Chapter 2
Network Centrality Measures: Role and Importance in Social Networks

Rahul Saxena and Mahipal Jadeja

Abstract Social Networks, in the twentieth century, have emerged as the greatest source of mass communication and possibly the best means of information propagation whether it is through Facebook, Twitter, Instagram, WhatsApp, or any other social platform. With the emerging use of ICT and digital globalization, social connections have increased at a rapid rate and are evolving faster with time. This calls for the need for the identification of entities that hold high importance in the network to strategize for information flow in the network. Network centrality or identifying central nodes in the network is a kind of study of this aspect only. There exist nodes of high value based on parameters like high reachability, high accessibility, closeness, etc. which makes the network traffic induced more toward them. Identification of such nodes helps in decision making for propagating information in short steps or in less number of communications in the network, avoiding information traverse from certain paths or curbing information flow, etc. For example, news published in New York Times will certainly be wildfire fast in comparison to when the news is being circulated in social ties of a community only. In this chapter, we come up with an interesting exploration of the centrality measures concept and theory for a network. The chapter focuses on how different centralities play a crucial role in determining the guiding nodes of any action in the network. Further, some existing prevailing works in the literature will be discussed indicating about how the information on the web can help us in identifying the nodes, set of nodes, or networks that are of prime importance based upon which the information flow in the network takes its shape. Experimental simulations over *SNAP* (*Stanford Network Analysis Platform*) are conducted to understand the concepts in a more applicative manner. The chapter serves as a basic exploration of the network centrality measures, their applications,

R. Saxena (✉) · M. Jadeja
Malaviya National Institute of Technology, Jaipur, India
e-mail: 2019rcp9153@mnit.ac.in

M. Jadeja
e-mail: mahipaljadeja.cse@mnit.ac.in

R. Saxena
Manipal University Jaipur, Jaipur, India

A. Biswas et al. (eds.), *Principles of Social Networking*,
Smart Innovation, Systems and Technologies 246,
https://doi.org/10.1007/978-981-16-3398-0_2

and viewpoints to further investigate the measures over more real-world networks to find out more interesting results.

2.1 Introduction

In this section, we will discuss about what a social network is, what are its components, and how a social network can be visualized as a graphical network. After discussing about these basic preliminaries, we will focus on the centrality analysis and its measures to understand how they can play a crucial role in understanding the evolution of social web graphs.

2.1.1 What is Social Networking?

Before the late 90s, people heavily relied on the information, mainly news-related content on newspapers, television, radios, etc. With the inception of social networking websites in the early 2000s, the paradigm shifted from hand-delivered news material in the form of newspapers, telegrams, etc. to electronic mode [1]. By 2005–06, Facebook and Twitter had entered the arena of the social networking web and are still highly popular social interacting platforms. Other sites like Tumblr, Spotify, Foursquare, and Pinterest tried to fill up specific social networking niches. Since the last decade, social networking websites have captivated the communication needs so well that now it has become an integral part of almost every human being's daily routine life. The impact has been so much powerful that there has been a drastic increase in the number of online apps, news apps, and channels. The news channels have expanded their horizon running from national television broadcasters to live updates on social sites and their own apps [2]. This growth in people's involvement over social networking websites has not constrained to this but has seen a rapid growth in the e-commerce trading [3] as well. The recent emergence of Amazon, Flipkart, and other E-commerce websites has seen a high jump in the digital and retail market. Similarly, YouTube, Netflix, Amazon Prime, etc. (offering various channels and web series) have created a whole new world of viewers. In simple words, social networking has made the world connected remaining at their places, still being able to do the most of what they can.

This discussion brings to the conclusion to define Social Networks as [4]*:* "*Social networking is the use of Internet-based social media sites to stay connected with friends, family, colleagues, customers, or clients. Social networking can have a social purpose, a business purpose, or both, through sites such as Facebook, Twitter, LinkedIn, and Instagram, among others. Social networking has become a significant base for marketers seeking to engage customers.*" Following this, we will now discuss some interesting insights and analyses over the social web graphs.

2.1.2 Social Networks as Graph

Social Networks analysis is stated as the study of investigating social structures, behaviors, and interactions between individuals using network analysis and graph theory. To perform analysis over edge links and nodes of the graph are used to model this interaction and relationship [5]. Each individual in the graph represents a *node* and the connecting edge between these individuals is drawn based upon some notion like *relationship as a friend, similar likes or recommendations,* etc. This connecting tie is of very high importance as this forms the basis of analysis of the network. There is a wide literature available on how to define social ties between the entities in the graph [6–8]. The connecting edges stores the information database which enables us to associate nodes referred to as generating graph embeddings [9].

2.1.3 Why Centrality Analysis?

Given a connected graph simulating a real scenario, it can be visualized that few node positions are more central while the rest are peripheral, in context to the view of the graph at that instance. These key positions help in identifying the nodes of prime importance and deciding upon their roles in the network. This notion was first discussed by Roethlisberger et al. [10]. This information can be of great significance in determining the flow of information, forming various strategies, and many more so that information can be channelized in the network as soon as possible and in a more appropriate manner. Consider a small example of a connectivity network as shown in Fig. 2.1.

In this small network, if *node* 1 is chosen as a central node, then, it may not be a good choice to circulate information as it is more distant to more number of nodes in the network. On the contrary, *node 2* will be a good choice for the node to be central as it is closer (just an edge away) to all the nodes in the graph. However, the

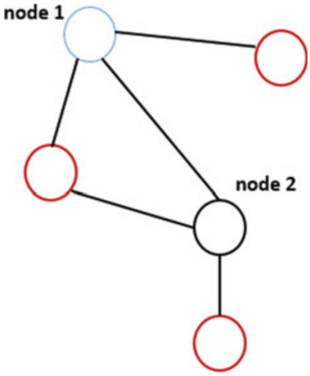


Fig. 2.1 Connectivity network

concept of centrality varies depending upon the objective and purpose. In the same manner, the applicability of the centrality measure may also vary. In the rest of the chapter, we discuss various centrality measures, their advantages, limitations, and their implication aspects.

2.2 Network Centrality: Measures and Concepts

Before delving into the details of network centrality measures and their types, let's first understand the categorization. The network centrality measures can be broadly classified into three classes:

- **Geometric Measures**
  In this class of measure, network centrality for a node is a *function of distances* to other nodes in the network. The importance of a node is determined based on how much a node is approachable to other nodes or how approachable a specific node is by other nodes.
- **Spectral Measures**
  In this class of centrality measure, a node's importance depends upon the eigenstructure of some graph-related matrix. In other words, a node is central depending upon its neighborhood nodes. Here the centrality is a function of the nodes associated with the node of interest.
- **Path-Based Measures**
  Here the centrality measure depends upon the fact that how often a node is visited between a defined source and destination. The concept originates from the idea of *edge betweenness* which gives the number of shortest paths passing over the edge.
  Apart from these, there exist other centrality measures but they may be considered as the variation to the basic versions. The deviation in the applicability procedure is problem and application specific. We will now put our focus on the centrality measures based upon these three categorizations.

2.2.1 Geometric Measures

- **Degree Centrality**

It is the simplest and historically first centrality measure that accounts for the count of the number of ties. It simply indicates the size of an individual's network. For a directed graph network, this centrality measure may have in-degree and outdegree centrality defined separately. For example, in the case of a web page navigation network, where nodes in the network represent the web pages. Incoming edge may be defined as the number of web pages referring to a particular page (say x). The

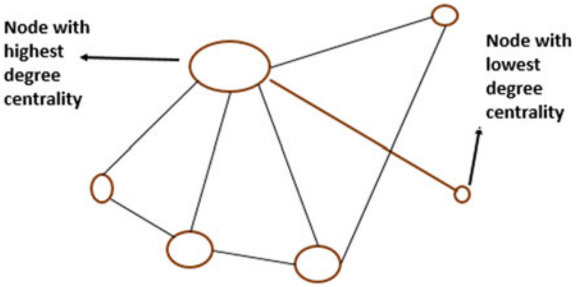

Node with highest degree centrality
Node with lowest degree centrality

Fig. 2.2 Graph Instance representing degree centralities; (i) Node having a large number of connections has high centrality measure and importance. (ii) Node with the smallest size has only one connection and hence has the lowest centrality score. (iii) Rest all nodes have the same centrality score owing to the same number of connections

number of outgoing edges from the web page 'x' stands for all the web pages which are being referred to by 'x'. Depending upon the situation or need, the centrality measure aspect is taken into consideration. Functionally, it can be defined as per the following equation (see Fig. 2.2 for an example).

$$c_{deg}\,(x) = d_{in}\,(x)$$

Liu et al. [11] have taken into account these centrality measures to study the effect of networked criterion-based community engagement on their performance. The in-degree centrality measure analysis in the study accounted for the popularity or measure of how much popularity index a student has in the network. Similarly, the outdegree centrality measure defines how actively a student links to other students in the network. Ergun et al. [12] used the concept of degree centrality to study the effect of social networking structure formed in an Online Learning Environment. Similarly, there are other implications of this centrality measure-based result mentioned in the reported literature from [13–15].

- **Closeness Centrality**

Alexander Bavelas (December 26, 1913 [16]–August 16, 1993) was an American psych sociologist credited as the first to define closeness centrality. Degree centrality only takes into account the connections and weight each link equally important. However, that may not be true for many real-world networks. For example, in a road traffic network, nodes which have high connectivity to many nodes may not be as equally important to the nodes which have reachability to the nodes in the least time. In these situations, nodes that are more central and have smaller distances from other nodes in the network are considered to have high significance. Based on this notion, the functional definition can be given as

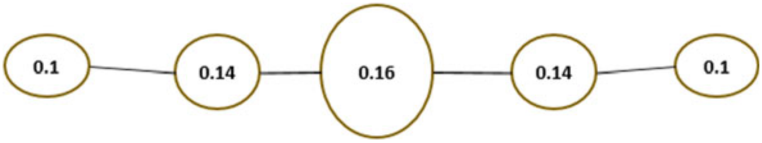


Fig. 2.3 Closeness centrality scores

$$\boldsymbol{c_{close}(x)} = \frac{\mathbf{1}}{\sum_y d(\boldsymbol{y}, \boldsymbol{x})}$$

Here $d(y, x)$ represents the shortest path from node$y$ to x. Let us consider a case as shown in Fig. 2.3.

Here the closeness centrality for the first node is calculated as

$$c_{close}(1) = \frac{1}{1+2+3+4} = 0.1$$

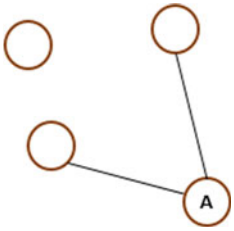
Similarly, for other nodes, the closeness centrality measures are calculated. Clearly, for the middle node, the centrality score will be highest as it has reachability to any node in the network in maximum 2 steps or can reach any node with maximum path length ($c_{close}(3) = {}^{1}\!/_{(2+1+1+2)} = 0.16$). The notion here is how much a vertex can communicate with other nodes without the help of in-between nodes to propagate the message. However, the problem that persists with this centrality measure is if the graph is disconnected, then this centrality measure fails. For example, in Fig. 2.4 shown, the centrality score calculation for any node will be undefined as the distance of any node 'x' with a disconnected 'y' will be defined as ∞.

The closeness centrality measure for node A will be

$$c_{close}\ (A) = \frac{1}{1+1+\ \infty} = 0$$

To counter this, the measure was remodeled by replacing the average distance with the harmonic mean of all the distances.

Fig. 2.4 Disconnected graph



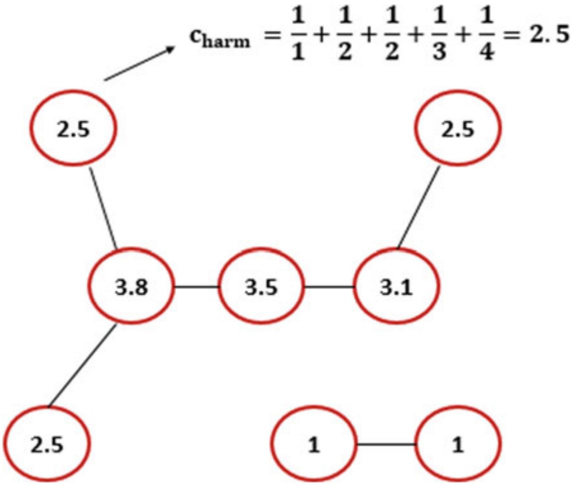

$c_{\mathbf{harm}} = \frac{1}{1}+\frac{1}{2}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4} = 2.5$
2.5
2.5
3.8
3.5
3.1
2.5
1
1

Fig. 2.5 Harmonic centrality scores

$$\mathbf{c_{harm}}\,(\mathbf{x}) = \sum_{\mathbf{d(y,x)}<\,\infty,\quad y\neq x} \frac{1}{\mathbf{d(y,\,x)}}$$

This modification helps in addressing the anomaly caused due to non-connected nodes and thus can be applied to graphs that are not strongly connected (Fig. 2.5).

Kas et al. [17] have proposed an incremental closeness centrality algorithm for dynamic social networks which has continuous addition and removal of edges and nodes. Mateusz et al. [18] used this centrality measure to identify the bus stops common to the several bus lines using the idea of Overlapping Community Structure. Likewise, there are various implications of this centrality measure [19–21].

Geometric measures discussed so far account for the node's importance based on the node's position in the network. In the next section, the discussion is focused upon how the centrality score of a node depends on the neighborhood nodes and how the centrality scores of the neighbor nodes too get influenced by central nodes.

2.2.2 Spectral Measures

The basic intuition of this class of centrality measure is that the nodes in contact with the central nodes have high centrality scores and those far away from these central nodes are considered to be low significance nodes.

- **Eigenvector Centrality**

Unlike degree centrality, the score calculation is done based on the fact that to which kind of nodes, the node 'x' is connected. It is better to be connected with a few popular (well connected) nodes than being connected to many nodes of low importance [22].

This measure of influence of a node proposed by *Phillip Bonacich*, in his 1986 paper *Power and Centrality: A Family of Measures* [23].

$$\boldsymbol{c_{eig}}\ (\boldsymbol{x}) = \frac{1}{\boldsymbol{\lambda}} \sum_{\boldsymbol{y}\to \boldsymbol{x}} \boldsymbol{c_{eig}}\ (\boldsymbol{y})$$

where λ is defined as normalization constant $= \left\| c_{eig} \right\|_2$.

Here c_{eig} converges to dominant eigenvector of adjacency matrix, λ converges to the dominant eigenvalue of adjacency matrix A. Initially, each node is assigned a centrality score of 1. Then, in each successive iteration, the score gets revised as per the formula mentioned above. The matrix formulation of the same can be given as

$$AX = \lambda X$$

To understand it more clearly, let us consider an illustration for the graph shown below.

Matrix A for this graph will be defined as $A = \begin{bmatrix} 0\ 1\ 1\ 0\ 0 \\ 1\ 0\ 1\ 1\ 1 \\ 1\ 1\ 0\ 0\ 1 \\ 0\ 1\ 0\ 0\ 0 \\ 0\ 1\ 1\ 0\ 0 \end{bmatrix}$ and initial centrality score, $c = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$. So for the first iteration, centrality scores will be evaluated as

$$\boldsymbol{\mathit{Iteration}\ 1}\text{:}\quad \mathbf{A.c}^{(0)} = \begin{bmatrix} 1\ 0\ 1\ 1\ 1 \\ 1\ 1\ 0\ 0\ 1 \\ 0\ 1\ 0\ 0\ 0 \\ 0\ 1\ 1\ 0\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \\ 1 \\ 2 \end{bmatrix} \underset{\equiv}{\underline{\boldsymbol{\mathit{def}}}}\ c^{(1)}$$

And, finally defining the normalized scores as

$$\mathrm{c}^{(1)}\big/\left\|\mathrm{c}^{(1)}\right\|_2 = \begin{bmatrix} \frac{2}{\sqrt{(2^2+4^2+3^2+1^2+2^2)}} \\ \frac{4}{\sqrt{(2^2+4^2+3^2+1^2+2^2)}} \\ \frac{3}{\sqrt{(2^2+4^2+3^2+1^2+2^2)}} \\ \frac{1}{\sqrt{(2^2+4^2+3^2+1^2+2^2)}} \\ \frac{2}{\sqrt{(2^2+4^2+3^2+1^2+2^2)}} \end{bmatrix} = \begin{bmatrix} 0.34 \\ 0.68 \\ 0.51 \\ 0.17 \\ 0.34 \end{bmatrix}$$

$$\textbf{\textit{Iteration 2}}\text{:}\quad \mathrm{A.\,c}^{(1)} = \begin{bmatrix} 0\,1\,1\,0\,0 \\ 1\,0\,1\,1\,1 \\ 1\,1\,0\,0\,1 \\ 0\,1\,0\,0\,0 \\ 0\,1\,1\,0\,0 \end{bmatrix} \begin{bmatrix} 0.34 \\ 0.68 \\ 0.51 \\ 0.17 \\ 0.34 \end{bmatrix} = \begin{bmatrix} 1.19 \\ 1.36 \\ 1.36 \\ 0.68 \\ 1.19 \end{bmatrix} \overset{\underline{\mathrm{def}}}{\equiv} \begin{bmatrix} 0.45 \\ 0.51 \\ 1.36 \\ 0.68 \\ 1.19 \end{bmatrix}$$

Progressing in this manner, the final convergence for the centrality scores attained for the example is $c = \begin{bmatrix} 1 \\ 1.41 \\ 1.27 \\ 0.52 \\ 1 \end{bmatrix}$

Carreras et al. [24] used this centrality measure to analyze the spread of the epidemic in a highly decentralized mobile network. Baldesi et al. [25] used this centrality measure to have a cooperative distribution of streamlined content efficiently. Determining the centrality scores help in having the idea of the topology of the network. Like this, there are a number of related articles which discuss the use of this centrality measure. However, this centrality measure has its limitations. Eigenvector centrality will only work for connected and undirected graphs. To counter these, the Katz centrality index was proposed by making a slight modification to the centrality calculation measure discussed.

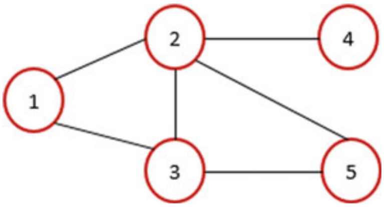
- **Katz's Centrality**

This centrality measure proposed by Leo Katz [26] defines a node's importance by taking into account the total number of walks between a pair of nodes, defined as

$$c_{katz}\,(x) = \beta \sum_{k=0}^{\infty} \sum_{x \to y} \alpha^k \left(A^k\right)_{xy}$$

where α is defined as the attenuation factor ranging from $\left(0, \frac{1}{\lambda}\right)$, λ being the largest eigenvalue of A. The attenuation factor penalizes the connection made with distant neighbors by factor k. A^{k} represents the path between nodes x and y with length k. β is to assign some importance to some particular nodes. Ideally, its value is kept one if none of the nodes in the network is to be assigned some special privilege. For the graph as per Fig. 2.6, the matrix A^{k} can be defined as

Fig. 2.6 Connected graph



$$A^1 = \begin{bmatrix} 0\,1\,1\,0\,0 \\ 1\,0\,1\,1\,1 \\ 1\,1\,0\,0\,1 \\ 0\,1\,0\,0\,0 \\ 0\,1\,1\,0\,0 \end{bmatrix}, \quad A^2 = \begin{bmatrix} 0\,1\,1\,0\,0 \\ 1\,0\,1\,1\,1 \\ 1\,1\,0\,0\,1 \\ 0\,1\,0\,0\,0 \\ 0\,1\,1\,0\,0 \end{bmatrix}^2 = \begin{bmatrix} 2\,1\,1\,1\,2 \\ 1\,4\,2\,0\,1 \\ 1\,2\,3\,1\,1 \\ 1\,0\,1\,1\,1 \\ 2\,1\,1\,1\,2 \end{bmatrix}$$

$$A^3 = \begin{bmatrix} 0\,1\,1\,0\,0 \\ 1\,0\,1\,1\,1 \\ 1\,1\,0\,0\,1 \\ 0\,1\,0\,0\,0 \\ 0\,1\,1\,0\,0 \end{bmatrix}^3 = \begin{bmatrix} 2\,6\,5\,1\,2 \\ 6\,4\,6\,4\,6 \\ 5\,6\,4\,2\,5 \\ 1\,4\,2\,0\,1 \\ 2\,6\,5\,1\,2 \end{bmatrix}$$

The entry in A^3 matrix in second row fifth column indicates there exists 6 paths of length 3 between vertices 2 and 5 [(2,1,3,5), (2,4,2,5), (2,3,2,5), (2,1,2,5), (2,5,3,5), (2,5,2,5)]**.** So, redefining Katz centrality as

$$c_{katz}\,(x) = \alpha \sum_{y \to x} (c_{katz}\,(y) + \beta)$$

This measure looks suitable for *directed acyclic graphs.* Since β is to assign a prioritized weightage to the nodes in the graph and is kept constant initially for a graph, it is α over which the centrality score of the node depends:

- For $\alpha \approx 0$, paths with length > 1 have low contribution and are less influential.
- For a large value of α, Katz scores are more influenced by topology and long paths are penalized gently.
- Measure diverges at $\alpha > \frac{1}{\lambda}$ and hence is the limit.

For the graph shown in Fig. 2.7, the initial centrality scores for the nodes are calculated for $\alpha = 0.85$ and $\beta = 1$ (for all nodes). For high α value, we have more paths greater than length 1 ending at node U than V. Changing the value of $\alpha = 0.15$ will revise the scores making node V's importance score closer to node U as longer paths will be penalized and shorter paths will be more important. Further, it can also be observed that increasing the β value for node B to 2 will make the centrality scores of node A, U, and all the nodes in contact with node B to rise [27].

Zhao et al. [28] used this centrality measure to rank the candidate disease gene and protein–protein interaction to predict the disease occurrence. Zhang et al. [29]

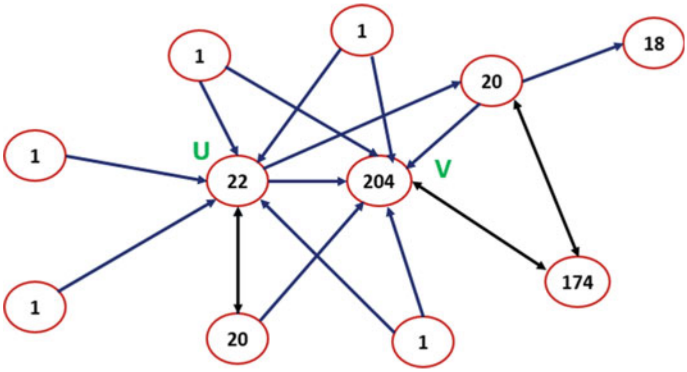


Fig. 2.7 Instance graph with Katz index for each node

use Katz's centrality measure to identify important nodes in a graph where each path has a different weightage. The results were found to have close coherence with the local path index. Similarly, there has been a lot of interesting research articles which have utilized Katz's centrality measure to identify nodes of importance and interest in a network. Landherr et al. [30] have given a comprehensive survey over the usage of various centrality measures and algorithm.

- **Page Rank and HITs Centrality Measure**

PageRank algorithm developed by Larry Page and *Sergey Brin* in *1996* at *Stanford University* is still used by Google to rank web pages. PageRank algorithm assign scores to the nodes in its simplest as

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i} \tag{2a}$$

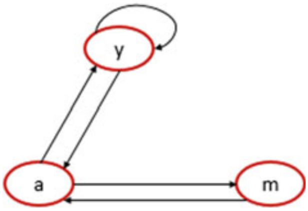
where r_j is the score for the node at time $t + 1$ and $\mathrm{r_i}$ is the importance contribution of node i to node j normalized by its outdegree $d_{i.}$ Normalization is done due to the fact that the same node i also makes a contribution to other nodes as well. The process assigns each node with an initial score *(say 1)* and the scores are updated for each node in every iteration till the time scores for the nodes do not converge, where the convergence criteria is given by

$$\sum_i \left| r_i^{(t+1)} - r^i \right| < \epsilon$$

Based on this, algorithmic steps can be defined as

- Set $r_j = \frac{1}{N'}$ where N are the total number of nodes in the graph.
- 1: $r_j' = \sum_{i \to j} j \frac{r_i}{d_i}$

Fig. 2.8 Graph Instance for PageRank algorithm



- 2: $\boldsymbol{r} \leftarrow \boldsymbol{r}'$
- If $\left|r - r'\right| > \epsilon$: ***goto*** 1.

Tracing the above algorithm over an example as shown in Fig. 2.8.
Score calculation equations over this graph can be defined as

$$r_y = {}^{r_y}\!/_2 + {}^{r_a}\!/_2 \tag{2.1}$$

$$r_a = {}^{r_y}\!/_2 + r_m \tag{2.2}$$

$$r_m = {}^{r_a}\!/_2 \tag{2.3}$$

Based on these flow equations, the algorithm can be run to get the final PageRank scores of the nodes as

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{array}{cccccc} 1/3 & 1/3 & 5/12 & 9/24 & & 6/15 \\ 1/3 & 3/6 & 1/3 & 11/24 & \cdots & 6/15 \\ 1/3 & 1/6 & 3/12 & 1/6 & & 3/15 \\ \textit{Iteration } 0 & \textit{Iteration } 1 & \textit{Iteration } 2 & \textit{Iteration } 3 & & \textit{Final Scores} \end{array}$$

Thus, we get the final scores for all the nodes once the algorithm converges. However, the algorithm may not converge under two conditions:

- The algorithm may get stuck up to *dead ends*, i.e., the flow equations get stuck up to the nodes having no out links. These pages cause the importance to leak out.
- Sometimes the flow equations stuck up, sending and receiving all the flow within a constrained group. This is known as the problem of *Spider traps*. These spider traps absorb all importance.

The solution to these problems was a slight modification to Eq. (2a) as per [31]

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta)\frac{1}{n}$$

where β being the probability of following a link randomly. Thus, (1 − β) is the probability of teleporting, i.e., jumping to a random page to get out of the stuck.

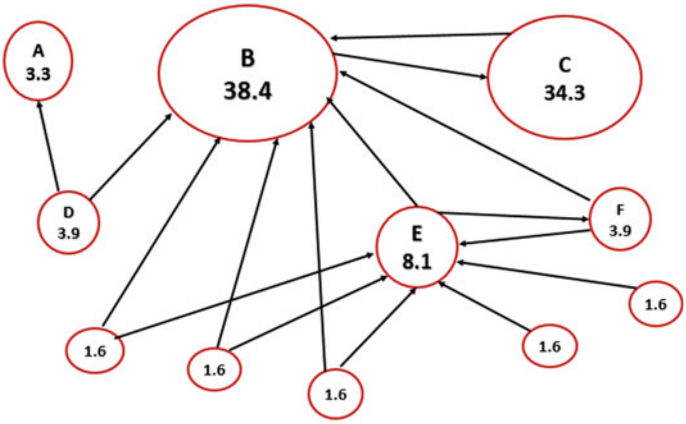


Fig. 2.9 Graph instance with PageRank scores of the nodes

Generally, the values of β range from 0.8 to 0.9. The above equation is equivalent to the dominant eigenvector:

$$r_j = \beta A_r + (1 - \beta)1^T \frac{1}{n}$$

Here A_r represents graph adjacency matrix, in which rows are normalized to row sum one. Figure 2.9 shows an instance of a graph with PageRank scores inside the nodes.

Node *B* with more in links has a more importance contribution from a greater number of nodes in comparison to others. Thus, it has the highest PageRank score. In contrast, node C although has one in link but it is being referred to by a node of high importance in the network; hence, its popularity score also becomes high. With the same explanation, node E although have a number of in links making a contribution in imparting and enhancing its popularity score but it is being referred to by the nodes of low importance in the network.

The above discussion gives rise to the concept of *Hubs* and *Authorities* in a social network and *HITS* centrality algorithm. The basic ideology behind the concept follows from what we have discussed for the PageRank algorithm so far. The pages of interest hold their importance based upon the kind of links (in links or out links) the node exhibit and thus are categorized into two classes:

- ***Authorities*** are nodes containing useful information (like the homepage of newspapers, course homepages, Wikipedia Web page, etc.). They have high incoming links or visits.
- ***Hubs*** are nodes that link to authorities (like List of newspapers, Course bulletin, etc.). These nodes have high outgoing links or visits made.

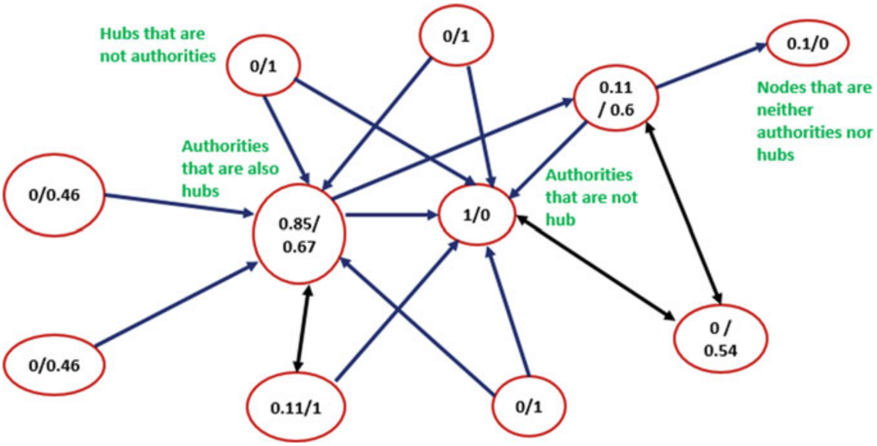

Hubs that are not authorities
0/1
0/1
0.1/0
0.11 / 0.6
Nodes that are neither authorities nor hubs
Authorities that are also hubs
0/0.46
0.85/ 0.67
1/0
Authorities that are not hub
0/0.46
0/ 0.54
0.11/1
0/1

Fig. 2.10 Graph instance with authority and hub scores of the nodes

These two notions of nodes have a mutually recursive definition given as: *A good hub links to many good authorities and a good authority is linked from many good hubs.* Based on this, the *authority* and *hub* scores for a node can be defined as

$$c_{aut}(x) = \sum_{y \to x} c_{hub}(y) \quad and \quad c_{hub}(x) = \sum_{x \to y} c_{aut}(y)$$

Each page *i* thus has two scores; *Authority score*: $\mathbf{a_i}$and *Hub score*: $\mathbf{h_i}$**.** Thus, HITs algorithm can be defined as

- Initialize: $a_j^{(0)} = {}^{1}/_{\sqrt{n}}, \;\; h_j^{(0)} = {}^{1}/_{\sqrt{n}}$
- Keep iterating till convergence:

$$\forall\, i: \; Authority: \;\; a_i^{(t+1)} = \sum\nolimits_{j \to i} h_j^{(t)}$$
$$\forall\, i: \; Hub: \;\; h_i^{(t+1)} = \sum\nolimits_{j \to i} a_j^{(t)}$$
$$\forall\, i: \; Normalize: \quad \sum\nolimits_i \left(a_i^{(t+1)}\right)^2 = 1, \quad \sum\nolimits_j \left(h_i^{(t+1)}\right)^2 = 1$$

In vector notation, these formulas can be expressed as per the following explanation:

- Vector $\boldsymbol{a} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n), \quad \boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n)$
- Adjacency matrix A(n x n): $A_{ij} = 1$ if i → j
- Can rewrite $h_i = \sum\nolimits_{i \to j} a_j \; as \; h_i = \sum\nolimits_j A_{ij} a_j$
- So: $\mathbf{h = A.a}$ and similarly: $\mathbf{a = A^T.h}$

An interesting result to note by combining the two expressions is that the ***authority score a is an eigenvector corresponding to the largest eigenvalue of*** $A^T A$. Similarly, ***hub score h is the eigenvector corresponding to the largest eigenvalue of*** AA^T**.**

Figure 2.10 shows the graphical instance of the nodes having authority and hub scores. Hub scores are accumulated based on the outgoing links to the node. Similarly, authority scores are based on the incoming links to the nodes [27]. Moreover, there are nodes that are acting both as hubs and authorities.

This proposed algorithm has found its importance in several fields. Coppola et al. [32] have used the concept of evaluating PageRank scores to evaluate and optimize the global performance of a swarm-based path evaluation for a robot. Zhao et al. [33] have proposed a motif-based PageRank mechanism to find out the top researchers in a citation network. Yin et al. [34] have proposed a variant of the PageRank algorithm, termed as *Signed PageRank* algorithm, to include both positive and negative recommendations from neighbors simultaneously for product recommendation.

De Blas et al. [35] used a weighted HITs centrality algorithm to identify and rank the most influential nodes by considering the impact of relations between the DMUs (Decision Making Units). There are few others reported in the literature [36, 37] which express high utility of the concept in social networks and varied fields. The centrality measure is highly popular in social networks analysis in the field of influence maximization, influencer detection, etc. and thus the class of algorithms belonging to it have a high significance in the current scenario.

2.2.3 Path-Based Measures

In this category of centrality measures, the centrality scores are defined based on the fact that how often a particular path or edge contributes for a node to make its information travel from one part of the network to other parts. This measure is often referred to as the *betweenness centrality measure* which has a close similarity to the *closeness centrality.* Betweenness centrality is the count of the number of times a given node is encountered in the shortest path between the two nodes. On the contrary, closeness centrality weighs the score based on the shortest path only. For example, if there are three shortest paths from node A to node Z, and node B is along two of them, B will be given two-thirds of a point for A to Z pair.

- **Betweenness Centrality**

The notion of *betweenness centrality*, proposed by *Freeman in 1977* [38], has two conjectures: *edge betweenness* and *node betweenness.* However, the notion of edge betweenness finally coincides with the latter, but provides a useful insight of path contribution or the number of paths through which a node ‘x’ can reach node ‘y’ [27]. Let us consider an example for the same as per Fig. 2.11. The figure shows the number of shortest paths from node A to all other nodes in the network. Based on this, the node flow can be defined as

$$node\ flow = 1 + \sum child\ edges$$

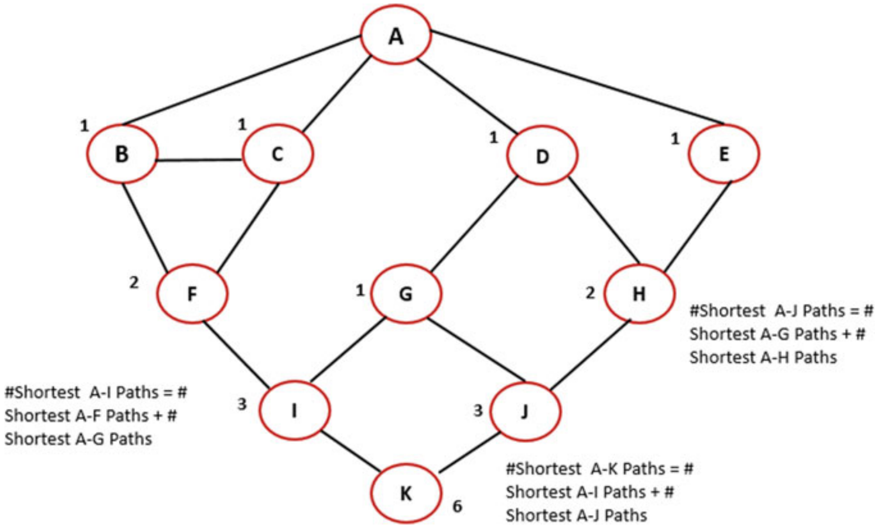

A
1
B
1
C
1
D
1
E
2
F
1
G
2
H
#Shortest A-J Paths = # Shortest A-G Paths + # Shortest A-H Paths
#Shortest A-I Paths = # Shortest A-F Paths + # Shortest A-G Paths
3
I
3
J
K
6
#Shortest A-K Paths = # Shortest A-I Paths + # Shortest A-J Paths

Fig. 2.11 Count of number of the shortest path from node

Further, the flow is split up based on the parent node's contribution. We have to keep exploring the path using BFS (Breadth First Search) mechanism. Multiple paths in between a given source and destination need to be counted fractionally as shown in Fig. 2.12.

This edge betweenness centrality can help us leverage the information to evaluate node betweenness centrality as well. The betweenness centrality for node x can be defined as the probability that the shortest path passes through x. Thus, we have node centrality measure defined as

$$c_{bet}(x) = \sum_{y,\ z\ \neq x,\ \sigma_{yz}\ \neq 0} \frac{\sigma_{yz}(x)}{\sigma_{yz}}$$

$$\sigma_{yz} : \quad \textit{number of shortest paths going from y to z}$$

$$\sigma_{yz}(x) : \quad \textit{number of such paths that pass through x}$$

Removal of nodes in betweenness order causes the network to disrupt as removal of a node with high centrality measure acts as a mediator between the nodes.

As per Fig. 2.13,

- A lies between no other two vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices (A, D), (A, E), (B, E)

There are no alternate paths for these pairs to take without C; thus, C has high betweenness centrality. Consider another example**.**

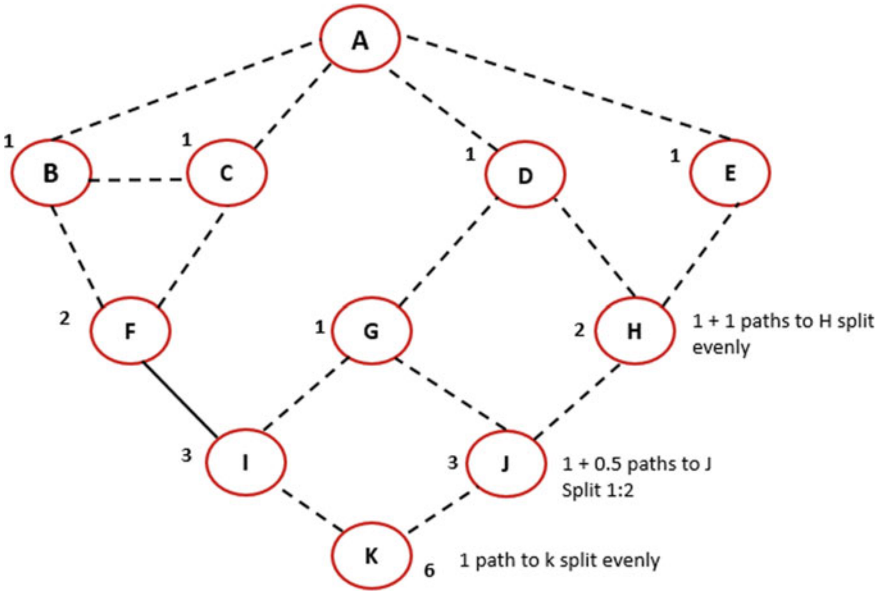


Fig. 2.12 Node flows to the path

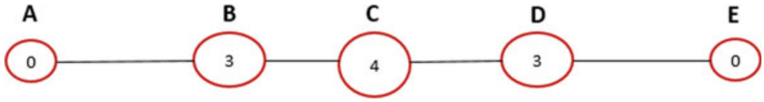


Fig. 2.13 Line graph with betweenness centrality scores of each node

Betweenness centrality score for the graph shown in Fig. 2.14 can be done as follows:

$$\begin{aligned} Betweenness(E) =& A \rightarrow F + A \rightarrow G + A \rightarrow D + A \rightarrow C \\ &+ B \rightarrow F + B \rightarrow G + B \rightarrow D + B \rightarrow C = 8 \end{aligned}$$

Similarly,

$$\begin{aligned} Betweenness(F) =& A \rightarrow G + A \rightarrow D + A \rightarrow C + B \rightarrow G + B \rightarrow D \\ &+ B \rightarrow C + E \rightarrow G + E \rightarrow C + E \rightarrow D = 9 \end{aligned}$$

In the same manner, betweenness centrality score calculations for every node of the graph can be done. Being one of the powerful centrality measure, a lot of applications have used this as a metric to develop a problem-solving approach where the interest is to find out the bridges of the network. Daly et al. [39] used this metric

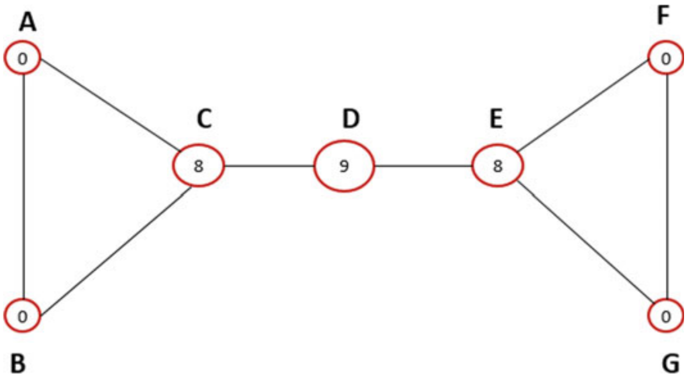


Fig. 2.14 Graph with Betweenness centrality scores of each node

to find out routes in a MANET environment by mapping the concept of small-world dynamics to find out the best message delivery routes. Kazerani et al. [40] discussed how betweenness centrality can be used to model the traffic flow of the cities. Haghir et al. [41] proposed a novel *k-path betweenness centrality* measure where start and endpoints are sampled for path evaluation until we have enough samples to converge. The method is found to have superior performance over the conventional algorithm. Likewise, there are many papers citing the importance of the metric to identify influential or highly important entities in a network that governs the flow of information.

Apart from this categorization of centrality measures, there exists modified versions like applying betweenness and PageRank centrality measure in combination. Then, there exists a notion of *Induced Centrality* measure which is explained at the end of *Katz Centrality* measure which suggests that the importance score of a node raises as soon as it comes in contact with an influential node. Likewise, there are derived versions and variations possible over these centrality measures which provide new evaluation metrics to judge for importance. In the next section, we will see the evaluation of these centrality metrics over real-world graph networks using **SNAP** (***Stanford Network Analysis Platform***).

2.3 Experimental Results and Analysis

To conduct experimental simulations, we have considered *gemsec_facebook_dataset* [42], which contains datasets of 8 different categories of Facebook Page network. The data was collected in November, 2017 through a framework *Graph Embedding with Self Clustering: Facebook* proposed in [43]. The dataset contains a network of various government websites, TV shows' actors, etc. Here the *nodes* represent the *individual entities* while the *edges* between the nodes represent the *mutual likes*. These edge networks have edge lists stored in CSV files where the nodes have been number from index value zero to maintain anonymity. For the purpose of comparative analysis, we considered the graphical network of TV shows where the file contains the edge list and the two TV shows are connected if they are mutually liked upon (undirected graph). Graph contains *3,892 nodes* and *17,662 edges.* The top-10 central nodes identified from various measures are as follows:

These results have been evaluated using SNAP centrality functions. From this score's table, few interesting facts can be determined:

- Node with *node id 2008* has high centrality scores rated by Degree centrality, Closeness centrality, Betweenness centrality, and PageRank centrality measure. Thus, it can be inferred that the TV show is being liked upon the most.
- Eigenvector Centrality scores and HITs centrality scores for the graph have the same top-10 nodes with identical scores. The obvious reason is due to the fact that the graph is undirected and the number of nodes in the shortest path coincides with the hub scores of the node.
- There are a number of nodes in closeness and betweenness centrality that appear in the top-10 central nodes. This is in relation to the first point where the nodes may be ranked.

Different centrality measures have different implications and meanings in the context of the network. In this case, high degree centrality refers to that the node has mutual liking with any other nodes, i.e., a TV show is being mutually liked with many other TV shows. Closeness centrality refers to the close association of the TV shows that have more likings together. Betweenness centrality refers to the shows that are more central in the graph and share likings from one kind of shows to other kinds of shows. In some cases, the centralities too may have a correlation with each other. However, this notion cannot be specific as it entirely depends upon the topology of the graphical network. However, to study upon a highly dense network like this, the centrality trends may be beneficial to identify influential nodes depending upon the objective to be attained. Like high degree nodes will transmit the information and cover the span of the graph. If we want to make the information to pass through particular nodes in maximum routes, betweenness centrality is to be weighted high. If we want to have information localization fast, closeness and eigenvector centrality measures are of high importance. Based upon the scores as per Table 2.1, a scatter plot of **Node ids versus centrality scores** can be determined as per Figs. 2.15 and 2.16.

The degree centrality distribution plot indicates that there are nodes in different regions of the graph having a high degree but are few that lies in the top region of the curve. The majority of the graph settles to the bottom. Closeness centrality seems to have uniform distribution as the closeness centrality takes into account the node's access in minimum distance to other nodes. The curve of the betweenness centrality measure has a smooth increasing trend which suggests there are nodes after every local structure to communicate information from one local region to another. The same is suggested by eigenvector centrality but the increasing trend is rapid as there are a high number of nodes with the shortest path to the majority of nodes in the

Table 2.1 Top-10 central nodes based on various centrality measures

Centrality measure	(Node id, Score)
Degree centrality	(2008, 0.03238), (3254, 0.03238), (3525, 0.02775), (1177, 0.0267), (1673, 0.02621), (3156, 0.02595), (1595, 0.02570), (3122, 0.02570), (2659, 0.02492), (1840, 0.02492)
Closeness centrality	(3254, 0.26098), (2008, 0.25938), (2895, 0.25618), (819, 0.25186), (2751, 0.24777), (211, 0.24425), (160, 0.24422), (3837, 0.24265), (2885, 0.24197), (2035, 0.24014)
Eigenvector centrality	(3525, 0.13852), (1673, 0.13848), (1840, 0.13726), (2659, 0.13690), (3156, 0.13683), (566, 0.13667), (1595, 0.13647), (2036, 0.13641), (1177, 0.13617), (1073, 0.13587)
Betweenness centrality	(3254, 798, 006.33), (2008, 707, 799.88), (819, 609, 148.19), (2170, 565, 442.70), (2751, 565, 010.67), (2895, 522, 963.48), (3038, 307, 027.24), (2682, 294, 900.79), (211, 286, 920.23), (2589, 259, 189.24)
PageRank centrality	(2008, 0.00306), (3254, 0.00305), (2170, 0.00271), (2589, 0.00164), (2076, 0.00161), (412, 0.00161), (2895, 0.001427), (2993, 0.00141), (819, 0.001405), (2524, 0.00136)
HITS centrality	(3525, 0.13852), (1673, 0.13848), (1840, 0.13726), (2659, 0.13690), (3156, 0.13683), (566, 0.13667), (1595, 0.13647), (2036, 0.13641), (1177, 0.13617), (1073, 0.13587)

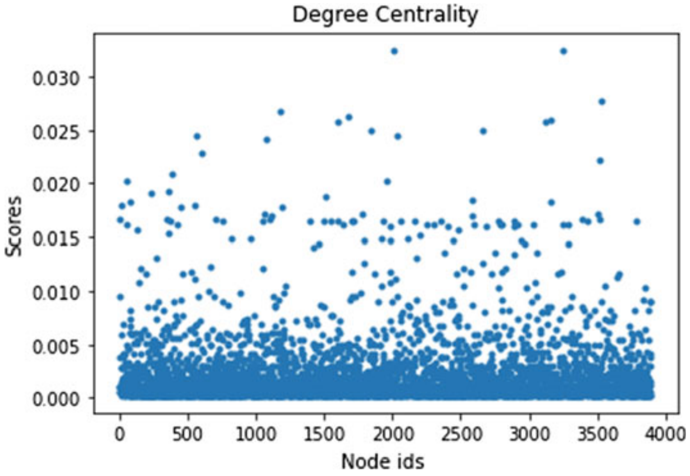


Fig. 2.15 Scatter plot for degree centrality

network. PageRank and HITs centrality have similar trends (Figs. 2.17, 2.18, 2.19, and 2.20).

Another analysis carried out over these centrality measures is how well they are correlated for this graph to each other. Table 2.2 represents the Spearman correlation matrix between the centrality measures. Each cell represents the correlation measure along with the p-value. Correlation between two factors under study is defined in the

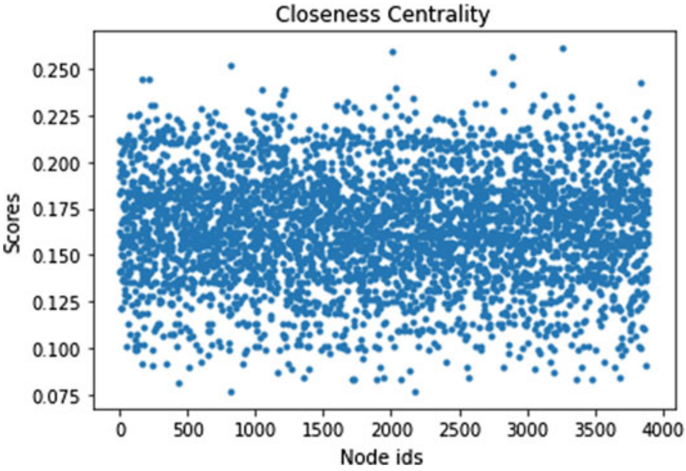


Fig. 2.16 Scatter plot for closeness centrality

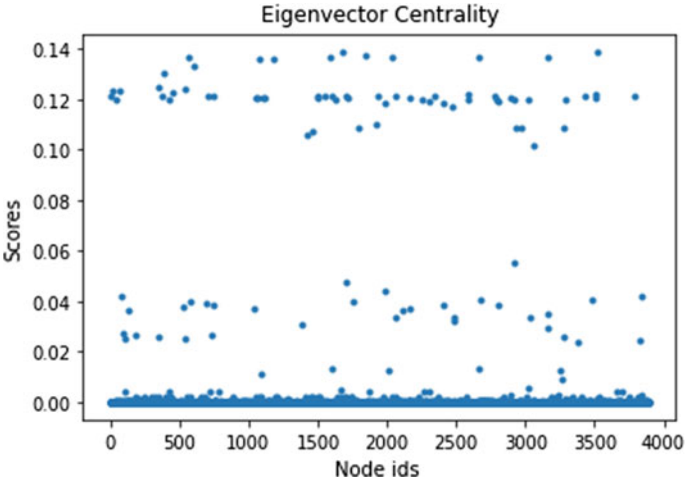


Fig. 2.17 Scatter plot for eigenvector centrality

range [–1, 1]. The strength of the correlation is defined as per the following rules [44]:

- **0.00–0.19—"*very weak*"**
- **0.20–0.39—"*weak*"**
- **0.40–0.59—"*moderate*"**
- **0.60–0.79—"*strong*"**
- **0.80–1.0—"very strong"**

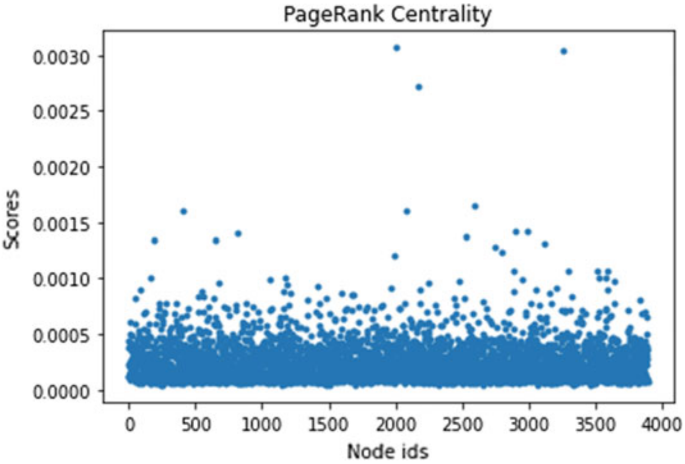


Fig. 2.18 Scatter plot for PageRank centrality

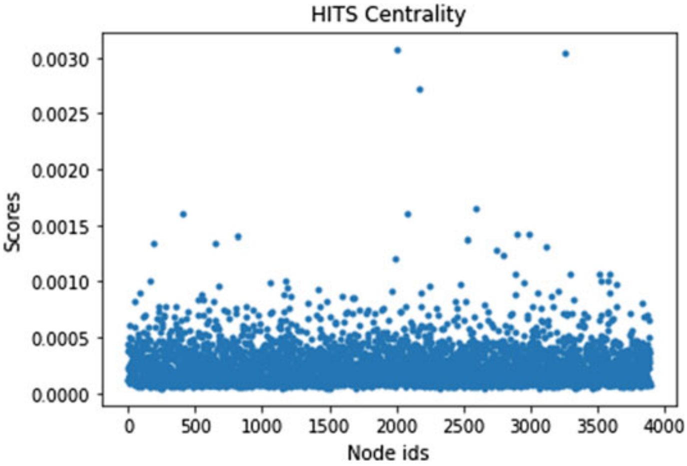


Fig. 2.19 Scatter plot for HITS centrality

The choice for Spearman correlation is due to the fact that it is observed that the centrality distributions are not necessarily normal. The matrix values have been evaluated with the p-value being zero or approximately zero. Degree Centrality has a strong association with the Eigenvector and PageRank Centrality matrix (in the case of undirected network). Similarly, Closeness Centrality has a very high correlation with HITs centrality which suggests that as more nodes accumulate closer, there are more chances of having more hits. There is a strong correlation between the Betweenness as well as Eigenvector Centrality which means that nodes having high betweenness in the network emerge out to be the most liked nodes. Being an

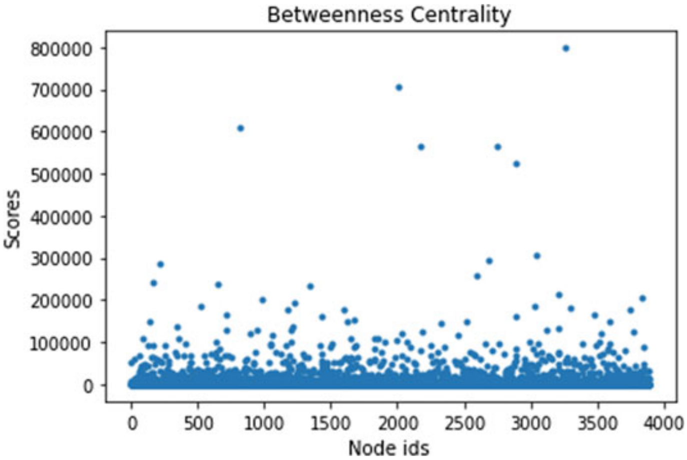


Fig. 2.20 Scatter plot for betweenness centrality

Table 2.2 Spearman correlation matrix

	Degree	Closeness	Betweenness	Eigenvector	PageRank	HITs
Degree		0.552	0.645	0.892	0.892	0.511
Closeness			0.418	0.327	0.327	0.901
Betweenness				0.746	0.349	0.349
Eigenvector					1.0	0.275
PageRank						0.275

undirected graph, Eigenvector and PageRank centrality stand out to be a similar concept as the in links and out links are equated. However, there exists a very weak correlation between the PageRank and HITs Centrality.

2.4 Conclusions

Social Networks being one of the prime sources of connecting real world virtually, the information over it is vast and can be utilized in various ways to earn value from it. The information flow in any network is governed by the number of high importance nodes in the network, and the importance of a particular node is measured on the basis of its position, linking, and its capacity to deliberate the information flow to maximum nodes in the network. This notion gives rise to the concept of network centrality.

This chapter focuses on various centrality measures and deciding criteria to certify a node's importance. Various centrality measures have been categorized into three categories depending upon the referential idea of importance. A detailed investigation has been presented with algorithms and examples for all centrality measures. Further, how a particular centrality measure has been investigated and used by various researchers to solve a particular problem of various domains is also mentioned as and when needed. To understand the concept and significance of centrality, the chapter takes into consideration real-world network's graph (edge list) over which each centrality measure is evaluated, and the results are analyzed over SNAP graphical simulation tool. This detailed analysis and description of the concepts motivate to utilize the knowledge in various domains like protein–protein interaction network, road traffic network, social networks, etc. to evaluate results of significance and identify hotspots of the network. Further, as discussed previously, various combinations of the centrality measures, variation in the conventional centrality measure, etc. can be exploited to identify nodes of high significance and help in building a decision model.

References

1. The Complete History of Social Media: Then And Now. https://smallbiztrends.com/2013/05/the-complete-history-of-social-media-infographic.html. Accessed 19 Apr 2020
2. The rise of Social Media-Our World in Data Homepage. https://ourworldindata.org/rise-of-social-media. Accessed 19 Apr 2020
3. Burt, S., Sparks, L.: E-commerce and the retail process: a review. J. Retail. Consum. Serv. **10**(5), 275–286 (2003)
4. Social Networking Definition-Investopedia. https://www.investopedia.com/terms/s/social-networking.asp. Accessed 19 Apr 2020
5. Graph Theory for skillted. http://blog.soton.ac.uk/skillted/2015/04/05/graph-theory-for-skillted/. Accessed 19 Apr 2020
6. Hogan, B.J.: Networking in everyday life. ON, Canada, University of Toronto, IGI Global, Toronto (2009)
7. Guo, J., Sun, J.: Link intensity prediction of online dating networks based on weighted information. In: 2010 International Conference On Computer Design and Applications, pp. 375. IEEE (2010)
8. Tang, W., Zhuang, H., Tang, J.: Learning to infer social ties in large networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 381–397. Springer, Berlin, Heidelberg (2011)
9. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: a survey. Knowl.-Based Syst. **151**, 78–94 (2018)
10. Roethlisberger, F.J., Dickson, W.J.: Management and the Worker. Psychology Press (2003)
11. Liu, C.C., Chen, Y.C., Tai, S.J.D.: A social network analysis on elementary student engagement in the networked creation community. Comput. Educ. **115**, 114–125 (2017)
12. Ergün, E., Usluel, Y.K.: An analysis of density and degree-centrality according to the social networking structure formed in an online learning environment. J. Educ. Technol. Soc. **19**(4), 34–46 (2016)
13. Zhao, X., Guo, S., Wang, Y.: The node influence analysis in social networks based on structural holes and degree centrality. In: 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), pp. 708–711. IEEE (2017)
14. Gaharwar, R.D., Shah, D.B.: Use of degree centrality principle in deciding the future leader of the terrorist network. Int. J. Sci. Res. Sci. Technol. **4**(9), 303–310 (2018)

15. Jiang, K., Ding, L., Li, H., Shen, H., Zheng, A., Zhao, F., Yu, S.: Degree centrality and voxel-mirrored homotopic connectivity in children with nocturnal enuresis: a functional MRI study. Neurol. India **66**(5), 1359 (2018)
16. Wikipedia. https://en.wikipedia.org/wiki/Alex_Bavelas. Accessed 21 Apr 2020
17. Kas, M., Carley, K.M., Carley, L.R.: Incremental closeness centrality for dynamically changing social networks. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1250–1258 (2018)
18. Tarkowski, M.K., Szczepański, P., Rahwan, T., Michalak, T.P.,Wooldridge, M.: Closeness centrality for networks with overlapping community structure. In: 30th AAAI Conference on Artificial Intelligence, pp. 622–629 (2016)
19. Bergamini, E., Borassi, M., Crescenzi, P., Marino, A., Meyerhenke, H.: Computing top-k closeness centrality faster in unweighted graphs. ACM Trans. Knowl. Disc. Data (TKDD) **13**(5), 1–40 (2019)
20. Wei, B., Deng, Y.: A cluster-growing dimension of complex networks: from the view of node closeness centrality. Phys. A **522**, 80–87 (2019)
21. Goldstein, R., Vitevitch, M.S.: The influence of closeness centrality on lexical processing. Front. Psychol. **8**, 1683 (2017)
22. Bonacich, P.: Some unique properties of eigenvector centrality. Soc. Netw. **29**(4), 555–564 (2017)
23. Neo4j. https://neo4j.com/docs/graph-algorithms/current/labs-algorithms/eigenvector-centrality/. Accessed 19 Apr 2020
24. Carreras, I., Miorandi, D., Canright, G.S., Engø-Monsen, K.: Eigenvector centrality in highly partitioned mobile networks: Principles and applications. In: Advances in biologically inspired information systems, pp. 123–145. Springer, Berlin, Heidelberg (2007)
25. Baldesi, L., Maccari, L., Cigno, R.L.: On the use of eigenvector centrality for cooperative streaming. IEEE Commun. Lett. **21**(9), 1953–1956 (2007)
26. Katz, L.: A new status index derived from sociometrist analysis. Psychometrical **18**(1), 39–43 (1953)
27. CS224W Analysis of Networks. http://snap.stanford.edu/class/cs224w-2018/data.html. Accessed 19 Apr 2020
28. Zhao, J., Yang, T.H., Huang, Y., Holme, P.: Ranking candidate disease genes from gene expression and protein interaction: a Katz-centrality based approach. PloS ONE, **6**(9) (2011)
29. Zhang, Y., Bao, Y., Zhao, S., Chen, J., Tang, J.: Identifying node importance by combining betweenness centrality and katz centrality. In: 2015 International Conference on Cloud Computing and Big Data (CCBD), pp. 354–357. IEEE (2015)
30. Landherr, A., Friedl, B., Heidemann, J.: A critical review of centrality measures in social networks. Bus. Inf. Syst. Eng. **2**(6), 371–385 (2010)
31. Berkhin, P.: A survey on PageRank computing. Internet Math. **2**(1), 73–120 (2005)
32. Coppola, M., Guo, J., Gill, E., de Croon, G.C.H.E.: The PageRank algorithm as a method to optimize swarm behavior through local analysis. Swarm Intell. **13**(3–4), 277–319 (2019)
33. Zhao, H., Xu, X., Song, Y., Lee, D.L., Chen, Z., Gao, H.: Ranking users in social networks with Motif-based PageRank. IEEE Trans. Knowl. Data Eng. (2019)
34. Yin, X., Hu, X., Chen, Y., Yuan, X., Li, B.: Signed-PageRank: an efficient influence maximization framework for signed social networks. IEEE Trans. Knowl. Data Eng. (2019)
35. de Blas, C.S., Martin, J.S., Gonzalez, D.G.: Combined social networks and data envelopment analysis for ranking. Eur. J. Oper. Res. **266**(3), 990–999 (2018)
36. Liu, C., Tang, L., Shan, W.: An extended hits algorithm on bipartite network for features extraction of online customer reviews. Sustainability **10**(5), 1425 (2018)
37. Ka-Wei Lee, R., Hoang, T.A., Lim, E.P.: discovering hidden topical hubs and authorities in online social networks. IEEE Trans. Knowl. Data Eng. 1–1 (2018)
38. Wikipedia. https://en.wikipedia.org/wiki/Betweenness_centrality. Accessed 20 Apr 2020
39. Daly, E.M., Haahr, M.: Social network analysis for routing in disconnected delay-tolerant Manets. In: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing, pp. 32–40. ACM (2018)

40. Kazerani, A., Winter, S.: Can betweenness centrality explain traffic flow. In: 12th AGILE International Conference on Geographic Information Science, pp. 1–9 (2018)
41. Haghir Chehreghani, M., Bifet, A., & Abdessalem, T.: Adaptive algorithms for estimating betweenness and k-path centralities. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1231–1240. ACM (2019)
42. Stanford University. http://snap.stanford.edu/data/gemsec-Facebook.html. Accessed 14 Apr 2020
43. Cornell University. https://arxiv.org/abs/1802.03997. Accessed 14 Apr 2020
44. Statstutor. http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf. Accessed 19 Apr 2020