Received: 31 October 2019 | Revised: 28 December 2019 | Accepted: 30 December 2019
DOI: 10.1002/cpe.5677

WILEY

RESEARCH ARTICLE

Influence maximization based on activity degree in mobile social networks

Min Gao[1,2] | **Li Xu**[1,2] | **Limei Lin**[1,2] | **Yanze Huang**[3] | **Xinxin Zhang**[1,2]

[1]College of Mathematics and Informatics, Fujian Normal University, Fuzhou, China
[2]Fujian Provincial Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou, China
[3]School of Mathematics and Physics, Fujian University of Technology, Fuzhou, China

Correspondence
Li Xu, College of Mathematics and Informatics, Fujian Normal University, Fuzhou, China.
Email: xuli@fjnu.edu.cn

Funding information
China Postdoctoral Science Foundation, No.2017M612107; The National Natural Science Foundation of China, No.61702100, No.61771140, U1905211

Summary
The problem of influence maximization (IM) has become an important research topic due to the rapid growth of mobile social networks. It attempts to identify a set of nodes, referred to as influencers, contributing to the spread of maximum information. In this article, we present the construction of social relation graph based on mobile communication data. And we propose a new centrality measure—activity degree to characterize the activity of nodes. By combining the local attributes of nodes and the behavioral characteristics of nodes to measure node activity degree, which can be used to evaluate the influence of users in mobile social networks, we introduce Susceptible-Infected-Susceptible model to simulate the dynamic spreading of information. We take advantage of the two indicators the degree centrality and the betweenness centrality to get a better ranking results. In comparison with spanning graph and initial graph, the results of comparison demonstrate that our algorithm has advantages in the scope of influence propagation.

KEYWORDS
activity degree, influence maximization, mobile social networks, SIS model

1 | INTRODUCTION

In the wake of the popularity of the cellular phone, dramatic changes have taken place in the way of communication of people's daily routine. SMS or MMS-based (simple message service or media message service) communication is the primary way of social interaction in mobile social networks.[1,2]

Mobile social network is composed of social network[3] and mobile communication network[4,5] (see Figure 1), where individuals with similar interests chat with each other through their devices. If you think of users as nodes and connections between users through short messages as edges, then users and user connections can form one mobile social network. According to graph theory, we can scientifically understand the interaction between users and accurately identify the important people in the social network. For mobile operators, they can identify important people in the real society, conduct marketing through the influence of important people, and maximize the promotion of commercial products.

The problem of user activity in a mobile social network can be abstracted into a problem on a social graph consisting purely of nodes and edges. The number and state of nodes and edges in social relationships are always in a dynamic change. The development of mobile social networks is a cascade reaction in which the growth of networks is accompanied by the addition of individual activity, which can be understood as a chain reaction. Real mobile social networks have a number of characteristics, each of which may affect individual activity.

How do we quantify activity degree and what is the relationship between activity and influence? These issues are undoubtedly interesting and meaningful for the theoretical study of complex networks and the practical operation of social groups.

The interaction of people in social life is called social influence.[6] Users are not only receivers of information, but also producers and disseminators of information. The influence factors include the individuals' friends number, the interaction frequency of users, the spread

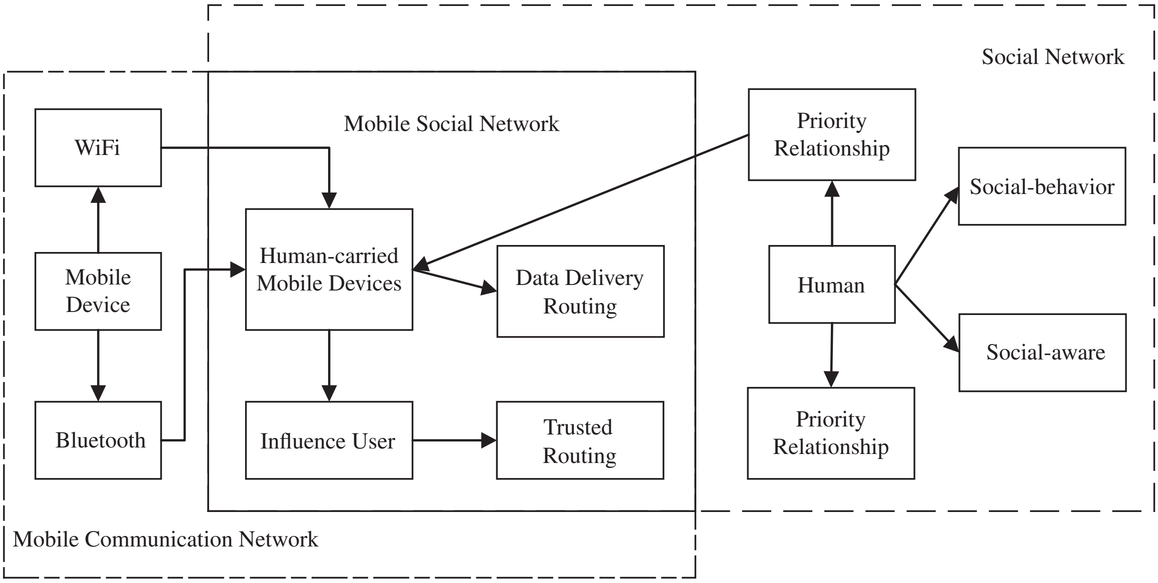


FIGURE 1 An illustration of the mobile social network

manner and the characteristics of networks, etc. Recently, the assessment of social influence has become the focal point, attracting more and more related researchers and interested groups to proceed further research. The critical and challenging issue is the problem of influence maximization (IM), whose top priority is to mine the initial *k* most influential nodes to disseminate influence as much as possible in a variety of ways and then spread to the entire network with appropriate propagation models. Social network analysis is inseparable from social influence research. In the past 10 years, with the rapid development of social network, researchers began to have the opportunity to model and analyze social based on a large number of real-world data, and achieved fruitful research results and broad applications. Domingos and Richardson[7] for the first time reflected on IM problem in terms of an algorithm problem. They used Markov random fields to model the problem and proposed heuristic solutions. Then, Kempe et al[8] expressed the problem as a discrete optimization problem and proved that the optimization problem was NP-hard. It is also proved that under the general framework of a base submodular function and a monotone function, an approximation algorithm can be established to maximize the influence of the optimization model. However, in the original greedy algorithm, the method of obtaining the diffusion range is inefficient and time-consuming. To solve this problem, Jure Leskovec[9] proposed the idea of lazy-forward to optimize the speed of the algorithm. Masahiro Kimura[10] proposed the SPM and SP1M based on the IC model, which calculated the diffusion range. Meanwhile, there are also some models for solving the scalability problem of the greedy algorithm and a series of diffusion models, problem of the greedy and heuristic algorithms have been studied.[11-16] In addition, there are ways to sort the node effects based on the structural attributes of a single node. For example, betweenness and distance centrality are simple measures. But they are affected by the intersection of the range of influence of the selected node. Therefore, many heuristic methods[17-20] were proposed to improve the results. Duan et al[21] considered the clustering coefficient and proposed the clustering rank algorithm. Suman Kundu et al[22] applied the degree centrality (DC) to the probability model (IC model[23]) to measure the top-*k* nodes' propagation ability. Ibrahima Gaye et al[24] proposed a new centrality of transmission degree, and designed a BRST algorithm to eliminate the influence of the ring formed in the process of transmission. Trusov et al[25] analyzed the relationship between user activity and user influence in social networks and found that if the number of users' friends increased rapidly when the user is active, then users' influence would be greater. Fei et al[26] identified important nodes by applying relative entropy to TOPSIS method. Under the premise of without considering time, Qiu et al[27] modified the distribution delays in LAIC model by taking power-law distribution into account. Aslay et al[28] extensively surveyed the research on social influence propagation and maximization, with a focus on the recent algorithmic and theoretical advances. Lü et al[29] systematically summarized the key node ranking methods, and compared the advantages and disadvantages of the main algorithms with the real network data. Then researchers have found that the modular nature of networks. Modularity is a way to identify communities. Ghalmane Li et al[30] implemented extensions of some common centrality indicators in modular networks. Taghavian et al[31] studied the influence of inter communities nodes on the spread of epidemic. Naveen et al[32] made full use of community structure to find influential nodes with a centrality indicator. Zhong et al[33] took differences and similarities of the structure into account and proposed ECDS centrality to measure the influence of nodes. Zhong et al[34] presented IIRA to accurately identify influential nodes and they also proposed comprehensively CI[35] to find the influential nodes.

However, this kind of model regards the behavior between a pair of users as irrelevant to the behavior of other users, without considering the complex relationship between users and their multiple connected users, and it does not make full use of the topological structure of social network, so there are inevitably some shortcomings of the current research. Therefore, it is one of the hot research trends of current researchers to comprehensively use the behavioral characteristics of users and the topological structure of social network to improve the accuracy and efficiency of the model for measuring users' influence.

For the purpose of effectively and efficiently spotting the top-k influential ones, we give an overall consideration of nodes' topological structure and behavior characteristics and it outperforms other approaches on influence propagation. More concretely, our contributions are as follows.

1. We map the mobile social networks into a mobile social graph, which vividly presents the interaction relationships of users and the propagation of information. We propose a new notion to evaluate social influence, which is measured by two indices based on local attributes and behavioral characteristics. The most common measure index based on local attributes is DC and the degree of a node refers as the number of its neighbor nodes. What's more, behavior-based user activity reflects users' influence to some extent.
2. By utilizing the HeapSort algorithm, we achieve the sort of the activity-calculated results to find out the most influential nodes. The experimental results show that using Susceptible-Infected-Susceptible (SIS) model to simulate influence propagation achieves better results.

2 | PRELIMINARIES

Usually, $G = (V, E, W)$ is an representation of a network, where V, E, and W are the set of nodes, edges, and weights on edges, respectively. The following is each centrality measure's definition.

Definition 1. (26) As for node i, its DC d_i is defined as:

$$d_i = \sum_{j}^{N} e_{ij}, \tag{1}$$

where e_{ij} represents the edge between node i and node j. The value of e_{ij} is 1 if there exists edge $e(i,j)$, and 0 under other circumstances.

Definition 2. (26) As for node i, its betweenness centrality (BC) b_i is defined as:

$$b_i = \sum_{j,k \neq i} \frac{g_{jk}(i)}{g_{jk}}, \tag{2}$$

where g_{jk} represents all possible shortest paths number between node j and node k, and $g_{jk}(i)$ refers to the paths number containing node i.

Definition 3. (26) As for node i, its closeness centrality c_i is defined as:

$$c_i = \frac{1}{\sum_j^N d_{ij}}, \tag{3}$$

where d_{ij} represents the distance between node i and node j.

With minor corresponding adjustments, the above measures can also be applied to weighted networks as follows.

Definition 4. (26) In a weighted network, node i's DC d_i^w is defined as:

$$d_i^w = \sum_{j}^{N} w_{ij}, \tag{4}$$

where w_{ij} denotes the edge $e(i,j)$'s weight, and it is a positive number if there exists edge $e(i,j)$.

Definition 5. (26) In a weighted network, node i's BC b_i^w is defined as:

$$b_i^w = \sum_{j,k \neq i} \frac{g_{jk}^w(i)}{g_{jk}^w}, \tag{5}$$

where g_{jk}^w is the number of all possible shortest paths between nodes j and node k, and $g_{jk}^w(i)$ refers to the paths number containing node i.

Definition 6. (26) In a weighted network, node i's closeness centrality c_i^w is defined as:

$$c_i^w = \left[\sum_{j}^{N} d^w(i,j) \right]^{-1} \tag{6}$$

where $d^w(i,j)$ refers to the weight on the corresponding edge, for instance, it can be said shortest distance or connection strength.

TABLE 1 Some related statements

Network	Description
DC(*i*)	the degree centrality of node *i*
BC(*i*)	the betweenness centrality of node *i*
AC(*i*)	the activity centrality of node *i*
TA(*i*)	the value of topology-based activity of node *i*
BA(*i*)	the value of behavior-based activity of node *i*
$A(i)$	the value of total activity of node *i*
deg(*i*), bedeg(*i*)	the degree of node *i* and the betweenness of node *i*, respectively
$\sigma(i,j)$	a sign of whether there is an edge $e(i,j)$
$d(i,j)$	the geodesic distance between node *i* and node *j*
$\omega(i)$	the number of communications with the user *i*
A	set of the total activity of each node
T	information spreading tree
K	set of the sorting result of the top-*k* nodes
t	a certain time at the simulation process
$S(t)$	the state of node is susceptible at time *t*
$I(t)$	the state of node is infected at time *t*
$<k>$	the average degree of network
α^t	the epidemic threshold
τ	the Kendall correlation coefficient
n_t, n_f	the number of concordant pairs and discordant pairs, respectively

3 | TOTAL ACTIVITY DEGREE

3.1 | Mobile social networks

Mobile social networks have the basic structural characteristics of complex networks. A complex graph can be used to describe a mobile social network which takes the people who constitute its basic elements as nodes and the relationships between people as connections or edges between nodes. Common models fall into two categories. One is network construction model, which constructs a network by explicitly setting the process of adding nodes and forming edges in the network. The advantage of this model is that the process is intuitive and can vividly simulate human social behaviors. Another kind of model is randomly generated by statistically modeling method, where the complex network structure generation process is simplified to a number of steps. The basic probability is obtained by statistical inference and model parameters in order to restore the generation process. This kind of model can explain the formation mechanism of the network structure, but it is not intuitive for the concrete structure model.

We use a graph $G(V, E, W)$, which is directed or weighted, to represent a mobile social network, where vertices set V is equivalently to users who use mobile devices, edges set E is equivalently to the interconnection or ties between nodes or users, and weight set W amounts to the interconnection totality between *i* and *j*. And Table 1 lists the used notations.

3.2 | The topology attribution of node activity degree

For characterization of nodes in topology from the macro level of activity, structural indicators in complex network topology are relatively mature, so using topology attributions to assess the activity degree of nodes becomes practical. By virtue of topological properties, the most common metric for the centrality, the out degree of a node is defined as the number of its neighbor nodes, which reflects the direct activity of the current node in the network as a whole. Betweenness centrality is defined as the shortest path between two nodes on the network through a number of the current nodes. It's a description of information in social networks spreading through the nodes of the frequency. The higher the value of the index, the busier the node in the network topology. If a section of the number of nodes with higher BC are removed, then it may cause network congestion, and it is not conductive for the spread of information.

As for a specific user, its location information is closely related to its activity. The direct influence not only depends on the number of his or her friends, namely the out degree of this node, but also has something to do with the betweenness of the node. If a node has a higher out degree, which means it has more neighbors, then it will spread more information to the network compared with other nodes in the mobile social networks, and therefore, its activity is relatively higher than the ones who have less neighbors. The more times a node acts as a “mediator”, the higher BC it has. Like the social talent around us, they have higher BC than others in a social network. We come up with the following formulas to represent a node's activity degree based on the topology attribution:

$$\mathrm{TA}(i) = \alpha \frac{\deg(i)}{\sum_{j \in V} \deg(j)} + (1-\alpha) bedeg(i), \tag{7}$$

$$\deg(i) = \sum_{j \in N_i} j, \tag{8}$$

$$bedeg(i) = \frac{1}{\sum_{i \in V, i \neq j} d(i,j)}, \tag{9}$$

where in the above formulas, $\deg(i)$ corresponds to the out degree of node i, bedeg(i) corresponds to the betweenness of node i in a social network influence graph $G(V, E, W)$, and parameter α ($\alpha \in [0, 1]$) denotes the weight adjustment parameter. What's more, N_i denotes the neighbors of node i and node j belongs to N_i. Node j represents any node pertaining to a mobile social network, and $d(i, j)$ is the geodesic distance between them.

3.3 | The behavior characteristic of node activity degree

Users in mobile social networks publish content information, and then disseminate contents through interactive behaviors. By analyzing these behaviors, we can not only measure the tie strength among users, but also can predict user information and behavior in mobile social networks. Here we regard the number of communications between users as the possibility of communication between users, that is, the weight of the edges between users.

All social communication behaviors (forwarding moments of friends, word of mouth) are that users express themselves in their own way. The purpose of all social interactions and social communications is not to know others, but to express the appeal of “who I am” to others. So if the number of communications between two users are high, we can assume that they have a close relationship. If this relationship is mapped to a graph, we will assume that there is a relatively high possibility to establish edges between nodes i and node j, characterized by the corresponding edge weights. So a node's activity degree based on the behavior can be acquired:

$$\mathrm{BA}(i) = \frac{\sum w(i)}{\sum_{j \in V} w(j)}, \tag{10}$$

where $w(i)$ or $w(j)$ refers to the number of communications between user i or j and its neighbors, and node j represents any node pertaining to a mobile social network.

3.4 | Total activity degree

Given a node i playing roles in an entire mobile social network, AD(i) refers to the total activity degree of i. Without loss of generality, we utilize $G(V, E, W)$ to represent a mobile social network influence graph, where nodes set V refers to the users of the mobile phones in a mobile social network, set E of directed edges or undirected edges indicates the information interaction between two devices, weight set W denotes the SMS or MMS messages totality sent from the two devices.

Based on the above content, we combine the topological information of nodes with the behavioral characteristics of nodes, and then comprehensively consider the activity of nodes according to the following formula and parameter β ($\beta \in [0, 1]$) denotes the weight adjustment parameter. We propose Algorithm 1 to calculate the total activity degree for all nodes.

$$A(i) = \beta \mathrm{BA}(i) + (1-\beta) \mathrm{TA}(i). \tag{11}$$

Algorithm 1. Total activity degree computing

Input: A network $G(V, E, W)$ and the number of initial nodes k.
Output: Total activity of nodes.

for i=1 to N **do**
  Initialize a tree T_i with only a leaf (the root);
  Compute node activity based topology for each node by Equation (1);
  Compute node activity based behavior for each node by Equation (5);
  Compute total node activity for each node by Equation (6);
  Calls the function of heap sorting algorithm;
end for
return K=the sorted former k nodes.

Algorithm 2. Information diffusion tree construction algorithm

Input: A network $G(V, E, W)$.
Output: Information spreading trees T.

Network pretreatment. Acquire the total activity degree of each node and obtain node subset K containing the top-k nodes according to Algorithm 1;
for each node i in K **do**
  Select the maximum weight of edge e which connected node i and construct the spreading edge e';
Add edge e' into E' if there does not exist cycles in tree T;
end for
return T

Algorithm 3. *SIS* algorithm

Input: A network $G(V, E, W)$, set K including the former k nodes, infected probablity α, susceptible probablity β, and the average degree $< k >$.
Output: Total number of infected node $I(t)$ at time t, total number of susceptible nodes $S(t)$ at time t.

Calculate total node activity degree by Equation (6);
Apply Algorithm 1 to the former k nodes, set $I(t) = k$;
Initialize the top-k nodes' states are I, and other nodes' states are S, $S(t) + I(t) = n$;
Statistical messages record the information of its or theirs friends information;
while $i \leq N$ **do**
  if j is i's neighbor and j's state is S **then**
    j becomes I with probability α, $I(t)$ plus 1;
  else
    j keeps original state;
  end if
  i goes to S with probablity β, $S(t)$ plus 1;
end while
if $\alpha \leq 1/ < k >$ **then**
  Steady state infection density will be 0, the entire network state finally will be S.
else
  Steady state infection density will be 1, the entire network state finally will be I.
end if
$t = t + \delta t$;
return $I(t)$, $S(t)$;

3.5 | Information diffusion tree construction algorithm

According to the above equations, we design Algorithm 2 which constructs information spreading tree for all nodes. The process of building the information spreading tree reflects the formation process of the propagation path.

3.6 | SIS model

There is an assumption that the infectious source of infectious diseases[36] infected individuals denoted as I (Infected), and other nodes are S (Susceptible) by default in SIS model. In the process of propagation, infected nodes and susceptible nodes can be converted to each other with a certain probability, which on the other hand, once susceptible nodes become infected, then they have new infection ability. The following is the infection mechanism:

$$\begin{cases} S(i) + I(j) \xrightarrow{\alpha} I(i) + I(j) \\ I(k) \rightarrow S(i) \end{cases}. \tag{12}$$

$$\begin{cases} \frac{dS(t)}{dt} = -\alpha I(t)S(t) + \beta I(t) \\ \frac{dI(t)}{dt} = \alpha I(t)S(t) - \beta I(t) \end{cases}. \tag{13}$$

Suppose that at time t the infected individuals and susceptible ones are $S(t)$ and $I(t)$ in a network, respectively. The model shows that individuals can be infected multiple times. The probabilities of transition between the two states are α and β. Then the differential equation shown in equation (13) embodies the dynamic equilibrium of the model, and Algorithm 3 depicts the SIS algorithm.

4 | PERFORMANCE EVALUATION

In this section, we make comparisons in terms of the top-10 nodes and present simulation results to evaluate the influence propagation of nodes. We use the total node activity that is involved in the topology feature and behavior information of nodes to determine the top-k nodes as initial active nodes. Figure 2 is the experiment diagram. Figures 3 to 8 depict, respectively, the influence spreading comparison results of our algorithm

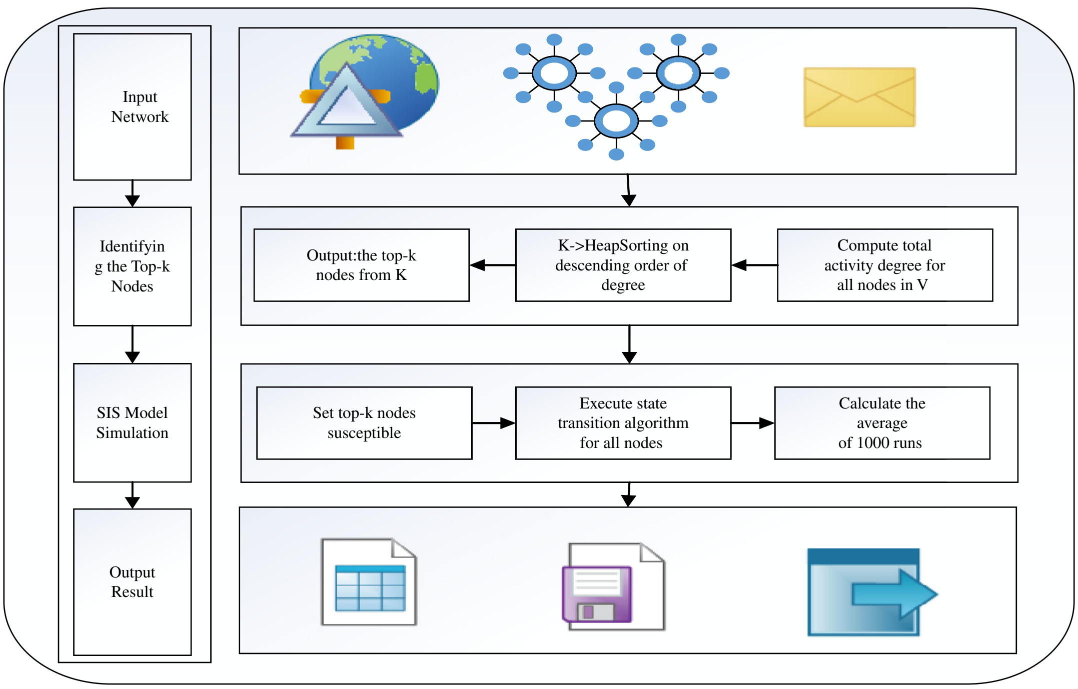
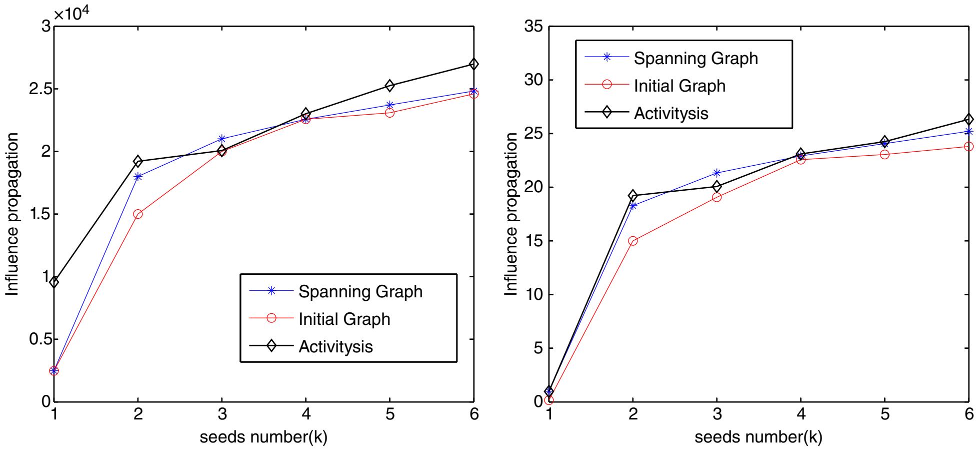
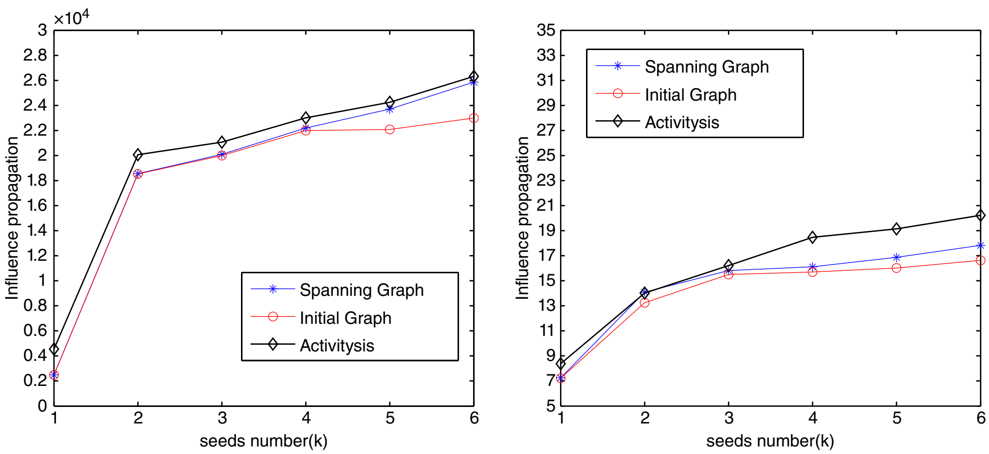


FIGURE 2 Diagram of experiment



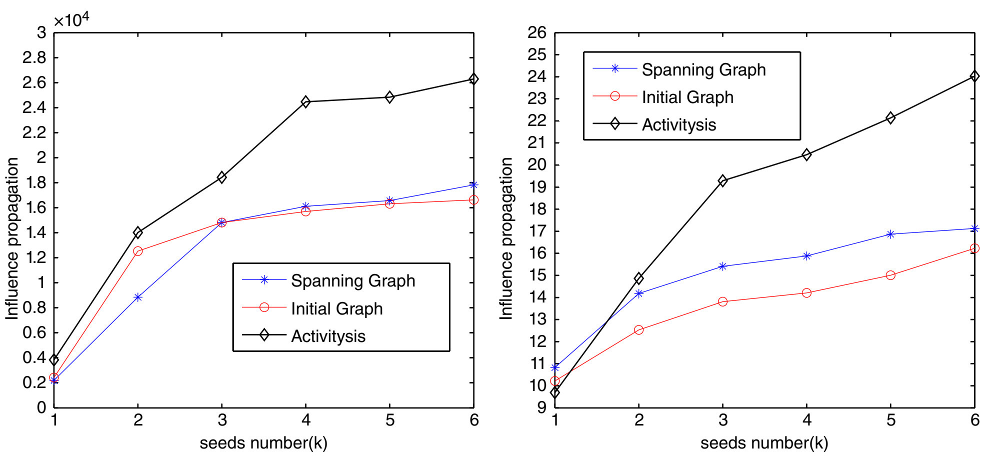
(A) EnronEmail (B) KarateClub

FIGURE 3 The influence of seed selection on information transmission



(A) EnronEmail (B) KarateClub

FIGURE 4 The influence of seed selection on information transmission

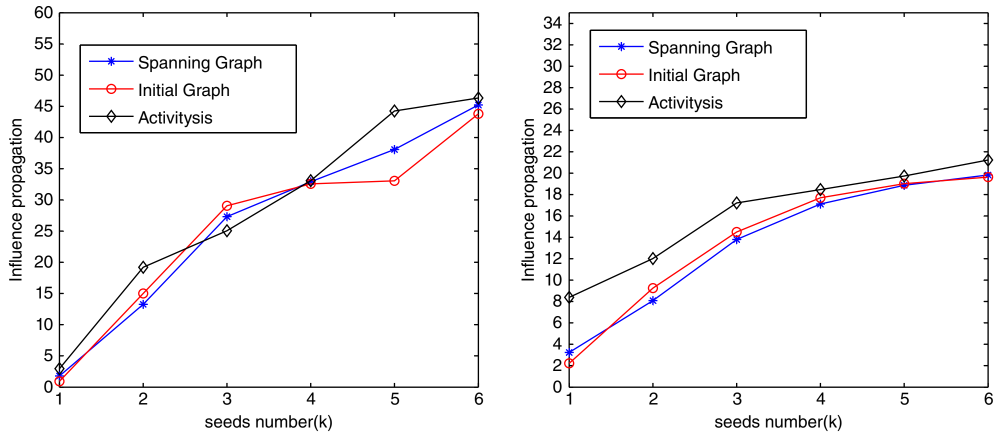


(A) EnronEmail (B) KarateClub

FIGURE 5 The influence of seed selection on information transmission

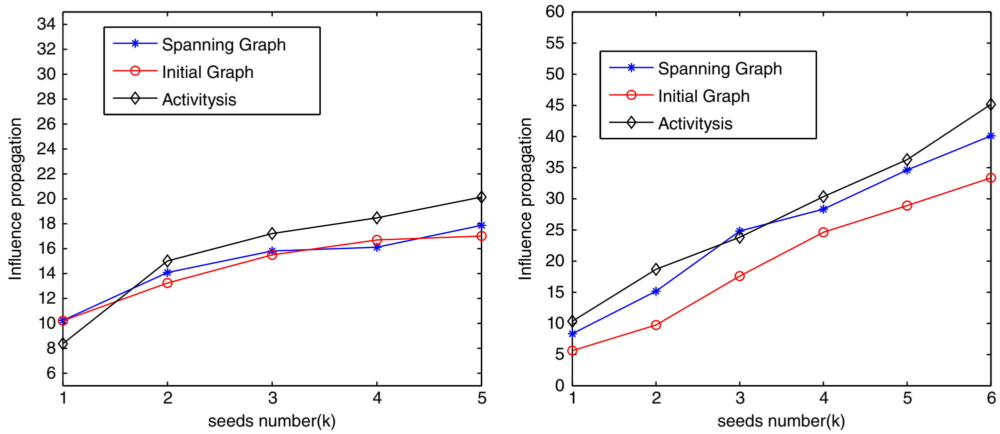
and spanning graph and initial graph, where in Figure 3, spanning graph and initial graph both select seeds according to DC, while Figures 4 and 5 find, respectively, the top-k seeds based on the degree discount and diffusion degree. The next three figures are similar to them. And (A) and (B) of each figure corresponds, respectively, to two different datasets. Figures 9 and 10 reveal the relationship of the influence propagation and two parameters in node state transition used in our algorithm. The relationships between spreading rate α and Kendall's τ in four networks have shown in Figures 11 and 12.

FIGURE 6 The influence of seed selection on information transmission



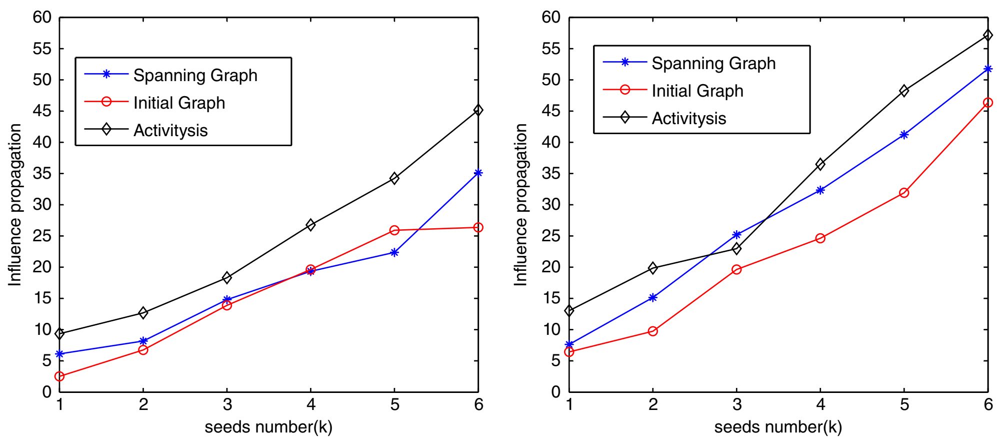
(A) Dolphins (B) LesMiserables

FIGURE 7 The influence of seed selection on information transmission



(A) Dolphins (B) LesMiserables

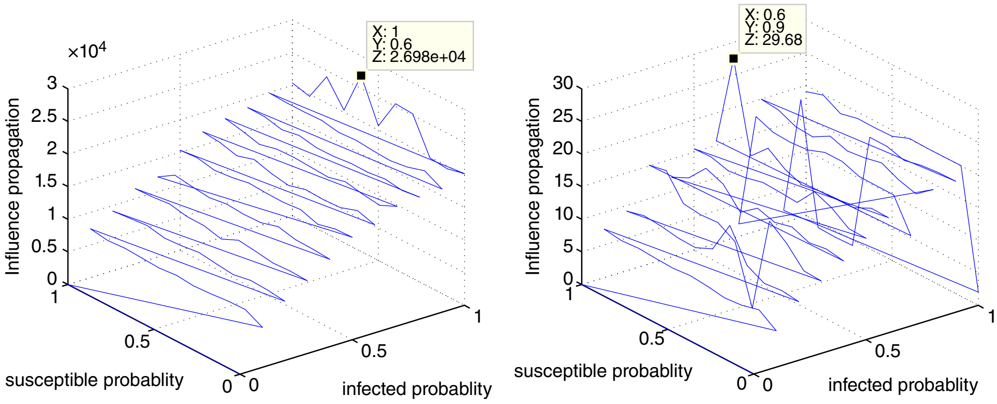
FIGURE 8 The influence of seed selection on information transmission



(A) Dolphins (B) LesMiserables

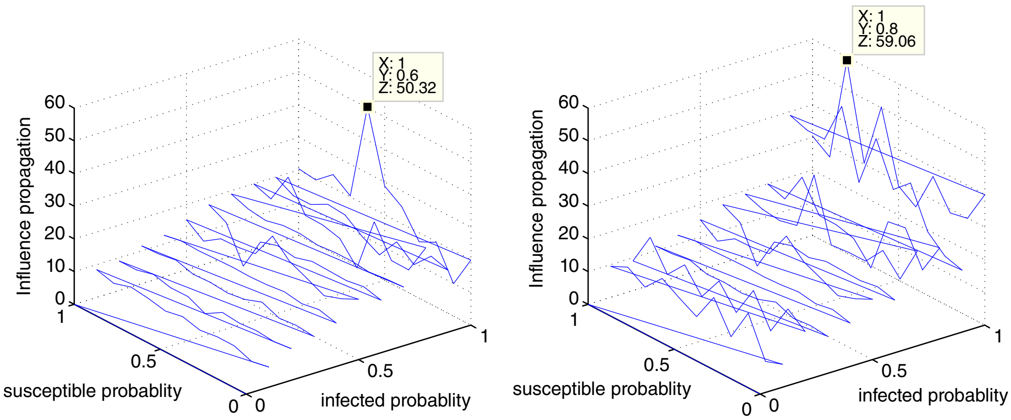
4.1 | Datasets

Four real-world networks, Zachary karate,[37] Enron email,[38] Dolphins,[39] and LesMiserables[40] are used in our experiment. The first one is a club network and there is no direction in its members' interconnection. The second one is a network of emails from senior employees of the Enron company. The Dolphins network is a dolphin community, and the last one derives from a well-known novel—LesMiserables. In Table 2, we list out some dataset information used for our experiment. DC, BC, and AC are acted on the selected four networks to determine the top-10 nodes in this experiment, and the results are listed in Tables 3 to 6. Table 7 shows the *tau* between different ranking lists.



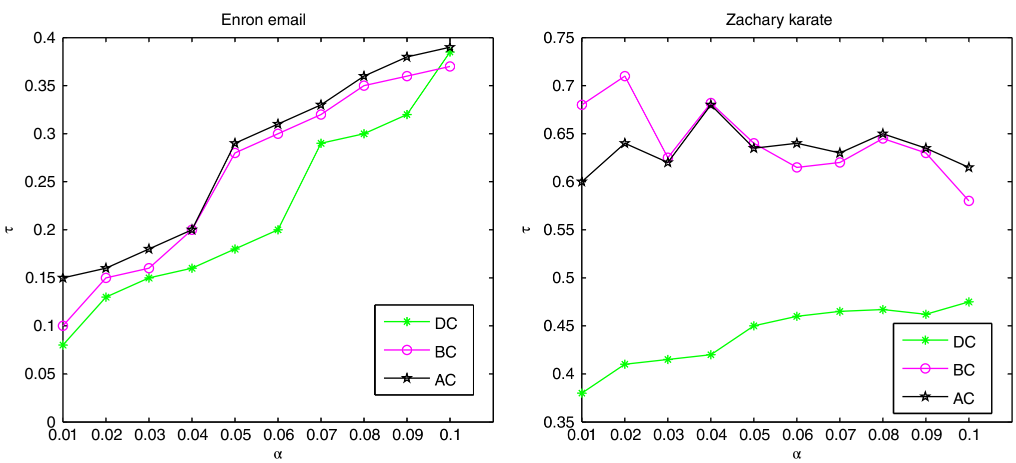
(A) EnronEmail (B) Karate Club

FIGURE 9 The impact of parameters α and β



(A) Dolphins (B) LesMiserables

FIGURE 10 The impact of parameters α and β



(A) Enronemail (B) Zacharykarate

FIGURE 11 The relationship between spreading rate α and Kendall's τ

4.2 | Performance comparison

Our comparsion algorithms use three centrality measures to determine the seed set: DC heuristic, Degree Discount heuristic, and Diffusion Degree heuristic. These algorithms are effective and we compared them with ours. In order to facilitate the experiment to reach a balance as soon as possible, the two parameters used in the SIS model are set as $\alpha = 0.5$ and $\beta = 1$, respectively, where α is also called spreading rate and generally its value is not less than the threshold α^t, which has to satisfy the following formula:[41]

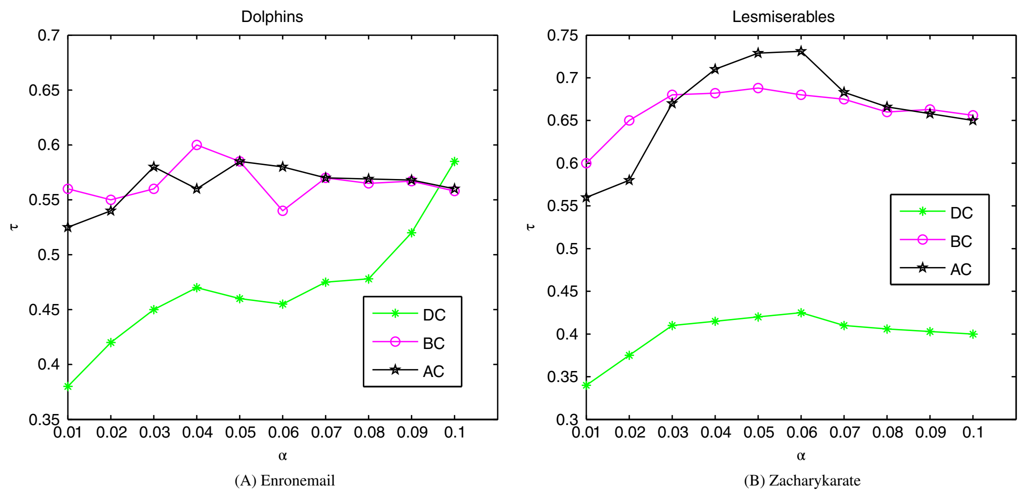


FIGURE 12 The relationship between spreading rate α and Kendall's τ

TABLE 2 The network information

Notation	Nodes	Edges	Descriptions	$< k >$
Karate Club	34	78	Connected	4.588
Enron email	36 692	367 662	Disconnected	8.502
Dolphins	62	159	Connected	5.129
LesMiserables	77	154	Connected	3.299

TABLE 3 The top-10 in Karate Club

Zachary karate			
Top-10	DC	BC	AC
1	34	34	34
2	1	1	1
3	33	32	32
4	3	33	33
5	2	24	3
6	4	31	2
7	32	23	24
8	7	8	31
9	9	5	23
10	14	9	4

Abbreviations: AC, activity centrality; BC, betweenness centrality; DC, degree centrality.

$$\alpha^t = \frac{\langle k \rangle}{\langle k^2 \rangle}, \tag{14}$$

where $< k >$ and $< k^2 >$ are related with the degree of the first two steps. In order to compare spreading rate αs impact on performance, we have set $\alpha \in [0.01, 0.1]$ in this part.

The results of the simulation showing the importance to take the behavior of node into account are showed by the below figures. Figure 3 shows that Spanning Graph and Initial Graph used the DC to select the initial nodes and we use total activity to select seeds. In Figure 3A, the dataset is the Enron email and the black line represents our algorithm. In Figure 3B, we choose Zachary karate as the dataset. As can be seen, our algorithm

TABLE 4 The top-10 in Enron email

Enron			
Top-10	DC	BC	AC
1	1	1	1
2	5	5	5
3	9	9	9
4	13	13	13
5	7	7	15
6	6	56	7
7	4	15	6
8	15	6	4
9	56	16	16
10	74	19	74

Abbreviations: AC, activity centrality; BC, betweenness centrality; DC, degree centrality.

TABLE 5 The top-10 in Dolphins

Dolphins			
Top-10	DC	BC	AC
1	15	37	15
2	38	2	38
3	46	41	46
4	34	38	52
5	52	8	34
6	18	18	18
7	21	21	21
8	30	55	58
9	58	52	30
10	2	58	2

Abbreviations: AC, activity centrality; BC, betweenness centrality; DC, degree centrality.

has a similar effect to the blue line, which represents the Spanning Graph. We assume the number of seeds k is from 5 to 30. In the beginning, when there are five seed nodes, Spanning Graph has similar information spreading with Initial Graph, and our method outperforms them at first. When k equals 10, the information propagation increases rapidly. As the value of k increases, the information propagation rises slowly. Our algorithm has good spreading performance because we take the activity of node into consideration.

Figure 4A,B shows that we used the Degree Discount Centrality to select the initial nodes. In Figure 4A, the dataset is the Enron email and the black line represents our algorithm, which influential ability is greater than the other two. We assume the number of seeds k is from 1 to 6. In Figure 4B, we do the experiment on Karate Club. As can be seen, when seeds number exceeds 10, the influence propagation is obviously accelerating, which means it has a clear advantage in the case of a large number of seeds. We set the number of seeds k to be the same with Figure 3B. In terms of overall trends, Figure 4A resembles Figure 3A and Figure 4B is similar to Figure 3B.

Figure 5 shows that we used the Diffusion DC to select the initial nodes. In Figure 5A, the dataset is the Enron email and the black line represents our algorithm, which influential ability is greater than the other two. In Figure 5B, we choose Karate Club as the dataset. We set the number of seeds k from 5 to 30. Obviously, Spanning Graph has a larger range of information spreading than Initial Graph, and our method outperforms Spanning Graph although when k is small, it has smaller propagation. As k increases, the information propagation of Spanning Graph and our method grows substantially, while Initial Graph rises slowly. As you can see, our algorithm has some advantage in information propagation. We assume the number

TABLE 6 The top-10 in LesMiserables

LesMiserables			
Top-10	DC	BC	AC
1	10	10	10
2	48	48	48
3	55	2	55
4	25	55	27
5	27	23	25
6	23	25	23
7	58	27	58
8	64	51	64
9	62	58	62
10	63	17	63

Abbreviations: AC, activity centrality; BC, betweenness centrality; DC, degree centrality.

TABLE 7 The Kendall correlation coefficient τ between various ranking lists

Networks	DC	BC	AC
Zachary karate	0.61	0.58	0.64
Enron email	0.36	0.37	0.45
Dolphins	0.54	0.51	0.63
LesMiserables	0.57	0.54	0.65

Abbreviations: AC, activity centrality; BC, betweenness centrality; DC, degree centrality.

of seeds k is from 1 to 6. When k varies from 1 to 3, Spanning Graph almost has same information spreading with Initial Graph, but our method outperforms them in the beginning. When k continues to increase, Spanning Graph shows its advantage over Initial Graph, and our algorithm continues to maintain its leading position.

With the increase of the number of seed nodes, the states of nodes of different influence are marked as I in turn, as a result, the range of influence propagation fluctuates relatively greatly in the Dolphins network as shown in Figure 6A. Although when the number of seeds k is 3, our method is slightly lower than Initial Graph and Spanning Graph, when seeds number k is greater than the other two. While in Figure 6B, as key figures often appear in the main chapters, so the overall trend is the same. Our method performs better than others at the beginning and it keeps higher influence propagation, and there is little difference in influence propagation with Initial Graph and Spanning Graph. Since we take local and global information into account, so our method performs better than the other two in terms of influence propagation.

It is observed that in Figure 7, the influence of the top-k nodes found by BC is lower than that found by DC. When seeds number k is greater than 1.5, influence spreading of our method is significantly greater than Initial Graph and Spaninng Graph in Figure 7A, and Initial Graph and Spanning Graph have little difference. No matter in Dolphins or LesMiserables, almost all of the trends are increasing, and the proposed method has a greater propagation range, which is particularly true in Figure 7B. Particularly, when the number of seeds k is 6, the difference in influence propagation of the proposed method over Initial Graph is about 10. Compared with the previous two Figures, Figure 8 shows the top-k nodes identified by AC can impact more nodes. The proposed method keeps the advantage in terms of influence propagation in Figure 8A. In Figure 8B, when seeds number is 3, influence propagation with our method is higher than Initial Graph, while its close to Spanning Graph. In addition, our method performs better than the other two methods.

Figure 9A shows the trend of the influence spreading of the Enron email dataset as the two parameters (α and β) of the SIS model are changing. As we can see, when infected probability α is 1 and susceptible probability β is 0.6, influence propagation achieves maximization 26 980, which means that in email communication systems, we are more susceptible to each other because we are in frequent contact. Figure 9B depicts the influence propagation of the Zachary karate dataset as the two parameters of the SIS model are changing. In these two figures, we can see that the parameters corresponding to the extreme values obtained by the two datasets are not completely consistent, which shows the difference of the proposed algorithm for different datasets.

Figure 10A shows the trend of the influence spreading of the Dolphins dataset as the two parameters (α and β) of the SIS model are changing. Obviously, when infected probability α is 1 and susceptible probability β is 0.6, influence propagation achieves maximization 50.32, which means that with a mass of long-time associations, Dolphin community gradually formed. Figure 10B depicts the influence propagation of the LesMiserables dataset as the two parameters of the SIS model are changing. It can be seen that when α is different, the range of influence propagation fluctuates greatly, which is because in each new chapter, the importance of supporting roles varies. In a word, Figure 10 shows the difference of the proposed algorithm for the two datasets.

The Kendall correlation coefficient τ[42] in statistics is used to measure the correlation between two random variables. We introduce Kendall's tau to compare and measure the rank correlation between the various ranking, one of which is generated by SIS model, and the other is the formation of centrality measures. We can calculate τ by the following formula:

$$\tau(U, V) = \frac{2\left(n_t - n_f\right)}{n(n-1)}, \tag{15}$$

based on aforementioned two ranking lists, n refers to the number of total nodes in a network, and n_t and n_f are the concordant pairs and discordant ones, respectively. The value of τ varies from −1 to 1, and when τ is 0, the two random variables are independent of each other. The higher the value of is, the better the performance is.

Figure 11 depicts the relationship between spreading rate α and Kendall's τ in Enron email and Zachary karate. A black line connected by pentacle represents the ranking list with AC, and the green one and the magenta one are DC and BC, respectively. When spreading rate is 0.04, the proposed AC equals to BC, and its better than DC in Enron email. Figure 11B shows in Zachary karate, DC performs worse, while DC and BC performs better. When α is 0.02, the value of τ corresponding to BC gets its maximization 0.71, and AC is inferior to BC. The presented method performs better when $\alpha \geqslant 0.05$. Figure 12A depicts the relationship between α and τ in Dolphins network, as we can see, the value of τ corresponding to proposed AC fluctuates between 0.5 and 0.6, and BC is slightly better than it. In LesMiserables, when spreading rate is 0.06, relative τ reaches 0.74, which is better than DC and BC. It is obvious that the ranking list based on BC and the list generated by the SIS are quite different in Figure 12B.

5 | CONCLUSION AND FUTURE WORK

In this article, we measured the activity of a node by combining its topology properties with its behavior characteristics, which not only makes full use of the topological structure of mobile social networks, but also takes the complex relationship between users and their multiple connected users into consideration. In addition, we selected the SIS model to achieve the propagation dynamics of information. Experimental results with four real world datasets achieved the purpose of the IM and demonstrate that the information can diffuse easily based the top-k active nodes in mobile social networks. Some future work can be done in quantification of node activity and exploring its significance is also an issue of thought.

ACKNOWLEDGMENTS

The authors would like to thank the National Science Foundation of China (No. U1905211, No. 61771140, No. 61702100, No. 61702103), China Postdoctoral Science Foundation (No. 2017M612107), Education Department of Fujian Province (JAT170397).

ORCID

Li Xu https://orcid.org/0000-0002-8972-3373

REFERENCES

1. Yu S, Liu M, Dou W, Liu X, Zhou S. Networking for big data: a survey. *IEEE Commun Surv Tutor*. 2016;19(1):531-549.
2. Lin L-M, Xu L, Huang Y-Z, et al. On exploiting priority relation graph for reliable multi-path communication in mobile social networks. *Inf Sci*. 2019;477:490-450.
3. Newman MEJ, Park J. Why social networks are different from other types of networks. *Phys Rev E*. 2003;68(3):036122.
4. Ioannidis S, Chaintreau A. Laurent Massoulié. optimal and scalable distribution of content updates over a mobile social network. *Proc IEEE Infocom*. 2009;1422-1430.
5. Kayastha N, Niyato D, Wang P, et al. Applications, architectures, and protocol design issues for mobile social networks: a survey. *Proc IEEE*. 2011;99(12):2130-2158.
6. Peng S-C, Yang A-M, Cao LH, Yu S, Xie D. Social influence modeling using information theory in mobile social networks. *Inf Sci*. 2017;379:146-159.
7. Domingos P, Richardson M. Mining the network value of customers. Paper presented at: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2001.
8. Kempe D, Kleinberg J, Tardos, É. Maximizing the spread of influence through a social network. Paper presented at: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2003:137; ACM.
9. Leskovec J, Krause A, Guestrin C, Faloutsos, C., VanBriesen, J., & Glance, N. Cost-effective outbreak detection in networks. Paper presented at: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2007:420.

10. Kimura M, Saito K. Tractable models for information diffusion in social networks. Paper presented at:Proceedings of the Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases; September 18-22, 2006; Berlin, Germany. Proceedings. 2006; Springer.
11. Wu H-H, Simge K. A two-stage stochastic programming approach for influence maximization in social networks. *Comput Optim Appl*. 2018;69(1):1-33.
12. Wang Y, Cong G, Song G, Xie, K. Community-based Greedy algorithm for mining top-K influential nodes in mobile social networks. Paper presented at: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 25-28, 2010; Washington, DC, ACM.
13. Goyal A, Lu W, Lakshmanan L V S. SIMPATH: an efficient algorithm for influence maximization under the linear threshold model. Paper presented at: Proceedings of the 11th IEEE International Conference on Data Mining, ICDM 2011; December 11-14, 2011; Vancouver, BC, Canada, IEEE.
14. Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. Paper presented at: Acm Sigkdd International Conference on Knowledge Discovery & Data Mining 2009:199.
15. Jiang Q, Song G, Cong G, et al. Simulated annealing based influence maximization in social networks. Paper presented at: Proceedings of the 25th AAAI Conference on Artificial Intelligence, AAAI 2011; August 7-11, 2011; San Francisco, CA, AAAI Press.
16. Jung K, Heo W, Chen W. IRIE. Scalable and robust influence maximization in social network. *IEEE INFOCOM*. 2011.
17. Lü L, Zhang YC, Yeung CH, Zhou T. Leaders in social networks, the delicious case. *PloS One*. 2011;6:e21202.
18. Hou B, Yao Y, Liao D. Identifying all-around nodes for spreading dynamics in complex networks. *Phys A Stat Mech Appl*. 2012;391(15):0-0.
19. Zeng A, Zhang C-J. Ranking spreaders by decomposing complex networks. *Phys Lett A*. 2013;377(14):1031-1035.
20. Yang F, Li X, Xu Y, et al. Ranking the spreading influence of nodes in complex networks: an extended weighted degree centrality based on a remaining minimum degree decomposition. *Phys Lett A*. 2018;382:S0375960118305632.
21. Chen DB, Gao H, Lü L, Zhou T. Identifying influential nodes in large-scale directed networks: the role of clustering. *PloS One*. 2013;8(10):e77455.
22. Kundu S K, Murthy C A, Pal S K. A new centrality measure for influence maximization in social networks. Paper presented at: Proceedings of the Pattern Recognition and Machine Intelligence - 4th International Conference, PReMI 2011; June 27-July 1, 2011; Moscow, Russia. Proceedings. DBLP, 2011.
23. Zhu T, Wang B, Wu B, Zhu C. Maximizing the spread of influence ranking in social networks. *Inf Sci*. 2014;278:535-544.
24. Gaye I, Mendy G, Ouya S, Seck D. Spanning graph for maximizing the influence spread in social networks; 2015: 1389-1394.
25. Reza KAM, Bakhshi M. Determining influential users in social networks (the case of: word of mouth on digikala company instagram page). *J Inf Technol Manag*. 2017;9:587-612. https://doi.org/10.22059/jitm.61727.
26. Fei L, Deng Y. A new method to identify influential nodes based on relative entropy. *Chaos Solitons Fract*. 2017;104:257-267.
27. Hu J, Meng K, Chen X, Lin C, Huang J. Analysis of influence maximization in large-scale social networks. *ACM SIGMETRICS Perform Eval Rev*. 2014;41(4):78-81.
28. Aslay C, Lakshmanan L, Lu W, Xiao X-K. Influence maximization in online social networks. Paper presented at: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining; 2018:775-776. https://doi.org/10.1145/3159652.3162007.
29. Lü L, Chen D, Ren X-L, Zhang Y-C, Zhou T. Vital nodes identification in complex networks. *Phys Rep*. 2016;650:1-63.
30. Ghalmane Z, El Hassouni M, Cherifi C, Cherifi H. Centrality in modular networks. *EPJ Data Sci*. 2019;8:15.
31. Taghavian F, Salehi M, Teimour M. A local immunization strategy for networks with overlapping community structure. *Phys A Stat Mech Appl*. 2017;467:1148-1156.
32. Gupta N, Singh A, Cherifi H. Centrality measures for networks with community structure. *Phys A Stat Mech Appl*. 2016;452:46-59.
33. Zhong L-F, Shang M-S, Chen X-L, Cai S-M. Identifying the influential nodes via eigen-centrality from the differences and similarities of structure. *Phys A Stat Mech Appl*. 2018;510:77-82.
34. Zhong L-F, Liu J-G, Shang M-S. Iterative resource allocation based on propagation feature of node for identifying the influential nodes. *Phys Lett A*. 2015;379:2272-2276.
35. Zhong L-F, Liu Q-H, Wang W, Cai S-M. Comprehensive influence of local and global characteristics on identifying the influential nodes. *Phys A Stat Mech Appl*. 2018;511:78-84.
36. Peng S-C, Wang G-J, Zhou Y, Wan C, Wang C, Yu S. An immunization framework for social networks through big data based influence modeling. *IEEE Trans Depend Secure Comput*. 2017;16:984-995.
37. Zachary WW. An information flow model for conflict and fission in small groups1. *J Anthropol Res*. 1976;33(4):452-473.
38. Klimt B, Yang Y-M. The enron corpus: a new dataset for email classification research; 2004, (217):3201.
39. David L, Karsten S, Boisseau OJ, Haase P, Slooten E, Dawson SM. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav Ecol Sociobiol*. 2003;54:396-405.
40. Knuth DE. *The Stanford Graph Base: A Platform for Combinatorial Computing*. Reading, MA: Addison-Wesley; 1993.
41. Ma Q, Ma J. Identifying and ranking influential spreaders in complex networks with consideration of spreading probability. *Phys A Stat Mech Appl*. 2017;465:312-330.
42. Kendall MG. The treatment of ties in ranking problems. *Biometrika*. 1945;33:239-251.

How to cite this article: Gao M, Xu L, Lin L, Huang Y, Zhang X. Influence maximization based on activity degree in mobile social networks. *Concurrency Computat: Pract Exper*. 2020;e5677. https://doi.org/10.1002/cpe.5677