A method based on k-shell decomposition to identify influential nodes in complex networks

Bakhtyar Rafeeq HamaKarim
University of Kurdistan
Rojiar Pir Mohammadiani
University of Kurdistan
Amir Sheikhahmadi (asheikhahmadi@iausdj.ac.ir)
Islamic Azad University
Bryar Rafeeq Hamakarim
University of Kurdistan

Research Article

Keywords: Influential Nodes, Diffusion Model, Community Detection, k-shell Decomposition, Complex Networks

Posted Date: November 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2242049/v1

Additional Declarations: No competing interests reported.

A method based on k-shell decomposition to identify influential nodes in complex networks

Bakhtyar Rafeeq HamaKarim[1], Rojiar Pir Mohammadiani[1], Amir Sheikhahmadi[2,*], Bryar Rafeeq Hamakarim[1]

[1] Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
[2] Department of Computer Engineering, Islamic Azad University, Sanandaj Branch, Sanandaj, Iran

Abstract

Finding the most influential nodes in complex networks is one of the open research issues. This problem can be divided into two sub-problems: (1) identifying the influential nodes and ranking them based on the individual influence of each node and (2) selecting a group of nodes to achieve maximum propagation in the network. In most of the previous articles, only one of these sub-issues has been considered. Therefore, this article presents a method to measure the spreading power of influential nodes in the network (the first sub-problem) and select the best group from them (the second sub-problem). In the proposed method, first, the input network is allocated to different communities. Then, the common neighbors and the degrees of the two end vertices of each edge are used to weigh the graph edges in each community. Next, in each of the communities, the nodes' propagation power is measured and ranked. Finally, a group of influential nodes is selected to start the propagation process.

Eight data sets collected from real networks have been used for evaluation. The proposed method is compared with other previously known methods based on ranking accuracy, assigning different ranks to nodes, and calculating the amount of diffusion created in the network. The results show the proposed method's significant superiority over other methods in all test datasets.

keywords: Influential Nodes, Diffusion Model, Community Detection, k-shell Decomposition, Complex Networks

1. Introduction

The rapid development of network science attracts much attention to complex networks in various fields, including society [1], biology, physics [2], time series [3], transportation [4], and immunization strategy [5]. In our daily life, we encounter many complex networks such as communication networks, social networks, biological networks, and the World Wide Web. Such networks are composed of many nodes with non-obvious characteristics, which are the source of various research problems. One of the most important research issues is identifying nodes in these networks with high propagation power. These nodes can play a critical role in dissemination. The phenomenon of diffusion in complex networks can be done in various fields, such as the spread of epidemic diseases [6], technical innovations [2], product promotion [7] , and behavior acceptance [8]. This can help better understand the mechanisms hidden in complex phenomena and guide human production and life.

* *Corresponding author: Amir Sheikhahmadi (asheikhahmadi@iausdj.ac.ir)*

The problem of finding the super spreader nodes divide into two parts: in the first part, the identification of influential nodes and their ranking based on the spread power of each node is discussed. While in the second part, the goal is to select a group of influential nodes to achieve maximum group propagation[1].

Many indicators have been presented to identify the most influential nodes, including degree, closeness, betweenness, k-shell, eigenvalue centrality, and page ranking. In all these methods, the propagation power is calculated according to the network structure and the position of each node in it. Then, the influential nodes are selected from among the nodes at the top of the ranked list. In order to identify a group of influential nodes with the maximum spread in the network, finding a seed set with a specific size is usually necessary so that the final influence obtained is maximized by activating the nodes in the set [9]. Finding a seed set with a given size is in the category of NP-Hard problems. A simple way to determine this set is to use identified influential nodes. In this method, first, all the nodes are ranked based on the amount of propagation power, and the top nodes are selected as the target set in the specified number. However, Kitsak et al. showed in [10] that selecting a group of nodes at the top of the ranked list is not efficient due to the considerable overlap of the nodes in this list. Therefore, there is a need to use another method with sufficient speed and proper efficiency.

For this purpose, it is possible to use the characteristics of communities in complex networks. In communities, nodes have the most connections with each other and the least connections with nodes in other communities. This feature is widely observed in real-world networks [11]. Communities greatly influence propagation in networks. Therefore, the motivation to select nodes from different communities as the origin of the start of propagation arose [3]. The next point is the significant difference between the nodes and edges in the communities regarding diversity and strength of communication. By considering these differences, a better method can be provided to calculate the spread power of nodes.

Therefore, this paper proposes a method to identify the top spreader nodes and select an optimal subset of them to maximize the propagation in the network. In the proposed method, the feature of the community structure is used to separate the network into different communities. Then, by measuring the spreading power of nodes and ranking the nodes of each community based on that, a diverse subset of nodes is selected as the initial core. The innovations made in the article are as follows:

- Selection of seed set members from different network communities to reduce overlap: dividing the network graph into communities and identifying influential nodes from each community can help increase the scope of propagation and spread in different parts of the network.
- Presenting a semi-local method for distinguishing nodes with similar local characteristics but different neighbors in terms of propagation power.
- Using the concept of the communication diversity of each node in measuring its spread power: in the proposed method, if a node has less communication diversity, it gets a lower score.
- Differentiating the edges of the graph by weighting the communication edges between the nodes in each community.

The structure of this article is as follows:

In Section 2, an overview of related works is presented. The proposed method and its parts are introduced in detail in Section 3. The data sets, evaluation parameters, and the obtained results are reviewed in section 4, and finally, in section 5, the conclusion and future works are discussed.

2. Related works

The problem of calculating the spreading power of nodes and selecting influential nodes in the network has been investigated in many researches. For this purpose, various indicators have been presented, some of which will be introduced below:

The degree centrality shown by DC determines the node's importance by comparing the node's degree. The degree centrality of node i is determined using equation 1:

$$DC(i) = \sum_{j}^{n} a_{ij} = k_i \quad (1)$$

, where k_i is the degree of node i. A node with a high degree also has a high influence [12].

Betweenness centrality [13], which is denoted by (BC), measures the importance of a node by the number of shortest paths that pass through it, and it is obtained by relation 2:

$$BC(i) = \sum_{j,k \neq i} \frac{N_{jk}(i)}{N_{jk}} \quad (2)$$

, where N_{jk} shows the shortest paths from node j to node k and $Njk(i)$ is the number of N_{jk} from node i. The greater the number of shortest paths that pass through node i, the more influential the node is.

Closeness centrality (CC) [12] calculates the influence of nodes by the inverse of the sum of the shortest paths between nodes, which is shown in equation 3:

$$CC(i) = \frac{1}{\sum_{j}^{n} d_{ij}} \quad (3)$$

, where d_{ij} represents the distance between node i and node j. The higher CC(i), the more critical node i is.

Eigenvector centrality [14], denoted by (EC), uses the importance of neighbors in addition to the number of neighbors to calculate the influence of a node. EC(i) can be calculated with equation 4:

$$EC(i) = \frac{1}{\lambda} \sum_{j=1}^{n} (a_{ij} x_j) \quad (4)$$

The largest eigenvalue of A is denoted by λ, and x_j is the input value of the *jth* eigenvector corresponding to λ.

PageRank [15], denoted by PC, uses an iterative approach to obtain the influence of nodes. PC(i) of node i is calculated by equation 5:

(5) $$PC(i)^q = \sum_{j=1}^{n} (a_{ij} \frac{PC(i)^{q-1}}{k_j})$$

The degree of influence of node *i* in step q is shown as $PC(i)^q$. The higher the PC score, the more influential the node is.

The k-shell decomposition method was proposed by Kitsak [16] to show the importance of nodes in the network. In this method, all network nodes whose degree is one are removed and placed in shell 1. The process of removing nodes whose degree is less than or equal to one is repeated until there are no nodes in the network with a degree less than or equal to 1. All the removed nodes are placed in shell 1 at this stage. Then this method continues to determine shell 2, shell 3, and so on. It should be noted that the node with a higher k-shell value is located in a more central position in the network [16]. In the k-shell method, it is assumed that the nodes located in a higher shell have a higher propagation power. Also, in this method, all the nodes in the same shell are given the same rank.

When Zheng and Zhang [17] used the k-shell method to measure the propagation power of nodes, they realized that in this method, only the remaining degree for each node is considered. To solve this problem, they proposed a method called mixed degree decomposition (MDD), in which the contribution of the remaining degree and the removed degree of each node was considered simultaneously to calculate the strength of that node. If k^r and k^e are the remaining degree and removed degree of node v_i, respectively, the MDD of node v_i is calculated as equation 6:

(6) $$k^m(v_i) = k^r(v_i) + \lambda * k^e(v_i)$$

Bai and Kim [18] used the balanced combination of degree and coreness of neighbors to solve the problem of assigning the same rank to a large number of nodes. Based on this, the coreness of the neighborhood of node *v*, which C_{nc} represents, can be calculated with equation 7:

(7) $$c_{nc}(v_i) = \sum_{v_{j \in \Gamma_{v_i}}} k_s(v_j)$$

, where Γ_{vi} is the set of neighbors of node *v*i and $k_s(v_j)$ is the k-shell value of its neighbor node v_j.

Next, the coreness value of the extended neighborhood C_{nc+} of node v_i is recursively calculated according to equation 8.

(8) $$c_{nc+}(v_i) = \sum_{v_{j \in \Gamma_{v_i}}} c_{nc}(v_j)$$

In the CLD measure [19], the effect of topological connections between neighbors is also considered on the node's spread power in addition to the number of neighbors. The more connections between the nearest neighbors of a node, the greater the influence of this node. Therefore, by combining the sum of the degrees of the nearest neighbors of a node and its clustering coefficient, the centrality of CLD is presented as equation 9:

(9) $$CLD(i) = (1 + C_i) \sum_{j \in N(i)} degree(j)$$

, where N(i) is the set of the nearest neighbors of node i and C_i is the clustering coefficiency of node i.

Ma et al. [20] used the gravity law to calculate the influence of one node on other nodes in spreading activity. They used the k-shell value of the node as the mass and the shortest path distance between any two nodes in the network as the distance related to Newton's gravity formula. Equation 10 shows how to calculate the gravity for each v_i node.

(10) $$G(v_i) = \sum_{v_j \in \Psi(v_i)} \frac{k_s(v_i) * (k_s(v_j)}{s^2(v_i, v_j)}$$

, where s(v_i, v_j) is the shortest distance between nodes v_i and v_j. $\Psi(v_i)$ is a set of nodes that are adjacent to node v_i up to the specified level r. The authors of this article consider the value of r to be 3; Therefore, in this case $\Psi(v_i)$ contains 3 levels of neighboring nodes vi.

Based on Newton's gravity formula, Li et al. [21] proposed another innovative method to measure the spread power of network nodes. In the proposed method, the degree of each node is used instead of the mass of the objects. The authors proposed the local gravity model to reduce the computational complexity and not reduce the accuracy of the final result. Only the effect of up to r levels of nodes is considered in this method. In the experiments, different values of r were tested. When the value of r became close to the half diameter of the network (r ≈ D/2), the diffusion estimate was obtained almost equivalent to the general method .

(11) $$S(v_i) = \sum_{v_i \neq v_j} \frac{k(v_i) * k(v_j)}{s^2(v_i, v_j)}$$

Namtirtha et al. [22] . proposed the gradient neighborhood method, weighted shell node by assigning weight to the edges using the degree and k-shell index of two end-point nodes. In the ksd^w method, the weight of all the edges connected to the respective node is added together and it is used to measure the spread power of each node. Equation 12 is used to calculate the weight of each edge, and Equation 13 is used to obtain ksdw of node v_i.

(12) $$w_{ij} = \left(c_1 * k(v_i) + c_2 * k_s(v_i)\right) * \left(c_1 * k(v_j) + c_2 * k_s(v_j)\right)$$

(13) $$ksd^w(v_i) = \sum_{v_j \in \Gamma(v_i)} w_{ij}$$

, where Γ(Vi) is the set of nodes that are in the neighborhood of node vi. c1 and c2 are two adjustable parameters. The authors used a set of c1 and c2 parameters whose values are between [1 and 0].

In [23], a method for selecting the initial seed set (set of nodes that are initially activated) among the super-spreaders is proposed. The proposed method is based on the k-shell decomposition method and attempts to reduce the overlap that exists among members with high k-shell. For this purpose, first the input network is separated from the existing communities by a community detection algorithm (here by improved version of Girvan-Newman's algorithm), then in each community the k-shell decomposition method is used and the nodes of that community are basically the number of shells is arranged from large to small. At the end of

this step, the side B sets are specified. To determine B sets by number, core nodes must be alternately selected from different communities to create an R ranking list. Finally, Seedset B is selected, which includes the nodes with the highest rank. A greedy quick strategy is then used to select the best seeds in the R-rated list. In the proposed CKS + method, R is specified as the initial rank and the R + list is specified as the final output.

3. Proposed method

The general framework of the proposed method of this thesis is shown in Figure 1. In the proposed method, the complex (social) network is entered as a list of edges. After partitioning the network into different communities, the edges of each community become weighty. Then, in each community, the spread power of the nodes is measured and ranked, and finally a group of influential nodes are selected to start the propagation process. In the following, each part of the proposed method is examined.

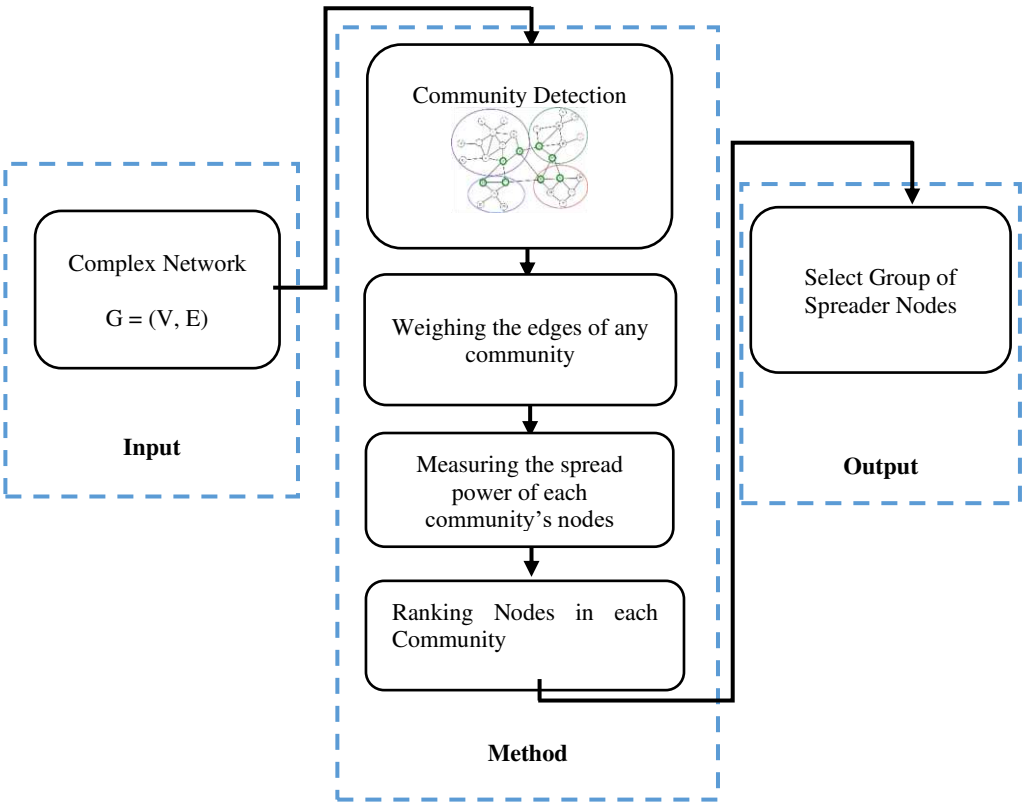


Figure 1 Steps of the proposed method

3.1 Partitioning the network graph into different communities

One of the main features of complex networks is the feature of the structure of communities, where network nodes exchange information with each other in the form of groups. Separating the input network into the communities can help select the spreading nodes from various network parts. In this way, instead of selecting the sources to start spreading from a particular part of the network, which will cause a limited

spread in the network, it is possible to start spreading in different communities in the hope of reaching many nodes in the network. In this paper, the method introduced in [24] is used to extract the input graph to communities. This method is proposed for the rapid extraction of community structure in large networks and is in the category of bottom-up methods. For example, the implementation of this algorithm of a small network method has a result like Figure 2.

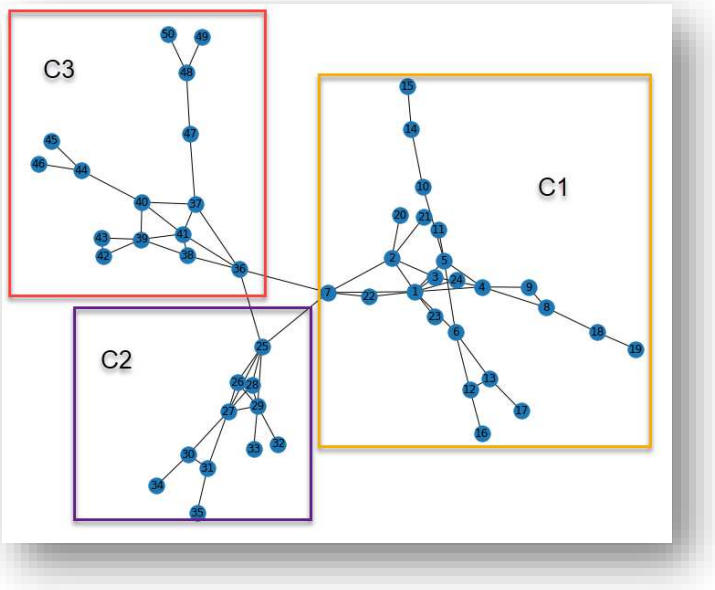


Figure 2. Extracted communities in sample network

3.2 Weighing the edges of each community

After identifying the communities, at this stage, we weigh them to differentiate the edges in each community. In most networks, there is no necessary information about the amount of real communication between network nodes. Therefore, to calculate the weight of the edges of the graph, the number of common neighbors and the degree of the nodes in the two ends of the edges are used. The reason for choosing these two criteria for weighting the edges is that the greater the number of common neighbors between two nodes, the greater the intensity of the connection edge between them. In addition, the more the communication edge connects two nodes with more friends, the more critical that edge is. Equation 14 is used to calculate the number of common neighbors of the two ends of the edge, and Equation 15 is used to weigh the communication edges between the nodes of each community.

(14) $$CN(v,w) = |N(v) \cap N(w)|$$

, where *N(v)* and *N(w)* denote the set of neighbors of nodes v and w.

(15) $$w_{v,w} = (CN(v,w) + 1) \times (deg_v + deg_w)$$

In equation 15, the value of the common neighbor is added by the number one so that if the two end vertices of the edge do not have a common friend, the weight of the edge is not zero. In pseudo-code 1, the method of calculating the weight of edges in each community is given.

Pseudo Code 1. Edges Weighting

```
Input: G=(V,E), communities
Output: edge_weight                    // list of edges' weights
    1.  edge_weight←[ ]
    2.  for each c∈communities do
    3.      for v ∈ V do
    4.          for w ∈ N(v) do
    5.              CN=|common_neighbors(v,w)|        //by eq. 14
    6.              weight=(CN+1) *(deg[v]+deg[w])    // eq. 15
    7.              edge_weight[v][w]←weight
    8.  return edge_weight
```

3.3 Calculate the spread power of the nodes of each community

After weighing the edges between the nodes of each community, it is necessary to calculate the spread power of each node in the community, and the nodes with higher power should be ranked higher. For this purpose, the following are calculated for each community node:

- Total weight of edges connected to each node:

  The total weight of the edges between that node and its neighbors is calculated for each community node. Equation 16 calculates the total weight of edges connected to node *v*.

  (16) $$Weighted\_sum(v) = \sum_{w \in N(v)} edge\_weight(v,w)$$

  , where N(v) is the neighbors of node v and edge_weight(v,w) is the weight of the edge between nodes v and w.

- Calculation of dispersion related to the presence of friends of a node in different cores:

  If a node has various friends in cores with different numbers, the ability of that node to spread information in different parts will increase. In other words, instead of the node spreading information in only a tiny part of the network, it can spread the information to large parts. For this purpose, entropy is used to measure the distribution of friends of a node in different cores. With the help of equation 17, the amount of entropy related to node v is calculated:

  (17) $$Diversity(v) = -\sum_{i=1}^{ks^{Max}} p_i \log_2 p_i$$

  , where p_i is the probability of the presence of friends of node *v* in shell number *i*. To calculate the value of p_i for node *v*, the formula $p_i = \frac{Number\ of\ Friends\ v\ in\ shell\ i}{Deg(v)}$ is used.

  Since friends in higher cores are likely to have a higher spreading power; Therefore, the shell number can be used as a coefficient in relation 17. Therefore, equation 18 can be used to calculate the distribution of friends of each node in different cores.

(18) $$W\_Diversity(v) = -\sum_{i=1}^{ks^{Max}} i * p_i \log_2 p_i$$

- Calculate the diversity of the presence of friends of a node in different communities:

  In different communities, there are nodes that, in addition to the connections they have in their community, have friends in other communities, which can promise their more effective and broader propagation. Therefore, in this section, the distribution of friends (neighbors) of each node in different communities is calculated. Here, entropy is used to calculate a node's friends distribution in different communities. In simple words, the higher the entropy obtained for a node, the greater the dispersion of its friends in different communities. Equation 19 is used to calculate this parameter.

(19) $$Com\_Diversity(v) = -\sum_{i=1}^{n} p_i \times \log_2 p_i$$

  In equation 19, n represents the number of communities, and p_i is the probability of the presence of friends of a node in community i. This probability is calculated as $\frac{|\text{friends of v in the i community}|}{\deg(v)}$.

After calculating these three indicators for each community node, it is time to measure their diffusion power. For this purpose, first, the indices are normalized to a number between [0,1] using the min-max method. In the min-max normalization method, each index's current value is subtracted from the maximum values; then, it is divided by the difference between the maximum and the minimum. Equation 20 shows how to calculate the normalized value.

(20) $$new\_value = \frac{current\_value - max}{max - min}$$

Therefore, at this stage, we have indicators as follows for each of the community nodes.

Node	$Normalized\ Weighted\_sum(v)$	$Normalized\ W\_Diversity(v)$	$Normalized\ Com\_Diversity$

Equation 21 can be used to calculate the spread power of each community node.

(21) $$SP(v) = \alpha \times Weighted_{sum(v)} + \beta \times W_{Diversity} + \gamma \times Com\_Diversity$$

, where parameters α, β , and γ are adjustable parameters in the range [0,1].

Finally, to increase the monotonicity of the nodes, the expanded power of each node v is calculated by equation 22:

(22) $$SP^{+}(v) = \sum_{w \in N(v)} SP(v)$$

, where $N(v)$ represents the set of neighbors of node v.

4.3 Choosing the best group of powerful nodes

After determining the propagation power of nodes in different network communities, pseudocode 2 is used to select the best group of powerful nodes.

Pseudo Code 2. Select Super-Group

```
Input: G(V,E) , θ, communities, k  //θ is threshold to delete small community, k is seedset size
Output: S                          //S is initial seed set
1.  pop←0
2.  for each c in communities:
3.     if len(c) <= θ:
4.        delete(c)
5.     else:
6.        pop←pop+|c|
7.  S←[]
8.  for c in communities:
9.     com_nodes←c
10.    rate_com←(len(com_nodes)/pop *seedsetsize)
11.    count←0
12.    for node in Inf⁺(c):    // Inf⁺(c) is order list of extended spread power for c community
13.       S←S ∪{node}
14.       count←count+1
15.       if count>=rate_com:
16.          break
17. return S
```

In pseudo-code 2, lines 2 to 4, communities less than the input threshold θ are removed. Then, according to the ratio of the population of each community to the total population (of course, minus the population of excluded communities), the number of candidates for each community to be a member of seed set is determined (lines 8 to 10). The more populous a community is, the more its candidates can be. After determining the number of candidate members of each community for membership in the identification team, the nodes in the community with the highest value of the expanded spread power calculated in the previous section are selected and placed in the initial seed set.

4. Evaluation

The proposed method of this article and other compared methods are implemented with Python language version 3.8. All methods have been evaluated using eight networks collected from real-world networks and the example network given in Figure 2. The characteristics of these datasets are listed in Table 1. Real-world networks can be obtained from http://konect.cc and some from the Snap project of Stanford University at https://snap.stanford.edu. The columns of Table 1-4, from left to right, are the name of the dataset, the number of network nodes (|V|), the number of its edges (|E|), the average degree of the graph nodes, and the highest degree.

Table 4-1. Characterization of the datasets used in the experiments

Network	$\|V\|$	$\|E\|$	Average Degree	Max Degree
Sample Network	50	70	2.8	9
Karate Club	34	78	4.588	17
Dolphins	62	159	5.1290	12
Jazz musician	198	2,742	27.6970	100
Netsciense	379	914	4.8232	34
Hamsterster	2,426	16,631	13.711	273
PowerGrid	4,941	6,594	2.669	19
PGP	10,680	24,316	4.554	205

4.1 Evaluation criteria

In the proposed method, which consists of two parts, first, the spread power of the nodes is calculated; Then, a subset of the super spreader is selected. Therefore, in this section, first, the method's ability to measure the nodes' spread power is evaluated. In the next section, the spread of the impact of the selected subset is measured. The following two criteria are used to evaluate the first part:

1- Measuring the power of the methods in distinguishing different nodes in terms of rank: The distinct power's meaning is the method's ability to differentiate the nodes in terms of their spread power. Therefore, the more a method can place the nodes in more distinct ranks, the better the method works. The Distinct Metric (DM) function is used to calculate the resolution of the methods. In relation 23, the method of calculating the resolution of the R ranking list is given:

(23) $$DM\ (R) = \frac{numder\ of\ distinct\ ranks}{|V|}$$

where, |V| The number of graph nodes and R is the ranking list of each method. If the method places all the nodes in the same rank, the DM of that list is zero, and if it gives each node a different rank, its DM becomes one.

2- Calculating the similarity of the list ranked by each method with the list ranked by the diffusion model:
To compare the accuracy of the methods, the nodes were first ranked using each method. Then, the obtained list is compared with the ranked list of the diffusion model. In this article, the IC model is used to calculate the spread power of the nodes. Kendall Tau correlation is used to calculate the similarity of these two lists. This correlation is usually a number between -1 and 1; The closer the obtained number is to 1, the higher the accuracy of the method in ranking. The IC propagation model was executed 1000 times, and the average obtained in these 1000 executions was used as the spread power of each node.

In the second part of the comparisons, the spread amount of influence obtained by the selected subsets is calculated with the IC diffusion model. The method by which the selected set had the highest spread of influence is selected as the superior method. In this section, the spread of influence is calculated by the IC diffusion model. In the IC diffusion model, each node is in one of the active or inactive states. When using the IC to calculate the spread power of a subset such as S, it is set to the active state, and other nodes, i.e., V-S, are set to the inactive state. In each time interval, the activated nodes are given a chance to activate their inactive neighbors so that they can activate them with a p probability. After this attempt, node v becomes inactive. The deactivated node cannot try to activate the nodes in subsequent intervals. This model is repeated until there is an active node left. Finally, the number of activated nodes is counted, and the spread power of S is obtained. The IC model is a probabilistic model, and it is necessary to execute

it in significant iterations. Therefore, the spread power is calculated based on the average number of runs. It should be noted that in all the evaluations, the values of α, β and γ parameters are set to 1.

4.2 Ability to assign a different rank to nodes

In the first part, the proposed method is compared with the k-shell, C_n, C_{nc+}, Closeness, Betweenness, and Gravity methods. Table 2 shows the ranks assigned to the sample graph nodes in Figure 2. Due to the table size, only the first 19 ranks are given. Among the compared methods, the k-shell method performs the worst; Because this method assumes that the nodes placed in the same number have the same spread power. The proposed method, which is called HKCD (Hybrid K-shell based methods using Community Detection), performs better than all other methods after the IC diffusion model and has a higher monotonicity. This differentiation can be seen in the high ranks.

Table 2 Comparison of methods in assigning distinct ranks in the sample graph

Rank	Degree	Betweenness	Closeness	k-shell	CNC+	Gravity	Proposed	IC
1	1	7	7	25, 36, 26, 27, 28, 29, 37, 38, 41, 40, 39	25	25	25	1
2	25, 27, 29	1	1	1, 2, 3, 4, 5, 6, 7, 22, 8, 9, 12, 13, 30, 31, 42, 43	36	27	27	7
3	2, 4, 5, 7, 36, 41, 39	36	36	23, 24, 20, 21, 18, 19, 10, 11, 14, 15, 16, 17, 34, 35, 32, 33, 47, 44, 46, 45, 48, 50, 49	27	41	29	25
4	6, 37, 40	25	25		1	36, 29	26, 28	36
5	3, 8, 12, 13, 26, 28, 30, 31, 38, 44, 48	37	2		7	39	7	2
6	22, 9, 18, 10, 14, 47, 42, 43	5	22		29	1	1	27
7	23, 24, 20, 21, 19, 11, 15, 16, 17, 34, 35, 32, 33, 46, 45, 50, 49	4	37, 41		41	37, 40	36	29
8		27	5		26, 28	26, 28, 38	4	4
9		6	4		39	7	5	41
10		40	6		37, 40	4	3	5
11		41, 47	38		38	2, 5, 6	6	3
12		2	3, 27		4, 5	3, 30, 31	41	26
13		39	29		2	8, 12, 13, 42, 43	40	28
14		29	26, 28		3, 6	22, 9	39	6
15		44, 48	23, 24		22	44	38	22
16		8, 10	40		30, 31	47	2	37
17		38	20, 21		42, 43	18, 10, 32, 33, 48	37	40

18		18, 14, 12, 13, 30, 31	39		8	23, 24, 20, 21, 11, 14, 16, 17, 34, 35	30, 31	38
19		3	47		23, 24	19, 15, 46, 45, 50, 49	22	39

Figure 3 shows the number of nodes placed in each rank. As mentioned, the lower the number of nodes with the same rank, the more acceptable the method is in terms of the monotonicity index. The data in Figure 3 shows the superiority of the proposed method, especially in high ranks. After the proposed method of Closeness and C_{nc+} method, they have excellent performance.

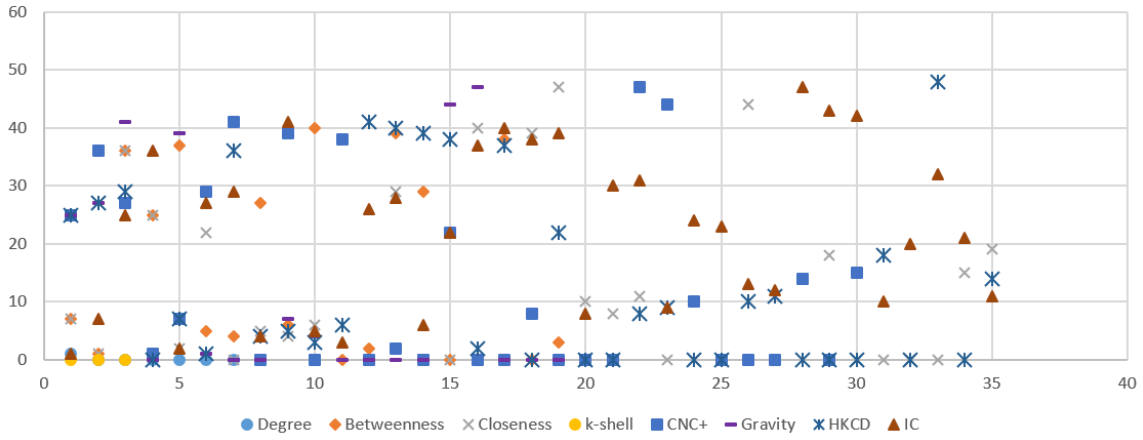


Figure 3. The number of assigned nodes in each rank by different methods

Table 3 shows the discriminating power of the compared methods on other datasets. The data in the table shows the superiority of the proposed method compared to other methods. Of course, in the small DLN dataset, the C_{nc+} method has similar results to the proposed method. However, in other datasets, the proposed method has shown better results in differentiating rank to nodes. The superiority of the proposed method is in selecting nodes from different parts of the network and paying attention to local and general measures in the graph, which makes the results of the proposed method more significant than other methods.

Table 3 Monotonicity of different methods

Data Set	Degree	Closeness	Betweenness	k-shell	Cnc	Cnc+	Gravity	HKCD
Sample Net	0.64	0.97	0.61	0.42	0.84	0.95	0.85	**0.98**
Karate Club	0.71	0.90	0.78	0.50	0.85	0.94	0.85	**0.95**
Dolphins	0.83	0.97	0.96	0.38	0.92	0.98	0.93	**0.99**
Jazz musician	0.965	0.987	0.988	0.794	0.992	0.998	0.998	**0.999**
Netsciense	0.81	0.96	0.94	0.65	0.91	0.97	0.91	**0.99**
Hamsterster	0.89	0.97	0.96	0.62	0.93	0.98	0.95	**0.99**

PowerGrid	0.67	0.93	0.88	0.44	0.85	0.96	0.84	**0.98**
PGP	0.62	0.97	0.61	0.48	0.89	0.98	0.91	**0.98**

4.3 Similarity of the lists ranked by each method and the diffusion model

This section uses Kendall's Tau correlation coefficient to calculate the correlation between the list ranked by each method and the actual ranking list. Each method was implemented on the sample graph in Figure 2, and the nodes were ranked based on their method. Figure 4 shows the correlation between the list ranked by each method and the list ranked by the IC diffusion model.

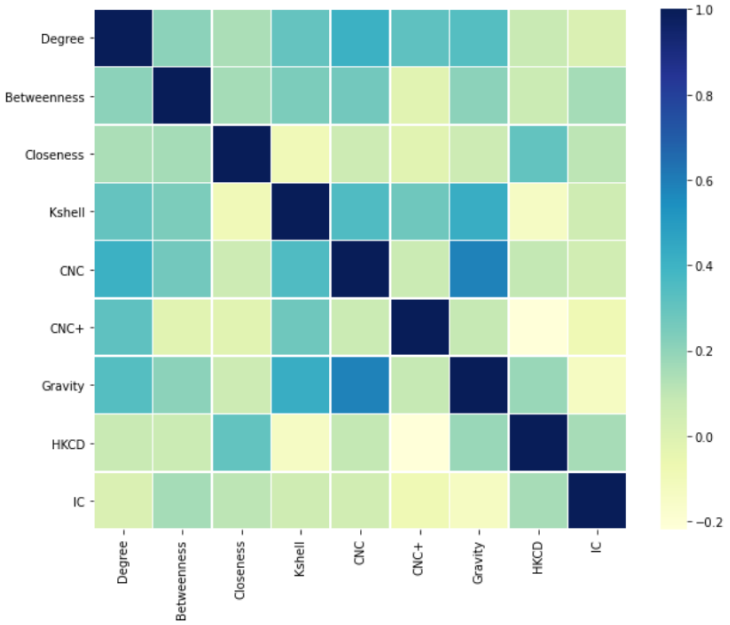


Figure 4 Correlation of the ranked list by each method and the ranked list by the IC model

Figure 4 shows that the proposed method has been able to calculate the spread power of each node with higher accuracy and is closer to the result of the IC diffusion model. Table 4 compares the ranking list obtained by implementing each method in different data sets with the list obtained by the IC diffusion model based on Kendall's Tau correlation.

Table 4 Correlation between lists ranked by each method and the IC diffusion model

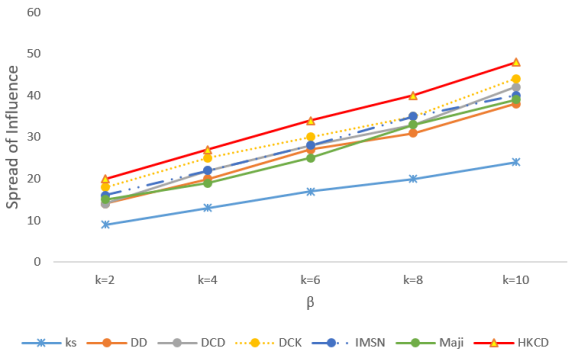
Data Set	**Degree**	**Closeness**	**Betweenness**	**KS**	**Cnc**	**Cnc+**	**gravity**	**HKCD**
KarateClub	0.70	0.73	0.76	0.55	0.60	0.71	**0.78**	0.76
Dolphine	0.80	0.84	0.88	0.58	0.78	0.81	0.91	**0.92**
JazzMusician	0.89	0.88	0.92	0.77	0.89	0.93	0.92	**0.94**
NetScience	0.59	0.82	0.83	0.50	0.55	0.62	0.88	**0.90**
Hamsterster	0.63	0.66	0.69	0.50	0.54	0.64	0.70	**0.73**

PowerGrid	0.74	0.78	0.81	0.53	0.72	0.74	0.84	**0.87**
PGP	0.73	0.81	0.84	0.55	0.73	0.79	0.83	**0.87**

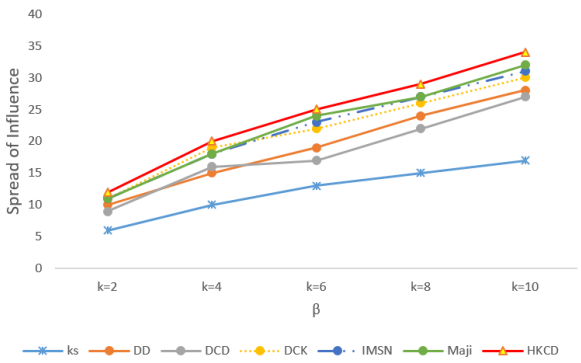
Table 4 shows that the proposed method has higher accuracy than other methods in the real data sets and sample network. The higher accuracy is due to paying attention to the total weight of the edges of each node and using the diversity of each node's neighbors in different communities and cores. Among the methods, Gravity has close results to the proposed method. This method has higher accuracy than ours only in the KarateClub data set, but the proposed method has higher accuracy in other data sets.

4.4 Evaluation based on the spread amount of influence

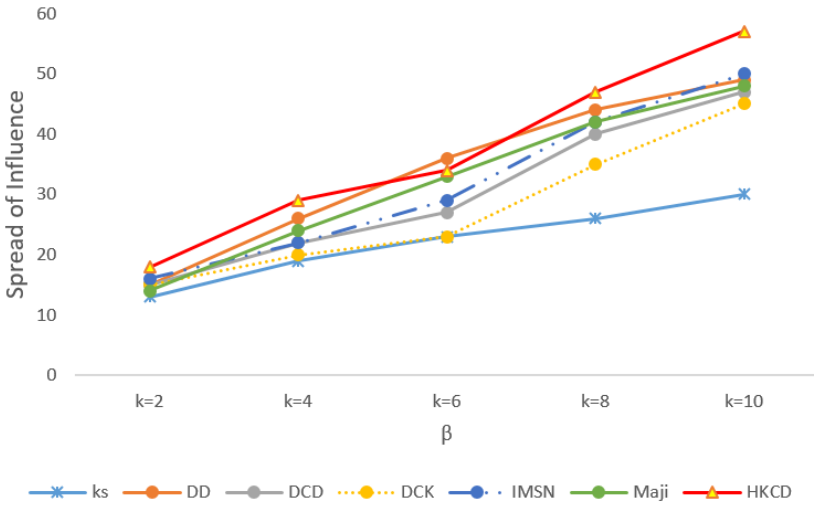
The second part compares the proposed method with the DegreeDiscount(DD), DCD, k-shell, DCK, Maji, and IMSN methods. IC model has been used to compare the spread power of the seed sets selected by different methods. The IC model is executed 1000 times, and the spread power of each set is considered the average of this number. Because the data sets are different, the member size of the selected set for the sample network in Figure 2, karate club, and dolphin network are considered 2, 4, 6, 8, and 10. The spread amount of influence of each set selected by different methods in these three data sets is shown in Figure 5. In Figure 5, the horizontal axis indicates the member size of the selected set (k), and the vertical axis indicates the spread amount of influence obtained by the IC model.



a. Sample network



b. Karate Club



c. Dolphins Network

Figure 5. spread power of the selected sets in small datasets

Figure 5 shows the superiority of the proposed method in all three datasets. As could be predicted, the choice of core members from different communities of the network based on the diversity of the presence of node friends in different communities and cores has increased the spread of the network.

In the Jazz Musician data set, the member size of the initial seed set is specified as 5, 10, 15, 20, and 25, respectively. In Figure 6, the results of the distribution of each set selected by different methods in this data set are given.

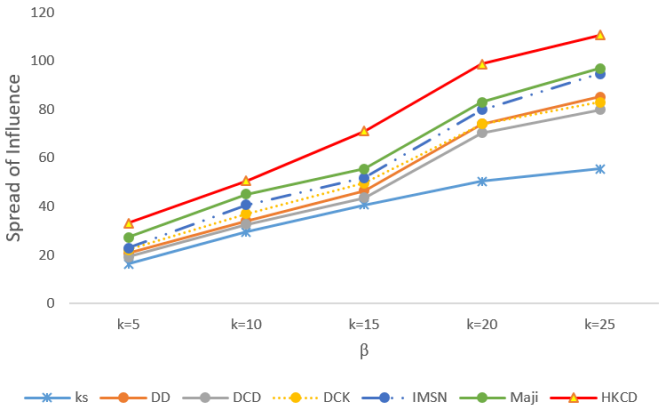
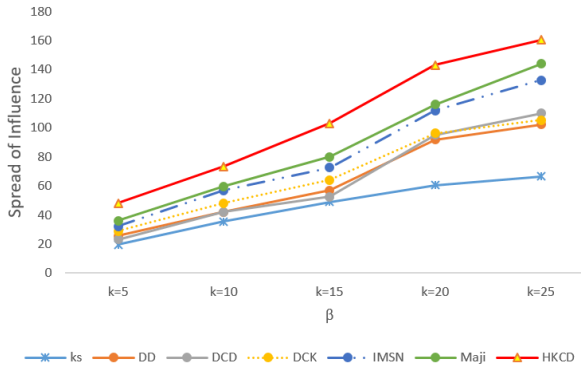
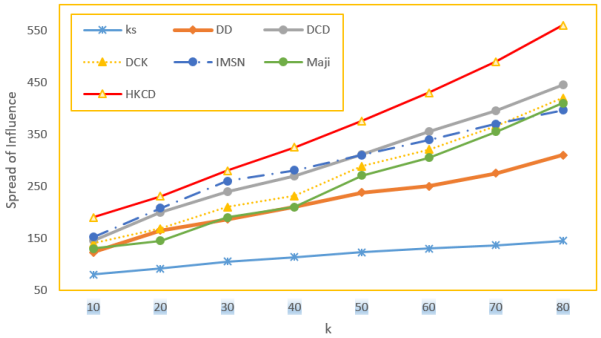


Figure 6. The spread amount of influence of the selected sets in the Jazz Musician dataset

Four other data sets with more nodes, 10, 20, 30, 40, 50, 60, 70, and 80 seed members, are considered. Then, the spread amount of influence obtained from each initial set of influential people selected by each method has been compared. Figure 7 shows the results from implementing the IC model starting from the seed sets determined by different methods.



a. NetScience



b. Hamsterster

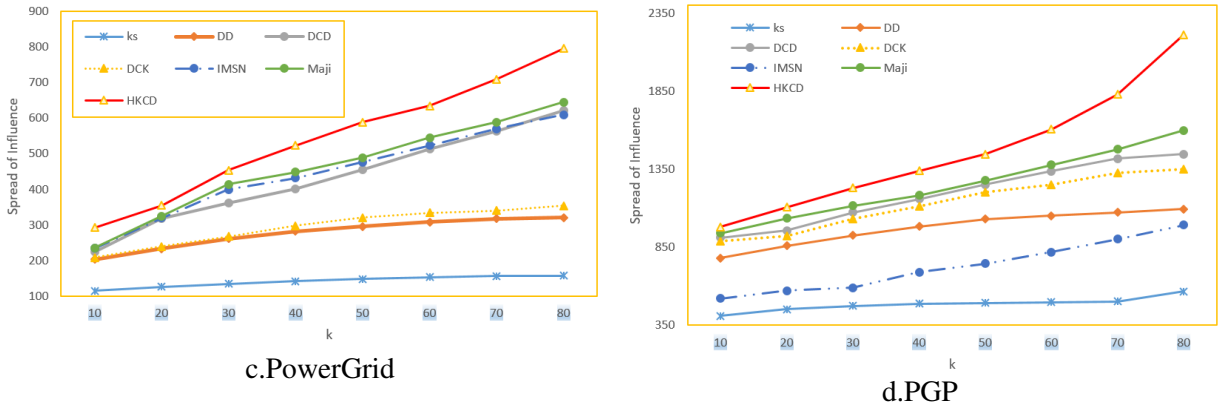


Figure 7. The results of the implementation of the IC model on the initial seed sets

Figure 7 shows that the proposed method, i.e., HKCD, has an outstanding performance in these data sets regarding the spread power compared to other methods. By increasing the member size of the initial seed set, the spread amount of influence obtained by this seed set is better than other methods. This remarkable performance of the proposed method is due to partitioning the network into communities and selecting the nodes of each community based on the diversity of their friends' presence in different communities and shells. Therefore, this process is done in all parts of the network instead of marketing only in limited parts.

5. Summary and future work

This article solved the problem of choosing the optimal set of influential people. For this purpose, a combined method based on the total weight of the edges connected to each node and the diversity of the presence of its neighbors in different shells and communities is proposed to select the best group of influential people in order to start the information propagation process. The proposed method was compared with other previously known methods in the two parts of ranking accuracy and the spread amount of influence. The methods were compared in the small data set and then in the large data set, and the results showed that the proposed method has more outstanding performance than other methods.

The significant results obtained by the proposed method compared to other methods can be seen in the intelligent selection of the initial seed set from different communities, which would spread to different parts of the network. In addition, selecting nodes based on the weight of the edges, the diversity of the presence of their friends in different shells, and communities was an additional reason for improving the method. Because the proposed method partitioned the network into different communities, the searches were limited from the whole network to the smaller communities, making it possible to use the method in more extensive networks. The proposed method introduces a framework for measuring the spread power and selecting a group of influential nodes in order to maximize the influence; Therefore, additional works can be done in each of its parts. For example, it is possible to improve the proposed method in such a way that it is suitable for weighted and directed networks.

Declarations

Ethical Approval
Not applicable.

Competing interests
The authors declare that they have no conflict of interest.

Authors' contributions
Bakhtyar and Amir wrote the main manuscript.
Bakhtyar, Rojiar, and Amir prepared All figures.
Rojiar, Amir, and Bryar edited the manuscript.
All authors reviewed the manuscript.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Availability of data and materials
The datasets analyzed during the current study are available in the konect repository, konect.cc.

6. References

[1] S. Pei and H. A. Makse, "Spreading dynamics in complex networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2013, no. 12, p. P12002, 2013.
[2] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.
[3] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Physics Reports,* vol. 650, pp. 1-63, 2016.
[4] S. H. Strogatz, "Exploring complex networks," *nature,* vol. 410, no. 6825, pp. 268-276, 2001.
[5] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of Internet topology using k-shell decomposition," *Proceedings of the National Academy of Sciences,* vol. 104, no. 27, pp. 11150-11154, 2007.
[6] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics,* vol. 74, no. 1, p. 47, 2002.
[7] F. Radicchi and C. Castellano, "Leveraging percolation theory to single out influential spreaders in networks," *Physical Review E,* vol. 93, no. 6, p. 062314, 2016.
[8] D. Centola, "The spread of behavior in an online social network experiment," *science,* vol. 329, no. 5996, pp. 1194-1197, 2010.
[9] C. De Bacco, D. B. Larremore, and C. Moore, "A physical model for efficient ranking in networks," *Science advances,* vol. 4, no. 7, p. eaar8260, 2018.
[10] M. Kitsak *et al.*, "Identification of influential spreaders in complex networks," *Nature physics,* vol. 6, no. 11, pp. 888-893, 2010.
[11] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences,* vol. 103, no. 23, pp. 8577-8582, 2006.
[12] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica a: Statistical mechanics and its applications,* vol. 391, no. 4, pp. 1777-1787, 2012.

[13] M. E. Newman, "A measure of betweenness centrality based on random walks," *Social networks,* vol. 27, no. 1, pp. 39-54, 2005.
[14] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," *Social networks,* vol. 23, no. 3, pp. 191-201, 2001.
[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab1999.
[16] M. Kistak and H. GALLOSLK, "Identifications of influential spreaders in complex network," *Natura Physics,* vol. 6, no. 11, pp. 888-893, 2010.
[17] A. Zeng and C.-J. Zhang, "Ranking spreaders by decomposing complex networks," *Physics Letters A,* vol. 377, no. 14, pp. 1031-1035, 2013.
[18] J. Bae and S. Kim, "Identifying and ranking influential spreaders in complex networks by neighborhood coreness," *Physica A: Statistical Mechanics and its Applications,* vol. 395, pp. 549-559, 2014.
[19] M. Li, R. Zhang, R. Hu, F. Yang, Y. Yao, and Y. Yuan, "Identifying and ranking influential spreaders in complex networks by combining a local-degree sum and the clustering coefficient," *International Journal of Modern Physics B,* vol. 32, no. 06, p. 1850118, 2018.
[20] L.-l. Ma, C. Ma, H.-F. Zhang, and B.-H. Wang, "Identifying influential spreaders in complex networks based on gravity formula," *Physica A: Statistical Mechanics and its Applications,* vol. 451, pp. 205-212, 2016.
[21] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, and T. Zhou, "Identifying influential spreaders by gravity model," *Scientific reports,* vol. 9, no. 1, pp. 1-7, 2019.
[22] A. Namtirtha, A. Dutta, and B. Dutta, "Weighted kshell degree neighborhood: A new method for identifying the influential spreaders from a variety of complex network connectivity structures," *Expert Systems with Applications,* vol. 139, p. 112859, 2020.
[23] P. G. Sun, Q. Miao, and S. Staab, "Community-based k-shell decomposition for identifying influential spreaders," *Pattern Recognition,* vol. 120, p. 108130, 2021.
[24] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment,* vol. 2008, no. 10, p. P10008, 2008.