Expert Systems With Applications 125 (2019) 249–258

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Groups make nodes powerful: Identifying influential nodes in social networks based on social conformity theory and community features

Wei Zhang*, Jing Yang, Xiao-yu Ding, Xiao-mei Zou, Hong-yu Han, Qing-chao Zhao

College of Computer Science and Technology, Harbin Engineering University, Heilongjiang 150001, China

ARTICLE INFO

Article history:
Received 11 November 2018
Revised 2 February 2019
Accepted 3 February 2019
Available online 4 February 2019

MSC:
00-01
99-00

Keywords:
Attractive power
Initiating power
Community feature
Community tightness
Conformity
Node selection strategy

ABSTRACT

Identifying a group of influential nodes in social networks help us understand the hierarchical structure of the network and make a better control the spread of information. Moreover, it can offer guidance in avoiding the breakdown of the power system and the Internet, identifying drug targets and essential proteins. Undoubtedly, most of the influence measures suffer the low resolution. The same score corresponds to multiple nodes. What's worse is that the effect of overlapping between nodes is not fully considered. It causes resource waste in node selection. The purpose of this paper is to identify a set of distributed nodes with the strong propagation ability. Inspired by the interplay between the individuals and groups from sociological and complex networks, we propose a node ranking method based on the social conformity theory and community feature based on VoteRank. This proposed method calculates the node influence capability from two points of view, one is the individual, the other is the group. From the point of the individual, it quantifies the attractive power of the nodes with the feature of their neighbors based on the theory of conformity. It can distinguish the nodes with the same degree and similar structure. From the other point, it measures the initiating power with the scale of the community and the relative location of the node. Furthermore, a node selection strategy based on information coverage and community tightness is proposed to solve the problem of overlapping. Finally, node attractive power, initiating power and the node selection strategy are combined to improve VoteRank. The experimental results on real-world networks show the effectiveness of our methods. The results also explains that the enormous energy from the groups makes the node powerful.

1. Introduction

With the rapid development of Internet technology, social networks have become the major channel for commercial activities. In order to achieve higher economic benefit, the influence maximization problem in social networks has become one of the most concerned issues in recent years. Enhancing the control of influential nodes provides new opportunities in accelerating information propagation, which is of great significance to viral marketing (Dinh, Nguyen, & Thai, 2012) and ad delivery (Leskovec, Adamic, & Huberman, 2005). In addition, its theory can be directly used to control the outbreak of infectious diseases (Pastor-Satorras & Vespignani, 2002), prevent the power system paralysis (Resende & Pardalos, 2006) and the Internet failure (Albert, Albert, & Nakarado, 2004), find the essential protein (Csermely, Korcsmáros, Kiss, London, & Nussinov, 2013), dispatch the transportation (Yan, Zhou, Hu, Fu, & Wang, 2005) and dismantle the criminal networks (Agreste, Catanese, Meo, Ferrara, & Fiumara, 2016; Grassi, Calderoni, Bianchi, & Torriero, 2019; Ren, Gleinig, Helbing, & Antulov-Fantulin, 2018). It can be seen that the study of this problem has a great theoretical and practical meaning. The existing research methods are divided into two major types.

One is the method based on the greedy algorithm (Cheng, Shen, Huang, Zhang, & Cheng, 2013; Kempe, Kleinberg, & Tardos, 2003) or heuristic algorithm (Narayanam & Narahari, 2010). The greedy algorithm is an approximate solution algorithm for this problem with the precision of 60%. Apparently, it is inefficient and not available for the large-scale networks. To improve the efficiency and speed, researchers have proposed some methods based on the heuristic algorithm. Although these methods improve the efficiency, they are lack of theoretical support. These algorithms are not stable and cannot balance the accuracy and efficiency.

The other is the node ranking method. The method based on the degree is easy and efficient. However, degree is not conducive

* Corresponding author.
E-mail addresses: zhangwei_jsj@126.com (W. Zhang), yangjing@hrbeu.edu.cn (J. Yang), B615060001@hrbeu.edu.cn (X.-y. Ding), zouxiaomei@163.com (X.-m. Zou), hanhongyu@hrbeu.edu.cn (H.-y. Han), zhaoqc418@hrbeu.edu.cn (Q.-c. Zhao).

https://doi.org/10.1016/j.eswa.2019.02.007
0957-4174/

to differentiate nodes with the same degree because of the low resolution. Considering this fact, scholars have proposed ClusterRank (Chen, Gao, Lü, & Zhou, 2013), K-shell (Kitsak et al., 2010), IC (Wang, Du, Fan, & Xing, 2017) and other methods. Although taking into account other factors, these methods cannot be used directly for the influence maximization problem. Because they all ignore the overlapping of the nodes' influence sphere.

Above all, there are three main problems in the previous studies:

- The node influence scores are so close because of the low resolution.
- The spheres of influence between the nodes overlap a lot.
- It is difficult to balance between accuracy and efficiency.

In recent years, some scholars try to solve the problem with new ideas (He, Fu, & Chen, 2015; Sheikhahmadi & Nematbakhsh, 2016; Zhao, Huang, Tang, Zhang, & Chen, 2014), but the effect is not very obvious. Since the existing methods are limited by the above three points, we try to propose a method to select a set of superior nodes to maximize the influence. Groups and communities are the same concepts in different subjects. On the basis of VoteRank (Zhang, Chen, Dong, & Zhao, 2016), we consider the information of the individual and the group to improve the resolution and solve the overlapping problem. While improving the accuracy, the efficiency of the method is guaranteed. The main contributions are as follows:

- Measuring the attractive power with neighbors based on the social conformity;
- Measuring the initiating power based on community size and node location;
- Proposing the community tightness based on the overlapping problem;
- Providing a node selection strategy from the view of individual and group.

The outline of this paper is as follows: Section 2 makes the overview of the related work; Afterwards, Section 3 describes the definition and the algorithm flow of our proposed methods; Experimental results are provided in Section 4; Section 5 summarizes the full paper and makes the conclusion.

2. Related work

2.1. Node ranking method

Most researchers focus on identifying the super spreader by sorting them in influence. A multitude of studies (Bonacich, 1972; Chen, Lü, Shang, Zhang, & Zhou, 2012; Gao, Wei, Hu, Mahadevan, & Deng, 2013) have used the degree centrality. These studies think the degree of a node represents its influence. However, there are quiet a few nodes with the same degree. This leads to the low resolution of influence. Moreover, some other important features are not considered. As the research moves along, scholars have noticed that. ClusterRank (Chen et al., 2013) got positive performance by considering the number of neighbors and the negative influence of local clustering on the network. Grassi et al. (2019) offered the relationship between different betweenness measures to find the criminal leaders. K-shell (Kitsak et al., 2010) measures the importance of the node from a global perspective. It is good at discovering the nodes at the core. But it faces the same problem as the degree ranking method does. IC (Wang, Du et al., 2017) improved K-shell, which used the iteration information of K-shell and the location to distinguish the nodes with same K-shell value.

With the rapid increasing of users and the explosive growth of data, HITs (Kleinberg, 1999), PageRank (Brin & Page, 1998), and LeaderRank (Li, Zhou, Lü, & Chen, 2014; Lü, Zhang, Yeung, & Zhou, 2011) were proposed for large-scale networks. PageRank has been used to study the importance of nodes in social networks (Fortunato, Boguna, Flammini, & Menczer, 2005; Heidemann, Klier, & Probst, 2010). Instead of calculating the degree, it regards the possibility that each neighbor jumps to the target node as the influence.

The above methods have a good performance to measure the spreading ability. Nevertheless, we cannot utilize these methods directly for the influence maximization problem. Because they do not take into account the overlapping issue. VoteRank (Zhang et al., 2016) used the local structural information of nodes and ranked the nodes by voting. This method provides a fast and convenient solution for finding important node set in large-scale networks. Unexpectedly, it is not fully considered the overlapping problem and other features.

2.2. The methods based on greedy and heuristic algorithm

Kempe et al. (2003) reported a hill climbing method based on the greedy algorithm. To get the accurate results, this method needs to be simulated thousands of times to reach the average level. Hence, the efficiency of this method is low. To figure out this problem, Cheng et al. (2013) calculated the influence of nodes based on the static propagation graph. They significantly improved the efficiency. But it still cannot agree with the large-scale networks. All in all, these methods face the high time complexity.

Few scholars tried to solve the influence maximization problem from other views. Ren et al. (2018) studied the node weighted Laplacian operator to optimization this problem with the minimum set of nodes. Zhao et al. (2014) extended the coloring problem to the complex networks. He et al. (2015) identified the influential nodes by the community structure. The IMSN algorithm (Sheikhahmadi & Nematbakhsh, 2016) selects seed nodes by considering the friends of the node. This strategy avoids the overlapping issue caused by the common friends. Sadly, this method still needs more parameters obtained by experience.

2.3. Social conformity theory

Conformity is an effect that the attitudes, beliefs, and behaviors of individuals tend to be same with the groups (Cialdini & Goldstein, 2004). Xu, Wang, and Zhang (2015) thought that social influence affected the decision-making process of individuals, which led to the herding effect. Wang, Jin, Cheng, and Yang (2017) revealed the influence of conformity on retweeting from emotional perspective. Studying the influence maximization problem with social conformity theory makes for understanding the influence propagation phenomenon.

2.4. Node similarity measure

A slice of studies measured the strength of edges by the similarity between nodes to determine the transmission of information between nodes (Pap, Jocić, Szakál, Obradović, & Konjović, 2017; Sun, Hu, Yang, Yao, & Yang, 2017). The main idea of node similarity measure is to determine the similarity between the two nodes by their common features. Its theoretical basis is the homogeneity in sociology. The most popular methods is Common Neighbor (Granovetter, 1977).

3. Methods

In this section, a node ranking algorithm named AIRank is proposed to improve VoteRank. In this method, the feature of the

node's neighbors captured based on the conformity is used to measure the node attractive power. The node's position and community size are taken into account to estimate the node initiating power. These measures enhance the resolution of node influence. By considering the community tightness and the feature of information coverage, a node selection strategy is introduced to solve the overlapping problem. Afterwards, the process of AIRank is provided to select a set of nodes with the strong spreading ability.

3.1. Attractive power

In essence, VoteRank measures node's influence through the local centrality. In VoteRank, the vote ability of a node is fixed to 1 at first. It denotes every neighbor is same to the node. Two nodes with the same degree exiting in the network is a common thing. Because of this rule, these two nodes cannot be distinguished in the same round of voting. This causes the low resolution. In social networks, the intimate degree of the relationships between individuals are different, the attractive power of individual is not necessarily the same. If we regard the voting process as the actual activity, then the voting ability that each nodes shows to others will vary with the attractive power. In order to improve the resolution, we propose the attractive power according to the node's neighbor, which makes the measurement accurate and reasonable. The attractive power of node i to node j can denote by the number of the followers that node i and node j's friends have.

Definition 1. Given a network $G(V, E)$, for any node $i, j \in V$, the attractive power that node i has to node j with respect to the feature of neighbors, denoted as $AP(i, j)$, is defined as

$$AP(i,j) = \begin{cases} \frac{|N_{out}(i)|}{\sum_{v \in N_{out}(j)} |N_{in}(v)|}, & \sum_{v \in N_{out}(j)} |N_{in}(v)| \neq 0 \\ \frac{1}{|N_{out}(j)|}, & \sum_{v \in N_{out}(j)} |N_{in}(v)| = 0, N_{out}(j) \neq \emptyset \end{cases} \tag{1}$$

where $N_{out}(i)$, $N_{out}(j)$ represents the successors of node i (the nodes that node i like) and node j (the nodes that node j like) respectively, node $v \in N_{out}(j)$, $N_{in}(v)$ refers to the followers of node v. When $N_{out}(j) = \emptyset$, $AP(i, j) = 0$.

Social conformity theory asserts that individuals tend to follow the choice of majority (Xu et al., 2015). The more followers node i has, the greater influence it has. This illustrates that more friends around node j like node i. The more likely node j will follow its neighbors to retweet from node i. The attractive power of node i to node j implies how much node i is attracted to node j by comparing with other neighbors of node j. The more neighbors that node i has, the more node i is attracted to node j.

3.2. Initiating power

Another feature that VoteRank does not consider is the location of the node. Fig. 1 shows an simple network. Node 1, node 2, node 3, and node 4 belong to community G1. Node 3 and node 7 are in community G2, node 4 and node 8 are in community G3. Node 1 and node 2 have the same number of neighbors. Hence, they get the same vote scores in VoteRank. The neighbors of node 2 are in the same community, the information is more likely spreaded in this community. However, the neighbors of node 1 are distributed in different communities. Obviously, the position of node 1 is more important than that of node 2. Node 1 is a bridge node. Although the degree of the bridge node is low, it links few communities and promote information spread between communities. Thus, it is more vital than other common nodes with the same degree.

Considering the above situation, we propose the node's detonate power based on the location and community size to highlight

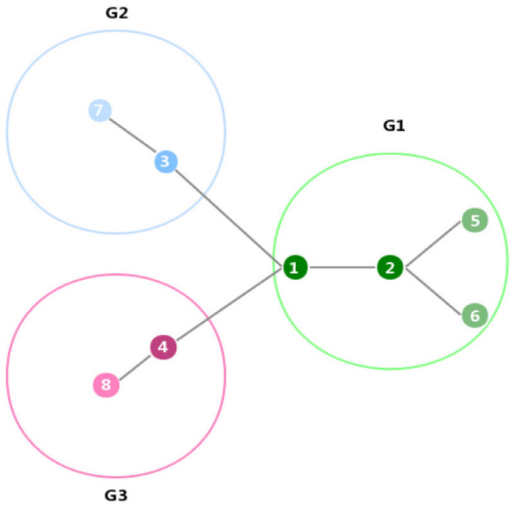


Fig. 1. The action diagram of the effect of the bridge node.

the effect of the bridge node. The initiating power also can improve the resolution. It is an additional vote ability. Node j is one of the neighbor of node i. In the process of voting, if node j and node i are in the different communities, node j is endowed with this ability. At this point, node j is not only representing itself, but also the community it belongs to. In order to measure this ability reasonably, we add a community with zero members into community set, and normalized the size of community. The value that we get is the node's initiating power. If node j and node i are in the same community, its initiating power is 0.

Definition 2. Given a network $G(V, E)$, for any node $i, j \in V$, the initiating power that node j gives to node i with respect to the feature of position and community, denoted as $IP(i, j)$, is defined as

$$IP(i,j) = \begin{cases} 0, & G(i) = G(j) \\ \frac{|G(j)|}{max(|G(v)|)}, & v \in V, G(i) \neq G(j) \end{cases} \tag{2}$$

where $G(j)$, $G(v)$ detnotes the community that node j and node v belongs to respectively, $|G(j)|$ refers to the size of the community that node j belongs to, $max(|G(v)|)$ represents to the size of the biggest community in the network.

Take Fig. 1 as an example. Because the neighbors of node 2 are in the same community, $IP(2, 1) = 0$, $IP(2, 5) = 0$, and $IP(2, 6) = 0$, the total initiating power that node 2 obtains is 0. Node 2 and node 1 are in the same community, $IP(1, 2) = 0$. However, the communities that node 3 and node 4 belong to is different from node 1, and the community size of these two nodes are both 4. According to the above Definition 2, $IP(1, 3) = 2/4 = 0.5$, $IP(1, 4) = 2/4 = 0.5$. The total initiating power that node 1 gets is 0+0.5+0.5 = 1. Thus, the initiating power emphasizes the importance of the bridge node, and node 1 and node 2 can be further distinguished.

Both attractive power and initiating power are important to the resolution of the method. By combing these two vote ability, we can calculate the vote score of a node.

Definition 3. Given a network $G(V, E)$, for any node $i \in V$, the vote score of node i, denoted as $VS(i)$, is defined as

$$VS(i) = \sum_{j \in N(i)} (AP(i,j) + IP(i,j)) \tag{3}$$

3.3. Node selection strategy

3.3.1. The view of individual

By weakening the vote ability of the selected node's neighbors, VoteRank reduces the effect of the overlapping problem. In social networks, when the original node publishes the message, all its neighbor can get its message. Whether the neighbors choice to retweet or not, they all have been affected. If a node is selected as a seed, its friends are not allowed to be in the seed set. Lower the neighbors' vote ability cannot make sure of it. On the micro level, the node has an impact on its neighbors at least. In order to further reduce the overlapping problem, when the node is selected as a seed, its friends will delete from the network. This strategy can further deal with the overlapping problem.

3.3.2. The view of group

Not only do the social networks consist of a number of interconnected nodes, but it also made up of abundant interconnected communities. The stronger relationships are, the higher frequency of interaction the nodes will be. It is true of communities, as well. The overlapping problem is still exits communities. The node similarity is used to solve the overlapping problem in the view of node in some methods. Similarly, we propose the community tightness to solve it in the view of communities.

Definition 4. Given a network $G(V, E)$, for any node i, $j \in V$, the community tightness between node i and node j, denoted as $CT(i, j)$, is defined as

$$CT(i,j) = \frac{|N(Com(i)) \cap N(Com(j))|}{min(|N(Com(i))|, |N(Com(j))|)} \quad (4)$$

where $N(Com(i))$, $N(Com(j))$ refers to the neighbor node set of the communities that node i and node j belong to respectively.

The higher the CT value is, the more tightly the communities connect. The message can be more easily spreaded between these two communities. In other words, when a node in community A is selected as a seed, the nodes in the community B which is tightly connected with community A will be influenced. Accordingly, these nodes will not be selected to reduce the impact of the overlap.

3.4. Algorithm description and complexity analysis

Now, we analyze the time complexity of Algorithm 1. Steps (1), (2) are the initialization. Step (3) uses the community detection method to get the community feature of the nodes. In our method, LPA is applied. The complexity is $O(n)$. Steps (4)–(5) calculate the attractive power and detonate power. The complexity of these steps is $O(n)$. Step (7) counts the vote score of the nodes. The complexity is $O(n)$. Steps (8)–(9) are the selection stage of the method, and the complexity is $O(n)$. Step (10) is the initialization. Thus, the total complexity of Algorithm 1 is $O(n)$.

Algorithm 2 adds the community tightness based on Algorithm 1. Step (6 calculates the community tightness, and the complexity is $O(n^2)$. Steps (9)–(20) are the selection rules for the community tightness. The CA value between the selected node and the nodes in S is higher than the threshold η, implies that the nodes in S can influence the selected node. Hence the node selected does not exist in S. The total complexity of the algorithm with community tightness is $O(n^2)$.

Algorithm 1 AIRank.

Input:
The network, $G = (V, E)$;
The total node number, n;
The size of influential spreaders set, r;
Output:
The top-r influential spreaders set, S;

1: $S = \emptyset$;
2: Set VS value as 0 for nodes in V;
3: Get the community results;
4: Calculate AP using formula (1) for nodes in V;
5: Calculate IP using formula (2) for nodes in V;
6: **while** $|S| < r$ **do**
7: Calculate $VS(i)$ using formula (3) for nodes in V;
8: Add node a which has biggest VS value into S;
9: Delete node a's neighbors from G, and a no longer votes;
10: Set VS value as 0 for nodes in V;
11: **end while**
12: **return** S;

Algorithm 2 AIRank': AIRank with community tightness.

Input:
The network, $G = (V, E)$;
The size of influential spreaders set, r;
The community tightness threshold, η;
Output:
The top-r influential spreaders set, S;

1: $S = \emptyset$;
2: Set VS value as 0 for nodes in V;
3: Get the community results;
4: Calculate AP using formula (1) for nodes in V;
5: Calculate IP using formula (2) for nodes in V;
6: Calculate CT using formula (4) for all node pairs in V;
7: **while** $|S| < r$ **do**
8: Calculate VS using formula (3) for nodes in V;
9: **if** $S \neq \emptyset$ **then**
10: Add node v with biggest VS value into S;
11: **else**
12: Select node v with biggest VS value;
13: Get the CT value list (CT_list) between node v and nodes in S;
14: Get the max value ($CA_{\max}$) from CA_list;
15: **if** $CA_{\max} \geq \eta$ **then**
16: Delete v from G and VS;
17: **else**
18: Add v into S, break;
19: **end if**
20: **end if**
21: Delete $N(v)$ from G, and v no longer vote;
22: Set VS value as 0 for nodes in V;
23: **end while**
24: **return** S;

4. Results and discussions

4.1. Data preparation

Three real world networks with various scales are used to evaluate the performance of different methods. Eu-core (Yin, Benson, Leskovec, & Gleich, 2017) is an email communication network from a large European research institution. Epinions (Richardson, 2003) is a general consumer review site in which users are nodes and the trust relationships are edges. In Notre Dame network (Albert, 1999), nodes represent pages from University of Notre Dame and edges represent hyperlinks between them. The information about the three datasets is shown in Table 1.

Table 1
Statistics of datasets.

Networks	Type	Nodes	Edges	$<k>$	k_{max}	$<c>$
Eu-core	directed	1005	25,571	20.0404	546	0.3994
soc-Epinions	directed	75,879	508,837	13.4118	3079	0.1378
Notre Dame	directed	325,729	1,497,134	4.5120	344	0.2346

$<k>$ is the average degree for the networks. k_{max} is the maximum degree. $<c>$ is the average clustering coefficient.

4.2. Evaluation metrics

SIR model (Tao, Zhongqian, & Binghong, 2006) is used to examine the spread speed of nodes. We conduct the experimenttal of the final affected scale based on SIR and IC model (Kempe et al., 2003). In SIR model, the infected rate $\lambda = \mu/\beta$ (μ is the infection probability, β is the recovery probability). In IC model, α represents the active probability.

We use the infected scale at time t to denote the spread speed of the compared method. The infected scale $F(t)$ (Zhang et al., 2016) is defined as follows:

$$F(t) = \frac{N_{I(t)} + N_{R(t)}}{N} \tag{5}$$

where $N_{I(t)}$, $N_{R(t)}$ are the amount of nodes in infecting and recovering state at time t respectively, N is the amount of nodes in the network.

The final affected scale (Zhang et al., 2016) is used to measure the final scale of the infection. The final affected scale $F(t_c)$ is defined as follows:

$$F(t_c) = \frac{N_{R(t_c)}}{N} \tag{6}$$

where $N_{R(t_c)}$ denotes the number of the affected nodes in SIR model and the activated nodes in IC model when the spread process stops.

4.3. Results

We firstly tune the value of the communtiy tightness to evaluate the influence on AIRank'. Then, AIRank and AIRank' have been compared to PageRank, ClusterRank, IMSN and VoteRank with the time complexity and the above evaluation metrics. The results are averaged over 50 independent runs.

4.3.1. Experiments for community tightness

In this experiment, the value of community tightness is set as $\eta \in [0.5, 0.95]$, the infected rate in SIR model is set as $\lambda = 1.5$, the active probability is set as $\alpha = 0.3$ and p represents the ratio of the number of seed nodes and that of nodes in network. The results are shown in Figs. 2 and 3.

Fig. 2 shows the curve of $F(t_c)$ with different η. In three networks, $F(t_c)$ shows tendency to ascend when $0.5 \le \eta \le 0.8$ both in SIR and IC model. For Eu-core, there are two peaks ($\eta = 0.8$, $\eta = 0.85$). For soc-Epinions and Notre Dame, $F(t_c)$ reaches a peak when η is 0.85 in both two models. Finally, $F(t_c)$ declines gradually when η>0.9.

The community tightness reveals that the number of common neighbors between two communities. The intergroup communication of information depends upon the common neighbors. The threshold of community tightness η is small, the number of common neighbors is relatively small. It is more and more possible that the information sticks in the community, which will block the intergroup diffusion. Thus, $F(t_c)$ is hard to expand. With the rising number of common neighbors, the intergroup communication is relatively easy to accomplish. Hence, $F(t_c)$ is continuously enlarging. When the common neighbor number increases to some point, two communities are too interconnected. Although we do not select seed nodes respectively from these two communities, the information will travels from one community to the other community. In other words, The influence of seed nodes from these two communities overlap and interweave. It is not only a waste of resources but also causes a down trend for $F(t_c)$.

The analysis and experimental results show that the algorithm has high performance within community tightness $\eta \in [0.8, 0.95]$. At the same time, the overlapping problem would be properly addressed.

Fig. 3 shows the curve of $F(t)$ over time with different η. As we can see from Fig. 3, there is a big influence on $F(t)$ when the community tightness threshold is the same value. First, $F(t)$ increases dramatically in the intervals [0,20] for Eu-core, [0,25] for soc-Epinions and [0,17] for Notre Dame. Then it presents a slow increase trend. Finally, it comes to a steady state when $t \le 75$. In the intervals [0,10] for Eu-core and [0,14] for soc-Epinions, the algorithm with $\eta = 0.5$ is mediocre. The rising rate of the curves with high value of η (0.75,0.85,0.90) above it is high. On the contrary, the rising rate of the curves with low value of η (0.6, 0.65, 0.7) under it is low. In the intervals [0,14] for Notre Dame, the algorithm with $\eta = 0.5$ is the worst. In the intervals [35, 60] for soc-Epinions and [20, 40] for Notre Dame, the curves with high value still remain high value of $F(t)$. When $\eta = 0.85$, the algorithm achieves the best results.

If η is too small, the influence gap between seed nodes will be so wide that the rising rate of $F(t)$ cannot ascend. The information cannot have an effective communication. And if η is too large, the influence of seed nodes will overlap. The information would propagate between the same throngs again and again. This causes the infected scale will not be able to expand.

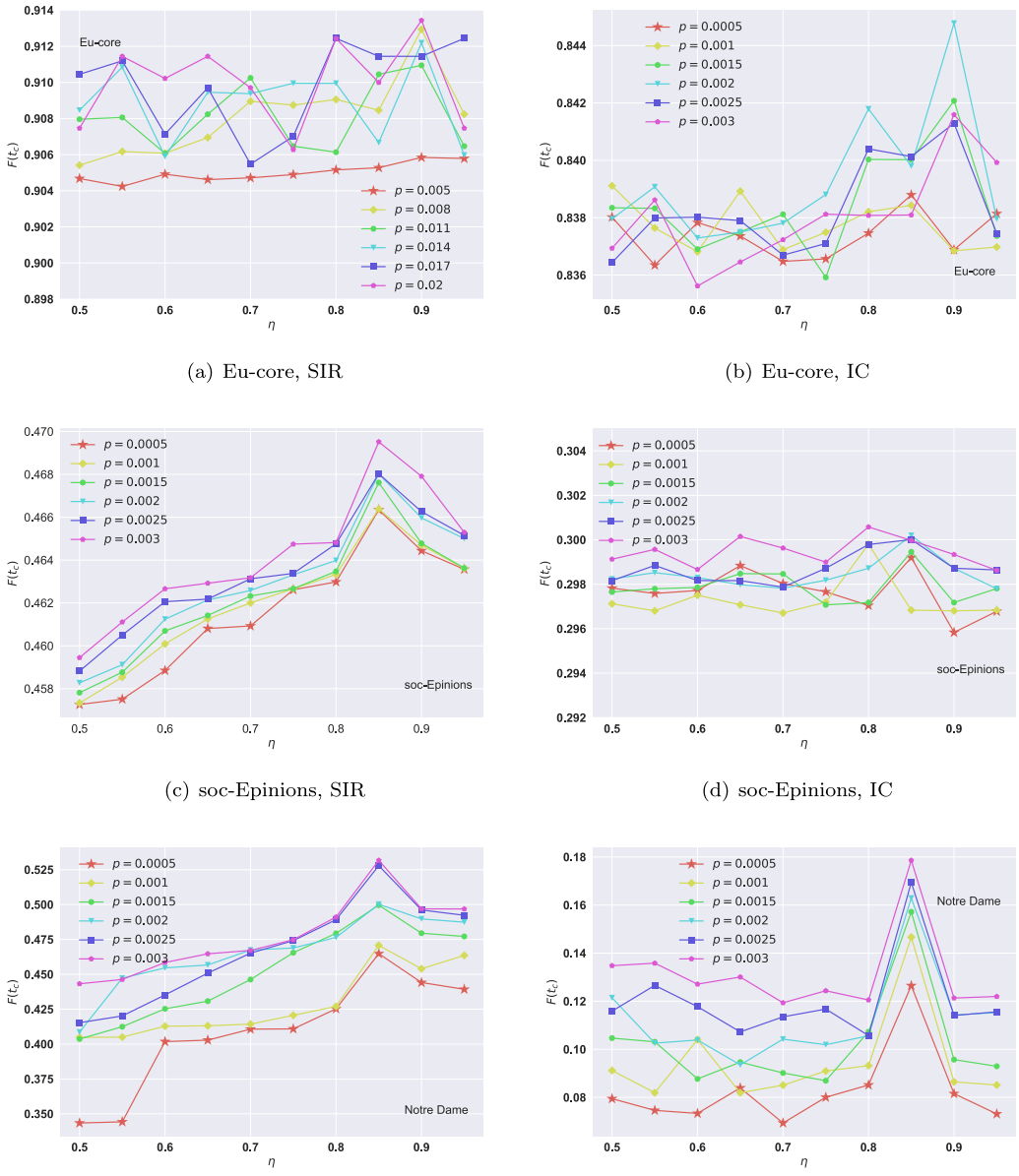
In this experiment, the algorithm with high η keeps a high propagation speed. Based on the results of Figs. 2 and 3, we set η to 0.85 in order to observe the results of the comparison test clearly.

4.3.2. Experiments for the time complexity and other parameters

To assess the performance of our proposed methods (AIRank and AIRank') effectively, we compare it with four ranking method, including PageRank, ClusterRank, IMSN, and VoteRank. First, we list the time complexity of these methods in Table 2. Then, we compare these methods with the evaluation metrics which are applied have introduced in Section 4.2. In this experiment, the vlaue of η in AIRank' is 0.85.

The infected scale $F(t)$ on three networks under different methods with infected rate $\lambda = 1.5$ are shown in Fig. 4. When time t is in the interval [0,10], all methods keep a high increasing rate in three networks. By contrast, the increasing rate of AIRank' is slower in soc-Epinions. In addition, AIRank' gains the largest infected scale and pulls away from the other methods in the period [55, 80] in soc-Epinions. In this interval, AIRank has the second infected scale only to AIRank'. Other methods have similar results. In Notre Dame and Eu-core, the increasing rate of AIRank' is the fastest in the interval [0,10]. The results of these two networks are slightly different for the second half of the diffusion. In Eu-core, $F(t)$ of AIRank' is not the biggest but is similar to the others. In Notre Dame, ClusterRank and AIRank' have the similar $F(t)$ in the period [20,40], followed by AIRank.

The identical vote ability of the nodes in VoteRank causes the low resolution. The attractive power in AIRank can effectively distinguish the nodes with same degree value and makes the seed node selection more scientific. As a result, the increasing rate of

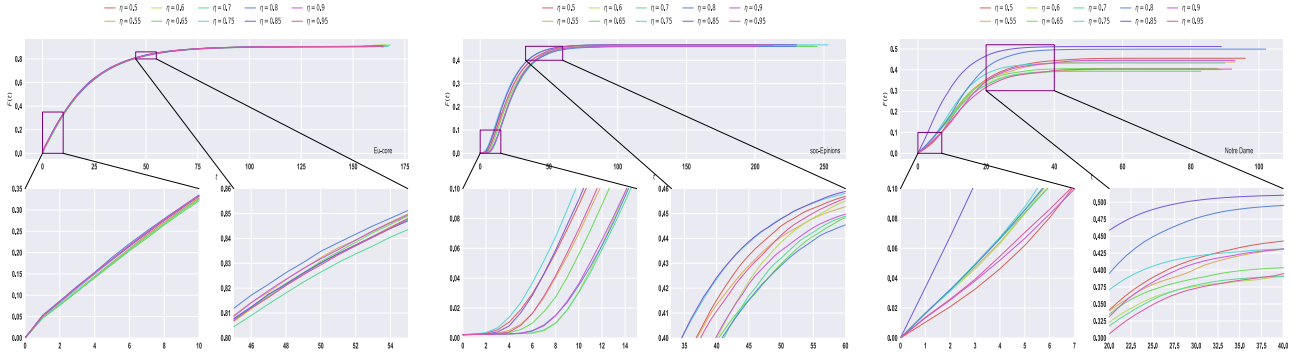


(a) Eu-core, SIR (b) Eu-core, IC

(c) soc-Epinions, SIR (d) soc-Epinions, IC

(e) Notre Dame, SIR (f) Notre Dame, IC

Fig. 2. The final infected scale $F(t_c)$ of AlRank' with different value of community tightness η.



(a) Eu-core, SIR (b) soc-Epinions, SIR (c) Notre Dame, SIR

Fig. 3. The infected scale $F(t)$ of AlRank' with different value of community tightness η ($p = 0.02$ for Eu-core and $p = 0.002$ for the other two networks).

Table 2
Time complexity of the methods.

Methods	Time complexity	Methods	Time complexity
AIRank	$O(n)$	IMSN	$O(n)$
AIRank'	$O(n^2)$	ClusterRank	$O(n^2)$
VoteRank	$O(n)$	PageRank	$O(n)$

AIRank is slightly higher than that of VoteRank at the initial stage of infection. Moreover, the initiating power in AIRank stresses the role of bridge nodes and makes the estimation of the interest sphere more accurate. Hence, the infected scale of AIRank is bigger than that of VoteRank at the end of infection. AIRank' adds the seed node selection strategy at the angle of groups based on AIRank. It further resolves the overlapping problem. Consequently, the infected scale of AIRank' is bigger than the other methods. Although, this seed node selection strategy alleviates the overlapping problem. Some nodes with large degree value are abandoned during the selection process. The increasing rate of nodes with high degree certainly is higher than that of the nodes with low degree. Then, AIRank' has a slower speed in soc-Epinions. Compared the attribute values of the two networks in Table 1, we can easily come to the following conclusion. Although the scale of soc-Epinions is smaller than Notre Dame, the maximum degree is much bigger than that of Notre Dame. In other words, the degree value of the nodes in Notre Dame has less gap than that of soc-Epinions. This leads to the less gap betweeen seed nodes in Notre Dame. Hence, AIRank' maintains a higher rate in Notre Dame.

The experiment reveals that the attractive power and the initiating power would help to distinguish the influence of the nodes, thus improving the increasing rate and the infected scale of the seed nodes. Besides, the seed node selection strategy is conducive to expand the infected scale. But it is unfavorable to obtain higher increasing rate.

Fig. 5 shows the final infected scale $F(t_c)$ of different seed node set size. The corresponding parameter settings are detailed as follows: $\lambda = 1.5$, $\alpha = 0.3$. As the size of the seed node set increases, the final infected scale of these methods ascend step by step basically for soc-Epinions and Notre Dame both in SIR an IC model. Due to the size of Eu-core, the size of the seeds with different p is so close. Hence, the change of $F(t_c)$ is small. In both three networks, the advantage of AIRank' is obvious. AIRank is appreciably better than VoteRank.

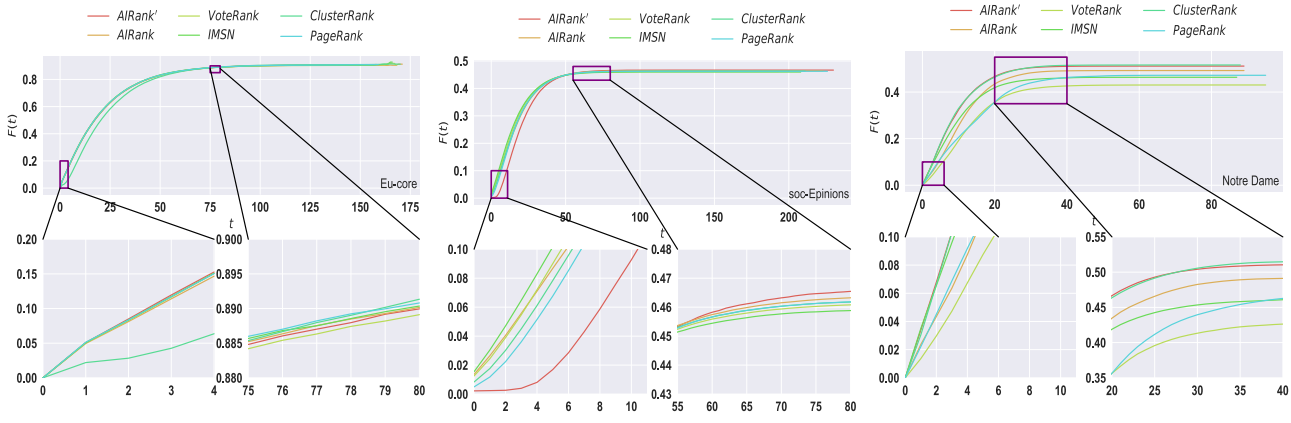
The main difference between AIRank and VoteRank is the assessment of the nodes' vote ability. We transit the same vote ability into the ability which depends on the node' location and the conformity. AIRank can differentiate the node with the same vote score in VoteRank. At the same time, it highlights the contribution of the bridge node to the community communication and reduces the probability the information spreads repeatedly within the community. Without doubt the probability the information spreads by intergroup communication can be increased. Therefore, the final infected scale of AIRank is slightly bigger than that of VoteRank and the other methods. The influence overlapping problem within the community and between communities are considered fully in AIRank'. The waste of resource is avoided in the seed node selection step. ClusterRank and PageRank do not reflect on this problem. VoteRank and IMSN do not go into the details of the perspective of communities. In a word, the final infected scale of AIRank' is noticeably larger than the other methods.

To summarize, the measurement of influence and the node selection strategy we proposed can be useful in further expanding outreach.

Fig. 6 illustrates the final infected scale $F(t_c)$ of six methods against λ ranging from 1.0 to 2.0, α ranging from 0.05 to 0.5. With gradually larger λ and α, the final infected of all six methods increase steadily in three networks, as shown in Fig. 6. In SIR model, AIRankćan achieve relatively large infected scale under the different value of λ and α. In most cases, AIRank is better than the other methods. In IC model, the advantage of these two methods still exits, but not so obvious.

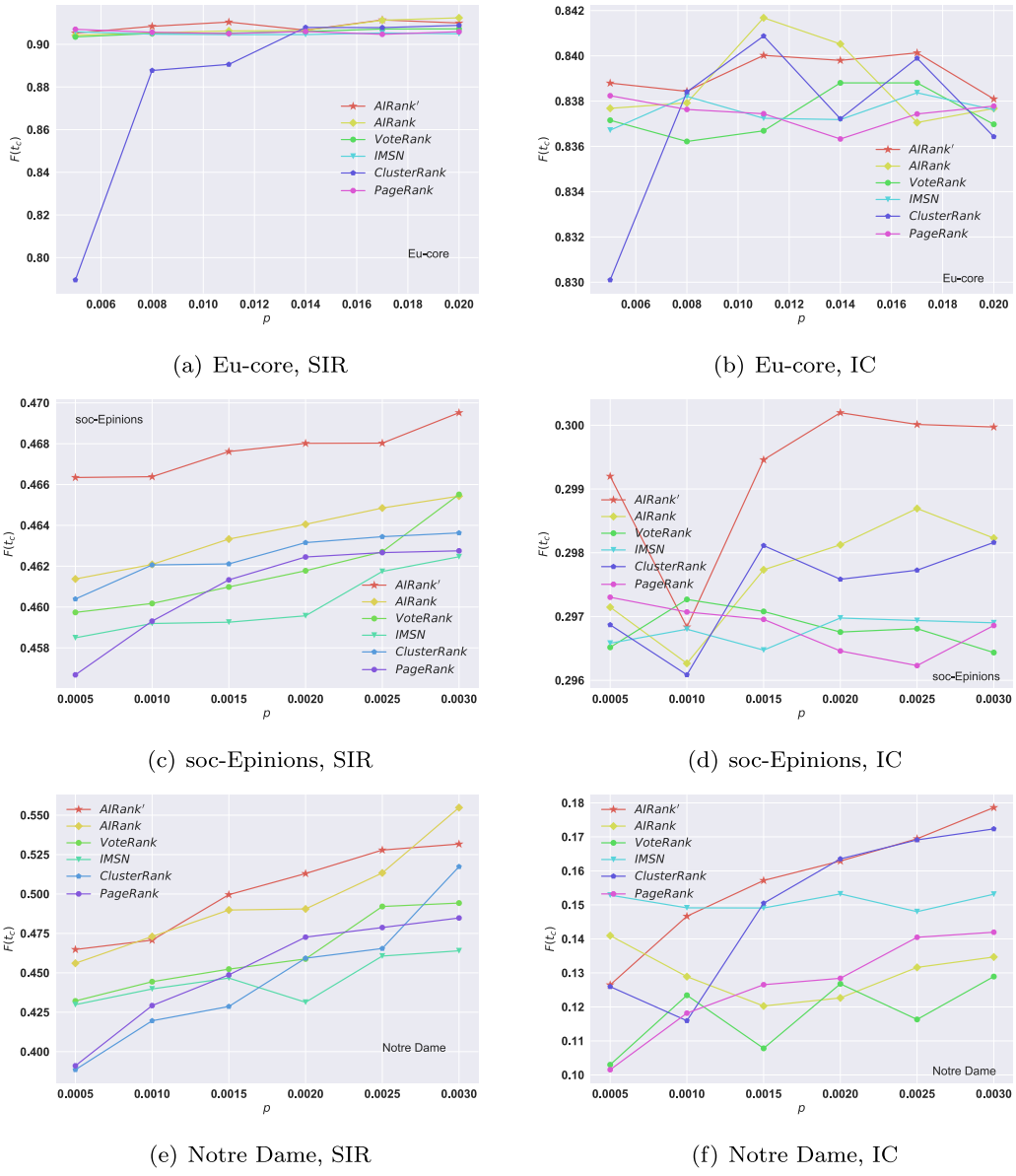
The influence of the node is not only related to the structure-property but also closely connected with the node location (Wang, Du et al., 2017). Our proposed method emphasizes the role of the bridge node in intergroup communication. Moreover, we try to avoid the influence overlap between seed nodes from the perspective of the community. These two steps both underlined the importance of the node location. To that extent, AIRank and AIRank' give a good account of themselves. VoteRank and IMSN just carry on the preliminary inquisition to the overlapping problem. They simply consider the neighbor of the seed nodes. In the view of the community, a community can be regarded as a node. Naturally, the neighborhood overlapping problem also exists between communities.

The results demonstrate that the final infected scale $F(t_c)$ is affected by λ and α as well as the node influence. Ulteriorly, the location property is useful in measuring the influence. Removing the neighbors and communities of the seed node is an effective way to deal with the overlap. Accordingly, the seed node selection strategy we proposed can improve the dissemination of results.



(a) Eu-core, SIR (b) soc-Epinions, SIR (c) Notre Dame, SIR

Fig. 4. The infected scale $F(t)$ of different ranking methods ($\lambda = 1.5$, $p = 0.02$ for Eu-core and $p = 0.002$ for the other two networks).



(a) Eu-core, SIR (b) Eu-core, IC

(c) soc-Epinions, SIR (d) soc-Epinions, IC

(e) Notre Dame, SIR (f) Notre Dame, IC

Fig. 5. The final infected scale $F(t_c)$ of different ranking methods with different scale of seed set($\lambda = 1.5$, $\alpha = 0.3$).

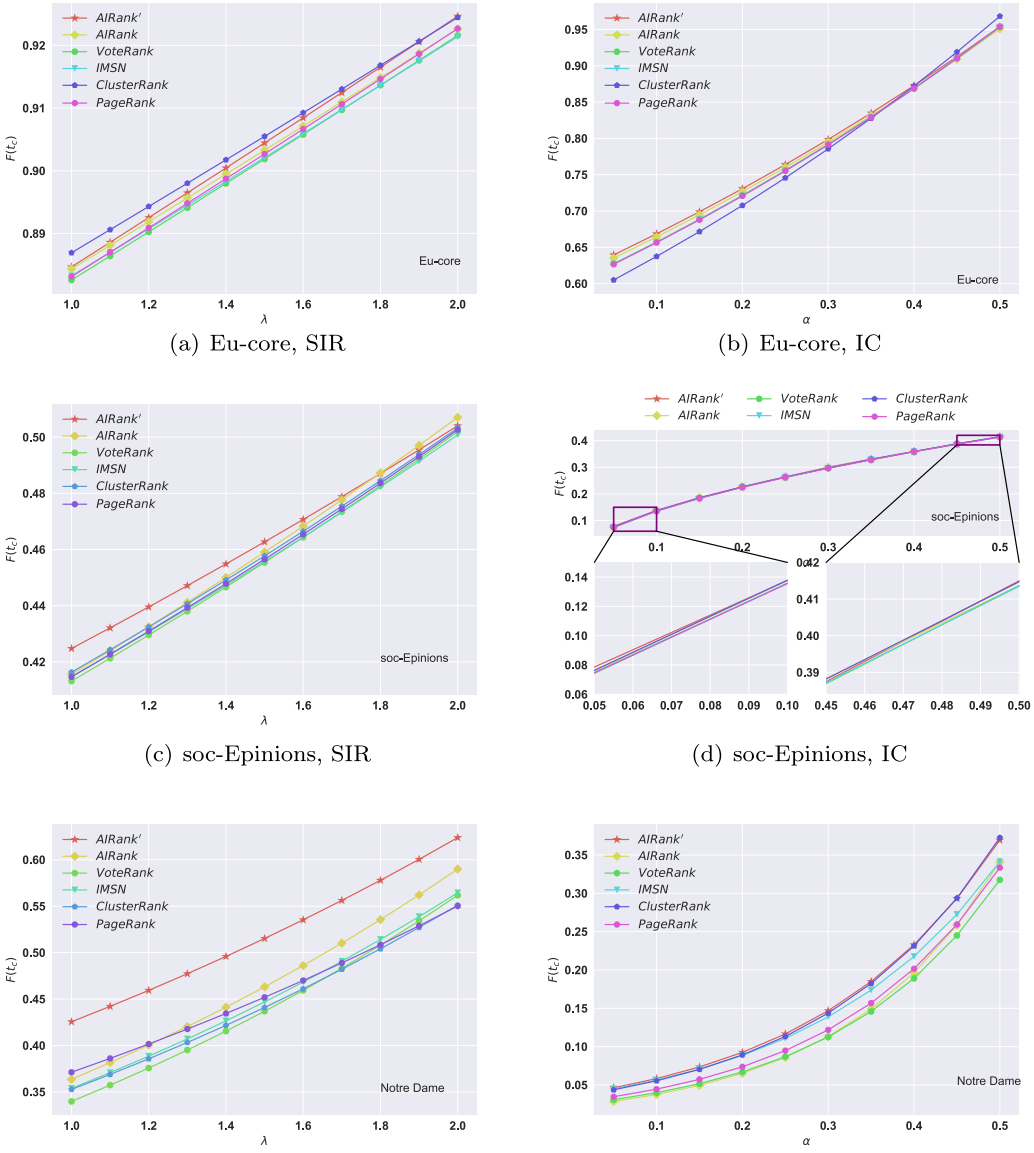
4.4. Discussion

We have identify a set of distributed nodes with strong propagation ability with our proposed methods. AIRank obtain larger final infected scale and higher infecting speed than VoteRank in most cases. This illustrates that the location and community properties have a bearing on the node influence. It help us find the seed nodes with better quality. To a certain extent, the attractive power and the initiating power improve the resolution. The social conformity theory has certain guiding significance for the node influence measurement. AIRank' performs outstandingly in the experiment of the final infected scale. It indicates that the seed node selection strategy can control the pull of distance between seed nodes. What's more, it makes the overlapping problem defect obsolete.

Although we overcome the influence overlapping problem, the infecting speed of our method is not always high as shown in Fig. 4 (b). Although the size of soc-Epinions is smaller than the others, the maximum degree is much bigger than the others according to Table 1. This means the gap between the seed nodes in soc-Epinions is much bigger. The message needs more steps to spread from one seed to another. And in these steps, the number of the infected nodes is smaller than the other networks. This leads the infecting speed of AIRank' is small in soc-Epinions.

In addition, the time complexity of AIRank' is high. The seed node selection strategy cost two much time, and the threshold of the community tightness is obtained by experience. Hence, we need to further study how to make the algorithm adapt to different kinds of networks and find the best threshold to identify the seeds quickly.

All in all, groups in sociology has an impact on decision making to individuals. The scale of community in complex networks implies the scope of the node influence. As consequence, groups make nodes powerful. They can help us identify the influential nodes in social networks. Our method provides the possibility for widely information spread and has a positive significance for high quality information.



(a) Eu-core, SIR (b) Eu-core, IC

(c) soc-Epinions, SIR (d) soc-Epinions, IC

(e) Notre Dame, SIR (f) Notre Dame, IC

Fig. 6. The infected scale $F(t)$ of different ranking methods with different value of λ and α ($p = 0.02$ for Eu-core and $p = 0.002$ for the other two networks).

5. Conclusions

Majority and community are the concepts of the group in different subjects. Individuals depend on groups and are affected by groups. On the basis of this fact, we propose two node ranking methods based on the social conformity theory and community feature, AIRank and AIRank'. AIRank starts with the neighbor's psychological feature and quantifies the attractive power. It differentiates the nodes with the same degree and improves the resolution of the ranking result. Then, AIRank combines location with community features to quantify the initiating power. It highlights the power of the bridge nodes and further deals with the resolution problem. Finally, AIRank adds the node selection strategy from the individual aspect. It reliefs from the overlapping problem. AIRank' adds the node selection strategy from group aspect based on AIRank. It solves the overlapping completely.

On the basis of the real social network datasets, the experimental results have shown that AIRank performs better than VoteRank, IMSN, PageRank, and ClusterRank. Because the time complexity of AIRank is $O(n)$, it balances between the relationship between accuracy and efficiency well. It could be applied on large networks. The spreading power of AIRank' is the strongest in the experiment. However, its time complexity is $O(n^2)$. It is not suitable for advertising which demands for rapid time-to-market. The results showed that groups make nodes powerful.

In the future, we will take into account the topic of the tweet and user's behavior to detect influential nodes.

Acknowledgments

Funding: This work was supported in part by the National Natural Science Foundation of China [grant numbers 61672179, 61370083, and 61402126], in part by the Natural Science Foundation of Heilongjiang Province [grant number F2015030], in part by the Science Foundation for Youths of Heilongjiang [grant numbers QC2016083], and in part by the Postdoctoral Foundation of Heilongjiang Province [grant numbers LBH-Z14071].

References

Agreste, S., Catanese, S., Meo, P. D., Ferrara, E., & Fiumara, G. (2016). Network structure and resilience of mafia syndicates. *Information Sciences, 351*, 30–47.

Albert, R. (1999). Diameter of the world wide web. *Nature, 401*, 130–131.

Albert, R., Albert, I., & Nakarado, G. L. (2004). Structural vulnerability of the north american power grid. *Physical Review E Statistical Nonlinear & Soft Matter Physics, 69*(2), 025103.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2*(1), 113–120.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Network and ISDN Systems, 30*(1–7), 107–117.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1777–1787.

Chen, D.-B., Gao, H., Lü, L., & Zhou, T. (2013). Identifying influential nodes in large-scale directed networks: The role of clustering. *PloS One, 8*(10), e77455.

Cheng, S., Shen, H., Huang, J., Zhang, G., & Cheng, X. (2013). Staticgreedy: Solving the scalability-accuracy dilemma in influence maximization. In *Proceedings of the 22nd ACM international conference on information & knowledge management*. In *CIKM '13* (pp. 509–518). New York, NY, USA: ACM.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1), 591.

Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., & Nussinov, R. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review.. *Pharmacology & Therapeutics, 138*(3), 333–408.

Dinh, T. N., Nguyen, D. T., & Thai, M. T. (2012). Cheap, easy, and massively effective viral marketing in social networks:Truth or fiction? In *ACM conference on hypertext and social media* (pp. 165–174).

Fortunato, S., Boguna, M., Flammini, A., & Menczer, F. (2005). How to make the top ten: Approximating pagerank from in-degree. *Physics*, 59–71.

Gao, C., Wei, D., Hu, Y., Mahadevan, S., & Deng, Y. (2013). A modified evidential methodology of identifying influential nodes in weighted networks. *Physica A: Statistical Mechanics and its Applications, 392*(21), 5490–5500.

Granovetter, M. S. (1977). The strength of weak ties. In S. Leinhardt (Ed.), *Social networks* (pp. 347–367). Academic Press.

Grassi, R., Calderoni, F., Bianchi, M., & Torriero, A. (2019). Betweenness to assess leaders in criminal networks: New evidence using the dual projection approach. *Social Networks, 56*, 23–32.

He, J.-L., Fu, Y., & Chen, D.-B. (2015). A novel top-k strategy for influence maximization in complex networks with community structure. *PloS One, 10*(12), e0145283.

Heidemann, J., Klier, M., & Probst, F. (2010). Identifying key users in online social networks: A pagerank based approach. In *International conference on information systems, ICIS 2010, Saint Louis, Missouri, USA, December* (p. 79).

Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACMSIGKDDinternational conference on knowledge discovery and data mining*. In *KDD '03* (pp. 137–146). New York, NY, USA: ACM.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*(11), 888–893.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604–632.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2005). The dynamics of viral marketing. *ACM Transactions on the Web, 1*(1), 228–237.

Li, Q., Zhou, T., Lü, L., & Chen, D. (2014). Identifying influential spreaders by weighted leaderrank. *Physica A: Statistical Mechanics and its Applications, 404*, 47–55.

Lü, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PloS One, 6*(6), e21202.

Narayanam, R., & Narahari, Y. (2010). A shapley value-based approach to discover influential nodes in social networks. *IEEE Transactions on Automation Science & Engineering, 8*(1), 130–147.

Pap, E., Jocić, M., Szakál, A., Obradović, D., & Konjović, Z. (2017). Managing big data by directed graph node similarity. In *IEEE international symposium on computational intelligence and informatics* (pp. 000025–000030).

Pastor-Satorras, R., & Vespignani, A. (2002). Immunization of complex networks. *Physical Review E Statistical, Nonlinear, and Soft Matter Physics, 65*(3 Pt 2A), 036104.

Ren, X. L., Gleinig, N., Helbing, D., & Antulov-Fantulin, N. (2018). Generalized network dismantling. arXiv: 1801.01357.

Resende, M. G. C., & Pardalos, P. M. (2006). *Handbook of optimization in telecommunications*. Springer US.

Richardson, M. (2003). Trust management for the semantic web. *International semantic web conference, 2003*.

Sheikhahmadi, A., & Nematbakhsh, M. A. (2016). Identification of multi-spreader users in social networks for viral marketing. *Journal of Information Science, 43*(3).

Sun, Q., Hu, R., Yang, Z., Yao, Y., & Yang, F. (2017). An improved link prediction algorithm based on degrees and similarities of nodes. In *2017 IEEE/ACIS 16th international conference on computer and information science (ICIS): 00* (pp. 13–18).

Tao, Z., Zhongqian, F., & Binghong, W. (2006). Epidemic dynamics on complex networks. *Progress in Natural Science:Materials International, 16*(5), 452–457.

Wang, Q., Jin, Y., Cheng, S., & Yang, T. (2017). Conformrank: A conformity-based rank for finding top-k influential users. *Physica A: Statistical Mechanics and its Applications, 474*, 39–48.

Wang, Z., Du, C., Fan, J., & Xing, Y. (2017). Ranking influential nodes in social networks based on node position and neighborhood. *Neurocomputing, 260*, 466–477.

Xu, B., Wang, J., & Zhang, X. (2015). Conformity-based cooperation in online social networks: The effect of heterogeneous social influence. *Chaos, Solitons & Fractals, 81*, 78–82.

Yan, G., Zhou, T., Hu, B., Fu, Z. Q., & Wang, B. H. (2005). Efficient routing on complex networks. *Physical Review E Statistical Nonlinear & Soft Matter Physics, 73*(2), 046108.

Yin, H., Benson, A. R., Leskovec, J., & Gleich, D. F. (2017). Local higher-order graph clustering. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 555–564).

Zhang, J.-X., Chen, D.-B., Dong, Q., & Zhao, Z.-D. (2016). Identifying a set of influential spreaders in complex networks. *Scientific Reports, 6*, 27823.

Zhao, X. Y., Huang, B., Tang, M., Zhang, H. F., & Chen, D. B. (2014). Identifying effective multiple spreaders by coloring complex networks. *Epl, 108*(6).