less complex centrality measure, but still useful in many studies and applications.

Degree centrality can be approximated by a normal distribution, with high accuracy, in random networks where the number of edges is at most 1/5 of the possible maximum (complete graph). Denser graphs are not well approximated because normal distribution does not have an upper bound while degree is at most *n*-1 (DONNINGER, 1986).

This measure is split into two measures for directed graphs, indegree and outdegree. Each of them can led to contrasting interpretations inside a specific application and are frequently used for different purposes. There are variations of degree centrality which consider

the values of the edges to compose the degree importance of a vertex, but they are rarely used in real world applications.

3.4 Eigenvector Centrality

Bonacich (1972) suggested a new centrality measure based on the eigenvector of the largest eigenvalue of an adjacency matrix. He justified his preference for the largest eigenvalue since it grants better accuracy. This is so because each eigenvector is a factor of the matrix (being that symmetric) and the associated eigenvalue measures the precision with which it can reproduce that matrix.

Unlike degree, which weights every neighbor equally, the eigenvector weights contacts according to their centrality (e.g. links with influential people makes you more powerful than links with powerless people). Moreover, eigenvector can be seen as a weighted sum of not only immediate contacts but also by indirect connections with every vertex of the network of every length. These characteristics rely on the idea that a central vertex is connected with vertices that are all also centered.

Bonacich (1987) generalized the eigenvector centrality adding a parameter called β to characterize inverse relationships, where being a friend of powerful elements makes you weak and vice-versa, and to control the weights of indirect connections.

The calculus of the largest eigenvalue and its respective eigenvector can be done directly via a series of mathematical operations and methods, but it suffers due to the precision needed that should be infinite along all the process involving its computation. To avoid that drawback, an iterative way known as "*power method*" is used instead (BONACICH, 1991; RICHARDS and SEARY, 2000), where the precision can be controlled as needed and the algorithm is feasible even for large networks.

The *power method* requires an infinite number of steps (worst case) to achieve the same or a multiple of the result achieved by the direct method but, as the number of steps increases, the precision of the measure increases. Therefore, the number of decimal places without value change can be used as a stopping criterion for early convergence and/or a maximum iteration value can be settled for the same purpose.

The complexity of both direct computation and iterative algorithm are based on the numerical precision required, being infinite in the worst case. However, an interesting fact analyzed in real world networks domains is that their sparsity generally causes a large difference between the largest eigenvalue to all other eigenvalues (remember that a matrix of order *n* can have a maximum of *n* distinct eigenvalues). Bonacich (2007) notices that this fact accelerates

the convergence of the *power method* because the larger is the differences between the eigenvalues the faster is the convergence of the *power method* to the largest one (the only eigenvalue needed for the metric).

$$C_E\,(p_k) = E_k^{\infty}\ where\ E^{\infty} = \sum_{it=1}^{+\infty} \frac{E^{it}A}{\sum_{i=1}^{n} E_i^{it-1}}$$

$$A = adjacency\ matrix\ with\ unitary\ values\ on\ the\ main\ diagonal$$

$$E^0 = vector\ complete\ with\ unitary\ values$$

$$\lim_{it \to +\infty} E^{it} = eigenvector\ respective\ to\ the\ largest\ eigenvalue\ of\ A$$

Eigenvector centrality measure is poorly defined for disconnected graphs (each component must be calculated separately) and for directed graphs where vertices with no outedges will accumulate the centrality importance entirely, leaving the other vertices equal to zero (known as the sink problem). This measure is applicable in weighted graphs, but its interpretation became unclear and its algorithm (*power method*) takes a larger number of iterations for convergence (RICHARDS and SEARY, 2000).

A similar centrality measure for directed networks proposed by PAGE et al. (1999) solved the sink problem and was highly relevant in the development of the Google engine. This newer measure serves as a ranking metric for webpages helping to order the results of a search in the engine.

3.5 Information Centrality

This centrality measure was proposed by Stephenson and Zelen (1989) and is based on the information contained in all possible paths between pairs of vertices. It is related to the closeness centrality, differing from it because it takes into consideration not only geodesic distances but also all other path distances. It has also the same restrictions of the closeness centrality measure, restricting its appropriate use for connected unweighted and undirected graphs.

The idea behind this measure is probabilistic. It gives uniform probability to each path that information can flow throw the network, considering these paths lengths and their probabilities to occur, to create a mean length distance from a vertex to all other vertices. The largest the length of a path, the lowest is the probability of it to occur and the least it will contribute for the centrality value of a vertex for the metric.

$$C_I\,(p_k) = \frac{1}{b_{kk} + (T - 2R)/n}$$

$$T = \sum_{j=1}^{n} b_{jj} \;\; and \; R = \sum_{j=1}^{n} b_{ij} \;\; for \; any \; fixed \; i$$

$$B = (D - A + U)^{-1}$$

$$D = diagonal \; matrix \; with \; degree \; values$$

$$A = adjacency \; matrix$$

$$U = matrix \; having \; all \; unitary \; elements$$

This measure has time complexity O(n^3) for all vertices due the inversion of a matrix (done once) using Gaussian elimination. There are optimization methods and more elaborated algorithms to calculate the inverse of a matrix that grants better bounds for the time complexity, but they are applicable mostly for huge entrance sizes. Simple optimizations that helping Gaussian elimination process were considered and applied in our algorithm.

3.6 Eccentricity as Centrality

The eccentricity of a vertex is the inverse maximum distance between that vertex and any vertex of the graph. The diameter of a graph is the maximum eccentricity of a vertex being the radius the minimum eccentricity of a vertex (HAGE and HARARY, 1995).

This centrality measure is closely related to closeness, but it ranks differently the vertices and clearly is less fine-grained. It also presents the same problems and restrictions that closeness centrality do. Vertex eccentricity and the diameter of a graph are properties also used for many purposes in network analysis and are considered to be important distance measures inside graph theory.

$$C_X\,(p_k) = \frac{1}{\max_{\,i \in V} d\,(p_i, p_k)}$$

$$d\,(p_i, p_k) = geodesic \; distance \; from \; p_i \; to \; p_k$$

This measure has complexity O(mn) for all vertices adapting the algorithm of Brandes (2001, 2008) and can be calculated simultaneously with closeness centrality by simply taking the largest path length (since closeness requires the calculus of all geodesic lengths).

3.7 Subgraph Centrality

Estrada and Rodríguez-Velázquez (2005) idea was to consider the number of closed walks, starting and ending at a vertex, to characterize its centrality value. Closed walks are weighted such that their influence on the centrality decreases as the order of the walk increases which gives more importance to smaller walks than longer ones.

Each order of closed walk is associated with a connected subgraph, which means that this measure counts the times that a vertex takes part in the different connected subgraphs (motifs) of the networks (trivial and non-trivial, acyclic and cyclic), with smaller subgraphs having higher importance for the metric. This metric is related with the idea of contribution and involvement (participation) in closely related groups (represented by the subgraphs).

$$C_S\,(p_k) = \lim_{e \to +\infty} \frac{\left(\frac{(A^e)_{kk}}{e!}\right)}{\sum_{i=1}^{n} A_{ii}}$$

$$A = adjacency\ matrix$$

The complexity of the iterative algorithm (similar to the eigenvector centrality) is based on the numerical precision required (convergence), being infinite in the worst case. Despite being very similar to the algorithm applied to the eigenvector centrality, this algorithm is quite more demanding of memory space than the one needed by the eigenvector centrality. While subgraph algorithm requests the allocation of two matrixes for the power-method multiplication step, the eigenvector algorithm only needs to allocate one matrix and one support vector. However, the successive multiplication of the adjacency matrix by itself can also be used to calculate the largest eigenvalue. Therefore, the algorithm applied to the subgraph centrality can simultaneously calculate both centrality measures in nearly the same time.

Since subgraph's algorithm is closely related to the one of eigenvector centrality, both measures suffer with the same restrictions due the kind of graphs they are applicable. Therefore, its appropriate use is restricted to connected unweighted and undirected graphs.

Subgraph and eigenvector are also present within the spectral graph theory area. Their application extends well beyond centrality in many other graph and linear algebra applications being studied and is important in many domains, especially in chemistry where the molecular connections spectral structure affects their behavior.

3.8 Walk-based Betweenness Centrality

The traditional betweenness centrality (FREEMAN, 1977) takes into consideration only geodesic paths, so Newman (2005) proposed a new centrality measure that considers all paths by using random walks.

His ideas were inspired by electrical circuits in which the current flows through the network in proportion to the resistance of each path. Resistance can be viewed in social networks as path distances and the current passing through each considering all paths between all pairs of vertices will be the centrality value.

$$C_W\ (p_k) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} I_{ij}(p_k)$$

$$I_{kj}(p_k) = 1\ and\ I_{ik}(p_k) = 1$$

$$I_{ij}(p_k) = \frac{1}{2}\sum_{t=1}^{n} A_{kt}|T_{ki} - T_{kj} - T_{ti} + T_{tj}|\ \ for\ k \neq i, j$$

$$T = ((D - A), remove\ a\ single\ row\ and\ its\ corresponding\ column)^{-1},$$

$$add\ back\ the\ removed\ row\ and\ column\ with\ values\ zero$$

$$D = diagonal\ matrix\ with\ degree\ values$$

$$A = adjacency\ matrix$$

The complexity of the measure is $O(mn^2)$ to evaluate the first equation for all vertices $O(n^3)$ for matrix inversion using Gaussian elimination, resulting in $O(mn^2 + n^3)$ for the complete algorithm. Brandes and Fleischer (2005) proposed a faster algorithm with complexity $O(n^3+mn \log n)$ where it is possible to calculate, simultaneously, the information centrality as well. The same considerations about matrix inversion presented for information centrality are also valid for the walk betweenness algorithm, so this metrics time complexity can be improved with a more elaborated algorithm of matrix inversion.

This metric has the highest time complexity between the ones analyzed in our work being also one of the most complex ones. Higher time complexity are not even considered because their application on real social and complex networks becomes too narrow. The networks analyzed by many studies consider graphs of several hundreds or thousands of vertices and/or dynamic networks (COHEN et al., 2014; EASLEY and KLEINBERG, 2010; CORREA and MA, 2011). This impose a severe computational problem for these high complexity kind of metrics.

Walk betweenness centrality also presents some of the restrictions of betweenness centrality. It is undefined for directed networks and each component of a disconnected graph needs to be calculated separately. The difference is that this measure, differently from classical betweenness, is perfectly defined for weighted graphs.

The next section will detail about the methodology followed in our research and our experimental set up characteristics together with related explanations that justify our choices.

4 EXPERIMENTAL METHODOLOGY

This section presents detailed information about our experiments and the main reasons that led us to choose the applied methodology. We carefully carry out preliminary tests to choose the best parameters and take the best decisions before the final experiments are set up and executed. Here, we only present the final experimental choices to keep it as simple and objective as possible.

First, we chose five complex network models (presented in Section 2), which are good approximations of real networks properties and behavior (BARABÁSI, 1999 and 2009; COSTA et al., 2008; EASLEY and KLEINBERG, 2010; NEWMAN, 2003). We planned the experiments to reflect real networks behavior, thus all parameters range and implementation choices to generate the random networks take this into consideration.

All complex network models were implemented by us in Java and validated by running a comparative analysis between the properties of the generated networks with the expected theoretical results presented in each paper (the ones that proposed each model).

We also used for our experiments all non-isomorphic[1] connected graphs of six (112 total) and seven vertices (853 total), available on a website dataset[2]. They are useful to illustrate an idea of extreme possible formations for centrality values (not probable in the random models). The sizes (six and seven vertices) were chosen because they are not so small (conceding more variability) and do not create too many networks (for example, there are 11,716,571 non-isomorphic connected networks with ten vertices).

The networks generation set up is summarized on Table 3. Each combination was represented by a hundred samples (trials) for the generated models. The number of trials takes in consideration the execution time required and the statistical relevance (confidence level) necessary to create generalizable and useful results.

A total of 5,765 synthetic networks were used in our experiments, summing 1,446,643 vertices and 39,655,102 edges. The main properties of the networks generated by each model are summarized in Table 5.

In addition, in our experiments, we used four real networks samples, which are: the Facebook ego network (MCAULEY and LESKOVEC, 2012), the USA airport connections (BATAGELJ and MRVAR, 2006), the Erdős collaboration network (BATAGELJ and

[1] Two isomorphic graphs can be drawn the exact same way despite having a distinct adjacency matrix.
[2] http://cs.anu.edu.au/~bdm/data/graphs.html

MRVAR, 2006) and the USA power grid (WATTS and STROGATZ, 1998). Table 5 presents information about these real networks.

For instance, M_{cs} represents Facebook and Erdős networks properties, M_{gr} represents USA Airlines and M_{sw} represents USA Power Grid (see Section 2.1 for complex network models main characteristics). The relation between the real network samples and the complex network models will be important for a complementary analysis of centrality measures behavior in distinct contexts and useful to reinforce the results obtained with the synthetic networks.

Table 4 – Summary of Networks Sample

Model	Parameters	Combinations
Community Structure (M_{cs})	p = {0.5, 0.7} c = {n/10, n/20, n/50} n = {100, 500}	12
Erdős and Rényi (M_{er})	p = {0.1, 0.3, 0.5} n = {100, 500}	6
Geographical (M_{gr})	k = {1.2, 1.5, 2} n = {100, 500}	6
Scale-free (M_{sf})	k = {2, 3, 5} n = {100, 500}	6
Small-world (M_{sw})	p = {0.1, 0.3, 0.5} k = {4, 8, 16} n = {100, 500}	18
Non-isomorphic Networks (N_{ni})	n = {6, 7}	2

Table 5 – Real Networks Main Properties

Network	**Vertices**	**Edges**	**Diameter**	**Clustering**
USA Airport Connections	332	2126	6	0.1031
Facebook Partial Snapshot	4039	88234	8	0.5192
USA Power Grid	4941	6594	46	0.3964
Erdős Collaboration	6927	11850	4	0.0357

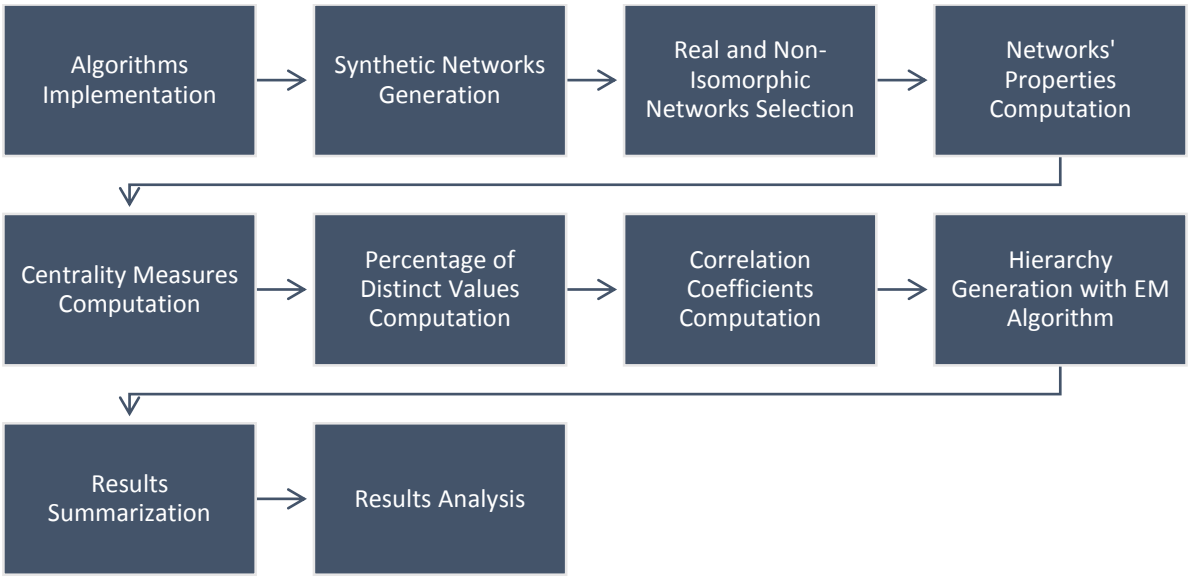
For each network, several properties were calculated: number of edges, minimum, maximum and mean degree, diameter, clustering coefficient, density and the mean distance

from all pairs of vertices. These properties, together with the complex network model, were used to create a simple guide for the selection of centrality measures based on the knowledge of simple properties of the networks. Our idea is that if one wants to apply and select a proper centrality measure, from the many ones that exists, one can use information about the networks of one's environment (application) as guidance to make a good choice, without losing valuable time testing and trying each one of them.

For each vertex, in each network, eight centrality measures (presented on Section 3) were calculated: betweenness (C_B), closeness (C_C), degree (C_D), eigenvector (C_E), information (C_I), subgraph (C_S), walk-based betweenness (C_W) and eccentricity (C_X). The algorithms for each centrality measure followed strictly the implementations proposed by the papers cited on Sections 3.1 to 3.8 to achieve the best known theoretical complexity. All the algorithms passed also by test sample graphs, in which the centrality values were previously published to check their accuracy and effectiveness. Some of the metrics and the networks were run in separate computers to speed up the experiments. It still took several weeks to generate all the synthetic networks using the complex network model algorithms, to calculate all the centrality measures in both synthetic and real network samples and finally to compute and generate elements for the analysis of the results, such as correlation values.

The main steps of the final experimental setup are summarized in Figure 6.

Figure 6 – Experimental Set-up Process



To keep all data organized, with easy access and proper backup, we used the database management system PostgreSQL. It also facilitated the process of parallel execution of experiments in several computers.

It is important to emphasize that the language and database system chosen will not interfere in the results (as well as the used hardware), since computation time will not be relevant to our analysis, therefore other similar technologies could be used to reproduce the same results presented in the next section.

Execution time and other temporal properties were not considered in any part of our analysis since we are not focused on performance constraints at this time. Even so, we pointed out some important characteristics about execution time and algorithm complexity because in real world applications they might be relevant.

After all the experimental data was gathered, we started the analysis process. Our analysis is divided into two moments: the comparison (similarities and differences) between centrality measures and their performance in different kinds of networks.

For the first part, we calculated the Kendall rank correlation coefficient between every combination of two metrics for each network. Recall that this coefficient evaluates the degree of similarity between two sets of ranks given to a same set of objects. It gives a score in a range [-1, 1], where the extreme values represent perfect correlation (+1) or perfect inverse correlation (-1) and values close to zero represent very low correlation.

Kendall's correlation is especially useful for centrality measures because normalization and distribution issues that can vary between metrics do not affect it. In addition, centrality measures values are used frequently as ranking factors where the absolute value is irrelevant. These characteristics can be a problem when using Pearson's correlation (most common correlation), which measures the difference between absolute values.

We also used the EM (Expectation-Maximization) algorithm (DEMPSTER et al., 1977) iteratively (starting with two and ending with eight clusters) to build a hierarchical relationship using a similarity criterion between the centrality measures. The EM algorithm tries to find a set of multivariate normal distributions (one for each cluster) that maximizes the likelihood of the data observed. The attributes provided for the algorithm in our experiments were the centrality values for each vertex. The number of clusters (parameter of the algorithm) was set iteratively from two to eight to create a hierarchy which indicates the metrics more distant to each other (separated early in the hierarchy) and the ones closer to each other (separated last).

The resulting hierarchy will give a clearer idea of the relationship between centrality measures, making it easier to have a notion of distance between groups of metrics (levels of the hierarchy) than correlation values and at the same time it helps to reinforce our conclusions.

For the second part of our analysis, we calculate the percentage of distinct centrality values for each metric in each network. For example, if we say that there are 50% distinct values

in a network of 500 vertices for a determined metric, it means that there are 250 distinct centrality values.

This information can be valuable in most applications where centrality measures are used. It is frequently required that the metric differentiates all vertices of the network, especially in social networks when we know in advance that each element of the network must be unique. Ties between centrality values of distinct vertices can be viewed as lack of information or incapability of the metric to differentiate properly the elements.

Moreover, correlation analysis between centrality measure values and domain specific ones are a common research methodology to measure the metric fitness for a given application. However, most correlation methods are impacted negatively by tied values, reducing their accuracy and distorting the analysis.

With this in mind, we can see that values close to 100% of distinct centrality values are always desirable. That is why we used the percentage of distinct centrality values as a generic performance gauge for the centrality analysis in our synthetic networks.

The final part of our analysis relates the results and behavior of centrality measures in the real network samples with the synthetic networks generated by the complex network models. This is important to reinforce our conclusions and exemplifies their applicability in real-world domains and applications.

The next Section will provide more detail about the networks generated for the experiments as well as the correlation, EM and performance results of the metrics. It will also present our developed guide (based on our experimental results) for the selection of centrality measures in new applications. Also, all the results are discussed and their implications explained and detailed.

5 EXPERIMENTAL RESULTS AND DISCUSSION

The first step of our experiments was the generation of the networks to provide sample graphs data to calculate the centrality measures. The main characteristics of the synthetic networks generated by the five complex models and the non-isomorphic networks samples are summarized in Table 6 grouped by size (number of vertices). Each cell of Table 6 represents all combinations of parameters for each model and that is why high standard deviations were expected.

Table 6 – Means and Standard Deviation of Networks' Main Properties

n = 100				
Model	Degree	Diameter	Clustering	Density
M_{cs}	29.58 ± 10.27	3.20 ± 0.57	0.49 ± 0.10	0.30 ± 0.10
M_{er}	29.71 ± 16.20	2.75 ± 0.93	0.30 ± 0.16	0.30 ± 0.16
M_{gr}	17.90 ± 12.14	4.36 ± 1.81	0.25 ± 0.10	0.18 ± 0.12
M_{sf}	6.51 ± 02.40	4.55 ± 0.78	0.11 ± 0.05	0.07 ± 0.02
M_{sw}	9.33 ± 04.99	5.34 ± 2.41	0.28 ± 0.15	0.09 ± 0.05
n = 500				
Model	Degree	Diameter	Clustering	Density
M_{cs}	142.01 ± 36.91	2.85 ± 0.36	0.43 ± 0.06	0.28 ± 0.07
M_{er}	149.72 ± 81.64	2.33 ± 0.47	0.30 ± 0.16	0.30 ± 0.16
M_{gr}	31.68 ± 26.99	6.59 ± 3.29	0.17 ± 0.04	0.06 ± 0.05
M_{sf}	6.63 ± 02.48	5.56 ± 1.20	0.03 ± 0.02	0.01 ± 0.00
M_{sw}	9.33 ± 04.99	7.42 ± 3.59	0.25 ± 0.16	0.02 ± 0.01
Model	Degree	Diameter	Clustering	Density
N_{ni} (n = 6)	2.83 ± 0.69	2.54 ± 0.68	0.45 ± 0.26	0.57 ± 0.14
N_{ni} (n = 7)	3.20 ± 0.70	2.68 ± 0.70	0.46 ± 0.20	0.53 ± 0.12

We can see through Table 6 that all networks have low diameter values, M_{cs} and M_{er} are denser than the others and, together with M_{sw} they are highly clustered. Both M_{gr} and M_{sf} have lower density and clustering coefficient, especially the latter. It is visible that each

complex network model has unique characteristics not present in the others. This fact becomes even clearer when each combination of parameter settings is analyzed separately. More detailed information about the networks generated, separated by each parameter combination, can be checked at Table 1 and Table 2.

Table 7 present the same properties from the real network samples used in our experiments. Further characteristics of these networks were presented in Table 5.

Table 7 – Real Networks Main Properties

Network	Mean Degree	Diameter	Clustering	Density
USA Airport Connections	12.8072	6	0.3964	0.0387
Facebook Partial Snapshot	43.6910	8	0.5192	0.0108
USA Power Grid	2.6691	46	0.1031	0.0005
Erdős Collaboration	3.4214	4	0.0357	0.0005

The next step of our analysis and the first one due to centralities similarities was calculate the Kendall rank correlation coefficient between every pair of metrics for each network. The mean correlation values grouped by each complex network model are shown in Table 8, Table 9, Table 10, Table 11, Table 12 and Table 13.

The highest overall mean correlation values between two metrics in all networks were dark in the table and the text color is white while the lowest values were grey shadowed.

Interesting patterns can be noticed looking at the Tables. Community structure and the random graphs of Erdős and Rényi were responsible from the highest correlation values among all metrics showing values above 0.8 for most pairs of metrics in the first model and above 0.9 for the second one. Eccentricity centrality measure is the only one opposing the trend, showing the lowest correlation values in the random graphs of Erdős and Rényi and the highest values in non-isomorphic networks. Later we present further details that explain this contrasting behavior of Eccentricity.

At the same time, the lowest overall correlation values were found in scale-free and small-world networks for most metrics. Still, several pairs of metrics do not shown correlation values lower than 0.5.

Interestingly, the geographic model presented median correlation values for all pairs of centrality measures, showing not a single top or bottom score for correlation between any metrics.

Table 8 – Mcs Mean Correlation Values

C_c							
0.82	C_b						
0.96	0.83	C_d					
0.89	0.74	0.91	C_e				
0.95	0.82	0.99	0.91	C_i			
0.89	0.74	0.91	1.00	0.91	C_s		
0.80	0.89	0.81	0.71	0.90	0.71	C_w	
0.35	0.32	0.33	0.31	0.33	0.31	0.32	C_x

Table 9 – Mer Mean Correlation Values

C_c							
0.88	C_b						
0.96	0.91	C_d					
0.92	0.81	0.93	C_e				
0.95	0.88	0.97	0.93	C_i			
0.92	0.81	0.93	1.00	0.93	C_s		
0.89	0.95	0.93	0.83	0.89	0.83	C_w	
0.13	0.11	0.12	0.12	0.12	0.12	0.11	C_x

Table 10 – Mgr Mean Correlation Values

C_c							
0.71	C_b						
0.75	0.74	C_d					
0.79	0.61	0.77	C_e				
0.78	0.73	0.95	0.80	C_i			
0.78	0.62	0.80	0.96	0.82	C_s		
0.69	0.84	0.85	0.64	0.81	0.65	C_w	
0.47	0.37	0.34	0.39	0.36	0.38	0.34	C_x

Table 11 – Msf Mean Correlation Values

C_c							
0.51	C_b						
0.53	0.81	C_d					
0.87	0.46	0.51	C_e				
0.61	0.72	0.88	0.59	C_i			
0.87	0.46	0.51	1.00	0.59	C_s		
0.35	0.77	0.87	0.31	0.67	0.31	C_w	
0.53	0.36	0.37	0.51	0.40	0.51	0.26	C_x

Table 12 – Msw Mean Correlation Values

C_c							
0.66	C_b						
0.53	0.62	C_d					
0.52	0.41	0.59	C_e				
0.69	0.66	0.80	0.61	C_i			
0.40	0.34	0.63	0.79	0.53	C_s		
0.59	0.82	0.70	0.41	0.71	0.36	C_w	
0.42	0.32	0.26	0.27	0.34	0.17	0.29	C_x

Table 13 – Nni Mean Correlation Values

C_c							
0.79	C_b						
0.94	0.78	C_d					
0.85	0.59	0.88	C_e				
0.88	0.76	0.86	0.76	C_i			
0.85	0.60	0.90	0.98	0.76	C_s		
0.88	0.86	0.82	0.63	0.78	0.63	C_w	
0.57	0.53	0.49	0.45	0.49	0.45	0.49	C_x

These results pointed out that most metrics present very similar results for most networks tested, while network properties and models considerably affect metrics correlation values, a clear tendency between several pairs of metrics high and low correlated is present on almost all networks.

To help in identifying the pairs of centrality measures that present a strong or weak relationship and redundant behavior, we take the mean correlation values over all synthetic networks used in our experiments. They are summarized in Table 14.

Table 14 – All networks Mean Correlation Values

C_c							
0.73	C_b						
0.75	0.75	C_d					
0.58	0.76	0.75	C_e				
0.75	**0.80**	**0.89**	0.75	C_i			
0.56	0.72	0.77	**0.93**	0.72	C_s		
0.85	0.69	**0.80**	0.57	0.76	0.56	C_w	
0.34	0.41	0.32	0.33	0.35	0.30	0.31	C_x

The highest correlation values between metrics are highlighted in bold in Table 14, they show five very redundant (above or equal 0.8) pairs of metrics: betweenness and information, degree and information, eigenvector and subgraph, closeness and walk betweenness, degree and walk betweenness. This is a strong indicative that the simultaneous use of the pairs of metrics that presented high correlation values will not be of much use since it will grant very similar results. Remember that a Kendall correlation coefficient value of 0.8 or more indicates roughly that at least 80% of the vertices rank produced by a pair of centrality measures values in order agree, i.e., the two metrics produce a centrality rank for the vertices of a given network 80% similar.

Table 15 presents the standard deviation related to the average values presented on Table 14. The lowest (smaller than or equal 0.20) standard deviation values are highlighted in bold. The highest correlation values also present the lowest standard deviations which evidences that these metrics are with a high confidence margin redundant (i.e. agree on their vertices ranking) in most networks. An overall high standard deviation value was expected already for all metrics correlation values due the variability of networks used in our experiments.

Table 15 – Standard Deviation of Correlation Values

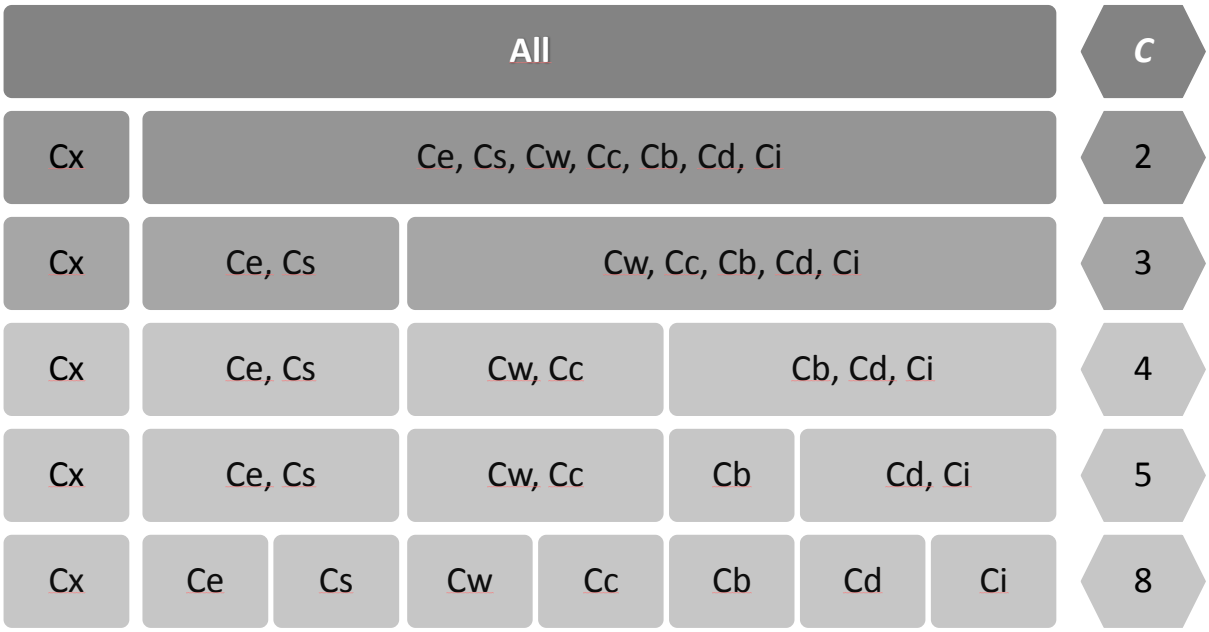
C_c							
0.15	C_b						
0.24	**0.17**	C_d					
0.21	0.23	**0.20**	C_e				
0.15	**0.14**	**0.13**	**0.19**	C_i			
0.29	0.26	**0.18**	**0.17**	0.24	C_s		
0.20	**0.11**	**0.14**	0.23	**0.13**	0.26	C_w	
0.26	0.22	0.22	0.23	0.22	0.25	0.21	C_x

Nakao (1990) and Goh et al. (2003) present sample networks in their experiments where betweenness centrality is strongly correlated with degree centrality but in our experiments, their correlation (0.75) was not that strong compared to the others. Moreover, we can see that just a few pairs of centralities have a correlation value below 0.7, which means that despite all centrality differences, they all have a certain amount of common agreement. Eccentricity presented the lowest correlation values by far, but this statement is mainly due to the fact that it evaluates many vertices as being equally central, i.e., they receive the same centrality value.

In addition, we use the EM (Expectation Maximization) algorithm (DEMPSTER et al., 1997) iteratively (starting with two and ending with eight clusters) to build a hierarchical relationship using similarity criterion between the metrics. Each metric is an instance and the attributes are the centrality rank for each node of all networks. Therefore, in our training set there were 8 instances (number of centrality measures) with 1,446,643 attributes each (number of vertices of all networks used in our experiments). The algorithm tries to find the best multivariate Gaussian distributions (one for each pre-defined clusters number C) that explain statistically better (maximum likelihood) using the attributes provided. There was a possibility that between each iteration, the groups found by the algorithm could be mixed, which will not result in a perfect hierarchy, but it did not happen in our experiments.

The resulting hierarchy is presented in Figure 7. The hierarchy is a good visual guide to easily identify how the centrality measures are related and at which order. It clearly shows an order of relationship and similarity between the centrality measures. The most distinct ones are separated first in an individual group while the most redundant are split later in the hierarchy.

Figure 7 – Metric's Hierarchical Clustering



Both techniques (correlation values and EM algorithm) agree in their results. Eccentricity (C_X) is the centrality measure less similar to the other metrics. It shows a low correlation value and it is separated first in the hierarchical clustering. We can also see three close pairs of centralities in both analyses: eigenvector and subgraph, walk betweenness and closeness, degree and information. Interestingly, these pairs (excluding eigenvector and subgraph) are not the expected ones as suggested by their underlying foundations, shown in Table 3.

The information presented on Table 14 and Figure 7 can be used for the proper selection of metrics that will be most useful at bringing distinct results. This is very important when the idea of centrality in the context worked is undefined and so, different points of view (provided by distinct centrality measures) are valuable for analysis and discussion. Another use of this study is to suggest metrics that will give very similar results despite their different formulations. Such information can be used to replace a metric with higher complexity or to make different relationships between the application environment and the metrics underlying ideas.

The second step of our analysis tries to identify the best metrics. Due to the fact that our experimental environment is general where the vertices and edges play no specific meaning (synthetic networks), there was not a real parameter to compare with the centralities measures to quantity their performance. However, we use the percentage of distinct values given to the nodes in a network by a centrality measure as a generic performance gauge or metrics

granularity. For example, if we say that a given metric achieved 75% of distinct values in a network with 500 vertices, there were 375 distinct centrality values given to the vertices.

It is frequently required that the metrics differentiate all vertices of the network as much as possible, especially in social networks when we know in advance that each element of the network must be unique and whenever the metrics are used as a ranking system (which is common in many applications). Ties between centrality values among different vertices from the same network can be viewed as lack of information or incapability of the metric to differentiate the elements. In addition, correlation with metrics are a common tool to the analysis of the metrics fitness in many applications. Most kinds of correlation are affected by tied values, reducing its accuracy. All these reasons reinforce the idea that values closer to a 100% of distinct values are always desirable for all centrality measures applications.

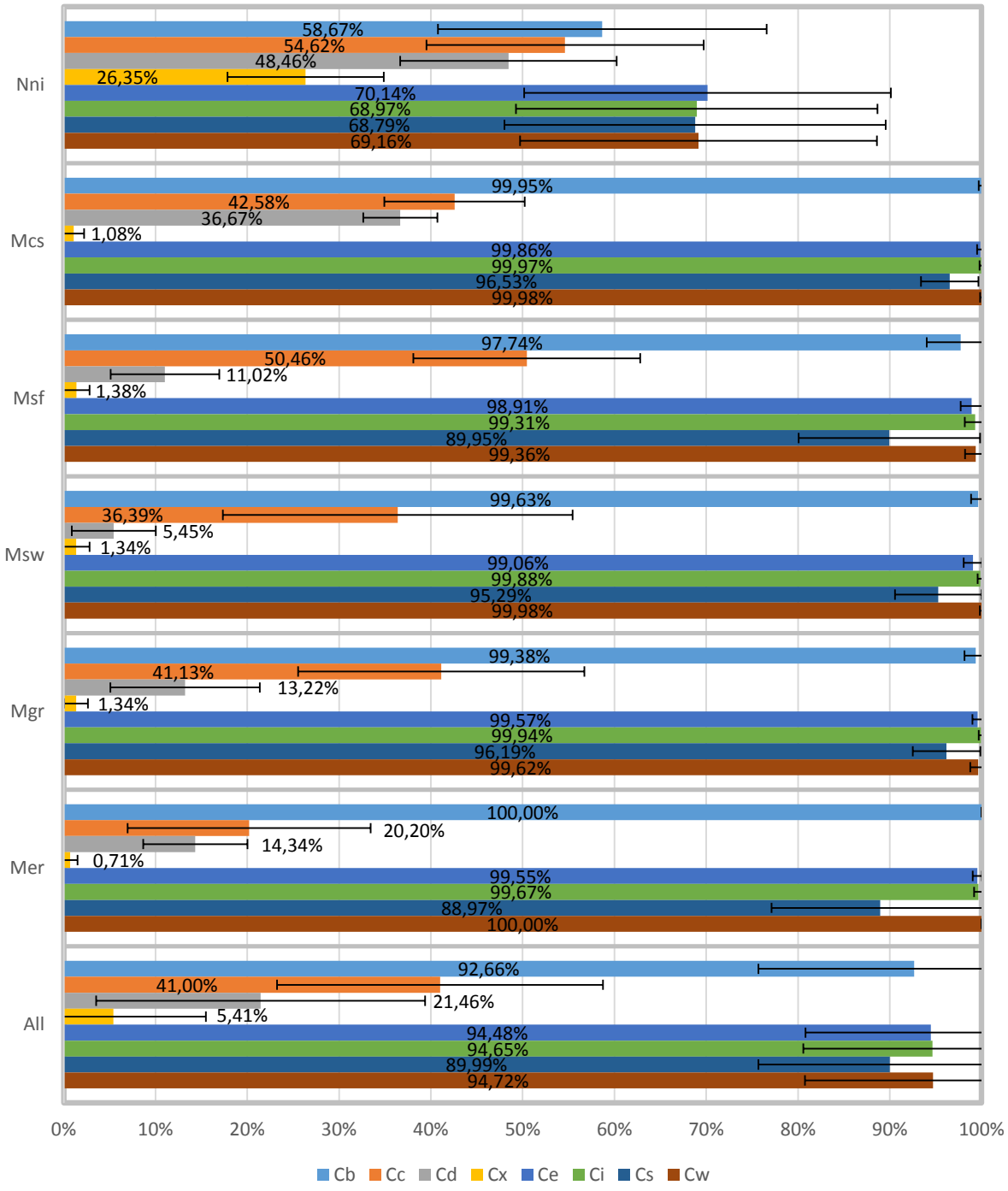
Figure 8 presents the mean percentage of distinct values with the standard deviations for each metric grouped by each kind of network analyzed in our experiments and by the average between all networks.

Our results show that eccentricity is the poorest metric with a high number of tied vertices, followed by degree and then by closeness centralities. These three measures are by far worse than the others in distinguishing vertices by their structural properties while walk betweenness has the best overall performance. It is always better than betweenness centrality (its simpler counterpart) but is worse than information centrality in geographic networks and eigenvector centrality in non-isomorphic networks.

Nonetheless, walk betweenness, subgraph, information, eigenvector and betweenness show more than 95% of distinct values in most of the bigger networks (complex networks models) and around 70% in the smaller networks (non-isomorphic networks set, which contains extreme cases such as complete graphs).

These results reinforce Freeman's (1978/79) experimental results, he also concluded that degree measure is less finer-grained than closeness and both are inferior to betweenness in this aspect. However, our results go against to Bonacich's (2007) conclusion that the eigenvector centrality is appropriate when centrality is ultimately driven by differences in degree in which a high degree position is connected to many low degree positions and vice-versa. If this was true, eigenvector should present better results in scale-free networks and worse ones in networks with community structure and that did not happen in our experiments.

Figure 8 – Mean Percentage of Distinct Values

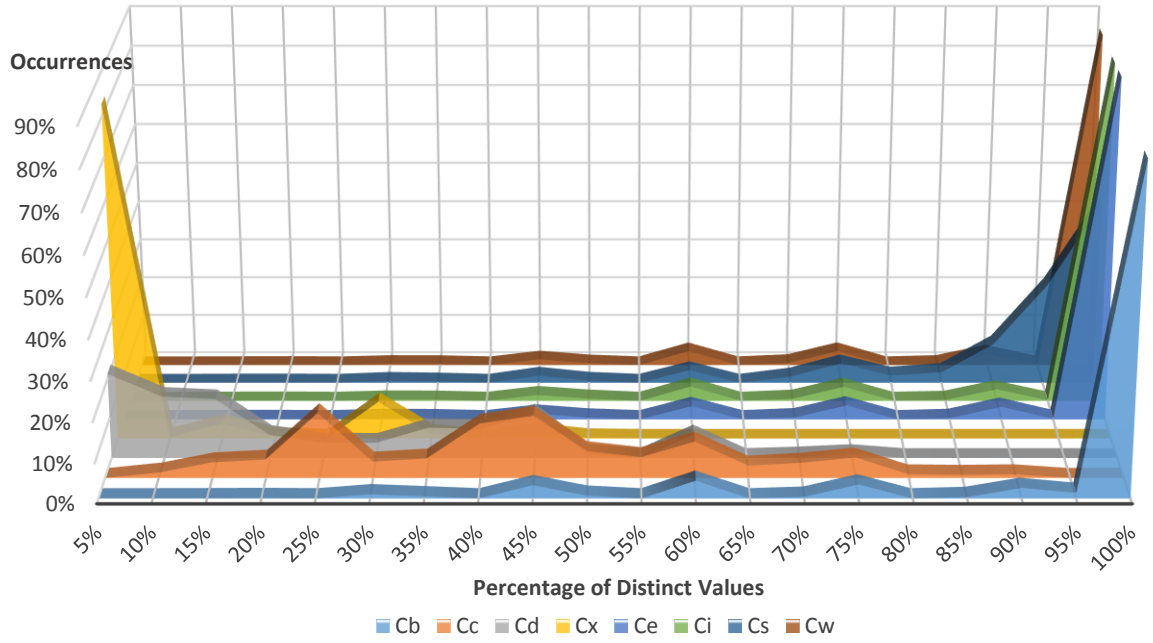


Regarding the standard deviation, *closeness* and *subgraph* present the highest relative values, indicating that their results are highly volatile. A high standard deviation was expected for non-isomorphic networks due to their high variability and smaller sizes. Further, the last block of Figure *8* composed by all networks data presented also higher standard deviations which reinforces the idea that each kind of network affects differently the behavior of the metrics. In addition, the centralities that present the highest scores also present the lowest

standard deviation values, offering a robust and reliable reference to the selection of centralities with high performance at distinguishing vertices.

Figure 9 shows the percentage of distinct values distribution of each metric over all networks. The first point comprehends [0%, 5%), the second [5%, 10%) and so goes on the subsequent points until the last one [95%, 100%]. It illustrates better the differences between the performances of the metrics at distinguishing vertices and strengthens the relevance of the averages shown in Figure 8 by showing little variance (high peaks) in centralities distribution of percentage of distinct values for the top four metrics (betweenness, eigenvector, information and walk betweenness).

Figure 9 – Metrics Distribution



Another interesting aspect of centralities performance (percentage of distinct vertices) analyzed in our experiments was the amount of times each metric achieved the best-known solution (highest number of vertices being distinguished) among all metrics. This information shows the amount of times one metric is better than the others are in distinguishing the vertices of the networks. The results are summarized in Table 16, the cells highlighted in bold present the highest values. The columns sum is higher than 100% because there are many networks where more than one metric achieves the best result.

We can check in Table 16 that despite a very close granularity performance between several metrics in Figure 8 and Figure 9, a clearer difference between the top five metrics is present.

Betweenness stayed way behind the other top metrics in non-isomorphic networks while keeping good scores for community structure and Erdős networks, but it always kept an equal or lower performance than its more complex counterpart, walk betweenness.

Closeness centrality did better than degree and eccentricity in non-isomorphic networks but all three measures are overly outmatched by the other metrics in all complex network model samples.

Eigenvector did really well for non-isomorphic networks with a considerable difference from the other metrics, but it lost several points in all other kinds of networks. It still showed considerable better scores than subgraph centrality in all samples despite their high similarity.

Information centrality is indeed a powerful metric at distinguishing vertices together with walk betweenness. However, the information centrality is quite inferior than walk betweenness and even betweenness in the random graphs of Erdős and Rényi.

When we analyze walk betweenness scores we realize that its higher algorithm complexity paid off. It is by a large amount superior on most networks, it only loses for information centrality in geographic networks and for eigenvector centrality in non-isomorphic networks (this time just for a few percentage points).

Table 16 – Percentage of time where best performance is achieved

Metric \ **Net**	N_{ni}	M_{cs}	M_{sf}	M_{sw}	M_{gr}	M_{er}	**All**
C_b	38.8%	97.6%	62.8%	78.3%	70.2%	**100.0%**	75.5%
C_c	33.7%	0.0%	0.0%	0.0%	0.0%	0.0%	5.6%
C_d	21.5%	0.0%	0.0%	0.0%	0.0%	0.0%	3.6%
C_x	4.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%
C_e	**98.4%**	87.7%	51.7%	47.9%	62.0%	55.0%	67.2%
C_i	90.6%	98.0%	92.8%	90.6%	**98.8%**	67.3%	90.8%
C_s	93.4%	32.4%	31.3%	34.2%	34.3%	32.0%	43.2%
C_w	92.0%	**99.8%**	**100.0%**	**99.9%**	76.5%	**100.0%**	**96.2%**

We constructed Table 17 aiming centrality measures with a high performance at distinguishing vertices by their structural properties and reduced redundancy by only using simultaneous metrics when they presented low correlation values. Its purpose is to serve as a guide for the best-suited centrality measure for an application, before real testing, taking into consideration the kind of networks used and the application goals (the centrality measure underlying idea most appropriate for the environment where it will be applied).

Table 17 – Centrality Measures Guide

Net \ Goal	Control	Independency	Visibility	General
$\boldsymbol{M_{cs}}$	C_b, C_w	C_i	C_e	C_b, C_w, C_i
$\boldsymbol{M_{sf}}$	C_w	C_i	C_e	C_w, C_i
$\boldsymbol{M_{sw}}$	C_w	C_i	C_e	C_w, C_i
$\boldsymbol{M_{gr}}$	C_b, C_w	C_i	C_e	C_i
$\boldsymbol{M_{er}}$	C_b, C_w	C_i	C_e	C_b, C_w
Unknown	C_w	C_i	C_e	C_w, C_i

The goals are the ones defined by Freeman (1978/79), which represent the underlying idea of what is centrality using a given metric (Table 3 explains better the goals and classifies the metrics). The general goal is applied whenever there is no defined centrality meaning in a given application or it did not fit well in any other category.

We picked the complex network model as the metrics main selection attribute because it was the most descriptive networks property among the ones tested (number of edges, minimum, maximum and mean degree, diameter, clustering coefficient, density and the mean distance from all pairs of vertices) to determine the metrics ability of distinguishing vertices. To get to such a conclusion we have run several attribute selection algorithms, used for rule learning and decision trees. They all selected as prior attributes the network model in which the networks generation were based as the most important attribute to determine the metrics performance. We also add an unknown line to be used when the network being analyzed for an application does not fit very well in any complex network model. The metrics suggested, for this case and for the general goal column, are the overall best ones considering all networks tested in our experiments.

For example, *walk betweenness* is the most appropriate measure if one has an application where networks present scale-free model properties and when the objective is to rank or identify objects by their control of communication or when the objective is generic/general, such as power. Whenever more than one measure is indicated in Table 17 in a determined cell, it means that the metrics will produce distinct results from each other and that they grant a high performance at differentiating vertices. Therefore, they can be used simultaneously to identify different aspects of the centrality concept for that application or another aspect of the centrality measures can be used to pick one over the other, such as algorithm complexity.

The last part of our experiments focused on four real networks samples (Table 5). The resulting centrality correlation values are summarized in Table 18.

Table 18 – Correlation Values

Network	Metric	Highest Value	Mean±Std.
USA Airport Connections (M_{gr})	C_b	0.80 (C_w)	0.59±0.15
	C_c	0.84 (C_e, C_s)	0.66±0.16
	C_d	0.91 (C_i)	0.70±0.18
	C_x	0.44 (C_c)	0.36±0.04
	C_e	1.00 (C_s)	0.68±0.22
	C_i	0.91 (C_d)	0.71±0.18
	C_s	1.00 (C_e)	0.68±0.22
	C_w	0.80 (C_b, C_d)	0.61±0.16
Facebook Partial Snapshot (M_{cs})	C_b	0.69 (C_w)	0.43±0.20
	C_c	0.52 (C_s)	0.41±0.08
	C_d	0.83 (C_i)	0.44±0.25
	C_x	0.47 (C_c)	0.10±0.21
	C_e	0.78 (C_s)	0.33±0.30
	C_i	0.83 (C_d)	0.50±0.21
	C_s	0.78 (C_e)	0.40±0.28
	C_w	0.69 (C_b)	0.34±0.19
USA Power Grid (M_{sw})	C_b	0.67 (C_d)	0.29±0.21
	C_c	0.74 (C_x)	0.38±0.22
	C_d	0.67 (C_b)	0.32±0.22
	C_x	0.74 (C_c)	0.37±0.23
	C_e	0.48 (C_x)	0.23±0.14
	C_i	0.53 (C_c)	0.41±0.10
	C_s	0.45 (C_i)	0.31±0.10
	C_w	0.66 (C_b)	0.29±0.20
Erdős Collaboration (M_{cs})	C_b	0.82 (C_d)	0.57±0.13
	C_c	0.85 (C_e, C_s)	0.63±0.21
	C_d	0.82 (C_b)	0.63±0.12
	C_x	0.54 (C_b)	0.41±0.08
	C_e	1.00 (C_s)	0.69±0.24
	C_i	0.82 (C_e, C_s)	0.66±0.18
	C_s	1.00 (C_e)	0.69±0.24
	C_w	0.80 (C_c)	0.60±0.15

The correlation results obtained in the real networks samples are similar to the results showed on Table 14 and Figure 7. We can see in Table 18 that eccentricity is still the lowest correlated centrality measure and that the strongest groups of correlated measures are still

present: degree with information and betweenness, subgraph with eigenvector. The exception is the relationship between walk betweenness and closeness that did not appear as much evident as it did in the synthetic networks. In addition, the overall correlation shown by all metrics are at similar amplitude, i.e. higher than 0.8 correlation in most cases.

Table 19 presents the metrics granularity performance on the real networks samples. They agree with the results obtained by the generation of synthetic networks using the complex network models (Figure 8 and Table 16). The order of the metrics granularity performance remains the same, eccentricity, followed by degree and closeness with the lowest values and subgraph, eigenvector and betweenness followed by information and walk betweenness with the highest ones.

Table 19 – Percentage of Distinct Values

Metric	**USA Airport Connections**	**Facebook Partial Snapshot**	**USA Power Grid**	**Erdős Collaboration**
C_b	55.42%	86.63%	59.28%	21.93%
C_c	57.83%	0.99%	0.22%	0.20%
C_d	17.47%	5.62%	0.32%	1.33%
C_x	1.20%	0.12%	0.49%	0.04%
C_e	82.83%	25.25%	2.25%	30.62%
C_i	83.13%	95.67%	88.54%	33.95%
C_s	62.65%	7.45%	3.97%	8.34%
C_w	74.40%	95.37%	63.75%	35.20%

Moreover, considering the fact that the community structure model represents Facebook and Erdős collaboration networks, we can see that Table 16, Table 19, and Figure 8 show higher values for betweenness, information and walk betweenness centrality measures. While, USA Airports (a geographic network) presented higher values for information centrality and USA Power Grid (related to the “small-world” model properties) presented higher values for information followed by walk betweenness centralities.

That is, both the synthetic generated networks based on the complex network models and the real networks samples analyzed in our experiments agree on their results due to the behavior presented by each metric, considering their granularity performance and correlation values.

We also can see that the suggested metrics showed in Table 17 (based on the synthetic networks results) were valid for all the real networks samples tested in our experiments, which strengthens the idea that one can use it as a useful and reliable guide for the selection of centrality measures based on their granularity and correlation values.

Final conclusions about our experiments and results, and possible future work are presented in the next section.

6 CONCLUSIONS AND FURTHER WORK

Today we have access to networks with millions of devices. Technological, biological or social networks data are available for study as never before, and they are just a small fraction of our connected world that we are trying to understand. Therefore, their analysis and understanding requires several tools and techniques.

We already know that network structure plays an important role in many applications and define many characteristics of the population being mapped by the network. That is the exact purpose of centrality measures applications. The analysis of the networks structural properties is relevant to a number of applications, in particular to artificial intelligence and computer science. Various measurements that are available in complex network analysis, such as vertex centrality, have the potential to provide useful knowledge about patterns and behaviors in complex and social networks.

The increasing availability of data on large networks and the greater variability of centrality applications have led to the creation and development of many centrality measures. Nonetheless, among all these centrality measures little is known or provided about them to help one choose the best metric for a specific environment or application. Most works in this area have focused only on showing in which kind of networks their metrics are unique or better applied when compared to the others. However, analyses of their differences and use as well as studies about when they can be best applied are still open issues. Thus, our work tries to fill part of this important gap.

Our work aims at analyzing the main centrality measures by using structural properties of the network, statistical methods such as correlation and machine learning. Using this methodology, we provided information that helps in the selection of centrality measures by complex network models properties as guidance. Our experiments show that the measures known as walk betweenness, information, eigenvector and betweenness centralities can distinguish vertices in all kind of networks with performance of at least 95% in most of the case studies (Figure 9 and Table 16). Further, each of these centralities achieve a better result in a determined kind of network, mainly defined by a structure model rather than by simpler properties (Figure 8).

While classifying a network to such models is not always an easy task, this classification can on its own define, with a high degree of precision, which centralities should be applied to

grant the best results. They can also be a strong index of centralities behavior about their granularity and their similarity with other metrics (Table 19).

We demonstrate also, by means of experimental evidence, that pairs of metrics achieve high correlation values, despite their theoretical foundations and underlying centrality concepts suggests otherwise (Figure 7, Table 14 and Table 18). Five pairs of metrics achieve very close results: betweenness and information, degree and information, eigenvector and subgraph, closeness and walk betweenness, degree and walk betweenness. Some of them are surprising by the fact that their theoretical foundations and algorithms are clearly distinct.

We can see that out of the eight centrality measures tested, half of them (closeness, degree, eccentricity and subgraph) are outmatched in every experiment by at least one of the others (betweenness, eigenvector, information and walk betweenness). Among the best four centralities only betweenness and walk betweenness, also eigenvector and walk betweenness present low redundancy if applied together.

The correlation values were surprisingly high even for very distinct measures considering their formulations and algorithms. The data presented in Table 8, Table 9, Table 10, Table 11, Table 12 and Table 13 showed a considerable difference in correlation values among metrics considering different kinds of networks (represented by the complex network models). However, the overall similarity among pairs of metrics followed a strict tendency as summarized in Table 14 and Table 15. In addition, Table 18 reinforced the trend presented by the synthetic network with real sample networks.

This helps to reduce even further the available options for many kinds of applications depending on their goals and network properties (as illustrated in Table 17). In addition, it suggests that the application of many centrality measures simultaneously can lead to fruitless results demanding more processing time to produce analogous results.

More importantly, we showed that the structural properties of the networks can be used as good predictors of centralities granularity and correlations. The guide presented in Table 17 can naturally be used in practice for the selection of the most appropriate centrality measures. They can be applied for a determined objective and network if one aims at high granularity of centrality values and for distinct results when applying more than one metric.

We also presented evidence that the use of the information retrieved with the synthetic networks would be useful and accurate about the behavior of the centrality measures in all real samples analysis despite the larger size of their networks. We uncovered that using only casual information about the networks' structural properties inherited from their application domain and their relation to the complex network models characteristics.

Centrality measures are a very useful tool for networks analysis. However, their application and proper selection requires analyses and information about their behavior. Our main contribution in this research is to provide information about centrality measures that helps in their selection for a given application domain. Our results can be used as a guide to select the best centrality measures if one knows characteristics of its network or which complex network model it fits in. Furthermore, centralities can be selected by excluding similar ones, reducing redundancy and optimizing resources.

Further research work includes the investigation of other network properties and their relationship with centrality measures, the application and comparison between metrics in directed and weighted networks, the study of parametric measures in other real world network domains and the analysis of the relationship between centrality measures and other network and graph measurements.

REFERENCES

ABBASI, Alireza; HOSSAIN, Liaquat. Hybrid centrality measures for binary and weighted networks. **Complex Networks:** Studies in Computational Intelligence. v. 424, p. 1-7, 2013.

ADAH, Sibel; LU, Xiaohui; MAGDON-ISMAIL, Malik. Deconstructing centrality: thinking locally and ranking globally in networks. In: **Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.** Niagara, Ontario, Canada, p. 418-425, 2013.

AHUJA, Ravindra K., MAGNANTI, Thomas L., ORLIN, James B.. **Network Flows:** Theory, Algorithms and Applications. Prentice-Hall Press, Upper Saddle River, NJ, USA, 1993.

BARABÁSI, Albert-László. **Linked:** the new Science of networks. Perseus Press, 2002.

BARABÁSI, Albert-László; ALBERT, Réka. Emergence of scaling in random networks. **Science.** v. 286, p. 509-512, 1999.

BATAGELJ, Vladimir; MRVAR, Andrej. **Pajek datasets.** 2006. Available at: <http://vlado.fmf.uni-lj.si/pub/networks/data/>.

BAVELAS, Alex. A mathematical model for group structures. **Human Organization.** v. 7, p. 16-30, 1948.

BELL, David C.; ATKINSON, John S.; CARLSON, Jerry W.. Centrality measures for disease transmission networks. **Social Networks.** v. 21, p. 1-21, 1999.

BOLLAND, John M.. Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks. **Social Networks.** North-Holland, v. 10, p. 233-253, 1988.

BONACICH, Phillip. Factoring and weighting approaches to status scores and clique identification. **Journal of Mathematical Sociology.** Birkenhead, England, v. 2, p. 113-120, 1972.

BONACICH, Phillip. Power and centrality: a family of measures. **American Journal of Sociology.** The University of Chicago Press, v. 92, n. 5, p. 1170-1182, 1987.

BONACICH, Phillip. Simultaneous group and individual centralities. **Social Networks.** North-Holland, v. 13, p. 155-168, 1991.

BONACICH, Phillip. Some unique properties of eigenvector centrality. **Social Networks.** v. 29, p. 555-564, 2007.

BORBA, Elizandro Max. **Medidas de centralidade em grafos e aplicações em redes de dados.** 77 p. Dissertação (Mestre em Matemática Aplicada) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

BORGATTI, Stephen P.. Centrality and network flow. **Social Networks.** v. 27, p. 55-71, 2005.

BORGATTI, Stephen P.. Identifying sets of key players in a network. **Computational and Mathematical Organization Theory**, v.12(1), p. 21-34, 2006.

BORGATTI, Stephen P.; CARLEY, Kathleen M.; KRACKHARDT, David. On the robustness of centrality measures under conditions of imperfect data. **Social Networks.** v. 28, p. 124-136, 2006.

BORGATTI, Stephen P.; EVERETT, Martin G.. A graph-theoretic perspective on centrality. **Social Networks.** v. 28, p. 466-484, 2006.

BRANDES, Ulrik. A faster algorithm for betweenness centrality. **Journal of Mathematical Sociology.** v. 25, p. 163-177, 2001.

BRANDES, Ulrik. On variants of shortest-pat betweenness centrality and their generic computation. **Social Networks.** v. 30, p. 136-145, 2008.

BRANDES, Ulrik; FLEISCHER, Daniel. Centrality measures based on current flow. **22nd Annual Symposium on Theoretical Aspects of Computer Science - STACS.** Volker Diekert and Bruno Durand Editors, Stuttgart, Germany, Springer Press, LNCS v. 3404, p. 533-544, 2005.

BRANDES, Ulrik; KENIS, Patrick; WAGNER, Dorothea. Communicating centrality in policy network drawings. **IEEE Transactions on Visualization and Computer Graphics.** v. 9(2), p. 241-253, 2003.

BUTTS, Carter T.. Exact bounds for degree centralization. **Social Networks.** v. 28, p. 283-296, 2006.

COHEN, E.; DELLING, D.; PAJOR, T.; WERNECK, R. F.. Computing classic closeness centrality, at scale. In: **Proceedings of ACM Conference on Online Social Networks.** Dublin, Ireland, p. 37-50, 2014.

CORREA, Carlos D.; MA, Kwan-Liu. Visualizing social networks. **Social Network Data Analytics.** Springer Press, Charu C. Aggarwal, Editor, p. 307-326, 2011.

COSTA, L. da F.; RODRIGUES, F. A.; TRAVIESO, G.; VILLAS BOAS, P. R.. Characterization of complex networks: a survey of measurements. **Advances in Physics.** v. 56, p. 167-242, 2008.

COSTENBADER, Elizabeth; VALENTE, Thomas W.. The stability of centrality measures when networks are sampled. **Social Networks.** v. 25, p. 283-307, 2003.

CUNNINGHAM, William H.. Optimal attack and reinforcement of a network. **Journal of the Association for Computing Machinery.** v. 32, p. 549-561, 1985.

DANOWSKI, James A.; CEPELA, Noah. Automatic mapping of social networks of actors from text Corpora: time series analysis. **Data Mining for Social Network Data / Annals of Information Systems.** Nasrullah Memon, Jennifer Jie Xu, David L. Hicks, and Hsinchun Chen Editors, Center for Applied Anthropology, Springer Press, v. 12, p. 31-46, 2010.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B.. Maximum likelihood from incomplete data via the EM algorithm. **Journal of Royal Statistical Society.** v. 39(1), p. 1-38, 1977.

DONNINGER, Ch.. The distribution of centrality in social networks. **Social Networks.** North-Holland, v. 8, p. 191-203, 1986.

DWYER, Tim; HONG, Seok-Hee; KOSCHÜTZKI, Dirk; SCHREIBER, Falk; XU, Kai. Visual analysis of network centralities. **Asia-Pacific Symposium on Information Visualization.** K. Missue, K. Sugiyama and J. Tanaka, Editors, Tokyo, Japan, v. 60, p. 189-197, 2006.

EASLEY, David; KLEINBERG, Jon. **Networks, crowds, and markets:** reasoning about a highly connected world. Cambridge University Press, New York, NY, USA, 2010.

ERDŐS, P.; RÉNYI, A.. On random graphs I. **Publicationes Mathematicae.** v. 6, p. 290-297, 1959.

ESTRADA, Ernesto; RODRÍGUEZ-VELÁZQUEZ, Juan A.. Subgraph centrality in complex networks. **Physical Review E.** v. 71:056103, 2005.

EVERETT, Martin G.; BORGATTI, Stephen P.. Induced, endogenous and exogenous centrality. **Social Networks.** v. 32, p. 339-344, 2010.

EVERETT, Martin; BORGATTI, Stephen P.. Extending centrality. **Models and methods in social network analysis.** Cambridge University Press, Carrington P., Scott J. and Wasserman S., Editors, p. 57-76, 2005.

FRATTA, L.; MONTANARI, U.. A vertex elimination algorithm for enumerating all simple paths in a graph. **Networks.** v. 5, p. 151-177, 1975.

FREEMAN, Linton C.. A set of measures of centrality based on betweenness. **Sociometry.** v. 40, p. 35-41, 1977.

FREEMAN, Linton C.. Centrality in social networks: conceptual clarification. **Social Networks.** Netherlands, v. 1, p. 215-239, 1978/79.

FREEMAN, Linton C.. The gatekeeper, pair-dependency and structural centrality. **Quality and Quantity.** Netherlands, v. 14, p. 585-592, 1980.

FREEMAN, Linton C.; ROEDER, Douglas; MULHOLLAND, Robert R.. Centrality in social networks: II. Experimental Results. **Social Networks.** Netherlands, v. 2, p. 119-141, 1979/80.

FRIEDKIN, Noah E.. Theoretical foundations for centrality measures. **American Journal of Sociology.** v. 96 (6), p. 1478-1504, 1991.

GOH, K. –I.; OH, E.; KAHNG, B.; KIM, D.. Betweenness centrality correlation in social networks. **Physical Review E.** v. 67:017101, 2003.

HAGE, Per; HARARY, Frank. Eccentricity and centrality in networks. **Social Networks.** v. 17, p. 57-63, 1995.

HEVENSTONE, Debra. Academic employment networks and departmental prestige. **Why Context Matters:** applications of social networks analysis. VS Research Press, Thomas N. Friemel, Editor, Germany, p. 119-140, 2008.

HUA, Guangying; SUN, Yingjle; HAUGHTON, Dominique. Network analysis of US Air transportation network. **Data Mining for Social Network Data / Annals of Information Systems.** Nasrullah Memon, Jennifer Jie Xu, David L. Hicks, and Hsinchun Chen Editors, Center for Applied Anthropology, Springer Press, v. 12, p. 75-89, 2010.

KAZA, Siddharth; CHEN, Hsinchun. Identifying high-status vertices in knowledge networks. **Data Mining for Social Network Data / Annals of Information Systems.** Nasrullah Memon, Jennifer Jie Xu, David L. Hicks, and Hsinchun Chen Editors, Center for Applied Anthropology, Springer Press, v. 12, p. 91-107, 2010.

KEARNS, Michael. Experiments in social computation. **Communications of the ACM.** North-Holland, v. 55, n. 10, 2012.

KISS, Christine; BICHLER, Martin. Identification of influencers: measuring influence in customer networks. **Decision Support Systems.** v. 46, p. 233-253, 2008.

KOSCHÜTZKI, D.; SCHREIBER, F.. Comparison of centralities for biological networks. In: **Proceedings of German Conference of Bioinformatics.** Bielefeld, Germany, p. 199-206, 2004.

MCAULEY, J.; LESKOVEC, J.. Learning to discover social circles in ego networks. **Neural Information Processing Systems.** n.4, 2012.

NAKAO, Keiko. Distribution of measures of centrality: enumerated distributions of Freeman's graph centrality measures. **The International Network for Social Network Analysis / Connections.** Alvin W. Wolfe Editor, Center for Applied Anthropology, University of South Florida, Tampa, FL, USA, v. 13, n. 3, p. 10-22, 1990.

NEWMAN, M. E. J.. A measure of betweenness centrality based on random walks. **Social Networks.** v. 27, p. 39-54, 2005.

NEWMAN, M. E. J.. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. **Physical Review E.** v. 64:016132, 2001.

NEWMAN, M. E. J.. The structure and function of complex networks. **SIAM review.** v. 45, n. 2, p. 167-256, 2003.

NEWMAN, M. E. J.; GIRVAN, M.. Finding and evaluating community structure in networks. **Physical Review E.** v. 69(2):026113, 2004.

NEWMAN, M. E. J.; PARK, Juyong. Why social networks are different from other types of networks. **Physical Review E.** v. 68:036122, 2003.

NIEMINEN, Juhani. On centrality in a graph. **Scandinavian Journal of Psychology.** v. 15, p. 322-336, 1974.

NOBLE, Diego; GRANDO, Felipe; LAMB, Luís da Cunha. The impact of centrality on individual and collective performance in social problem-solving systems. In: **Proceedings of Genetic and Evolutionary Computation Conference.** Madrid, Spain, p.1-8, 2015.

ORTIZ-ARROYO, Daniel. Discovering sets of key players in social networks. **Computational Social Network Analysis:** trends, tools and research advances. Springer Press, Ajith Abraham, Aboul-Ella Hassanien and Václav Snášel, Editors, p. 27-47, 2010.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T.. **The Page Rank citation ranking:** bringing order to the web. Technical Report SIDL-WP-1999-0120, Sanford University, 1999.

PUZIS, Rami; TUBI, Meytal; ELOVICI, Yuval. Optimizing targeting of intrusion detection systems in social networks. **Handbook of Social Network Technologies and Applications.** Springer Press, Borko Furht, Editor, p. 549-568, 2010.

RICHARDS, William; SEARY, Andrew. Eigen analysis of networks. **Journal of Social Structure.** v. 1, 2000. Available at: <http://www.cmu.edu/joss/content/articles/volume1/RichardsSeary.html>.

SABIDUSSI, Gert. The centrality index of a graph. **Psychometrika.** v. 31, p. 581-603, 1966.

SHAW, Marvin E.. Group structure and the behavior of individuals in small groups. **Journal of Psychology.** v. 38, p. 139-149, 1954.

STEPHENSON, Karen; ZELEN, Marvin. Rethinking centrality: methods and examples. **Social Networks.** North-Holland, v. 11, p. 1-37, 1989.

THILAGAM, P. Santhi. Applications of social network analysis. **Handbook of Social Network Technologies and Applications.** Springer Press, Borko Furht, Editor, p. 637-649, 2010.

TUTZAUER, Frank. Entropy as a measure of centrality in networks characterized by path-transfer flow. **Social Networks.** v. 29, p. 249-265, 2007.

WATTS, Duncan J.; STROGATZ, Steven H.. Collective dynamics of ‘small-world’ networks. **Nature.** v. 393(6684), p. 440-442, 1998.

YAN, Erjia; DING, Ying. Applying centrality measures to impact analysis: a coautorship networks analysis. **Journal of the American Society for Information Science and Technology.** Wiley Press, New York – NY, USA, v. 60(10), p. 2107-2118, 2009.

ZEMLJIČ, Barbara; HLEBEC, Valentina. Reliability of measures of centrality and prominence. **Social Networks.** v. 27, p. 73-88, 2005.