Centrality Measures in Biological Networks

Mahdieh Ghasemi¹, Hossein Seidkhani¹, Faezeh Tamimi¹, Maseud Rahgozar² and Ali Masoudi-Nejad*¹

¹Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
²Database Research Group (DBRG), Control and Intelligent Processing Center of Excellence (CIPCE), School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

Abstract: Many complex systems such as biological and social systems can be modeled using graph structures called biological networks and social networks. Instead of studying separately each of the elements composing such complex systems, it is easier to study the networks representing the interactions between the elements of these systems. A commonly known fact in biological and social networks analysis is that in most networks some important or influential elements (e.g. essential proteins in PPI networks) are placed in some particular positions in a network. These positions (i.e. vertices) have some particular structural properties. Centrality measures quantify such facts from different points of view. Based on centrality measures the graph elements such as vertices and edges can be ranked from different points of view. Top ranked elements in the graph are supposed to play an important role in the network. This paper presents a comprehensive review of existing different centrality measures and their applications in some biological networks such as Protein-Protein interaction network, residue interaction and gene–gene interaction networks.

Keywords: Centrality measures, graph structures, network analysis, network structures.

1. INTRODUCTION

Complex systems such as biological and social systems can be modeled mathematically using graph structures [1]. Studying these mathematical models is usually easier than studying separately the elements of such complex systems. These models are called (biological or social) networks (e.g. protein-protein interaction networks). A widely accepted fact in biological and social network analyses is that in most networks some vertices or edges are more important or influential than others. This importance can be quantified using centrality measures. Having a large network with thousands of vertices, it is always useful to find those vertices that play a more important role in the network. Several types of centrality measures are defined by scientists, which can be used for different application goals. Centrality measures can be categorized into different groups according to the method they are being calculated. These measure are used to rank the vertices or the edges of a graph from different points of view. They are also used as a global graph measure to assign a numerical value to the whole graph. This paper presents a study of existing different centrality measures and their applications in biological network analysis. The paper is structured as follows: In section 2, we explain the applications of centrality measures; the focus of this section is on explaining some of the many researches in bioinformatics area that used centrality analysis to study different biological networks. In section 3, we study and explain different centrality measures; we also analyze Zachary’s well-known “karate club” network [2] with each centrality measure. We show visually the top ranked vertices for each centrality measure and discuss briefly the differences between these measures. In section 4, some issues about complexity of computing different centrality measures are discussed. Finally, in section 6, we make some overall conclusion for the study.

2. CENTRALITY MEASURES’ APPLICATIONS: ANALYZING BIOLOGICAL NETWORKS

As mentioned earlier, network’s elements can be ranked using centrality measures. This ranking is based on the structural features of the studied network. This information is used in different areas, for example, in social network analysis, to find opinion leaders who are able to affect a lot of people [3] and to discover communities in networks [4, 5], in communication network flow analysis, to estimate network traffic and congestion on vertices [6], In the network approach to psychopathology, to study how symptoms dynamically interact over time in a network architecture [7], in biometric systems [8] and in bioinformatics in many researches [9-14]. In this section, we explain some of these researches.

Hahn et al. [15] used Closeness centrality [16], Degree centrality [17] and Shortest-Path Betweenness centrality [18] examine the evolution of protein-protein interaction networks. They realized that proteins that have high centrality score evolve more slowly and are more likely to be essential for survival. Özgür et al. [9] introduced data mining and network structure analysis based approach in order to predict good candidate disease-related genes. They used centrality analysis on gene-interaction network using Degree centrality, Eigenvector centrality [19], Closeness centrality and Shortest-Path Betweenness centrality. Centrality measures have been used in many researches in order to find...
essential proteins in protein-protein interaction networks using topological characteristics of the network [10-13]. For example, Del Rio et al. [11] analyzed 18 different reconstructed metabolic networks using 16 different centrality measures and Estrada [10] utilized 6 different measures to find essential proteins in Yeast protein-protein interaction network. Hsu et al. [20] investigated the relationship between miRNA regulation and protein-protein interaction network of human using some centrality measures such as Degree centrality, Local Clustering Coefficient centrality [21] Shortest-Path Betweenness centrality, and Closeness centrality in order to understand how miRNAs influence the protein interaction network. To perform the analyses, they used the predicted miRNA targets from TargetScan [22] and found that by targeting the hub proteins (usually those with high Degree centrality) and bottleneck proteins (usually those with high Shortest-Path Betweenness centrality), miRNA may regulate the protein interaction network in a wider scope. Sengupta et al. [23] analyzed TF-miRNA-gene network in order to find out topological structures that are associated with a number of common diseases and those structures that facilitate disease progression. Jovelin et al. [24] studied the evolution of the yeast transcription factors in the context of the structure of the gene regulatory network. They used two datasets of the Yeast transcriptional network [25, 26] and protein sequences of orthologous genes from Yeast [27]. For each dataset, they calculated In-degree centrality, Out-degree centrality and shortest-path betweenness centrality in their centrality analyses. Two main findings of the paper is that central transcription factors tend to evolve faster and the effect of centrality on protein sequence evolution is independent of other genomic variables. Amitai et al. [28] used Closeness centrality to Identify active site residues from residue interaction graphs that vertices are amino acid residues and edges are the interactions between amino acid residues. Joy et al. [29] analyzed Shortest-Path Betweenness centrality and Degree centrality scores of network vertices in the Yeast protein-protein interaction network. They found that proteins that have high Shortest-Path Betweenness centrality score, but low Degree centrality score were found to be abundant in the yeast proteome. Their analyzing results suggested that high-Betweenness, low-Degree proteins may act as important links between different functional modules. Fierst et al. [30] analyzes the Yeast genetic interaction network to understand how the Yeast cell maintains across genetic and environmental variation. They used Shortest-Path Betweenness centrality, Closeness centrality, and Degree centrality in their analyses. Vaggi et al. [31] studied the protein-protein interaction network of fission Yeast using graph theoretical approaches. In order to find certain proteins that function as bridges between diverse cellular processes, they introduced a new network measure, linkurity, and compared the measure with Shortest-Path Betweenness centrality. Pandey et al. [32] prioritized genes for pathway enrichment in genome-wide association studies (GWAS) of bipolar disorder (BD) suggesting a four-step approach. Their approach aggregates gene–gene interaction information with main effect associations through a machine learning feature selection and epistasis network centrality analysis. In their approach, they used Eigenvector centrality algorithm to prioritize SNPs/gene in the network. Ran et al. [33] tried to identify the key proteins and the biological regulatory pathways involving in Essential Hypertension (a complex disease) and to investigate the molecular connectivities between these pathways by the topological analysis of the protein-protein interaction network. In their topological analysis, they used Shortest-Path Betweenness centrality, Closeness centrality and Degree centrality.

3. CENTRALITY MEASURES

In 1948, the idea of applying centrality measure to human communication was presented by Bavelas [34]. In 1963, by publishing the book “Applications of Graph Theory to Group Structure” by Flament [35], the interest of applying graph theory to the social sciences was increased. Since then many centrality measures have been introduced in different areas. In this section we study these measures.

3.1. Mathematical Definition

Let \( G(V_1, E_1) \) and \( H(V_2, E_2) \) be two weighted or unweighted, directed or undirected graphs where \( V_1 \) and \( V_2 \) are vertex sets and \( E_1 \) and \( E_2 \) are edge sets in these graphs respectively. If \( G \) and \( H \) are isomorphic and \( M \) is a mapping function from \( V_1 \) to \( V_2 \) then a real-valued function \( S \) will be called structural index if and only if the following condition is satisfied: for each \( v \in V_1 \rightarrow S(v) = S(M(v)) \) where \( S(v) \) denotes the value of \( S(v) \) in \( G \). A real-valued function \( C \) is called a centrality index if the following two conditions are satisfied: I) \( C \) is a structural index II) \( C \) derives an order on the set of vertices or edges [17], so the vertices or edges of the network can be ranked based on this index. By this order we can say between two vertices one of them is more important than the other one.

In this section we present a study of existing different centrality measures. As mentioned earlier, different centrality measures are used to rank the vertices or the edges of a graph. They also are used as a graph measure and assign a numerical value to the whole graph. In this section, we study these measures separately.

3.2. Karate Club Network: A Small Sample Data to Study Different Centrality Measures

For each vertex measure we briefly introduce the measure and visually show each measure’s result on Zachary’s well-known karate club network [2]. We bold the top 20% vertices (≈7 vertices) in rank. The karate club network was collected from the members of a university karate club by Wayne Zachary in 1977 and shows social relationships between members. Two main characters in the club were the club president and a part-time karate instructor (represented by vertices 1 and 33). At the beginning of the study there was an argument between these two characters over the price of karate lessons. Zachary observed the club to explain the split-up of this group following conflicts among the members for a period of three years, from 1970 to 1972. The Zachary club network has become in clustering literatures a well-known benchmark for the use of graphs in finding social groups. Two main groups (clusters) are shown in Fig. (1) with different colors. As mentioned before two vertices 1 and 33 are the most influential vertices in this
Degree centrality can be normalized as in equation (2) where \( n \) denotes the number of vertices in the graph.

\[
\tilde{C}_\text{degree}(v_j) = \frac{C_\text{degree}(v_j)}{n-1}
\]

3.3.2. Leverage Centrality

Joyce et al. introduce the concept of Leverage centrality to find highly essential vertices in the brain network [41]. In contrast with degree centrality, which shows the importance of a vertex based on its degree, Leverage centrality considers the extents of connectivity of a vertex in relation to the connectivities of its neighbors [41]. This measure determines the extent to which the neighbors of a vertex rely on that vertex to access the network resources. By considering neighbors degrees, Leverage centrality gives different information about connectivity of a vertex in contrast with simple Degree centrality. In fact, a vertex with high Degree centrality has not high Leverage centrality if its neighbors have also high degrees. These differences can be inferred from Fig. (3). The result of applying Leverage centrality on Zachary’s karate club network is shown in Fig. (3b). The 20% top-ranked vertices are colored in black.

Leverage centrality is a measure of the relationship between the degree of a given vertex \( k_i \) and the degree of each of its neighbors \( k_j \), averaged over all neighbors \( N_j \).

\[
L(v_j) = \frac{1}{k_i} \sum_{N_j} \frac{k_i - k_j}{k_i + k_j}
\]

Just like Degree centrality, an extension of Leverage centrality could easily be applied to directed graphs by using in-degree and out-degree and compute In-Leverage and Out-Leverage respectively.

3.3.3. Local-Leader and Strict-Leader

Blondel et al. [42] introduced two new concepts based on the degree of a vertex and also the degree of all its neighbors. The simple idea behind these measures is that a vertex is important if it is rich (i.e. has many links) and is among poor vertices (i.e. with low number of links).

A vertex whose degree is larger than or equal to the degree of their neighbors are Local-Leaders and vertices whose degree is strictly larger than the degree of their neighbors are Strict-Leaders. These two measures are calculated using the following equations (4) and (5) respectively. In these equations \( d_i \) and \( d_j \) are the degree of \( v_i \) and \( v_j \) respectively and \( v_j \) is belonged to all \( v_i \)’s neighbors.

Local-Leader \( (v_i) \) = \[
\begin{cases} 
1, & d_j \geq d_i \text{ For all } v_j \\
0, & \text{O.W}
\end{cases}
\]

Strict-Leader \( (v_i) \) = \[
\begin{cases} 
1, & d_j > d_i \text{ For all } v_j \\
0, & \text{O.W}
\end{cases}
\]

Simply a vertex \( v_i \) with degree \( d_i \) is Local-Leaders if all of its neighbors have degrees smaller than or equal to \( d_i \) and is Strict-Leader if its entire neighbors have degrees smaller than \( d_i \). These two concepts are not originally defined as centrality measures, but they can be used as centrality measures.
Fig. (2). A classification tree of vertex measures. Some measures may belong to more than one category. See 3.3.20 for more details.
(a) Degree centrality

(b) Leverage centrality

Fig. (3). The results of Degree centrality (a) and Leverage centrality (b) on karate club network. Top-ranked vertices in each measure are colored in black. This figure shows the differences between simple Degree centrality and Leverage centrality. For example, two vertices 33 and 2 are given a high rank by Degree centrality, but not by Leverage centrality. These two vertices have neighbors with higher degrees. For example, vertex 33 has three vertices 34, 32 and 3 in its neighborhood.

As shown in Fig. (4b) three black vertices are Local-Leaders and only two of them (4 and 5) are Strict-Leaders. These two measures are entirely different from simple Degree centrality. A vertex with a large number of neighbors has a top rank in Degree centrality regardless of its neighbors’ degrees, but if this vertex has a neighbor with a

(a) Local-Leader and Strict-Leader

(b)

Fig. (4). The results of Local-Leader and Strict-Leader (a) in karate club network. In this specific network, both of the Local-Leader and Strict-Leader do the same. As can be extracted from (a) the two most influential vertices 1 and 33 (the club president and a part-time karate instructor) are considered as Local-Leaders and Strict-Leaders as well. These two measures are completely different from Degree centrality and Leverage centrality (see Fig. 3). Local-Leader and Strict-Leader’s definitions are too strict and a few vertices (only two) could fulfil the criteria. (b) shows the result of Local-Leader and Strict-Leader in another random network. In this graph, two vertices 4 and 5 are Strict-Leaders and vertex 3 is Local-Leader. This sample network shows the difference between Local-Leader and Strict-Leader.
higher degree than itself, it is not a Local-Leader and therefore is not a Strict-Leader. These two measures are somehow similar to Leverage centrality in the way they compute the importance of a vertex. To compute the centrality of a given vertex they consider the degree of that vertex and the degrees of all of its neighbors. The result of applying Local-Leader and Strict-Leader on karate club network is shown in Fig. (4a). Only two vertices (1 and 33) met these measures criteria and are colored in black. In this network both measurer's results are the same.

3.3.4. Lobby Index

Hirsch in [43] proposed the concept of Hirsch-Index or h-index as a useful index to characterize both the productivity and impact of a researcher’s or scholar’s published work. This index can also be applied to the productivity and impact of a group of scientists or a scholarly journal. Korn et al. in [44] proposed Lobby Index as a new vertex centrality measure based on Hirsch Index. They argued that this measure considers the efficiency of communication between a vertex and its neighbors. Lobby Index is defined as the largest integer \( k \) such that a vertex \( v_i \) has at least \( k \) neighbors with a degree of at least \( k \). Lobby Index can be calculated using equation (6):

\[
C_{\text{Lobby}}(v_i) = \max\{k: \text{degree}(y_j) \geq k\} \tag{6}
\]

where all the neighbors of \( v_i \) are considered so that degree \( (y_j) \geq \text{degree}(y_2) \).

This measure is completely different from Degree centrality and vertices getting scores by these two measures in a completely different manner. For example, in a network with \( n \) vertices and with star topology the central vertex is the most important vertex and top-ranked in Degree centrality, but in Lobby Index this vertex is not considered powerful because it has not any high-performance neighbors. Usually, but not always, a vertex with high Degree centrality has more chance to have a more Lobby Index score too. Lobby Index is more similar to Leverage centrality and Local-Leader than Degree centrality, because to compute a vertex score it considers its neighborhood. The Lobby Index 20% top-ranked vertices of the karate club network are colored in black and are visualized in Fig. (5a).

3.3.5. Local Clustering Coefficient

Watts and Strogatz introduced the concept of Local Clustering Coefficient of a vertex in order to quantify the local density around each vertex’s neighborhood [21]. This measure calculates by dividing the number of actual edges between \( v_i \)’s neighbor \( (E) \) by the maximum possible edges between \( v_i \)’s neighbors when they all are connected. This measure is calculated based on equation (7), where \( k_i \) is the number of \( v_i \)’s neighbors.

\[
C(v_i) = \frac{E_i}{k_i(k_i-1)/2} \tag{7}
\]

If all neighbors of a vertex connect to each other via an edge, they make a complete graph called clique and then that vertex’s score would be equal to 1. The result of applying this measure on karate club network is shown in Fig. (5b). In this Fig. (11) vertices with a score equal to 1 are colored in black.

Fig. (5). The result of applying Lobby Index on karate club network. The vertices whose neighbors have higher degrees usually get a high Lobby Index score. For example, vertex 14 has vertices with higher degrees such as 34, 1, 3 and 2 in its neighborhood and therefore it is a top-ranked vertex in this network. Vertex 9 is another example. Its neighbor’ set consist of vertices 34, 33, 1, 3. This measure is entirely different from Degree centrality (see Fig. 3). The Top-ranked vertices in Degree centrality (1, 33 and 34) do not get a high score in Lobby Index, because their neighbors are not strong enough. (b) Shows the results of Local Clustering Coefficient on karate club network. All these black-colored vertices get a score equals to 1. Almost all these vertices (except the vertex 8) belong to maximal cliques with three vertices and the vertex 8 is belonging to a maximal clique with five vertices. In other hand, all these vertices make a complete graph with their neighborhood. If we consider two different clusters that are showed in Fig. (1), none of these vertices are placed on the border between two main clusters, which means they do not have any neighbors in different clusters.
3.3.6. Closeness Centrality and Wiener Index

The idea behind Closeness centrality [16] is that a vertex $v_i$ is central if it can interact with all the other vertices in the graph quickly. Closeness centrality uses the length of the shortest paths between all pairs of vertices and computes the centrality score using equation (8):

$$C_{\text{Closeness}}(v_i) = \frac{1}{\sum_{j=1}^{n} \text{dis}_{ij}}$$

where $\text{dis}_{ij}$ denotes the length of the shortest paths between $v_i$ and $v_j$. For example, if a city is modeled as a graph in where the potential places for building a shopping center are vertices of that graph and road between these places are the edges of the graph, the best place to build shopping center is a vertex with maximum Closeness centrality score. The high Closeness centrality score means among all the places in the graph this place can be reached more quickly by other vertices of the graph. It is not important if there are some particular places with a long way to that place, and the best place is a place that can interact with the majority of vertices quickly.

Wiener Index is the inverted form of Closeness centrality. The Wiener Index score of a vertex $v_i$ is equal to the sum of the lengths of the shortest paths between $v_i$ and all the other vertices in the graph. The results of applying Closeness centrality and Wiener Index on karate club network are shown in Fig. (6a, b) respectively. The 20% high-ranked vertices are colored in black.

3.3.7. Eccentricity

Eccentricity centrality [45] computes the score of each vertex by considering the maximum distance, which is the length of the longest shortest path, to all the other vertices in the graph. Simply if all of the shortest paths from a given vertex to all other vertices are computed, the shortest path (i.e., distance) with the longest size is considered as the maximum distance from that given vertex. For example, if we consider a problem of finding the best place in a city to build a hospital and model the problem just like arged in section 3.3.6 the best place to build a hospital is a vertex with the largest Eccentricity centrality score. Unlike Closeness centrality which considers the sum of all distances from the majority of other vertices in the graph for a given vertex (i.e. do not pay attention to the few number of vertices with long distances from that vertex), in Eccentricity centrality the distances between that given vertex to every other vertex in the graph are important. Therefore, one specific vertex can affect the Eccentricity centrality score of entire vertices. Therefore, we can say this measure is sensitive to outlier (noisy) vertices. Eccentricity centrality score of vertex $v_i$ is computed using equation (9):

$$C_{\text{Eccentricity}}(v_i) = \frac{1}{\max \{\text{dist}(v_i,v_j) : v_j \in V\}}$$

The result of applying this measure on Zachary’s karate club network is shown in Fig. (7a). The 20% high-ranked vertices are colored in black. Two vertices, 17 and 5 in this network have long distances to other vertices and have the most effect on Eccentricity scores of all the other vertices.

3.3.8. Radiality

Radiality is a global measure introduced by Thomas et al. [46]. A vertex that has direct edges to those vertices that are not connected directly to each other has a more Radiality score than a vertex that connects to already directly connected vertices. For example, in a graph with star topology a center vertex has largest Radiality. The formal
The definition of Radiality is given by equation (10), where \( D_G \) denotes the diameter of the graph and \( \text{dist}_{ij} \) denotes the length of the shortest paths between \( v_i \) and \( v_j \):

\[
C_{\text{Radiality}}(v_i) = \frac{\sum \text{dist}_{ij} = \text{Diam} + 1}{n-1}
\]

Radiality score of \( v_i \) shows the average reduction in sum of the shortest distances between \( v_i \) to other vertices due to \( v_i \). This is what \( D_G + 1 - \text{dist}_{ij} \) computes. The longest distance between two vertices \( v_i \) and \( v_j \) can potentially be the diameter of the graph (since the diameter of the graph is the maximum longest distance between every two vertices in the graph), and \( D_G + 1 - \text{dist}_{ij} \) show the different between this potential longest distance and the real shortest distance between \( v_i \) and \( v_j \). The result of applying Radiality on karate club network is shown in Fig. (7b). The 20% high-ranked vertices in this network are colored in black.

3.3.9. Stress

In general, many different equal or unequal-length paths may exist between the two vertices \( v_i \) and \( v_j \) in graph \( G \). One or more of these paths from \( v_i \) and \( v_j \) will be the shortest paths. Stress centrality, introduced by Shimbel [47], is based on enumeration of these shortest paths. The goal of defining this measure is to find the amount of "work" done by each site in a communication network. It is supposed that communications are done via shortest paths and proposed counting all the shortest paths in \( G \), which pass through vertex \( v_i \) measures the amount of "stress" which \( v_i \) must handle during the communications. The result of applying Stress centrality on karate club network is shown in Fig. (8a). The 20% high-ranked vertices in this network are colored in black. This measure can be computed by the equation (11):

\[
C_{\text{Stress}}(v_i) = \sum_i \sum_j \sigma_{ij}(v_i)
\]

where \( \sigma_{ij}(v_i) \) denotes the number of shortest paths between \( i \) and \( j \) which pass through vertex \( v_i \). Fig. (9), clearly shows that Stress centrality does not evaluate the amount of communication control each vertex has and therefore a vertex with high Stress centrality score is not necessarily a bottleneck in communicating vertices. In Fig. (9a) three vertices in the middle layer of graph have \( C_{\text{Stress}}(v_i) = 16 \). These vertices are the most central vertices in Stress centrality, but none of them is the bottleneck because by omitting any of these three vertices, the rest of the network still can communicate. Furthermore a vertex in the middle layer of the graph in Fig. (9b) has the same Stress centrality score equals to 16, but this vertex in Fig. (9b) is a critical vertex that can observe all the communication in the graph. This vertex is really the bottleneck of the graph because by omitting this vertex destroys the entire Communication links in the graph. Unfortunately, Stress centrality cannot distinguish these situations.

3.3.10. Shortest-Path Betweenness

Shortest-Path Betweenness centrality is considered as relative Stress centrality [18]. Shortest-Path Betweenness can be computed using the equation (12):


\[
C_{\text{Shortest-Path Betweenness}}(v_i) = \sum_{v \neq v_i} \sum_{\delta ij(v_i)} \delta ij(v_i) 
\]

where \(\delta ij(v_i)\) denotes the proportion of shortest paths between \(i\) and \(j\) that pass through vertex \(v_i\) which shows the probability that a randomly selected geodesic path between \(i\) and \(j\) contains vertex \(v_i\) and can be calculated as:

\[
\delta ij(v_i) = \frac{\sigma ij(v_i)}{\sigma ij}
\]

where \(\sigma ij(v_i)\) denotes the number of shortest paths between \(i\) and \(j\) which pass through vertex \(v_i\). Shortest-Path Betweenness centrality supposes that information almost always passes through geodesic paths and then the Shortest-Path Betweenness of a vertex measures how much information will pass through that vertex, but information such as news, rumour, messages does not know the best path (Shortest path) to spread in the network and it does not always spread only through this path. In real cases, to go from one place to the destination, more likely a message selects the next vertex randomly and checks if it is the destination or not [48]. The result of applying Shortest-Path Betweenness centrality on karate club network is shown in Fig. (8b). The 20% high-ranked vertices in this network are colored in black. Fig. (9) shows that Shortest-Path Betweenness centrality covers the drawback of Stress centrality in distinguishing two different situations argued in the previous section. In Fig. (9a) all three vertices in the middle layer of the graph have

\[
C_{\text{Shortest-Path Betweenness}}(v_i) = \frac{1}{3}
\]

and the vertex in the middle layer of the graph in Fig. (9b) has

\[
C_{\text{Shortest-Path Betweenness}}(v_i) = 1.
\]

Therefore, Shortest-Path Betweenness can truly evaluate the amount of communication control related to each vertex and give more scores to vertices that are bottlenecks of the graph.

3.3.11. Random Walk Betweenness

Random Walk Betweenness assumes that to send information from source to the destination the message is spread randomly in the network until it reaches the destination. In this model, any vertex including source does not know the optimal path to send information to destination. Newman introduced Random Walk Betweenness of a vertex \(v_i\) in [48] as the number of times that a random walk starting at source vertex \(s\) and ending at destination vertex \(t\) contains \(v_i\), averaged over all \(s\) and \(t\). The result of applying Random Walk Betweenness on karate club network is shown in Fig. (10a). The 20% high-ranked vertices in this network are colored in black.

3.3.12. PageRank

The structure of World Wide Web is completely different from a stack of documents. The World Wide Web is Hypertext and related pages are connected to each other via hyperlinks. These links between web pages provides useful structures such as link structure and link text. The PageRank algorithm introduced by Page et al. [49] tries to use preferences of the link structure of the web and assigns a numerical value to each of the web pages to measure its relative importance within the set of all pages. The PageRank algorithm computes the relative importance of the web page based on the graph of the web and has applications in search and browsing. This measure is used by the Google Internet search engine [49]. A slightly simplified version of PageRank is given by equation (14):
Fig. (9). compare Stress centrality and shortest path betweenness.

(a) Random Walk Betweenness centrality

(b) PageRank

Fig. (10). The results of Random Walk Betweenness centrality (a) and PageRank (b) on Zachary's "karate club" network. The high-ranked vertices in each measure are colored in black. Both measures identify the vertices 1 and 33 as the most important vertices.

PageRank \( (v_i) = \frac{1-d}{N} + d \sum_{v_j \in B(v_i)} \frac{PageRank(v_j)}{N_{vj}} \) (14)

where \( B(v_i) \) is a set of pages that point to \( v_i \) and \( N_{vj} \) is the number of pages to which page \( v_j \) points. According to equation (14), each page evenly divides its importance between its entire children (pages to which it is pointing) and this iterative procedure is repeated until convergence occurs. As mentioned in [49] there is a problem in this simplified PageRank which is called rank sink. When two or more pages connect to each other and make a loop and none of them point to any other vertices in the graph and there is at least one other vertices in graph pointing to at least one of the vertices participating in the loop, then rank sink occurs. This loop accumulates weights and does not spread weight at all, because it has no outgoing edges. To solve this problem the original PageRank is changed as below in equation (15):

PageRank \( (v_i) = \frac{1-d}{N} + d \sum_{v_j \in B(v_i)} \frac{PageRank(v_j)}{N_{vj}} \) (15)

This original PageRank models the real behavior of web surfers. Real surfers or web visitors not only follow the existing link in the current page, but also they may choose another page that does not have link from current page. In equation (15) \( d \) is a damping factor that means the surfer follows the existing link in the current page with the probability \( d \) and randomly chooses another page to surf with probability \( (1-d) \). Fig. (10b) shows the result of applying PageRank on karate club network. The 20% high-ranked vertices in this network are colored in black.

3.3.13. Weighted PageRank

Although the usefulness of the original PageRank algorithm has been shown by applying the algorithm by Google search engine, Xing et al. [50] mentioned a problem with it. Not all links in the actual web have the same importance and some links may be more important than the other ones. Some web pages are very popular and other ones always try to get a link from them or give them a link. The Weighted PageRank (WPR) [50] assigns score to a pages based on their popularity instead of dividing the score of each page evenly among all the pages to which it is pointing.

To calculate the score of \( v_i \) in addition to considering the number of in-links of \( v_i \), Weighted PageRank also considers the importance of the collection of pages pointing to \( v_i \) and collection of pages which \( v_i \) is pointing to. Weighted PageRank assigns two different weights \( W^{in}_{(v_i, v_j)} \) and \( W^{out}_{(v_i, v_j)} \) to each link \( (v_i, v_j) \) as bellow:

\[
W^{in}_{(v_i, v_j)} = \frac{I_{v_i}}{\sum_{k \in R(v_j)} I_k} \]
\[
W^{out}_{(v_i, v_j)} = \frac{O_{v_i}}{\sum_{k \in R(v_j)} O_k} \]

(16)  (17)

where \( I_{v_i} \) and \( O_{v_i} \) are the number of in-links and out-links of the page \( v_i \) respectively and \( R(v_j) \) is the collection of pages that \( v_i \) is pointing to them. \( W^{in}_{(v_i, v_j)} \) is the importance of link \( (v_i, v_j) \) based on the collection of pages pointing to \( v_i \) and \( W^{out}_{(v_i, v_j)} \) is the importance of link \( (v_i, v_j) \) based on the collection of pages that \( v_i \) is pointing to them. Based on two weights \( W^{in}_{(v_i, v_j)} \) and \( W^{out}_{(v_i, v_j)} \) for each link \( (v_i, v_j) \) the Weighted PageRank is given by equation (18):
Similar to original PageRank $d$ is a damping factor to model the behavior of real random surfers or real web visitors.

### 3.3.14. Eigenvector Centrality

Eigenvector centrality was proposed by Bonacich [19]. This measure is based on the idea that existing edges between $v_i$ and other vertices that are powerful and influential makes $v_i$ more powerful than existing edges between $v_i$ and less influential vertices. This measure can be easily calculated based on the eigenvectors of the adjacency matrix ($A$). The Eigenvector centrality of $v_i$ is given by equation (19):

$$\lambda e(v_i) = \sum_j A_{ij} e(v_j)$$  \hspace{1cm} (19)

In matrix notation $\lambda e = A e$, where $e$ denotes an eigenvector of $A$ and $\lambda$ is its eigenvalue. In general, many different eigenvalues $\lambda$ may exist for each associated eigenvector solutions since it is necessary that all the entries in the eigenvector be positive the largest eigenvalue is the preferred one. Fig. (11a) shows the results of applying Eigenvector centrality on karate club network.

### 3.3.15. Power Centrality

Fifteen years after proposing Eigenvector centrality Bonacich proposed a more flexible measure named Power centrality [51]. The idea behind this measure is that existing connections between $v_i$ and other vertices that are already well connected without considering $v_i$ does not lead to increase in $v_i$’s power, since $v_i$ does not have any role in communication of these vertices. Instead, if $v_i$ have edges to other vertices that are not well connected by themselves, $v_i$ will have a more important role in communicating these vertices and this makes $v_i$ more powerful. On the other hand, being powerful and having weak neighbors makes someone more powerful. In this way, the idea behind Power centrality is similar to the idea on which Local-Leader is based. Fig. (11) visually shows the differences between Power centrality and Eigenvector centrality on karate club network. Fig. (11b) shows the 20% high-ranked vertices in Power centrality. This measure can be defined by the equation (20):

$$C_{Power}(\alpha, \beta)(v_i) = \sum_j A_{ij}(\alpha + \beta C_{Power}(v_j))$$  \hspace{1cm} (20)

where the parameter $\beta$ determines the effect of $v_i$’s neighbors’ degree and connectivity on Power centrality score of $v_i$. If $\beta = 0$ and $\alpha$ be a constant, then the Power centrality is transformed to the Degree centrality. If $\beta > 0$, then Power centrality is the same as Eigenvector centrality (since a vertex is powerful when its neighbors are powerful too). $\beta < 0$ leads to the original definition for Power centrality and makes a vertex powerful when its neighbors are not well connected.

### 3.3.16. HITS

HITS, also known as hubs and authorities, introduced by Kleinberg is a link analysis algorithm that considers the link structure of the web and assigns scores to each web page and ranks web pages [52]. Pages that are widely cited by the other pages are good authorities and pages that cite many other pages are good hubs. After an iterative process, this algorithm assigns two scores, hub-score and authority-score,
to each page. The hub-score of each page is calculated based on the authority-scores of the entire pages that are pointed by it and the authority-score of a page is computed based on the hub-scores of the entire pages pointing to it. The main idea behind this process is if a page is a good authority then good hubs have links to it and if a page is a good hub then it has links to good authority pages. If $i$ is a vertex of the graph of the web, hub and authority values of $i$ are iteratively updated by the update rules in equations (21) and (22):

$$h_i = \sum_{j \in N} a_j$$  \hspace{1cm} (21)
$$a_i = \sum_{j \in N} h_j$$  \hspace{1cm} (22)

where $N$ is the collection of pages connected to $i$. That is, the authority-score of a page is the sum of all the hub-scores of the pages that point to it.

where $N$ is the collection of pages $j$ connect to $i$. Thus, a page's hub-score is the sum of the authority-scores of pages to which that page points. HITS algorithm was originally defined for directed graphs and in an undirected case, the hub and authority scores of each vertex in the graph are the same. The result of applying the HITS algorithm on karate club network is shown in Fig. (12).

HITS has been used to rank the transactions in association Rule Mining. For example, inspired by the HITS, Sun et al. introduced $w$-support and $w$-confidence, the new weighted support /confidence measures of item sets in databases, and using these measures they propose d a fast new mining algorithm [53]. In weighted Association Rule Mining [54-56], all the transactions are weighted. The classical model of association Rule Mining [57] is a data mining method to explore interesting relations between variables in large databases, which treats every transaction equally. Numerous algorithms have been published to efficiently find association rules [58-60]. Association Rule Mining has been used for tumour prediction [61].

3.3.17. Markov Centrality

This measure was introduced by White et al. [62]. This algorithm finds the importance of a vertex relative to a set of root nodes R. This measure can be calculated using equation (23):

$$I(t|R) = \frac{1}{|R|} \Sigma_{r \in R} m_t$$  \hspace{1cm} (23)

The result of applying Markov centrality on karate club network is illustrated in Fig. (13a).

HITS has been used to rank the transactions in association Rule Mining. For example, inspired by the HITS, Sun et al. introduced $w$-support and $w$-confidence, the new weighted support /confidence measures of item sets in databases, and using these measures they propose d a fast new mining algorithm [53]. In weighted Association Rule Mining [54-56], all the transactions are weighted. The classical model of association Rule Mining [57] is a data mining method to explore interesting relations between variables in large databases, which treats every transaction equally. Numerous algorithms have been published to efficiently find association rules [58-60]. Association Rule Mining has been used for tumour prediction [61].

3.3.17. Markov Centrality

This measure was introduced by White et al. [62]. This algorithm finds the importance of a vertex relative to a set of root nodes R. This measure can be calculated using equation (23):

$$I(t|R) = \frac{1}{|R|} \Sigma_{r \in R} m_t$$  \hspace{1cm} (23)

The result of applying Markov centrality on karate club network is illustrated in Fig. (13a).

![Fig. (12). The result of applying HITS algorithm on karate club network. Since the network is undirected, the hub scores and authority scores of each vertex are equal. The 20% top-ranked vertices are colored in black.](image1)

![Fig. (13). The results of Markov centrality (a) and Sphere Degree (b) on karate club network. The 20% high-ranked vertices in each measure are colored in black.](image2)
3.3.18. Sphere Degree

Sphere Degree score of a given vertex is defined as the number of reachable vertices up to a distance of 2 [11]. This measure can be calculated using the equation (24):

\[ C_{\text{sphere}}(v_i) = \text{out degree}(v_i) + \sum_{w \in \text{neighbors}(v_i)} \text{out degree}(w) \]  

(24)

Fig. (13b) illustrates the result of applying Sphere Degree on karate club network. This measure considers a wider neighborhood than Degree centrality, but usually a vertex with high Degree centrality score has high Sphere Degree score too.

3.3.19. Measures Based on Vitality

Vitality measures are induced measures that can be calculated using any centrality measures introduced in earlier sections of this paper. The goal of a Vitality measure of vertex \( v_i \) in graph \( G \) is to quantify the difference between the centrality of a given graph with and without a given vertex \( v_i \). These measures indicate the effect of removing a vertex of a given graph [17]. For example, let consider Degree centrality. If removing a vertex \( v_i \) of the graph \( G \) causes many changes in the structure of \( G \) and reduces \( G \)'s Degree centrality a lot, then \( v_i \) has a large Vitality score.

3.3.20. Categorizing Vertex Centrality Measures

Centrality measures can be categorized into different groups according to the method they are being calculated [17]. Here, we categorize the introduced vertex measures. Readers must consider that there is an overlap between the categories and some measures may belong to several categories.

3.3.20.1. Measures related to Vertex Neighborhood

To compute the score of the vertex \( v_i \), the measures in this category only consider \( v_i \) and its direct neighbors. Because of the limited scope, these measures are very fast and their computation time is linear. Measures such as Degree centrality, Leverage centrality, Local-Leader, Strict-Leader, Lobby Index and Local Clustering Coefficient belong to this category.

3.3.20.2. Measures Related to Distance

Some measures compute the score of the vertex \( v_i \) based on the distance between \( v_i \) and all the other vertices \( v_j \) in a graph. In general, several paths may exist from \( v_i \) to \( v_j \) and the distance between \( v_i \) and \( v_j \) is the length of the shortest one. Closeness centrality, Wiener Index, Eccentricity centrality and Radiality are those that belong to this category.

3.3.20.3. Measures Related to Betweenness

Some measures such as Stress centrality and Shortest-Path Betweenness centrality compute the score of a vertex \( v_i \) based on the basic assumption that communication between vertices only uses shortest paths. For example, if the graph is modeled as a communication network, and each vertex is a site that tries to communicate (for example send a message) to other vertices in the graph, this message passes through the shortest path from a source to a destination. This assumption is not always true in real networks. For example, spreading gossip in a friendship network can be done via any paths [63].

3.3.20.4. Measures Based on Random Walk

The assumption behind this category of measures is that two vertices communicate by spreading information, message (or anything else) randomly in the graph. For example, if source vertex \( v_i \) wants to send a message to a destination vertex \( v_j \), \( v_i \) randomly sends the message to one of its neighbors (for example \( v_k \)), if \( v_k \) is the destination vertex \( v_j \), then the communication occurs, otherwise \( v_k \) randomly sends the message to one of its neighbors and this process continues until the destination vertex \( v_j \) gets the message. One of the benefits of measures that belong to this category is that they consider more paths than measures like Shortest-Path Betweenness centrality that only consider shortest paths. Random Walk Betweenness, PageRank and Weighted PageRank belong to this category.

3.3.20.5. Measures Based on Feedback

In measures belonging to this category, the importance of a vertex comes from the importance of its neighbors, which are computed themselves based on the importance of their neighbors too. The importance values are propagated through the network and are updated until a convergence occurs. This is a recursive definition; in the first step a predefined default score is assigned to each vertex and then each vertex’s score is iteratively computed based on the scores of its neighbors. PageRank, Weighted PageRank, Eigenvector centrality, Power centrality and HITS are those measures that belong to this category.

Table 1. Centrality measures on graph.

<table>
<thead>
<tr>
<th>Measures on Graph</th>
<th>Extend Equation on Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree centrality</td>
<td>[ C_{\text{degree}}(G) = \sum_{i=1}^{n} \left( C_{\text{degree}}(v_i)<em>{\text{max}} - C</em>{\text{degree}}(v_i) \right) \div (n-1)(n-2) ]</td>
</tr>
<tr>
<td>In-degree centrality</td>
<td>[ C_{\text{in-degree}}(G) = \sum_{i=1}^{n} \left( C_{\text{in-degree}}(v_i)<em>{\text{max}} - C</em>{\text{in-degree}}(v_i) \right) \div (n-1)(n-2) ]</td>
</tr>
<tr>
<td>Out-degree centrality</td>
<td>[ C_{\text{out-degree}}(G) = \sum_{i=1}^{n} \left( C_{\text{out-degree}}(v_i)<em>{\text{max}} - C</em>{\text{out-degree}}(v_i) \right) \div (n-1)(n-2) ]</td>
</tr>
<tr>
<td>Closeness</td>
<td>[ C_{\text{closeness}}(G) = \sum_{i=1}^{n} \left( C_{\text{closeness}}(v_i)<em>{\text{max}} - C</em>{\text{closeness}}(v_i) \right) \div (n-1) \div (n-2) \div (2n-3) ]</td>
</tr>
<tr>
<td>Shortest-path betweenness</td>
<td>[ C_{\text{betweenness}}(G) = \sum_{i=1}^{n} \left( C_{\text{betweenness}}(v_i)<em>{\text{max}} - C</em>{\text{betweenness}}(v_i) \right) \div (n-1) ]</td>
</tr>
</tbody>
</table>

3.4. Edge Measures

Two measures (Shortest-Path Betweenness and Random Walk Betweenness) are generalized on edges to sort the edges of a network. These two measure are used in researches to find clusters in graphs [37, 38].

3.5. Graph Measures

Some measures such as average clustering coefficient of the network, the diameter of the network, average distance between any pair of vertices in the network can be computed and can be used to compare different networks. Vertex centrality measures also can be generalized as a graph measure. These aggregated measures can be calculated using
equation (25). In this equation $C(v^*)$ in the numerator is the centrality value of a vertex with maximum centrality in $G$ and $C(v^*)$ in the denominator is the centrality value of a vertex with maximum centrality on all possible graphs with the same number of vertices as $G$.

$$C(G) = \frac{\sum_{v \in V} (C(v^*) - C(v))}{\max \sum_{v \in V} (C(v^*) - C(v))}$$

For example, we listed five generalized measures in Table 1. Other measures can be calculated in the same manner.

4. DISCUSSIONS

In this section, first, we summarize the discussions made on Zachary's "karate club" network and explain the statistical analysis we have done on the results then, we discuss complexity of computing different measures.

4.1. Analyzing Zachary's "Karate Club" Network

As it can be extracted clearly from the Figs. (3-13), different measures by using different ideas and policies make some different vertices important. For example in Fig. (4a), two most influential vertices 1 and 33 are considered as Local-Leaders. All the other measures except Clustering Coefficient, Wiener Index and Lobby Index put these two vertices among 20% important vertices in Zachary’s karate club network. All measures except Clustering Coefficient, local-Leader and Strict-Leader find some edgy vertices among the 20% top ranked vertices.

We did a statistical analysis on the results of different centrality measures on Zachary's "karate club" network. Our results showed that in general, approximately at 60% of the measures for this network two influential vertices 1 and 33 belong to 20% importance top rank, and as well as the worst measures, which didn’t identify any of the influential vertices 1 and 33 were Clustering Coefficient, Lobby Index and Wiener Index. For a study of the significant relationship between ranks of the vertices in these different centralities, first we arrange the vertices in different centralities then based on the rank correlation coefficient by Kendall’s statistics [64], possibility of relationships between them at 5 and 1 percent significant level were tested. Based on this test, three different groups at different centralities were detected:

1- Metrics in this group have a positive (direct) significant relationship at 5% (or 1%) significant level. The increase of rank of the one vertex will increase the rank of the other related vertex at 95% (or 99%) confidence level and vice versa (Supplementary Table 1).

2- Metrics in this group have a negative (reverse) significant relationship at 5% (or 1%) significant level. By increasing the rank of one of these vertices, rank of another vertex will decrease at 95% (or 99%) confidence level (Supplementary Table 2).

3- Metrics in this group don’t have any significant linear relationship at 5% (or 1%) significant level (Supplementary Table 3).

We found that none of these centralities can be an adequate criterion for the clustering of observations. Based on the Kendall's correlation coefficient we also found that between ranks of the vertices by Eigenvector Centrality and PageRank is a complete direct relationship. Same relationship exists between ranks of vertices by Local-Leader and Strict-Leader. Contrary to these, between ranks of vertices by Closeness Centrality and Wiener Index is a complete reverse relationship.

4.2. Complexity of Calculating Different Centrality Measures

In calculating some measures such as Closeness centrality, Eccentricity centrality, Shortest-Path
Betweenness, Wiener Index the distances between all pairs of vertices in the input network must be computed. Computing these distances can be done using Floyd–Warshall algorithm [65, 66]. The algorithm’s complexity is \( \Theta(n^3) \). Therefore the running times of these measures are high. Fig. (14) and Fig. (15) show the running times of different measures in millisecond on Yeast and Human protein-protein interaction network [67, 68] respectively. The Yeast network contains 2324 vertices which are connected via 4376 edges and the Human Network contains 3989 vertices and 7465 edges.

Some centrality measures are relatively stable when the network is sampled [69]. These measures such as In-degree and Eigenvector centrality can be still be able to use even when some network data is missing. Some centrality measures that are introduced in this paper can be calculate using existing popular centrality tools such as PAJEK [70, 71], UCINET [72], CentiBin [73], CytoScape [74] and CentiLib [75]. We also implemented a new tool, NetCentra, which is freely available to download on LBB website\(^1\). NetCentra covers many early-published measures such as Local-Leader and Strict-Leader, Leverage centrality, Lobby Index and Weighted PageRank that are not covered by similar tools.

5. CONCLUSION

We have studied and compared different centrality measures. We have also analyzed a well-known small sample network (karate club) using different measures. We visually showed top ranked vertices for each centrality measure and briefly discussed the differences between the measures. The analyzing results showed that different measures by using different ideas and policies rank different vertices as important. For example, for the sample karate club network the two most influential vertices, 1 and 33, are considered as Local-Leaders. All the other measures except Clustering Coefficient, Wiener Index and Lobby Index put these two vertices among the 20% most important vertices in karate club network. All measures except Clustering Coefficient, Local-Leader and Strict-Leader find some vertices that are located on the border between two main clusters among the 20% top ranked vertices. We also analyzed the complexity of the different centrality measures and showed the running time of different measures on Yeast protein-protein interaction network. We hope this information would help researchers to choose the suitable measures to analyze their particular biological or social networks.

CONFLICT OF INTEREST

The authors declare no conflict of interest in the publication of this manuscript.

ACKNOWLEDGEMENTS

We would like to appreciate help from LBB's member during this study. Part of this research was supported by Iranian INSF (insf.gov.ir)

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher’s web site along with the published article.

REFERENCES


\(^1\)http://lbb.ut.ac.ir/Download/LBBsoft/NetCentra


Agrawal R, Imielitkis T, Swami A, eds. Mining association rules between sets of items in large databases. ACM SIGMOD Record 1993: ACM.


