Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **7**, Number **4**, November **2022**

Identify A Group of Influential Nodes in Social Networks Based on Overlapping Community Detection

Jing-Dong Wang

School of Computer Science
Northeast Electric Power University
Jilin 132000, China
School of Computer Science
Guangdong Atv Academy For Performing Arts
Hujing Road, Dongguan City,Guangdong Province, 523710, China
707569380@qq.com

Qi-Zi Mu

School of Computer Science
Northeast Electric Power University
Jilin 132000, China
965068552@qq.com

Yan-Kun Gao

Beijing Institute of computer technology and Application
Beijing 100000, China
jlugyk@163.com

Fan-Qi Meng*

School of Computer Science
Northeast Electric Power University
Jilin 132000, China
Corresponding Author: mengfanqi@neepu.edu.cn

Li-Na Zhou

Beijing Institute of computer technology and Application
Beijing 100000, China
1764413771@qq.com

Shuang Guo

Science and Technology Institute of Jilin City
Jilin 132000, China
52850658@qq.com

*Corresponding author: Fan-Qi Meng

Received December 14, 2021, revised March 24, 2022, accepted August 27, 2022.

Abstract. *Most of the existing influential nodes identification methods are prone to produce "rich-club" effects due to the possibility of nodes being highly clustered. In this paper, we propose a method to identify a group of influential nodes in social networks based on overlapping community detection. The method firstly detects overlapping communities in the network, then optimizes the K-shell decomposition to mine central nodes within communities, and simultaneously exploits the bridging influential nodes belonging to multiple communities. Finally, the two are combined into one set. Its innovation lies in considering the actual factors that nodes are in multiple interest groups, introducing overlapping community detection technology to divide nodes into communities, and overcoming the shortcomings of traditional community-based methods accurately classify nodes into a single community; meanwhile, it scatters the influential nodes throughout the network, so that the information can spread as wide as possible. The performance evaluated by the susceptible-infected-recovered(SIR) model shows that our method comprehensively considers the topology of the entire network, and is more accurate and stable, at the same time, the connections among influential nodes obtained by our method are more dispersed, which provides a new solution to avoid "rich-club" effect.*
Keywords: complex network, "rich-club" effect, susceptible-infected-recovered model, overlapping community

1. **Introduction.** Identifying a group of influential nodes in complex social networks is of great significance for epidemic prevention, rumor control, word-of-mouth marketing, and advertising [1–3]. Based on the network topology, the key to this research is to allow all nodes to spread information as much as possible, to find those nodes which can accomplish influence maximization of the entire network. Nowadays, topology-based methods have attracted more and more researchers' attention [4–8]. Most of these methods start from the node's attributes and location, sort the influence of all nodes and select the top-ranked nodes as the influential nodes in the network. This kind of method takes relatively little time and can be roughly divided into two types. Those are local-attributes-based methods and global-attributes-based methods.

The local-attributes-based methods measures the influence of nodes through the nodes' local information. Such as degree centrality (DC) [9]. It utilizes the information of the node and its direct neighbors, and has the characteristics of simple calculation and low time complexity. However, due to the incomplete consideration of nodes' information, it often fails to accurately identify influential nodes. As an improvement, Xu et al. [10] proposed a local clustering H-index (LCH) centrality measure considering the neighbourhood topology, the number and quality of neighbouring nodes. Zhu et al. [11] divided nodes into neighbouring layers and measured nodes based on the distance between surrounding nodes, and proposed a degree centrality method combining its own and neighbouring layer nodes' information of node influence identification method. The Global-attributes-based methods utilize the global topology of the entire network. Such as betweenness centrality (BC) [12] and closeness centrality (CC) [13]. Due to the computational complexity of both, they still cannot be applied to large-scale networks. Kitsak et al. [14] first found that a node's influence depends on its location in the network. They proposed K-shell decomposition to better describes nodes' spreading ability. However, the results obtained by the method are often too coarse-grained, and existing pseudo influential nodes (nodes that are only connected to the highest and second-highest layer). To improve the method resolution, Wang et al. [15] distinguished multiple nodes with the same shell value by considering nodes' iteration number, degree, and their neighbors' information. Namtirtha et al. [16] comprehensively considered the nodes' shell value, degree, proximity, and neighbors' influence and proposed a new K-shell hybrid method. In recent

years, with the deepening of research, novel results continue to emerge. Liu et al. [17] proposed a general tightness index GCC, and approximately found nodes with the highest GCC as the influential nodes through the K-means method. Bian et al. [18] proposed a sorting method based on the analytic hierarchy process. They used the multi-attribute decision-making technology AHP to aggregate several centrality methods to evaluate the influence of each node, and select nodes with the highest influence as influential nodes in the network. Zhang et al. [19] comprehensively measured the influence of nodes through the local bridging ability and the global connection strength.

However, because social networks follow a degree assortativity [20] and the nodes' degree follows a heavy-tailed distribution, the key point of most methods is to avoid the "rich-club" effect [21]. For example, in DC, nodes with a high degree are usually connected to many identical neighbors, causing a large number of nodes' influence overlaps during the propagation process. Liu et al. [22] avoid the "rich-club" effect by screening the nodes whose degree is greater than all surrounding neighbors' in the network. Since only the local information of the node is considered, the obtained results still have deviations.In recent years, with the development of community detection techniques, many researchers have considered combining community partitioning techniques to identify influential nodes in the network [23, 24]. These methods reduce the "rich-club" effect to a certain extent by controlling the influential nodes to individual communities. However, they classify all nodes into a particular community precisely, and thus contradict the real-life reality that nodes naturally belong to multiple communities due to having multiple identities.In addition, the above-mentioned methods are mainly carried out based on a single node. But the fact is that the spread of some diseases, rumors, or advertisements is often carried out under the joint action of multiple sources of infection, and there are often multiple influential nodes in the network. Given this, we propose a new method for identifying a group of influential nodes in social networks, which applies overlapping community detection technology. It finds the central nodes of each community through an optimized K-shell decomposition method and forms this set of influential nodes together with the bridging nodes that belong to most communities. Since nodes are closely connected within communities and sparsely connected between communities, nodes have an advantage in local information dissemination. Bridging nodes that connect multiple communities have a very important role in information dissemination between communities, and the distribution is relatively scattered. Thus, while ensuring that the node itself is sufficiently important, our method can make the spread range as wide as possible.

The innovation of the method lies in the division of communities through overlapping community discovery technology, which overcomes the shortcomings of traditional community-based influential node identification methods that accurately classify nodes into a single community, which is more inconsistent with the fact that nodes belong to multiple interest groups in real situations. At the same time, it comprehensively considers the internal community and the overall network structure and improves the accuracy of the algorithm.

The main contributions are as follows:

1) A method for identifying a group of influential nodes in social networks is proposed. In the method, the important role of community-centric nodes and bridging nodes for information dissemination is comprehensively considered;

2) Optimize the K-shell decomposition. Through nodes' local information, the influence of the pseudo-core nodes is reduced and the resolution of the impact of the K-shell decomposition on the nodes in each layer is improved;

3) Through the overlapping community detection theory, the possibility of highly clustered influential nodes is reduced, and the "rich-club" effect is avoided.

The main contents of this paper organizes as follows: The second chapter introduces the definition and algorithm flow of the proposed method. The third chapter presents the experimental model and results. And the fourth chapter summarizes the full paper and gives the conclusion.

2. **Overlapping Community based Influential Nodes Identification Method.** Social networks are naturally divided into multiple communities. Based on the characteristics that nodes are closely connected within communities and sparsely connected between communities, information spreads faster inside communities. From the perspective of the global network topology, bridging nodes that connect multiple communities play an important role in information spreading between communities. Based on the above factors, we propose a method CbKS+ (Community-based K-Shell+) for identifying a group of influential nodes based on overlapping community detection technology. The research block diagram is shown in Figure. 1. First, consider the individuals or organizations in the social network as nodes, and treat their relationships as edges to build the network. Then, the network is divided into communities through overlapping community detection technology. Utilize the optimized K-shell decomposition to obtain influential nodes inside each community; at the same time, mine the bridging influential nodes belonging to the most communities, and finally, put the two into on set to form a group of influential nodes. The community-centric nodes can ensure the rapid spread of information inside the community, and the bridging influential nodes highly connected to communities are not only scattered but also can ensure normal communication across communities. Therefore, while ensuring the propagation performance, the method disperses the influential nodes to the whole network as much as possible.

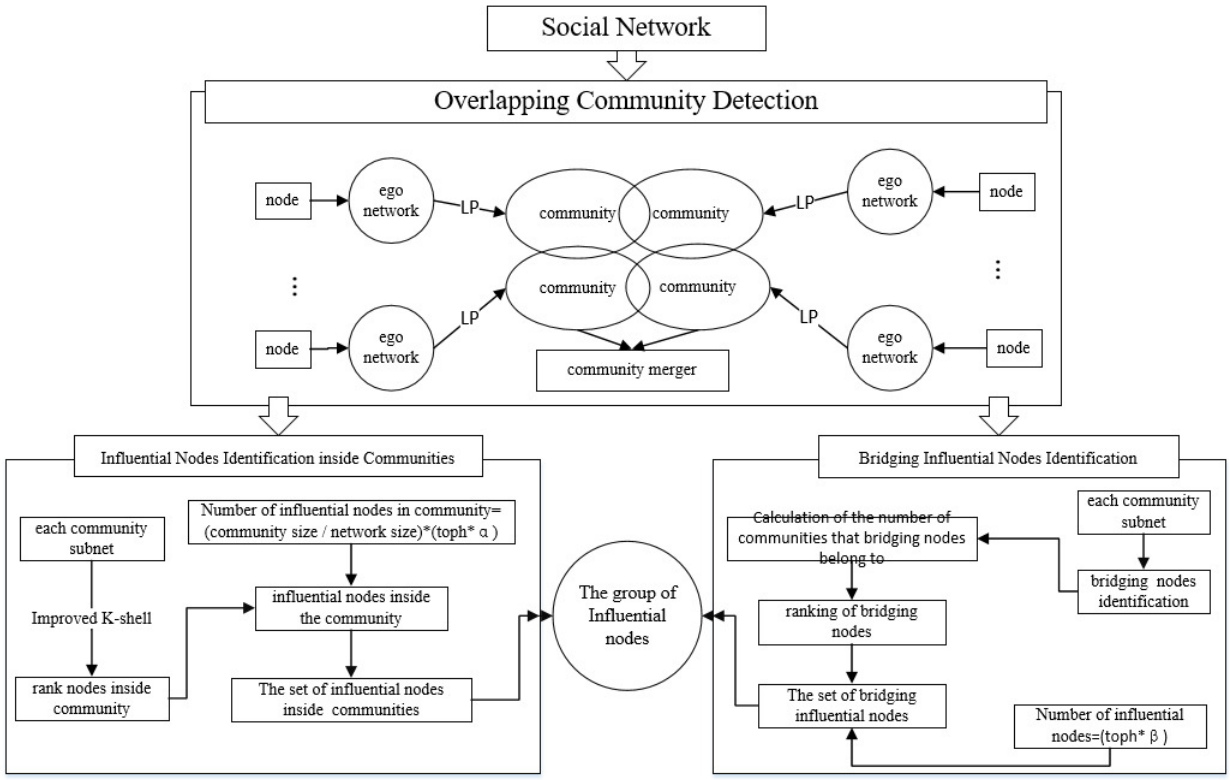


FIGURE 1. Research block diagram

2.1. **Overlapping community detection.** Traditional community detection methods accurately classify nodes into a certain community, but in the real world, nodes naturally belong to multiple interest groups based on their multi-attribute characteristics. Therefore, based on the actual situation and considering the efficiency of the method, in this paper, we divide the network into several overlapping communities through DEMON [25], a locally extended overlapping community discovery method. The method uses all nodes in the entire network as seeds, expands the community structure through nodes' neighbors, and then merges similar community structures. This makes the method have low time complexity while ensuring the quality of received communities. The steps for dividing overlapping communities in social networks are as follows:

Step 1: For all nodes in the network, the local ego network is found according to its neighbors;

$$EgoMinusEgo(v, G) = -g(v, EN(v, G)) \tag{1}$$

Where $EgoMinusEgo(v, G)$ is a defined function used to identify the local ego network of node v. $EN(v, G)$ is a subgraph extraction operation, representing the subgraph $G'(V', E')$, in where V' is the combination of node v and all its neighbors in the original graph $G(V, E)$, and E' is a subset of E , including all edges of (μ, v) ($\mu \in V'$,v belongs to V'); $-g$ represents a graph vertex difference operation function, $-g(v, G)$ produces a copy of G that remove node v and all edges of v.

Step 2: For the ego network formed by each node, through the label propagation algorithm LP to make each node vote for its surrounding communities when it has limited observation of the global system. So as to obtain the communities $c(v)$ contained in the ego network;

$$c(v) = \left\{C | C \in LP(EgoMinusEgo(v, G))\right\} \tag{2}$$

where C is a community in the community set $c(v)$ which obtained after the ego network community is divided.

Step 3: Merge the communities in the ego network. The large community C and the small community I are merged when they are highly overlapped. Then delete C and I, and add the merged community in $c(v)$.

2.2. **Mining the set of influential nodes inside communities.** Nodes in the community center have inherently better communication capabilities than nodes at the border. Kitsak [14] uses the K-shell decomposition to find nodes at the center of the network. They calculate the K-shell value of every node by stripping the node with the smallest degree layer by layer. However, because the method is too coarse grained and ignores the nodes' local attributes, some pseudo core nodes that are only connected to the core layer and the next core layer are added to the diffusion seed set during the identification process. In view of this, we propose an improved K-shell decomposition method KL, which optimizes the K-shell decomposition by considering the direct and indirect effects of the node s' neighbors and second-neighbors. And then we use KL to identify the influential nodes inside each community. Compared with K-shell decomposition, KL considers more neighborhood information of nodes, the resolution is significantly improved, and the influence of pseudo influential nodes is also reduced. It is worth pointing out that the inspiration comes from the LocalRank algorithm proposed by Chen et al [26].

The influence of node v0 in the ith community calculated by the following formula:

$$KL_{v_0}^i = \frac{ks_{v_0} + \sqrt{\frac{\sum_{v_1 \in \Gamma_{v_0}} \sum_{v_2 \in \Gamma_{v_1}} R(v_2)}{c_i}}}{\max\left(ks_i\right)} \tag{3}$$

Where ks_{v_0} represents the K-shell value of node v_0 inside the community, Γ_{v_1} is the set of nearest neighbors of the node v_0, $R(v_2)$ is the number of the nearest and the next nearest neighbors of node v_2. $\sum_{v_2 \in \Gamma_{v_1}} R(v_2)$ calculates the number of third-ordered neighbors of node v_1. c_i represents the whole number of nodes in the ith community, and $\max(ks_i)$ represents the max K-shell value in the ith community.

The influential nodes inside each community are aggregated to form the set of influential nodes inside communities S_1 with the size of r. The number of seeds is allocated to each community according to the size of the community, k_i, which represents the number of seeds is calculated by the following formula:

$$k_i = \frac{c_i \times r}{n} \tag{4}$$

The method flow of the influential node set identification method inside communities is shown in Algorithm 1. First, enter the complex network $G = (V, E)$, use formula(4) to determine the size of the influential seed set, allocate a corresponding number of seeds k_i to each community, and obtain the ranking result of nodes' influence inside each community according to formula (3). The top k_i nodes in each community are selected to join S_1, the set of influential nodes inside communities.

Algorithm 1

Input:
The network, $G = (V, E)$;
The size of influential spreaders set S_1, r;
The number of interactions T and threshold l for DEMON;
Output:
The $top - r$ influential spreaders set inside communities, S_1;

```
S = ∅;
Get the community results;
while |S_1 < r| do
  for i from 0 to len(community) do
    Calculate the size of influential spreaders set in community_i, using formula (4);
    Calculate KL^i_{v_0} using formula (3) for nodes in community_i;
    sort the KL^i_{v_0} value of each node and add the top − k_i nodes to S_1;
  end for
end while
return S_1;
```

2.3. **Bridging Influential Nodes Identification.** From the perspective of the global network topology, the nodes belonging to multiple communities often live on the boundaries of each community. However, for the information dissemination across communities, these nodes have inherently better spreading capabilities. For example, if a person lives in multiple communities, he/she should have friends in different communities and can play an important role in receiving and disseminating information among communities. Moreover, the more communities he/she belongs to, the stronger his/her ability to spread information. Therefore, in order to realize the widespread dissemination of information throughout the network, we first identify all bridged community nodes in the social network to join the candidate set, and then sort these nodes according to the number of crossed communities, and select the $top - p$ nodes to join S_2, the bridging influential node set. For the network $G = (V, E)$, we combine S_1, the influential node set inside communities with S_2, the bridging influential node set to obtain the group of influential nodes

S with size h ($S = S_1 \cup S_2$). Where the size of S_1 is $r = \alpha h$; p, the number of seeds allocated to S_2 is calculated as follows $p = \beta h$ ($\alpha + \beta = 1$).

The overall process description of CbKS+ is shown in Algorithm 2.

Algorithm 2

Input:

The network, $G = (V, E)$;
The size of influential spreaders set S_1, r;
The size of influential spreaders set S_2, p;
The number of interactions T and threshold l for DEMON;

Output:

A group of influential spreaders in the social network, S;

$S = \emptyset$; $S_1 = \emptyset$; $S_2 = \emptyset$;
Get the community results;
while $|S_1 < r|$ **do**
  for i from 0 to $len(community)$ **do**
    Calculate the size of influential spreaders set in $community_i$, using formula (4);
    Calculate $KL^i_{v_0}$ using formula (3) for nodes in $community_i$;
    sort the $KL^i_{v_0}$ value of each node and add the $top - k_i$ nodes to S_1;
  end for
end while
Set overlap value as 1 for nodes in V
for each node v in V **do**
  for i from 0 to $len(community) - 1$ **do**
    if node v in $community_i$ **then**
      $overlap(v) = overlap(v) + 1$
    end if
  end for
end for
while $|S_2 < p|$ **do**
  add node v with biggest overlap value to S_2;
end while
add S_1 to S, add S_2 to S
return S;

3. **Experiment results and analysis.** A good influential node group identification algorithm should prove its robustness in networks with different structures, node sizes, initial number of infected nodes, infection rate and recovery rate. To evaluate the performance of the proposed algorithm, we compared it with other six algorithms. The data set used for the experiment is shown in Table 1. According to the table, Karate [27] is a real social network constructed by scholars by observing an American university karate club containing 34 nodes and 78 edges. Email [28] is a communication network for email users, including 1133 nodes and 5451 edges. Hamster [29] is a network that reflects friendship and family links between users, where nodes and edges respectively represent the users and the relationship between them. In the facebook network [30], nodes and edges represent users and friendships between users. HepPh [31] is an authors' collaborative network of arXiv scientific papers. A node in the network represents an author, and an edge represents the co-autorship in a paper between the nodes. Among the table, n represents the size of the network, m represents the number of edges, $\langle k \rangle$ represents the average degree, k_{max} represents the maximum degree, and $\langle c \rangle$ is the average clustering coefficient of the network.

TABLE 1. The basic topological features of all synthetic and real networks used for experiment.

Networks	n	m	$\langle k \rangle$	k_{max}	$\langle c \rangle$
karate	34	78	4.59	17	0.57
email	1133	5454	9.62	71	0.22
hamster	2426	16631	13.711	273	0.538
facebook	4039	88234	43.691	1.45	0.606
HepPh	12008	237010	8.083	281	0.633

In networks of different sizes, by analyzing the dynamics of information dissemination over time, the final spreading scale under different infection rates and the initial number of nodes, the accuracy of the algorithm in identifying the group of influential nodes and its robustness on the network structure and scale are measured. At meanwhile, the effectiveness of the proposed algorithm in avoiding "rich-club" is characterized by the average distance.

3.1. **Evaluation index.** The best way to measure the influence of a group of nodes is to use them as initial nodes to spread information in the real network. We use the SIR model [32] to verify the proliferation ability of the node. The SIR model was originally used to simulate the dynamic spread of diseases, and was later widely used in similar scenarios, such as word-of-mouth marketing. At the beginning of spread, the nodes in the SIR model are divided into three states, susceptible state, infected state and recovery state. First, the influential nodes obtained through the method is regarded as infected nodes, and the other nodes are set to a susceptible state. In each propagation iteration, each infected node infects its randomly selected neighbors with a probability of μ. At the same time, each infected node will recover with a probability of β and will not be infected again. The infection rate λ is defined as: $\lambda = \frac{\mu}{\beta}$. Since if the value is less than the threshold value, it will cause the spreading range to be too small or not spreading. If the value is too large, almost all the influential nodes identified by the algorithms can spread the information throughout the entire network. Here, We set μ to 1.5 times to its threshold so that the information can be spread widely in the network.(The threshold of μ is defined as $\mu_{max} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$), where $\langle k \rangle$ represents the average degree in the network).

The infection scale $F(t)$ is used to compare the spreading ability of nodes at time t. The $F(t)$ at time t is defined as:

$$F(t) = \frac{n_{I_{(t)}} + n_{R_{(t)}}}{n} \tag{5}$$

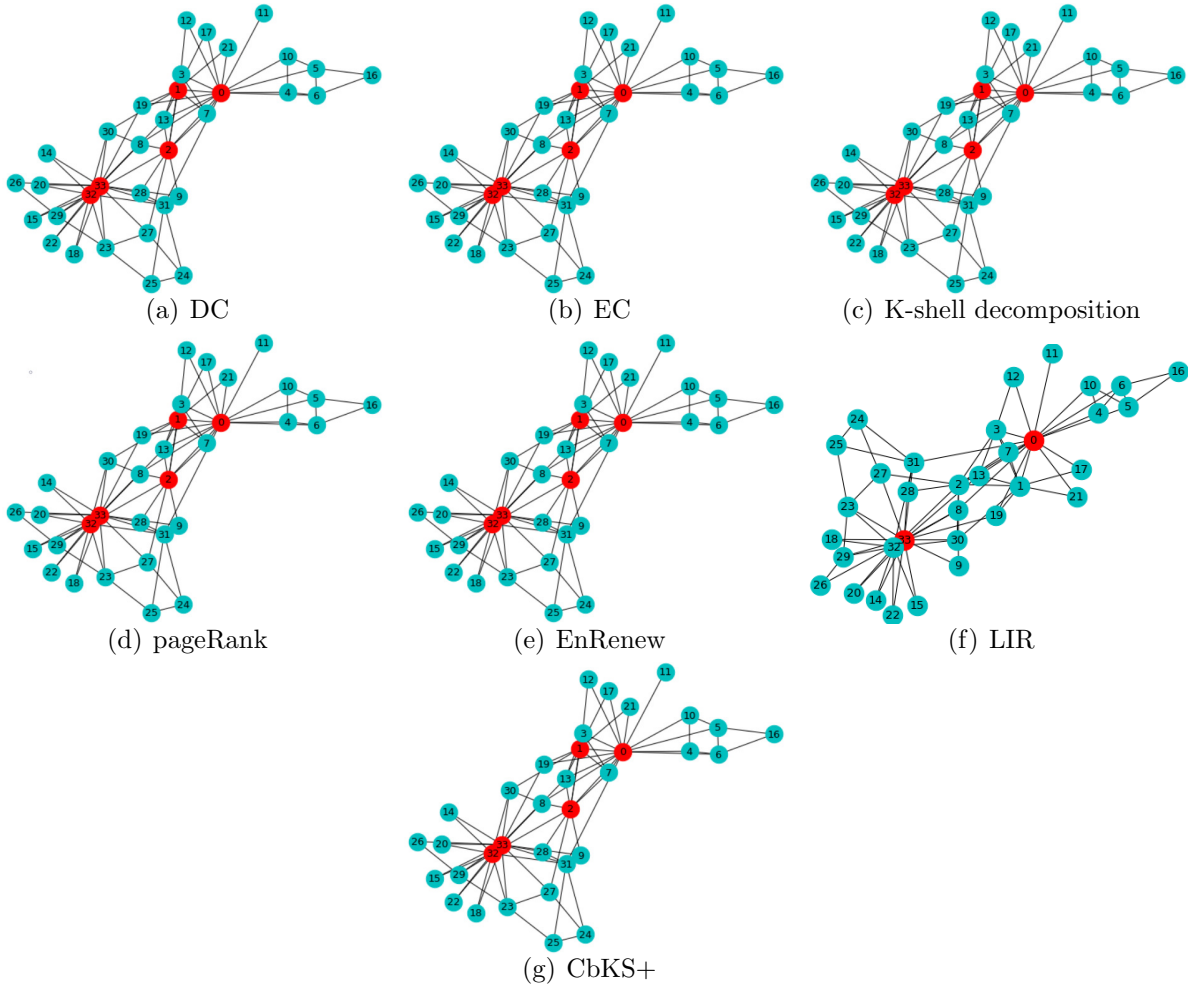
Where $n_{I_{(t)}}$ and $n_{R_{(t)}}$ represent the number of nodes in infection state and recovery state at time t respectively. n is the total number of points in the network. At time t, the larger the number of $F(t)$, the more nodes are infected by the initial influential nodes. As for the same $F(t)$, the more obvious the t is, the faster the node influence spreads in the network. F_{t_c} is the final affected scale. The larger the F_{t_c} value is, the stronger the propagation ability of the initial nodes are. F_{t_c} is defined as:

$$F_{t_c} = \frac{n_{R_{t_c}}}{n} \tag{6}$$

Where t_c represents the time when the propagation reaches the stable state, and $n_{R_{t_c}}$ is the number of infected nodes.

3.2. **Analysis of experimental results on small-scale data sets.** Firstly, use Karate to verify the feasibility of the proposed method on small-scale networks. The influential nodes extracted by each method are shown in Figure.2. Where the blue nodes are ordinary nodes, and the red nodes represent influential nodes. Table 2 lists the top-5 nodes obtained by the seven comparison methods degree centrality (DC), eigenvector centrality (EC) [33], K-shell decomposition, PageRank [34], EnRenewRank (EnRenew) [35], LIR and CbKS+) in detail. In the table, each row represents the 5 most influential node IDs in the network obtained by each algorithm (in no particular order).

According to Table 2, in LIR, only two nodes with LI value equal to 0 are identified, whose IDs are 33 and 0. Although these two nodes have high influence, it is found that when the number is specified, the LIR is not necessarily able to give the full number of of influential nodes; The node IDs of the top-5 influential nodes calculated by DC, EC, K-shell decomposition, EnRenew, pageRank, and CbKS+ are all 33, 0, 32, 2, 1. It can be seen that the results of these algorithms in small-scale networks are exactly the same, indicating that CbKS+ can accurately identify influential nodes in small-scale networks.



(a) DC (b) EC (c) K-shell decomposition

(d) pageRank (e) EnRenew (f) LIR

(g) CbKS+

FIGURE 2. Influential nodes obtained by different algorithms in karate club network

3.3. **Analysis of experimental results on large-scale data sets.** The proposed algorithm CbKS+ is compared with degree centrality, eigenvector centrality, K-shell decomposition, pageRank and EnRenew in the real network through the SIR model to verify the algorithms' performance in large social networks. (To simplify the experiment while ensuring high performance, the experiments use the following configurations: $\alpha = \beta = 0.5$,

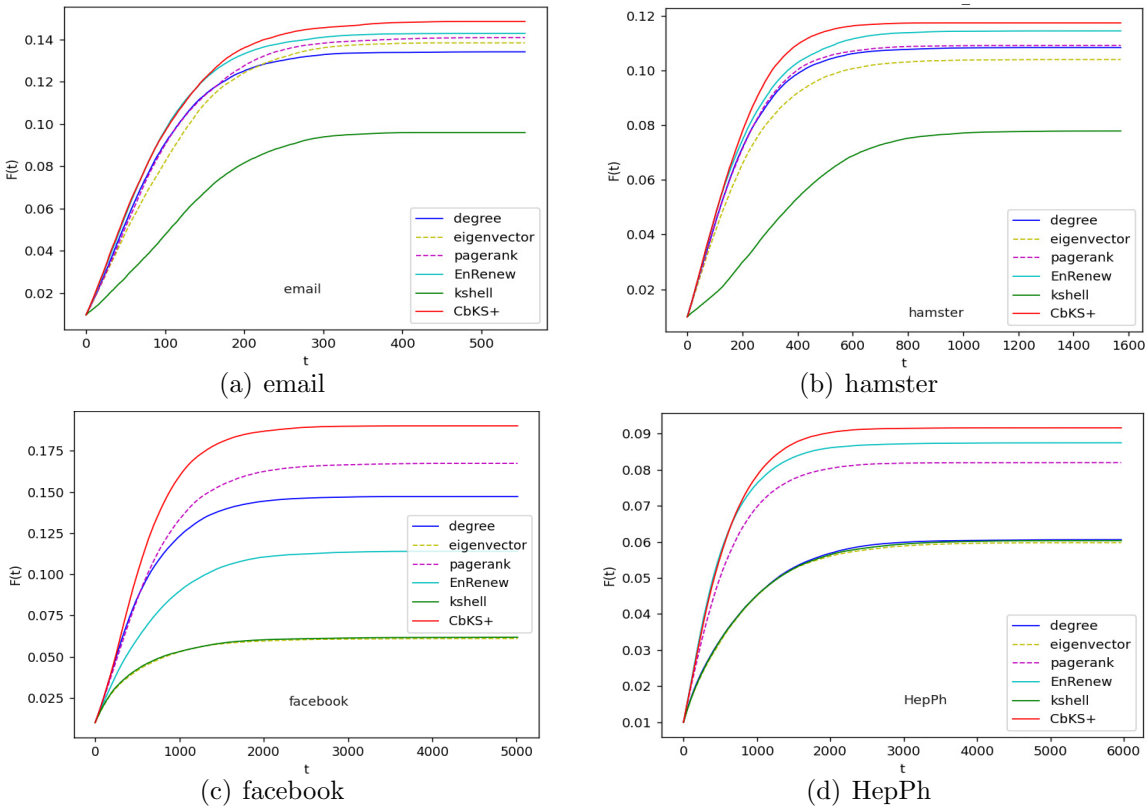
TABLE 2. Top-5 influential nodes in karate club.

Algorithms	Node 1	Node2	Node 3	Node4	node 5
DC	33	0	32	2	1
EC	33	0	2	32	1
K-shell decomposition	33	0	32	2	1
PageRank	33	0	32	2	1
EnRenew	33	0	32	2	1
LIR	33	0			
CbKS+	33	0	32	2	1

which are defined in 2.3). Take the $top-1\%$ nodes in each comparison method as the initial infection nodes, and conduct information dissemination analysis based on the SIR model in the four networks. These nodes, which only appear in the $top-1\%$ of one method, just reflect the different focus of methods.

The steady state is reached by taking $t = 100$, and 200,000 independent experiments are conducted to take the average value. The total number of infected and recovered nodes during this time is taken as the propagation influence of the nodes. Figure.3 shows the simulation results of the infection scale $F(t)$ obtained by each algorithm over time in four networks. It can be seen that K-shell decomposition performs the worst in all networks starting from early propagation, and the gap between K-shell decomposition and other algorithms is larger in networks with larger average degree, which is inseparable from the presence of pseudo influential nodes in the core set. CbKS+ has reached the highest peak value in all four networks, and most of the time, it maintains a leading position in terms of both the infected scale and the speed of propagation. Especially in the facebook network, with the growth of time, the infected scale of CbKS+ has been achieved a surprising result. This shows that the proposed method is more accurate, and has well optimized the K-shell decomposition algorithm. In the early propogation stage of the email network, the information dissemination speed of CbKS+ is slightly lower than that of EnRenew, which is inseparable from the slow functioning of the bridging influential nodes, which are located in the community boundary.

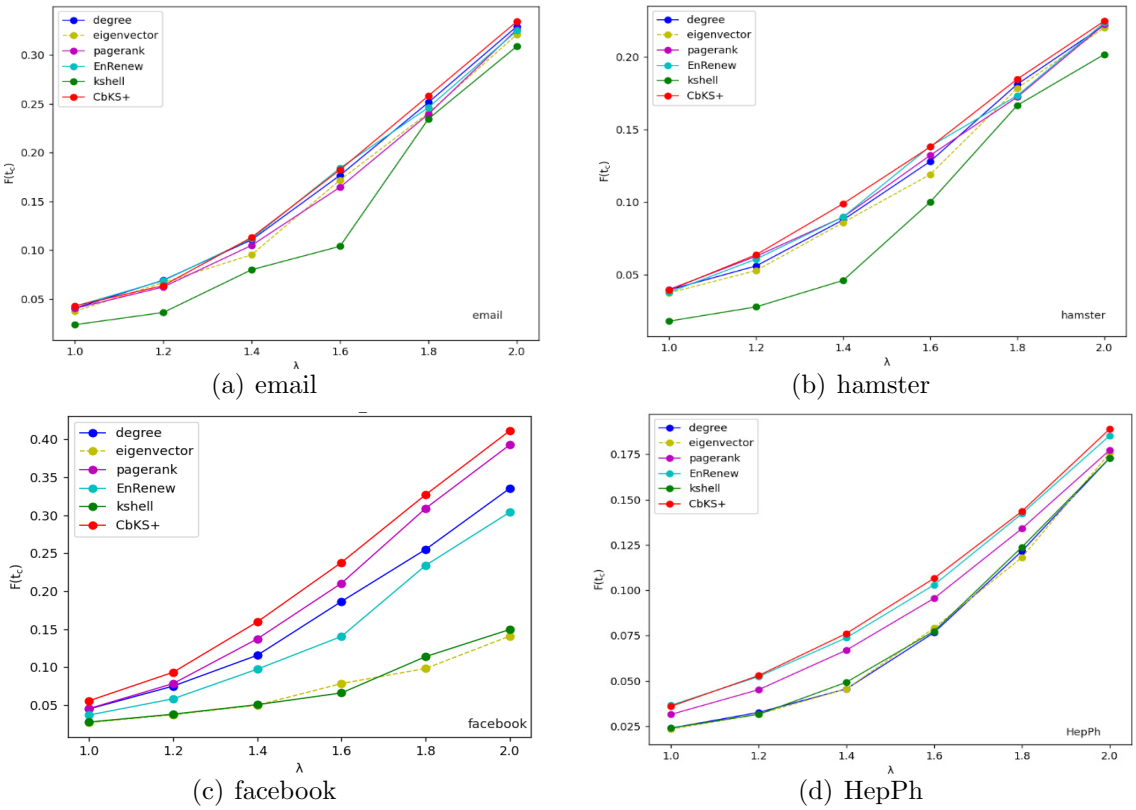
Secondly, a good influential nodes identification method should be robust to network structure, the number of initial nodes, and the infection rate. Figure.4 shows the variation of the final affected scale with infection rate for six algorithms in different networks at $1.0 < \lambda < 2.0$. According to the figure, CbKS+ performs very similarly to the other algorithms when the infection rate is low. In the email network, where both the maximum and average degrees are small, the proposed algorithm outperforms the other comparative algorithms in general, and only slightly underperforms EnRenew when λ equals 1.2. However, the difference between the six algorithms is not significant as the results of the information dissemination are averaged over 200,000 experiments conducted. In the rest of the networks, the final affected size of the proposed methods steadily increases as λ increases, while the final infection size of CbKS+ exceeds that of the other methods, and the larger the λ, the superior its performance. This also demonstrates the ability of the method to generalize for different propagation probabilities. The ability of a node to spread is not only related to its own structural properties, but also to the location of the node. Our approach not only emphasizes the important role of bridging nodes in the dissemination of information between communities, but also avoids overlapping influence of nodes from a community perspective. This allows the method to achieve better results than other methods in the case of scaling infection rates.



(a) email (b) hamster

(c) facebook (d) HepPh

FIGURE 3. Comparison of the spreading scale $F(t)$ as a function of infected time t of six methods on four networks.

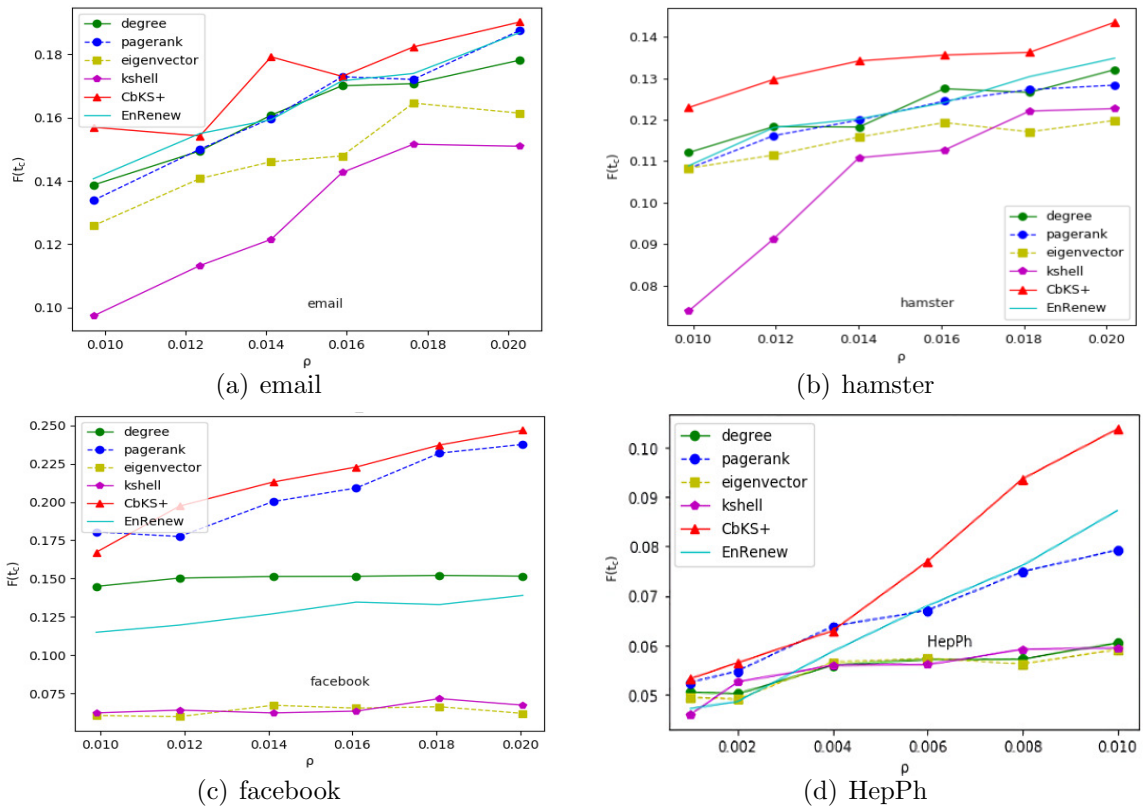


(a) email (b) hamster

(c) facebook (d) HepPh

FIGURE 4. Comparison of the final affected scale F_{t_c} on different λ of six methods on four networks

To access the impact of the initial number of nodes on the final spread scale, we simulated the spreading process on the four networks by changing the ratio of the initially infected nodes from 0.1 to 0.2. Figure.5 shows how the final impact changes with the increase of initially infected nodes. It can be seen that CbKS+ basically presents better results than the other tested methods. It illustrates that in most cases, the influential nodes obtained through the CbKS+ can spread information faster and ultimately affect a larger range. The performance of degree centrality is not the worst in any of the remaining three networks except Facebook, due to the fact that the gap between the average and maximum degrees is not particularly large in these networks, the nodes with the largest degrees are not particularly close, and the nodes do not have a high degree of overlap in their spread. It can be manifested that the DC can still play a certain role in small networks, but overall, it does not have a great impact on the final affected scale. Besides, the K-shell decomposition does have a large number of pseudo influential nodes, resulting in the worst performance in the final affected scale of all networks. And synthesize the results in Figure.3, the spread range of the k-shell decomposition grows slowly, which shows that the "rich-club" effect for the k-shell decomposition is significant, and it tends to consume more time zones to affect the same number of nodes. In the email network, the proposed method achieves similar results with PageRank and EnRenew, and to some extent, it has only a slight advantage. This situation suggests that in a small-scale network, where the differences in community structure are not particularly pronounced, the absolute advantage of the proposed method in identifying the removed nodes in terms of position is reduced. Overall, however, CbKS+ achieves a superior performance compared to the other algorithms, suggesting that the group of influential nodes mined by CbKS+ have a more important role in the networks than these algorithms.

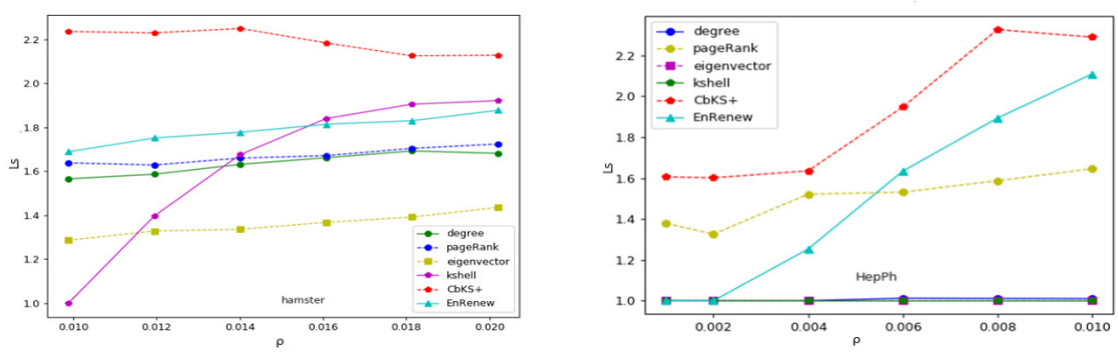


(a) email (b) hamster

(c) facebook (d) HepPh

FIGURE 5. Comparison of the final affected scale F_{t_c} on different λ of six methods on four networks

A comprehensive analysis of the differences between CbKS+ and other comparative algorithms for different network structures and sizes shows that the proposed algorithm is more robust to network structure, initial number of nodes and infection rate than other algorithms, and is more suitable for large-scale social networks. At last, the effectiveness of CbKS+ in avoiding the overlap of influence between nodes was verified in the hamster network and the largest network HepPh by analyzing the average distance of influential nodes. The change of the average distance Ls between influential nodes obtained by different methods with the initially infected nodes ρ is shown in Figure. 6. As can be seen from the figure, the Ls value obtained by CbKS+ is much larger, and in large-scale social network, the average distance tends to increase as the number of selected initial infected nodes increases. It indicates that the influential nodes acquired by CbKS+ are relatively scattered among each other, which provides a new solution to avoid the "rich-club" effect.



(a) hamster (b) HepPh

FIGURE 6. Average distance of influential nodes in six methods

4. **Conclusions.** Since social networks follow a degree assortativity, the key point of most current influential node identification methods is to avoid the "rich-club" effect. Moreover, these studies are mainly carried out based on a single influential node, which is not in line with the reality that diseases, rumors, or advertisements work together with multiple sources of infection. Based on this fact, in this study, we propose a method for identifying a group of influential nodes based on overlapping community detection. By combining the set of bridging influential nodes and the set of center nodes inside communities, the group of influential nodes in the entire network is obtained. Since the influential nodes is controlled in the center of each community, and the connections between the bridging influential nodes are relatively scattered, the proposed method effectively avoids the "rich-club" effect. Compared with the traditional DC, EC, PageRank, K-shell decomposition, and the new method EnRenewRank based on information entropy, our method can better distinguish the influence of nodes with higher accuracy and stability, and can be applied to public opinion control, epidemic prevention, and control, and advertising and marketing in large-scale social networks.

The research result show that the influence of nodes in social networks not only depends on the local influence of the nodes themselves but also depends on the structural superiority of the nodes. However, which of the two plays a key role in information dissemination needs further research. In addition, due to the continuous expansion of social networks, their network topology changes are becoming more and more complicated. In this case, how to efficiently and accurately mine influential node groups in the dynamic social network to achieve emergency public opinion control will be of more practical significance. The key to this problem lies in how to accurately describe the dynamic change process of the network topology and aggregate the information between them. Therefore,

the construction of the dynamic evolution model of social networks is the focus of our next work.

Acknowledgment. This research was funded by the Jilin City Science and Technology Innovation Development Project "Sentiment Classification Research of Jilin Tourism Online Review Text" (No. 20200104108).

REFERENCES

[1] E.K. Wang, C.M. Chen, S.M. Yiu, M.M. Hassan, Incentive evolutionary game model for opportunistic social networks, *Future Generation Computer Systems*, vol. 102, pp. 14–29, 2020.

[2] T.Y. Wu, X. Guo, L. Yang, Q. Meng, C.M. Chen, A Lightweight Authenticated Key Agreement Protocol Using Fog Nodes in Social Internet of Vehicles, *Mobile Information Systems*, vol. 2021, 3277113, 2021.

[3] C.T. Li, T.Y. Wu, C.M. Chen, A provably secure group key agreement scheme with privacy preservation for online social networks using extended chaotic maps, *IEEE Access*, vol. 6, pp. 66742–66753, 2018.

[4] L. Qiu, J. Zhang, X. Tian, Ranking influential nodes in complex networks based on local and global structures, *Applied Intelligence*, vol. 51, no. 7, pp. 4394–4407, 2021.

[5] X. Yang, F. Xiao, An improved gravity model to identify influential nodes in complex networks based on k-shell method, *Knowledge-Based Systems*, vol. 227, 107198, 2021.

[6] X.H. Yang, Z. Xiong, F. Ma, X. Chen, Z. Ruan, P. Jiang, X.Xu, Identifying influential spreaders in complex networks based on network embedding and node local centrality, *Physica A: Statistical Mechanics and its Applications*, vol. 573, 125971, 2021.

[7] G. Maji, Influential spreaders identification in complex networks with potential edge weight based k-shell degree neighborhood method, *Journal of Computational Science*, vol. 39, 101055, 2020.

[8] M. Wang, W. Li, Y. Guo, X. Peng, Y. Li, Identifying influential spreaders in complex networks based on improved k-shell method, *Physica A: Statistical Mechanics and its Applications*, vol. 554, 124229, 2020.

[9] L.C. Freeman, Centrality in social networks conceptual clarification, *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[10] G.Q. Xu, L. Meng, D.Q. Tu, P.L. Yang, LCH: a local clustering H-index centrality measure for identifying and ranking influential nodes in complex networks, *Chinese Physics B*, vol. 30, no. 8, 088901, 2021.

[11] J. Zhu, L. Wang, Identifying Influential Nodes in Complex Networks Based on Node Itself and Neighbor Layer Information, *Symmetry*, vol. 13, no. 9, 1057, 2021.

[12] G. Sabidussi, The centrality index of a graph, *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[13] A. Bavelas, A mathematical model for group structures, *Applied anthropology*, vol. 7, no. 3, pp. 16–31, 1948.

[14] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L.Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, *Nature physics*, vol. 6, no. 11, pp. 888–893, 2010.

[15] Z. Wang, Y. Zhao, J. Xi, C. Du, Fast ranking influential nodes in complex networks using a k-shell iteration factor, *Physica A: Statistical Mechanics and its Applications*, vol. 461, pp. 171–181, 2016.

[16] A. Namtirtha, A. Dutta, B. Dutta, Identifying influential spreaders in complex networks based on kshell hybrid method, *Physica A: Statistical Mechanics and its Applications*, vol. 499, pp. 310–324, 2018.

[17] H.L. Liu, C. Ma, B.B. Xiang, M. Tang, H.F Zhang, Identifying multiple influential spreaders based on generalized closeness centrality, *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 2237–2248, 2018.

[18] T. Bian, J. Hu, Y. Deng, Identifying influential nodes in complex networks based on AHP, *Physica A: Statistical Mechanics and its Applications*, vol. 479, pp. 422–436, 2017.

[19] D. Zhang, Y. Wang, Z. Zhang, Identifying and quantifying potential super-spreaders in social networks, *Scientific reports*, vol. 9, no. 1, pp.1–11, 2019.

[20] J. Leskovec, C. Faloutsos, Sampling from large graphs, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 631-636, 2006.

[21] V. Colizza, A. Flammini, M. A. Serrano, A.Vespignani, Detecting rich-club ordering in complex networks, *Nature physics*, vol.2, no.2, pp. 110–115, 2006.

[22] D. Liu, Y. Jing, J. Zhao, W. Wang, G. Song, A fast and efficient algorithm for mining top-k nodes in complex networks, *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.
[23] W. Zhang, J. Yang, X.Y. Ding, X.M. zou, H.Y. Han, Q.C.Zhao, Groups make nodes powerful: Identifying influential nodes in social networks based on social conformity theory and community features, *Expert Systems with Applications*, vol. 125, pp. 249–258, 2019.
[24] M.M. Tulu, R. Hou, T. Younas, Identifying influential nodes based on community structure to speed up the dissemination of information in complex network, *IEEE Access*, vol. 6, pp. 7390–7401, 2018.
[25] M. Coscia, G. Rossetti, F. Giannotti, D. Pedreschi, Demon: a local-first discovery method for overlapping communities, *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 615–623, 2012.
[26] D.B. Chen, L. Lü, M.S. Shang, Y.C. Zhang, T. Zhou, Identifying influential nodes in complex networks, *Physica A: Statistical Mechanics and its Applications* vol. 391, no. 4, pp. 1777–1787, 2012.
[27] W.W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.
[28] J. Kunegis, Konect: The koblenz network collection, *Proceedings of the 22nd international conference on world wide web*, pp. 1343–1350, 2013.
[29] R. Rossi, N. Ahmed, The network data repository with interactive graph analytics and visualization, *Twenty-ninth AAAI conference on artificial intelligence*, pp. 4292–4293, 2015.
[30] J.J. McAuley, J. Leskovec, Learning to discover social circles in ego networks, *Advances in neural information processing systems*, pp. 548–56, 2012.
[31] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: Densification and shrinking diameters, *ACM transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no.1, pp. 2–es, 2007.
[32] T. Zhou, Z. Fu, B.H Wang, Epidemic dynamics on complex networks, *Progress in Natural Science*, vol. 16, pp. 452–457, 2006.
[33] C.F.A. Negre, U.N.Morzan, H.P.Hendrickson, R. Pal, G.P. Lisi, J.P. Loria, I. Rivalta, J. Ho, V.S. Batista, Eigenvector centrality for characterization of protein allosteric pathways, *Proceedings of the National Academy of Sciences*, vol. 115, no. 52, pp. E12201-E12208, 2018.
[34] D.F. Gleich, PageRank beyond the web, *Siam Review*, vol. 57, no. 3, pp. 321–363, 2015.
[35] C. Guo, L. Yang, X. Chen, D.B. Chen, H. Hao, Influential nodes identification in complex networks via information entropy, *Entropy*, vol.22, no. 2, 2020.