DOI: 10.22067/cke.2023.74745.1047

P-Centrality: An Improvement for Information Diffusion Maximization in Weighted Social Networks*

Research Article

Najva Hafizi[1], Mojtaba Mazoochi[2], Ali Moeini[3], Leila Rabiei[4], Seyed Mohammadreza Ghaffariannia[5]
Farzaneh Rahmani[6]

Abstract: Online social networks (OSNs) such as Facebook, Twitter, and Instagram have attracted many users all around the world. Based on the centrality concept, many methods are proposed in order to find influential users in an online social network. However, the performance of these methods is not always acceptable. In this paper, we proposed a new improvement on centrality measures called P-centrality measure in which the effects of node predecessors are considered. In an extended measure called EP-centrality, the effect of the preceding predecessors of node predecessors are also considered. We also defined a combination of two centrality measures called NodePower (NP) to improve the effectiveness of the proposed metrics. The performance of utilizing our proposed centrality metrics in comparison with the conventional centrality measures is evaluated by Susceptible-Infected-Recovered (SIR) model. The results showed that the proposed metrics display better performance finding influential users than normal ones due to Kendall's τ coefficient metric.

Keywords: Centrality Measures, Influential Users, Online Social Networks, Susceptible-Infected-Recovered Model

1. Introduction

Online social networks (OSNs) have attracted many users all around the world. Online social networks such as Facebook, Twitter, and Instagram are a framework to exchange and disseminate information [13]. Social graphs are used to describe social networks, in which the nodes are users and the edges indicate the relationship between them such as friendship, retweet, etc. Influence Maximization (IM) is the problem of finding the most influential users in a network to spread out the information to a wider population [1]. An influential user is a super spreader who spreads the information among larger groups of people [15]. Different methods have been proposed to find these influential users. Centrality measures play an important role in finding an influential user for a large amount of data [16]. In these methods, the centrality of the node is used to calculate the importance of any given node in a network. Centrality methods find the most important nodes through the value function on the nodes and the nodes are ranked by these values to identify the influential users. In this paper, we proposed new centrality measures to find influential users in social networks. The importance of the predecessors of a node in a directed graph impacts the importance of the node. K nodes with the highest scores are considered as seed nodes to spread the information through the social network. The layout of the paper is as follows: Section 2 reviews centrality methods that are used to find influential users. In Section 3, the proposed method is described. Then the experimental results are presented in Section 4. Finally, the conclusion is summarized in Section 5.

2. Centrality measures

Social networks are networks in which the vertices are people and the edges represent the social interactions between them, such as friendship, retweet, etc. [14]. These networks are assumed as a graph $G = (V, E)$ in which V is the set of users and E represents the set of edges showing the relationships between the users. A subset S of all users V, $S \subseteq V$, is a set of seed nodes that can spread the influence to a larger group of users. A diffusion model indicates the spreading of information for S through social network G. The influence spread (or influence function) of seed set S, sigma is the expected number of users which are influenced by S at the end of diffusion process [19]. Influence spread is a non-negative set function denoted as $\sigma_{G,M}(S)$, in which its input is the set of seed nodes and its output is the number of final influenced users.

According to the influence spread, the influence maximization problem is defined. Considering a social graph G, a diffusion model M and the number of seed nodes k, influence maximization is an optimization problem of selecting seed set S^* of k users from V maximizing $\sigma_{G,M}(S)$ [19]:

$$S^* = argmax_{S \subseteq V, |S|=k} \sigma(S) \qquad (1)$$

The most common proposed centrality measures to find influential users are listed below:

2.1. Degree centrality

The degree of a node is the number of edges or neighbors the node has in a network. In this method, the user with the highest degree is an influential user. Freeman [3] developed a mathematical model based on the edges connected to a

* Manuscript received, 2022 January 15, Revised, 2022 July 26, Accepted, 2023April 9.

[1] MSC, Department of Algorithms and Computation, Faculty of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran.

[2] Corresponding Author, Assistant Professor, ICT Research Institute (ITRC), Tehran, Iran. **Email:** mazoochi@itrc.ac.ir.

[3] Professor, Department of Algorithms and Computation, Faculty of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran.

[4] Innovation and Development Center of Artificial Intelligence, ICT Research Institute (ITRC), Tehran, Iran.

[5] MSC Student, Department of Algorithms and Computation, Faculty of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran.

[6] Innovation and Development Center of Artificial Intelligence, ICT Research Institute (ITRC), Tehran, Iran.

node. The degree of each node is calculated as follows:

$$C_D(\mathrm{v}_i) = \deg_{in}(\mathrm{v}_i) = \sum_{\mathrm{v}_j \in \mathrm{neighbor}(\mathrm{v}_i)} \mathrm{e}(\mathrm{v}_i, \mathrm{v}_j) \qquad (2)$$

where $\mathrm{e}(\mathrm{v_i}, \mathrm{v_j})$ is the edge between $\mathrm{v_i}$ and $\mathrm{v_j}$. It's equal to 1 when there is an edge between v_i and v_j, otherwise, it is 0 [5]. The time complexity of this centrality is $O(m)$, where $m$the number of edges in the network is. In directed graphs, in-degree or out-degree can be calculated as:

$$C_{D_{in}}(\mathrm{v}_i) = \deg_{in}(\mathrm{v}_i) = \sum_{\mathrm{v}_j \in \mathrm{neighbor}(\mathrm{v}_i)} \mathrm{e}(\mathrm{v}_i, \mathrm{v}_j) \qquad (3)$$

$$C_{D_{out}}(\mathrm{v}_i) = \deg_{out}(\mathrm{v}_i) = \sum_{\mathrm{v}_j \in \mathrm{neighbor}(\mathrm{v}_i)} \mathrm{e}(\mathrm{v}_i, \mathrm{v}_j) \qquad (4)$$

For normalization, degree centrality is calculated as below:

$$\mathrm{C}'_D(\mathrm{v}_i) = \frac{deg(\mathrm{v}_i)}{n-1} \qquad (5)$$

where n is the size of the network. The limitation of degree centrality is that only local information of a vertex in the network is considered to find the importance of the nodes [2].

2.2. Closeness centrality

Closeness centrality considers the distances between a node and all others to calculate the score of the node [9]:

$$C_c(v_i) = \frac{1}{\sum_{v_j \in neighbor(v_i)} d(v_i, v_j)} \qquad (6)$$

where $d(v_i, v_j)$ is the distance between v_i and v_j. The time complexity of this centrality is $O(n^3)$, where n represents the number of nodes and is not applicable to most large networks [10]. For normalization, closeness centrality is calculated as follows:

$$C_c(v_i) = \frac{n-1}{\sum_{v_j \in neighbor(v_i)} d(v_i, v_j)} \qquad (7)$$

2.3. Betweenness centrality

Betweenness centrality is based on the shortest path. It counts the number of shortest paths that pass through a node to rank the given node [6].

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\partial_{st}(v)}{\partial_{st}} \qquad (8)$$

where $\partial_{st}(v)$ is the number of shortest paths from node s to node t passing through node v, and ∂st is the total number of shortest paths from node s to node t. The best time complexity for this method is $O(mn)$ for unweighted networks [7]. Betweenness centrality changes significantly with the variation of the network structure [8].

2.4. PageRank centrality

PageRank is a global measure based on iterative calculation [11]. PageRank centrality is calculated as below:

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in P(u)} PR(v)/L(v) \qquad (9)$$

where N is the count of nodes, $P(u)$ is the list of predecessors of u, $L(v)$ is the number of out-degree links from v, and d is a damping factor usually set to *0.85* [12]. To calculate the PageRank centrality a random walk continues on the edges of the graph until convergence. The time complexity of this centrality is *O(n+m)*.

2.5. K-shell centrality

In k-shell decomposition, the graph is pruned so that all the nodes in the k^{th} shell have a minimum k-degree [4]. In this method, a network is split into a k-shell structure and at the end of the process, each node is placed in a shell. In the first step, all nodes with degree=1 are pruned and placed in $k_s = 1$. This procedure continues until there is no node with degree=1. All of these nodes are placed in the first shell. In the next step, all the nodes with degree=2 are removed and the process is prolonged until all the remaining nodes have a degree larger than 2. The pruning process will continue until all the nodes are placed in the shells. The shell with the highest shell number is called the core of the graph and the nodes that are placed in the core are influential users. The complexity of this centrality is $O(n)$ and it is suitable for large networks.

2.6. Katz centrality

This centrality considers the total number of walks between a pair of nodes to find influential users [20]. In other words, all network paths are considered in this method. Katz centrality $C_K(i)$ of a node i is calculated as follows:

$$C_k(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji} \qquad (10)$$

Element (α_{ij}) of A is equal to 1 if there is a link between i and j and *0* otherwise. Katz centrality has high computational complexity $O(N^3)$.

2.7. Eigenvector centrality

This centrality is based on eigenvector of the greatest eigenvalue of the adjacency matrix. User influence is proportional to the sum of neighbor's influence score [21]. Eigenvector centrality for node i is calculated as below:

$$Ax = \lambda x, \lambda x_i = \sum_{j=1}^{n} a_{ij} x_j, , i = 1, 2, \ldots, n \qquad (11)$$

where A is the adjacency matrix and λ is a constant.

3. Proposed method

In this article, two improvements in centrality measures are proposed called P-centrality measure and an Extended P-centrality measure that is called EP-centrality measure to find influential users in the social network. The idea behind these improvements is the property of "Page rank" which is "Page is more important if it has more links" and "Links from important pages count more", the influence of each node in the graph depends on the influence of previous nodes of that node [23]. It leads to the importance of the predecessors to calculate the node's score in social networks by considering the number of interactions between users. High number of interactions between two users shows a close relationship between them. The closer a relationship, the more impact the predecessor has on the node. To quantify this impact, we first scale the edge weighs with logarithmic scale and then the results are normalized to [0.1, 0.9] range to use them as coefficients for representing the influence of the predecessor. After converting edge weights to coefficients, the score of each node in P-centrality measure is calculated as follows:

$$\mathrm{P-centrality}_i = (\mathrm{centrality}_i * |p_i|) + \sum_{\mathrm{v} \in P_i} (\mathrm{centrality}_v * c_{vi}) \qquad (12)$$

where p_i is the set of predecessors of node i and c_{vi} is the coefficient obtained from the edge weight of v to i. Since the influence of a node itself and its predecessors is not equal, there are two weights in this equation, one of them is $|p_i|$ and the other one is c_{vi}. c_{vi} reduces the value of the predecessor's centrality, because it has a value between $[0.1\ and\ 0.9]$ and $|p_i|$ increases the value of the node's P-centrality. This means that a centrality measure that belongs to the node has a higher impact on the score of the node, but the centrality measures of the predecessors are also effective. In an extended P-centrality measure, called EP-centrality measure, the predecessors in paths where lengths are 2, are also considered. In other words, the predecessor of the predecessor in the graph impacts the score of the node. The impact of the predecessor with a path length of 2 is less than the impact of a direct predecessor. To find a coefficient for the predecessor with length 2, two normalized edge weights are multiplied or simply:

$$c_{wv} = c_{wj} * c_{jv} \tag{13}$$

Because c_{wj} and c_{jv} both have values between $[0.1, 0.9]$, the value of cwv is small; which means a predecessor of a predecessor has less impact on the node compared to a direct predecessor. The score of the node i in the EP-centrality measure is calculated as below:

$$\mathrm{EP-centrality}_i = \mathrm{centrality}_i * (|\mathrm{p}_i| + |\mathrm{q}_i|) + \sum\nolimits_{v \in P_i} (\mathrm{centrality}_v * c_{vi}) + \sum\nolimits_{w \in q_i} (\mathrm{centrality}_w * c_{wi}) \tag{14}$$

where p_i is the set of predecessors of node i, q_i is the set of predecessors of the predecessors for node i and c_{vi} is the coefficient obtained from the edge weight of v to i.

After calculating the scores of all nodes, k nodes with the highest scores are called influential users. The time complexity of the proposed measures depends on the time complexity of the centrality measure which is used.

To improve the effectiveness of P-centrality and EP-centrality measures, a combination of centrality measures can be used as the centrality measure in computing the score of the nodes. By studying the features of basic centrality measures, the combination of degree centrality and K-shell centrality is chosen. This combination is called Node-Power (NP) which is the summation of normalized degree and normalized shell number in k-shell decomposition of a node. The NP of node v is calculated as below:

$$NP_v = \frac{|\mathrm{p}_v|}{\sqrt{\sum_{\mathrm{u \in N(G)}} (|\mathrm{p}_u|)^2}} + \frac{shellNumber_v}{\sqrt{\sum_{\mathrm{u \in N(G)}} (shellNumber_u)^2}} \tag{15}$$

where $N(G)$ is a list of nodes in the graph. Degree centrality is used instead of in-degree in undirected networks. The reason behind normalizing is that the value of having a neighbor and being in a higher shell aren't the same. There are two reasons for choosing shell and degree centrality as NP:

1. K-shell and degree centrality have the highest correlation among pairs of basic centrality measures;
2. K-shell and degree centrality have the lowest time complexity among all basic centrality measures.

P-centrality and EP-centrality using NP for node i are calculated as below:

$$P - NP_i = (NP_i * |P_i|) + \sum_{v \in P_i} (NP_v * c_{vi}) \tag{16}$$

$$EP - NP_i = NP_i * (|P_i| + |q_i|) + \sum_{v \in P_i} (NP_v * c_{vi}) + \sum_{w \in q_i} (NP_w * c_{wi}) \tag{17}$$

where NP_i is the summation of normalized shell number and normalized degree centrality, P_i is the set of predecessors of node i, q_i is the set of predecessors of the predecessors for node i, and c_{vi} is the coefficient obtained from the edge weight of v to i.

4. Experimental results

To demonstrate the effectiveness of the proposed measures, we applied them on four datasets. Table 1 shows the datasets.

Table 1. Some statistical properties for datasets

Dataset	#Nodes	#Edges	Average degree	type
Twitter_mention	1587	11179	7.04	directed
Seventh-graders	29	376	12.96	directed
Residence hall	217	2672	12.31	directed
Twitter_retweet	7073	441927	62.48	directed

4.1. Twitter_mention

This is a real-world network consisting of 1587 users and the edges indicate the mentioned relationship between them. To construct the graph the members of a community found by a community detection algorithm are utilized. All the tweets and replies of this group of users are gathered in a month. Then the mentioned relationships are extracted from the texts. If user A mentions user B in a tweet, there is an edge from A to B. After constructing the graph, the strongest connected component is extracted. The edge weights show how often the user on the left mentioned the user on the right in this month.

4.2. Seventh-graders

In this directed network, nodes represent the seventh-grade students from a school in Victoria and directed edges show the students' preferred classmates [18]. The edge weight shows how often the student on the left chose the student on the right as his favorite. In this graph, the edge weights consist of three values 1, 2, or 3. Figure 1 shows the structure of seventh-graders' network with 29 nodes and 376 weighted edges visualized by the software Gephi. The size of the node shows the value of its in-degree. The color of the edges represents the value of the weights. The distribution of the edge weights shows that most of the edges have a weight of 1 or 3. Normalized values of the edge weights 1, 2, and 3 in seventh-graders network are equal to 0.1, 0.604, and 0.9 respectively.

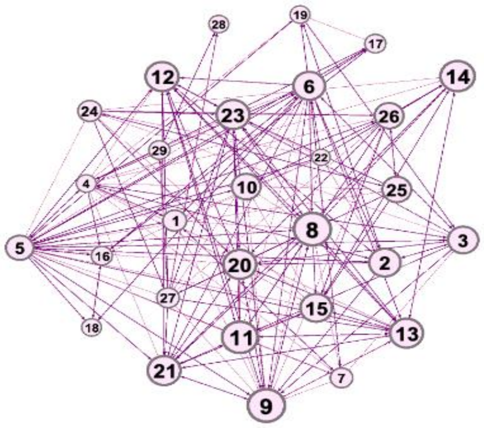


Figure 1. Structure of seventh-graders network

4.3. Residence hall

It is a directed network containing 217 residents, living at a residence hall located on the Australian National University campus, as nodes and the edges represent the friendship between them [18]. Edge weights indicate the strength of each friendship tie.

4.4. Twitter-retweet

This is a real-world network in which the edges indicate the retweet relationship between the users. To construct the graph, the members of a community are utilized. The retweets of the users in a month are used to create the graph. If user A retweets a tweet shared by user B, there is an edge from A to B. The most strongly connected component is extracted as the graph. The edge weights show how often the user on the left retweets the user on the right in this month.

Figure 2 shows the process of finding coefficients to apply the impact of the predecessors on the nodes. Here, 30 random edge weights of the Twitter-mention network are chosen and the coefficients are calculated. The final coefficients which are obtained by normalizing the logarithmic scaled values are shown in black in Figure 2. To compare the effectiveness of centrality, P-centrality, and EP-centrality measures, four most popular centrality measures are chosen: degree centrality, betweenness centrality, PageRank centrality, and k-shell centrality. Some centrality measures like closeness and Katz have high time complexity and are not applicable for most large networks, so we ignored them in the experiments.

Table 2 shows the influence of each node measured by four different centrality methods in Seventh-graders network. The results of degree centrality, betweenness centrality, PageRank centrality, and k-shell centrality are determined for each node of the network. The most influential user found by degree centrality is the node with label 6, but the most influential user found by PageRank centrality is number 21. As shown in the table, the most influential users found by different centrality measures are not the same. Table 3 shows the 5 most influential nodes of Seventh-graders network in descending order. In this comparison, the top 5 users are identified by the normal centrality, P-centrality, and EP-centrality measures. Five centrality measures such as degree centrality, betweenness centrality, PageRank centrality, and k-shell centrality are also used in this experiment.

There is a close similarity between the epidemic spread and information spread in social networks. To evaluate the spread power of the seed nodes, found by different measures, Susceptible-Infected-Recovered (SIR) [22] model is utilized. SIR model is an epidemic diffusion model that is widely used to examine the spreading influence of top ranked nodes. It divides the population into three classes. Each node is placed in one of the Susceptible (S), Infected (I), or Recovered (R) states. In the first step, the seed nodes are infected and all other nodes are placed in the susceptible state. In each step, infected nodes convert susceptible nodes into infected nodes with probability β and the infected node enters into the recovered state with probability γ. The recovered nodes cannot infect others anymore. Figure 3 shows the structure of SIR model.

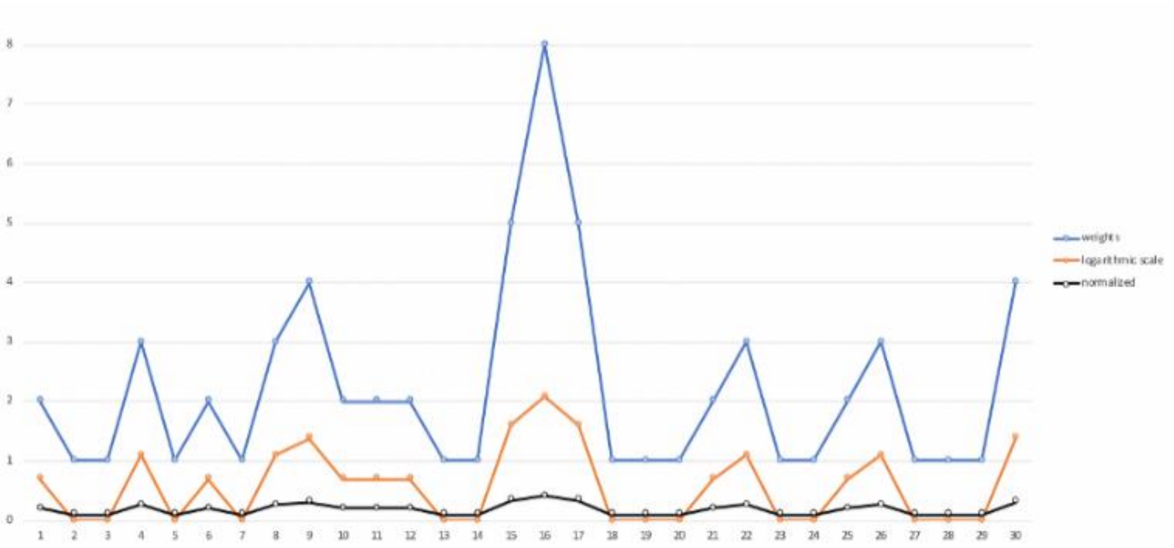


Figure 2. Converting edge weights of 30 random edges in Twitter-mention network to coefficients to determine the impact of the predecessors

Table 2. The influence of each node measured by different centrality measures in Seventh-graders network

Node	Degree centrality	Betweenness centrality	PageRank centrality	K-shell centrality
1	0.86	5.38	0.02	20
2	1.11	31.50	0.03	20
3	1.14	24.05	0.03	20
4	1.07	14.28	0.02	20
5	1.43	59.56	0.03	20
6	1.54	74.52	0.04	20
7	0.75	6.57	0.02	18
8	1.21	12.41	0.06	20
9	1.29	33.10	0.06	20
10	0.93	11.71	0.03	20
11	1.18	15.51	0.05	20
12	0.96	8.87	0.05	20
13	1.07	9.48	0.06	20
14	0.93	2.68	0.04	20
15	0.89	4.65	0.04	20
16	0.46	31.27	0.03	11
17	0.5	6.60	0.02	11
18	0.29	0	0.02	8
19	0.46	5.51	0.02	12
20	1.07	7.20	0.05	20
21	1.14	17.61	0.06	20
22	0.89	27.40	0.02	16
23	1.04	7.31	0.05	20
24	0.93	17.18	0.02	20
25	0.89	22.17	0.03	20
26	1.07	10.53	0.04	20
27	0.86	10.93	0.03	20
28	0.43	3.96	0.02	12
29	0.46	7.04	0.02	12

Susceptible
β
Infected
γ
Recovered

Figure 3. Susceptible-Infected-Recovered (SIR) model

In Table 2, the centrality of "seventh grades" graph nodes is shown with four different measures, and in three of these measures, node number six has the highest score. In Table 3, the top five nodes are ranked using the combination of P-centrality and EP-centrality with the previous four measures.

A SIR model is applied to the Seventh-graders network. The number of susceptible, infected, and recovered nodes are shown in Figure 4 for each step. The results are obtained by averaging 500 iterations on SIR model. The value of β and γ are considered 0.08 and 0.8, respectively. In this experiment, top 5 nodes obtained by degree centrality method are considered as initially infected nodes. At the end of the spreading process with 10 steps, there are 16.7 recovered nodes out of 29 nodes in Seventh-graders network.

To verify the effectiveness of the measures, the correlation between the ranked list generated by SIR model and the ranked list generated by each centrality measure should be discussed. To generate a ranked list using the SIR model, the influence spread of each node should be calculated. To do this, for each node of the graph, we consider the node as an infected user and all other nodes are placed in the susceptible state. The number of recovered users at the end of the diffusion process is the spreading influence of the given node. The spreading influence of the nodes is creating a ranked list generated by SIR. In the next step, the scores assigned to the nodes by the centrality measures result in creating a ranked list generated by the measures. The correlation between these two ranked lists should be calculated. A well-known correlation coefficient, called Kendall (Kendall's τ coefficient) [17], is a criterion for determining the correlation between two same-sized random variables. It is used here to determine the correlation between these two ranked lists. The Kendall's τ considers a set of pairs from two random variables A and B which are two ranked lists here. Any pair $(A_i\,, B_i)$ and $(A_j\,, B_j)$ are either concordant or discordant. They are said to be concordant if both $(A_i > A_j)$ and $(B_i > B_j)$ or both $(A_i < A_j)$ and $(B_i < B_j)$. The pairs are called discordant if $(A_i > A_j)$ and $(B_i < B_j)$ or $(A_i < A_j)$ and $(B_i > B_j)$. Pairs where $(A_i = A_j)$ and $(B_i = B_j)$ are neither concordant nor discordant. The Kendall's τ coefficient is calculated as follows:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{18}$$

where n_c is the number of concordant pairs and n_d is number of discordant pairs, respectively. τ is a number between -1 and $+1$. The closer τ is to $+1$, the higher the correlation between the ranked lists, and therefore the more efficient the measure.

Table 3. The Five most influential nodes identified by centrality measures, P-centrality measures, and EP-centrality measures in Seventh-graders network

Centrality	Measure	A	B	C	D	E
Degree centrality	Centrality	6	5	9	8	11
	P-Centrality	9	8	11	6	21
	EP-Centrality	9	8	11	21	13
Betweenness centrality	Centrality	6	5	7	2	16
	P-Centrality	6	5	9	2	3
	EP-Centrality	6	9	5	2	21
PageRank centrality	Centrality	21	8	13	9	12
	P-Centrality	8	9	21	13	12
	EP-Centrality	8	21	9	13	11
K-shell centrality	Centrality	27	26	25	24	23
	P-Centrality	9	8	12	23	11
	EP-Centrality	8	11	9	21	12

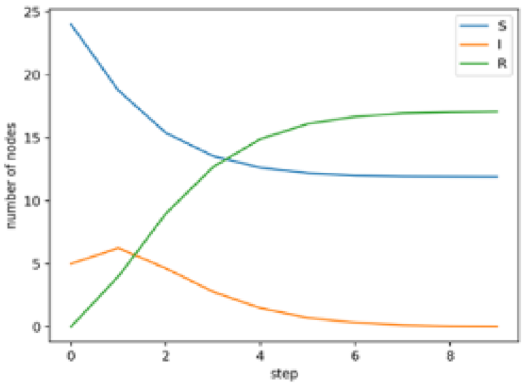


Figure 4. SIR model applied on Seventh-graders network with $\beta\ =\ 0.08$ and $\gamma\ =\ 0.8$. Top 5 nodes found by degree centrality measure are considered as initial infected nodes. The results were obtained by averaging 500 iterations.

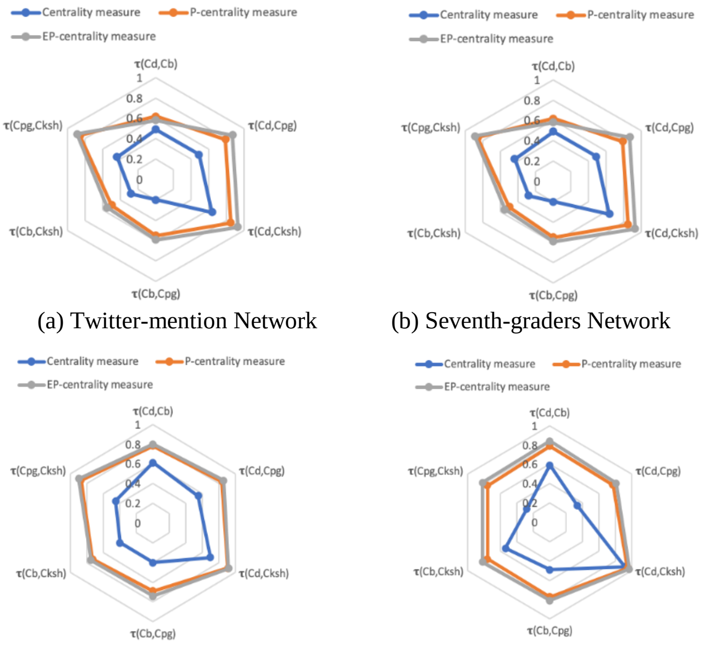
Table 4 shows the correlations between four centralities on normal centrality, P-centrality, and EP-centrality measures. Degree centrality, betweenness centrality, PageRank centrality, and k-shell centrality are compared with each other in this experiment. To find the correlation between the measures, Kendall's τ is used. The average correlations for all the pairs in four networks show that the pair of degree centrality and k-shell centrality has the strongest correlation on all three normal centrality measures, P-centrality measures, and EP-centrality measures. The average correlations show that the weakest correlation belongs to the pair of betweenness and PageRank centrality for normal centrality. In P-centrality and EP-centrality measures the betweenness and k-shell pair has the weakest correlation. The results of Table 4 are shown in Figure 5. The Kendall's τ for six pairs of centrality measures in three groups of centrality measure, P-centrality measure, and EP-centrality measure are shown for four networks. As displayed in the figure, the correlation between the centrality measures in P-centrality and EP-centrality is greater than normal centrality. This means that considering the predecessors of the nodes increase the correlation between the measures. The highest increasing rank for correlation between centrality measures from normal centrality to P-centrality or EP-centrality belongs to the pair of PageRank and k-shell. In this pair, τ value has increased by more than 102% from normal centrality to P-centrality and more than 115% from normal centrality to EP-centrality. The pair of degree and k-shell centrality causes the minimum increase in which τ value has increased by 14.91% and 21.08% from normal centrality to P-centrality and from normal centrality to EP-centrality, respectively. Figure 6 shows the comparison between normal centrality, P-centrality and EP-centrality using Kendall's τ. Experiments are done on datasets using four centrality measures: Degree centrality, betweenness centrality, PageRank centrality, and k-shell centrality. The τ values are compared on different values of β. On all the datasets, EP-centrality measures have the highest correlation with real spreading process SIR. Moreover, the improved centrality measures which consider direct predecessors to calculate the score of the nodes (P-centrality) are more effective than the simple centrality measures but less effective than the extended ones on all the datasets. Betweenness centrality has the lowest τ value and k-shell using EP-centrality has the highest τ value among centrality measures on all the datasets. Betweenness centrality causes the maximum increase in τ in which τ value has increased by more than 158% from normal centrality to EP-centrality and the minimum increase belongs to PageRank centrality in all the networks. Table 5 shows the increasing rate for different centrality measures based on equation 18. In the Seventh-graders and Residence hall data set, the transformation of normal betweenness centrality to EP-centrality has the greatest improvement in τ value. In the Twitter-mention data set, the same conversion for the PageRank centrality and in the Twitter-retweet for the same conversion were the maximum in the k-shell centrality. The lowest amount of improvement for the three datasets of the Seventh-graders, Residence hall, and Twitter-retweet was for the conversion of the normal PageRank centrality to the PageRank in the EP-centrality mode, and for the Twitter-retweet dataset, the same conversion was minimal in the k-continuous centrality. This is in case that the transformation of centrality from P-centrality to EP-centrality is not considered and only the transformation from normal centrality is considered.

Figure 7 shows the comparison between centrality measures in normal centrality, P-centrality, and EP-centrality measures. The Kendall's τ values are obtained by averaging 100 iterations on SIR model. As displayed in the figure, the best and the worst centrality measures are not the same in different networks. K-shell centrality has the highest τ in Twitter-mention and the lowest τ in Twitter-retweet network. Moreover, as mentioned, the increasing rate of different centrality measures from normal centrality to P-centrality and EP-centrality are not the same.

Figure 8 shows Kendall's τ values obtained by the ranked lists generated by SIR model and the ranked lists generated by centrality measures, P-NP, and EP-NP measures. Using NP in P-centrality and EP-centrality increases the value of τ. P-NP and EP-NP have the highest τ among all the centrality measures in all the networks. Moreover, EP-NP outperforms P-NP on all the datasets. The average τ values on all datasets show that the betweenness centrality has the lowest τ and EP-NP has the highest τ values. The average τ values of EP-NP is 280% higher than τ value of betweenness centrality in all the networks. Figure 9 shows the Kendall's τ values obtained by the ranked lists generated by SIR model and the ranked lists generated by centrality measures, P-centrality, and EP-centrality measures with degree, k-shell and NP. In this experiment, the impact of using a combination of k-shell and degree centrality is shown. The combination of k-shell and degree centrality compared to using them separately in P-centrality and EP-centrality shows that using NP increases the value of τ in all the networks. NP-P increases τ values of degree and k-shell in P-centrality with 4.8% and 2.8%, respectively. Moreover, the increase of τ in EP-NP compared to degree and k-shell centrality is 3.6% and 0.2%.

Table 4. The correlations between centrality measures in four networks

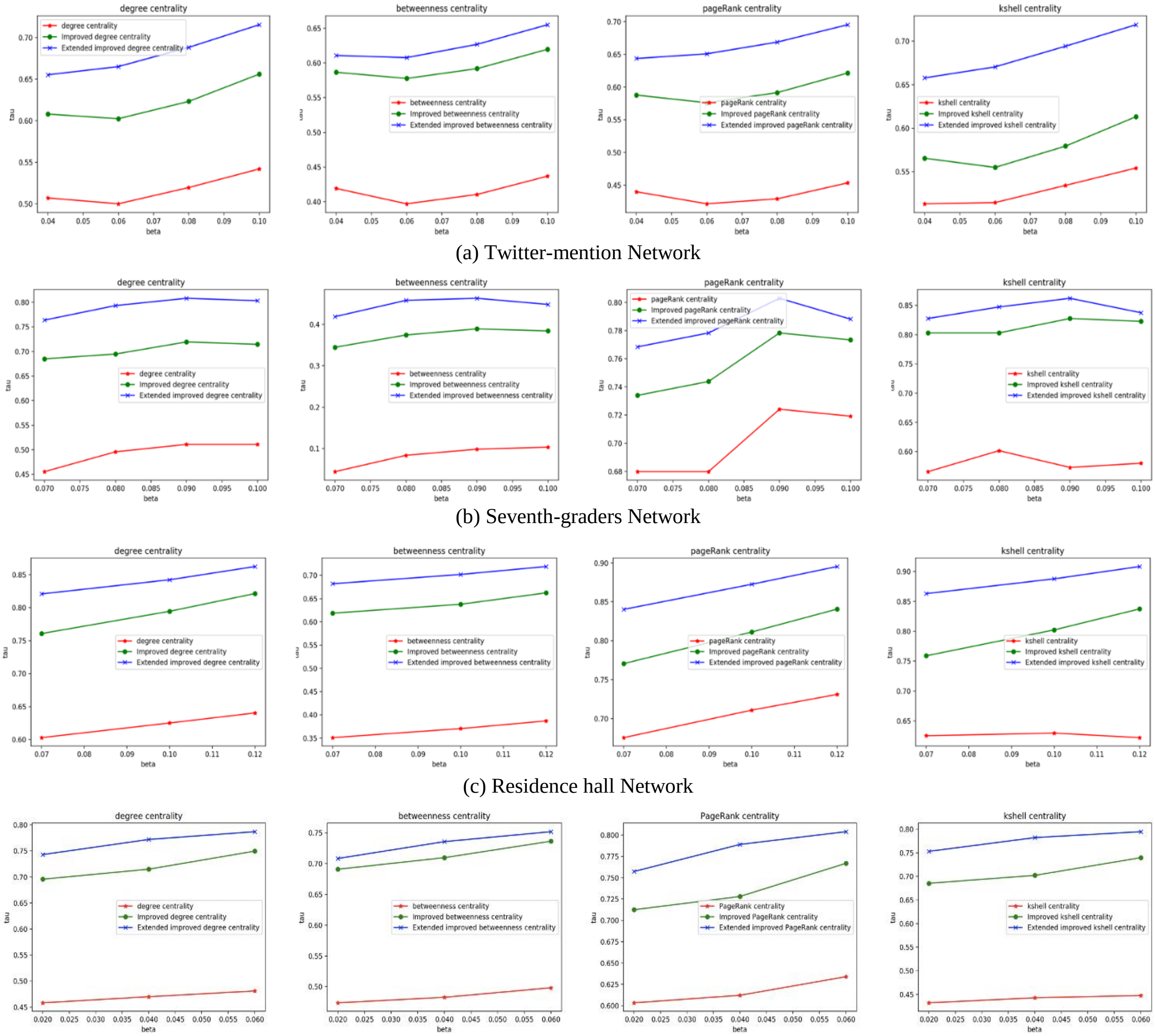
Network	**Measure**	$\boldsymbol{\tau(C_D, C_B)}$	$\boldsymbol{\tau(C_D, C_{pg})}$	$\boldsymbol{\tau(C_D, C_{ksh})}$	$\boldsymbol{\tau(C_B, C_{pg})}$	$\boldsymbol{\tau(C_B, C_{ksh}}$	$\boldsymbol{\tau(C_{pg}, C_{ksh}}$
Twitter-mention	Centrality	0.73	0.53	0.93	0.53	0.67	0.50
	P-Centrality	0.81	0.81	0.87	0.77	0.69	0.74
	EP-Centrality	0.84	0.84	0.95	0.80	0.80	0.83
Seventh-graders	Centrality	0.49	0.49	0.64	0.20	0.28	0.44
	P-Centrality	0.62	0.79	0.85	0.55	0.50	0.86
	EP-Centrality	0.58	0.87	0.93	0.59	0.55	0.89
Residence hall	Centrality	0.61	0.55	0.70	0.40	0.40	0.45
	P-Centrality	0.78	0.84	0.91	0.69	0.73	0.87
	EP-Centrality	0.80	0.86	0.92	0.74	0.75	0.90
Twitter-retweet	Centrality	0.59	0.34	0.92	0.49	0.54	0.28
	P-Centrality	0.79	0.77	0.95	0.77	0.76	0.76
	EP-Centrality	0.84	0.81	0.97	0.81	0.82	0.82



(a) Twitter-mention Network (b) Seventh-graders Network

(c) Residence hall Network (d) Twitter-retweet Network

Figure 5. Radial plot of the Kendall τ for six pairs of centrality measures in three groups of centrality measure, P-centrality measure, and EP-centrality measure analyzed in four social networks



(a) Twitter-mention Network

(b) Seventh-graders Network

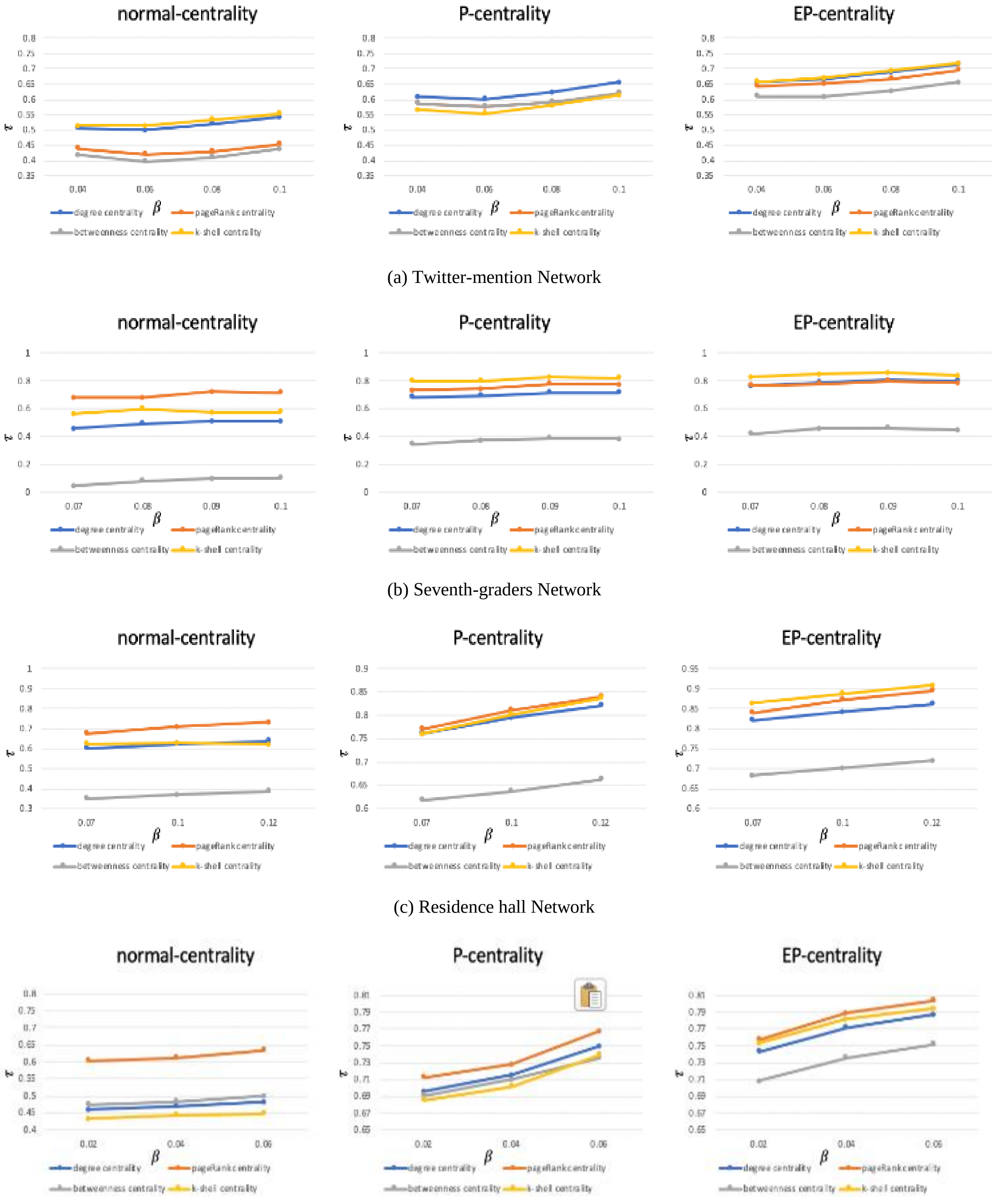
(c) Residence hall Network

(d) Twitter-retweet network

Figure 6. The Kendall's τ values obtained by the ranked lists generated by centrality measures and the ranked lists generated by SIR model. The results were obtained by averaging 100 iterations on the SIR model, γ = 0.8, with different values of β.

Table 5. Increasing rate for different centrality measures

Centrality	**Measure**	**Twitter-mention**	**Seventh-graders**	**Residence hall**	**Twitter-retweet**
Normal → P-centrality	Degree	20.37%	42.58%	27.20%	53.28%
	Betweenness	42.83%	352.24%	73.09%	46.87%
	PageRank	36.29%	8.08%	14.42%	19.36%
	k-shell	9.32%	40.30%	27.81%	60.80%
Normal → EP-centrality	Degree	31.71%	60.56%	35.17%	63.34%
	Betweenness	50.33%	441.79%	89.68%	50.93%
	PageRank	52.45%	11.95%	23.19%	27.08%
	k-shell	29.60%	45.39%	41.68%	76.12%
P-centrality → EP-centrality	Degree	9.42%	12.61%	6.26%	6.56%
	Betweenness	5.25%	19.80%	9.59%	2.76%
	PageRank	11.86%	3.58%	7.67%	6.47%
	k-shell	18.55%	3.63%	10.85%	9.53%



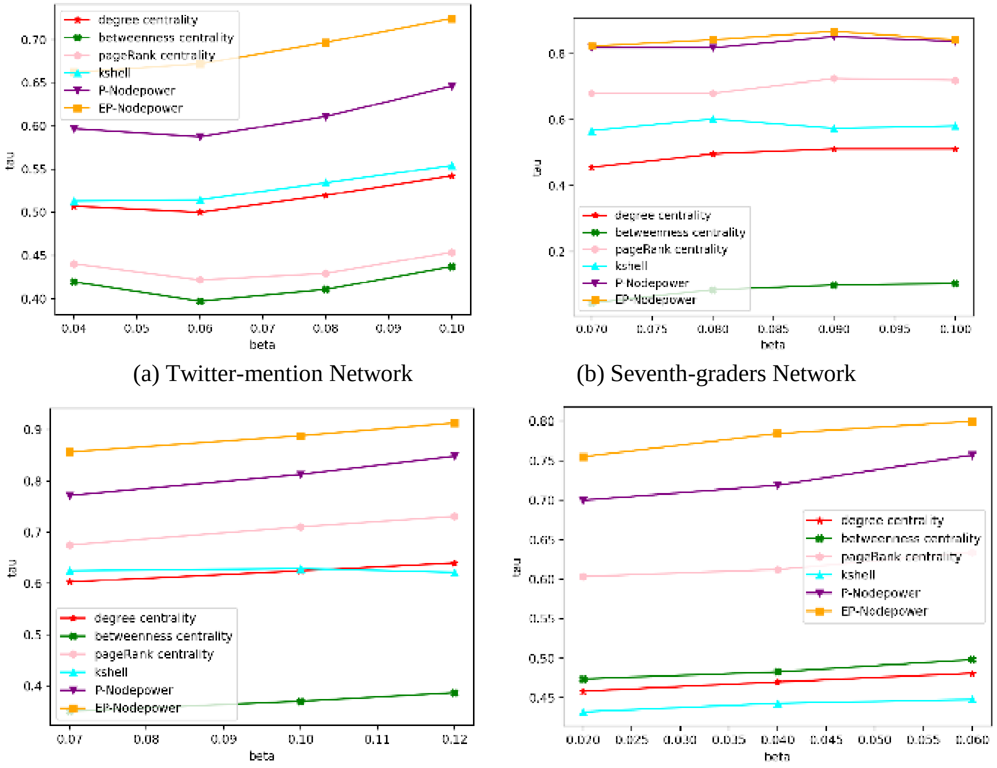
(a) Twitter-mention Network

(b) Seventh-graders Network

(c) Residence hall Network

(d) Twitter-retweet network

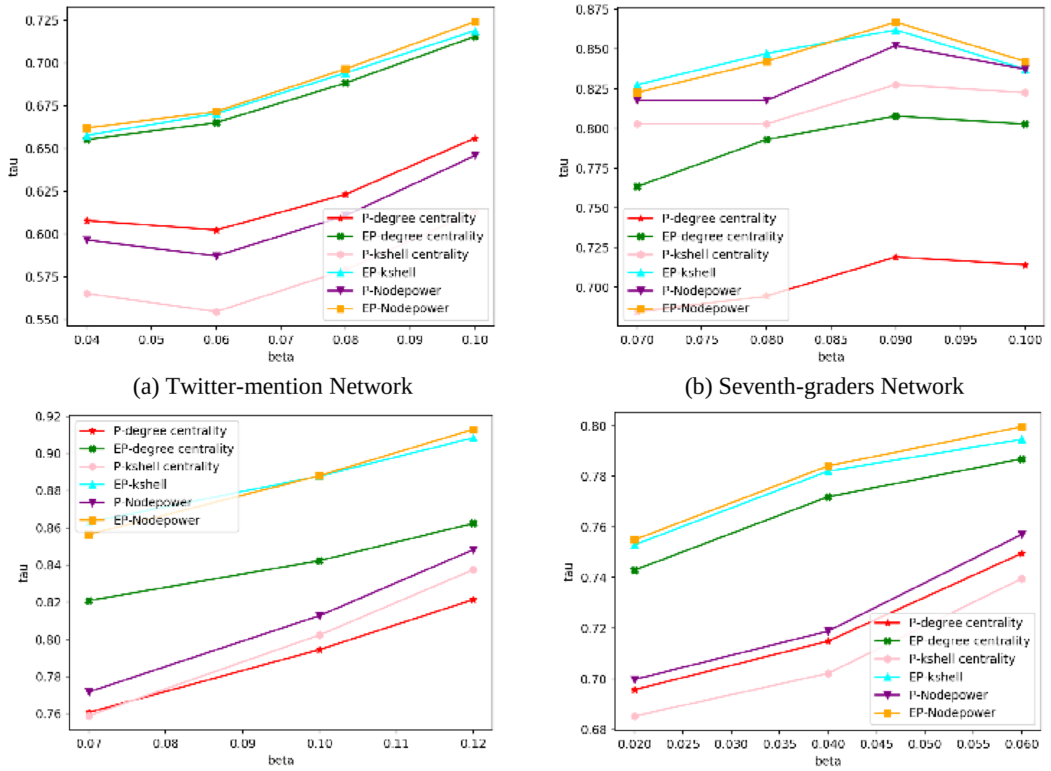
Figure 7. The Kendall's τ values obtained by the ranked lists generated by SIR model and the ranked lists generated by centrality measures, P-centrality measures, and EP-centrality measures. The results were obtained by averaging 100 iterations on the SIR model, γ = 0.8 with different values of β.



(a) Twitter-mention Network (b) Seventh-graders Network

(c) Residence hall Network (d) Twitter-retweet Network

Figure 8. The Kendall's τ values obtained by the ranked lists generated by SIR model and the ranked lists generated by centrality measures, P-NP, and EP-NP measures. The results were obtained by averaging 100 iterations on the SIR model, γ = 0.8 with different values of β.



(a) Twitter-mention Network (b) Seventh-graders Network

(c) Residence hall Network (d) Twitter-retweet Network

Figure 9. The Kendall's τ values obtained by the ranked lists generated by SIR model and the ranked lists generated by centrality measures, P-centrality, and EP-centrality measures with degree, k-shell and NP. The results were obtained by averaging 100 iterations on the SIR model, γ =0.8 with different values of β.

5. Conclusion

In this paper, we proposed an improvement on centrality measures, called P-centrality measure, to find influential users in a social network. This measure uses the centrality measure of the node and its predecessors to calculate the score of the node. To calculate the impact of the predecessors, we defined a coefficient obtained by the number of interactions between the users for the centralities of the predecessors. The nodes with higher scores are considered as influential users to spread the information in the network. We also proposed an extended measure called EP-centrality measure in which the predecessors of the predecessors also impact the score of the node. A combination of two centrality measures is defined to improve the performance of the proposed metrics. This combination, called NodePower (NP), is a summation of normalized degree centrality and normalized shell number in k-shell centrality which is used as the centrality in P-centrality and EP-centrality metric. The time complexity of the proposed measures depends on the time complexity of the centrality measure which is used. To evaluate the effectiveness of the measures, Susceptible-Infected-Recovered (SIR) model is used. The correlation between ranked list obtained by the measures and ranked list obtained by the SIR model is calculated using Kendall's τ coefficient. Four centrality measures are chosen for the experiments called degree centrality, betweenness centrality, PageRank centrality, and k-shell centrality. The results show that EP-centrality measure outperforms normal centrality and P-centrality measures. Moreover, P-centrality measure performs better than the normal centrality measure. In future work, Independent Cascade Model (ICM) model can be used instead of SIR model and the results of these two models can be compared.

6. References

[1] Kempe, David, Jon Kleinberg, and Éva Tardos. ”Maximizing the spread of influence through a social network.” In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 4 137-146. 2003.

[2] Al-Garadi, Mohammed Ali, Kasturi Dewi Varathan, Sri Devi Ravana, Ejaz Ahmed, Ghulam Mujtaba, Muhammad Usman Shahid Khan, and Samee U. Khan. ”Analysis of online social network connections for identification of influential users: Survey and open research issues.” ACM Computing Surveys (CSUR) 51, no. 1 (2018): 1-37.

[3] Freeman, Linton C. ”Centrality in social networks conceptual clarification.” Social networks 1, no. 3 (1978): 215-239.

[4] Kitsak, Maksim, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. ”Identification of influential spreaders in complex networks.” Nature physics 6, no. 11 (2010): 888-893. 11

[5] Wei, Daijun, Xinyang Deng, Xiaoge Zhang, Yong Deng, and Sankaran Mahadevan. ”Identifying influential nodes in weighted networks based on evidence theory.” Physica A: Statistical Mechanics and its Applications 392, no. 10 (2013): 2564-2575.

[6] Freeman, Linton C. ”A set of measures of centrality based on betweenness.” Sociometry (1977): 35-41.

[7] Brandes, Ulrik. ”A faster algorithm for betweenness centrality.” Journal of mathematical sociology 25, no. 2 (2001): 163-177.

[8] Estrada, Ernesto. The structure of complex networks: theory and applications. Oxford University Press, 2012.

[9] Sabidussi, Gert. ”The centrality index of a graph.” Psychometrika 31, no. 4 (1966): 581-603.

[10] Chen, Duanbing, Linyuan Lü, Ming-Sheng Shang, Yi-Cheng Zhang, and Tao Zhou. ”Identifying influential nodes in complex networks.” Physica a: Statistical mechanics and its applications 391, no. 4 (2012): 1777-1787.

[11] Brin, Sergey, and Lawrence Page. ”The anatomy of a large-scale hypertextual web search engine.” Computer networks and ISDN systems 30, no. 1-7 (1998): 107-117.

[12] Al-Garadi, Mohammed Ali, Kasturi Dewi Varathan, Sri Devi Ravana, Ejaz Ahmed, and Victor Chang. ”Identifying the influential spreaders in multilayer interactions of online social networks.” Journal of Intelligent & Fuzzy Systems 31, no. 5 (2016): 2721-2735.

[13] Al-garadi, Mohammed Ali, Kasturi Dewi Varathan, and Sri Devi Ravana. ”Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method.” Physica A: Statistical Mechanics and its Applications 468 (2017): 278-288.

[14] Newman, Mark. Networks. Oxford university press, 2018. 30

[15] Zhu, Zhiguo. ”Discovering the influential users oriented to viral marketing based on online social networks.” Physica A: Statistical Mechanics and its Applications 392, no. 16 (2013): 3459-3469

[16] Gao, Zhenxiang, Yan Shi, and Shanzhi Chen. ”Measures of node centrality in mobile social networks.” International Journal of Modern Physics C 26, no. 09 (2015): 1550107.

[17] Kendall, Maurice G. ”A new measure of rank correlation.” Biometrika 30, no. 1/2 (1938): 81-93.

[18] Kunegis, Jérôme. ”Konect: the koblenz network collection.” In Proceedings of the 22nd international conference on world wide web, pp. 1343-1350. 2013.

[19] Li, Yuchen, Ju Fan, Yanhao Wang, and Kian-Lee Tan. ”Influence maximization on social graphs: A survey.” IEEE Transactions on Knowledge and Data Engineering 30, no. 10 (2018): 1852-1872.

[20] Katz, Leo. ”A new status index derived from sociometric analysis.” Psychometrika 18, no. 1 (1953): 39-43.

[21] Bonacich, Phillip. ”Factoring and weighting approaches to status scores and clique identification.” Journal of mathematical sociology 2, no. 1 (1972): 113-120.

[22] Anderson, Roy M., and Robert M. May. Infectious diseases of humans: dynamics and control. Oxford university press, 1992.

[23] Leskovec, J., Rajaraman, A. and Ullman, J.D., 2020. *Mining of massive data sets*. Cambridge university press.