IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 28, 2019, accepted September 14, 2019, date of publication September 27, 2019, date of current version October 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944350

CBIM-RSRW: An Community-Based Method for Influence Maximization in Social Network

FENG CAI, LIRONG QIU, XINKAI KUAI, AND HONGSHUAI ZHAO
School of Information Engineering, Minzu University of China, Beijing 100081, China

Corresponding author: HongShuai Zhao (zhaohs_muc@163.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 61672553, in part by the Project of Humanities and Social Sciences, Ministry of Education of China 16YJCZH076, and in part by the Postgraduates Independent Research Project, Minzu University of China under Grant SSZZKY-2019116.

ABSTRACT Influence maximization is an important problem, which seeks a small set of key users who spread the influence widely into the network. It finds applications in viral marketing, epidemic control, and assessing cascading failures within complex systems. The current studies treat nodes in social network with equal weights, and the influence possibility mainly decide by node degree. In this paper, we study the influence maximization problem in social networks and we improve the independent cascade model to realize the goal of different weights for different users, and the differentiation of influence probability. Meanwhile, We take advantage of the community structure to speed up the algorithm. Then, we propose a method called the reverse reachable index method based on random walk (RSRW) to select potential high-impact nodes from those communities. The experimental result on four actual data set shows that these improvements can greatly reduce the calculation time while ensuring the accuracy of the results.

INDEX TERMS Influence maximization, community discovery, random walk, diffusion model.

I. INTRODUCTION

In recent years, social networks, such as Weibo, Twitter and Wechat, have connected all web users in the Internet and received great attention from all over the world. It provides the opportunity to analyze the spread of product adoption, ideas, and news through the population, and exploit the results to do more things. But it is difficult to convery the infomation(product,news) to all users due to the limited resource. So many researchers choose to select some seed nodes to spread the information to their neighbors and friends, ultimately achieving the goal of maximum the influence spread, which is the core concept of influence maximization.

There are usually two core parts in influence maximization: (i) propagation models and (ii) selection of the initial nodes. And the influence maximization can be divided into greedy algorithm and heuristic algorithm.Domingos and Richardson *et al*. [1] first study the influence maximization problem as an algorithm problem. Tardos *et al*. [2] treat the influence maximization problem as a discrete optimization problem for the first time and officially proposed two types of influence diffusion models: independent cascade model (IC) and linear threshold model (LT) and they proved that the IM is an NP-hard problem.

Many scholars try to optimize the influence maximization from different perspective. For example, Cui *et al*. [3] propose a more efficient algorithm called degree-descending search evolution (DDSE), using the idea of eliminating repeated simulations and replacing it with rough estimation, to improve the low efficiency of the greedy approaches. YunYong *et al*. [4] propose Hybrid-IM to address the expensive simulations and significant computational overhead problems in SimpleGreedy by combining PBIM (Path Based Influence Maximization) and CBIM (Community Based Influence Maximization). Some scholars try to make their models more realistic. Jie and Jing [5] consider the cost differences of activating the individual among the seed set, and propose a new model called influence maximization-cost minimization (IM-CM) to obtain the maximum influence with minimum cost, or acceptable cost. Ali *et al*. [6] take time constraints into consideration and propose a novel nested Q-learning (NSQ) algorithm to reduce the complexity of exploring the large action space while maintaining the performance so as to maximize profits before the respective deadlines.

Angell and Schoenebeck [7] propose a heuristic dynamic approach on a hierarchical decomposition of the social

The associate editor coordinating the review of this manuscript and approving it for publication was Lu An.

network to leverage the relation between the spread of cascades and the community structure of social networks. Bozorgi *et al*. [8] propose an efficient algorithm for finding the influential nodes in a given social graph under the proposed propagation model which exploits the community structure. Ye *et al.* [9] propose two novel and robust community-based approximation algorithms, basic community-based robust influence maximization (BCRIM) and improved community-based robust influence maximization (ICRIM).

Although the influence maximization has been explored, the current proposals lack an efficient and accurate scheme to deal with the diffusion model. We notice that in previous work, they treat the nodes with same weight, and the influence possibility usually decided by the node's in-degree. Therefore, in this paper, we improve the traditional independent cascade model to realize the probability of differentiation between users. Meantime, based on the community structure of network, we propose a reverse reachable method based on random walk(RSRW) to select the high influential node from the community join to the candidate seed set, and then we use classical greedy algorithm to select the top-k influential node as seeds and output.

Compared to other community-based methods for influence maximization which pay attention to the influence spread in divided community, we take into account the influence of the nodes in the community, and the global influence of the nodes.

The main contributions of this paper are summarized as follows.

1) We propose a method for influence maximization in social network, which utilizing the community structure of social network and combine the benefit of greedy algorithm and heuristic algorithm.
2) We improve the traditional independent cascade model, Different weights are given to nodes according to their importance in social network. We also realize the probability of differentiation between users, making the influence probability between users can change with the influence of the two users and the number of interactions, making the impact process more realistic.
3) We propose a method callde RSRW to select the potential high influence nodes from the community or social network.
4) On the four real dataset, we compare the method we proposed with other excellent baseline methods, the experimental results show that the proposed method have good performance in influence spread with relative acceptable time.

The remainder of this paper is organized as follows. In Section II, we introduce the current research status of the influence maximization problem at home and abroad. The related technologies used in this paper is described in Section III. Ihe influence maximization algorithm proposed in this paper and our improvements on IC model are illustrated in Section IV. In Section V, the experiments conducted in this paper are introduced. Section VI conclues this paper.

II. RELATED WORKS

Borgs *et al*. [10] propose a polling process, randomly selecting a node, and then determining the set of nodes that can activate the node according to the selected propagation model. After repeating the process multiple times, it can be found that some nodes appear often; thus, these nodes can be considered a high-influence node, and the nodes can be used as alternative nodes for high-impact nodes. Yang *et al*. [11] propose an information maximization strategy based on online and offline double-layer communication schemes.

Aggarwal *et al*. [12] propose that social networks are highly dynamic, and the connections between the nodes may be established quickly and may disappear quickly. On this basis, a forward tracking method is proposed. Liqing *et al*. [13] analyze influence maximization problem in temporal social networks, and present a greedy-based GLAIC algorithm enhanced by Cost-Effective Lazy Forward optimization based on Latency Aware Independent Cascade model proposed in [14] to capture the dynamic aspects of real-world social networks. Wang *et al*. [15] study the IM problem over a social action stream. They define the influence between users in the sliding window model and propose the Stream Influence Maximization (SIM) query to continuously track a seed set maximizing the influence with the current window. Tong *et al*. [16] take the uncertainness in real-world social due to high speed data transmission into account and propose dynamic independent Cascade (DIC) model to capture the dynamic aspects of real-world social networks, and formally define the adaptive seeding strategy with the concept of seeding pattern. Azaouzi and Romdhane [17] believe that the way users influence social networks can be modeled as sharing, forwarding, commenting, etc. They divided these user behaviors, gave the behaviors different impact weights, and used the PageRank algorithm to calculate the influence of each behavior. Jaouadi and Romdhane [18] proposed an algorithm called DIN, which reasonably utilizes the social network structure and the semantics of information transport by the network.

Kim *et al*. [19] propose a pruning method based on random walk and rank merging, where the core idea is to find the nodes in the network that are not high-impact nodes. Jendoubi *et al*. [20] study the problem of maximizing influence on Twitter, therein calculating the influence of users through indirect functions. Lu *et al*. [21] proposed a new algorithm that enables greedy strategies to achieve better performance on large-scale social networks to influence maximization problems. Shang *et al*. [22] propose a community-based framework for large-scale social networks, CoFIM, which performs node expansion in the first phase and influence propagation in the community in the second phase. Hosseinipozveh *et al*. [23] incorporate distrust relationship into information diffusion model and propose two schemes

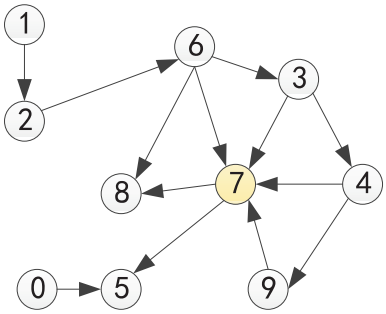

FIGURE 1. Degree diagram.

to find out how influence may propagate through distrust relationships.

He *et al*. [24]–[27] proposed a Two-stage Iterative Framework for the Influence Maximization in social networks(TIFIM) to ensure efficiency and accuracy of the proposed schemes at the same time, and a new Influence Power-based Opinion Framework (IPOF). They focus on the area of opinion formation and propose a new Interative 2-hop algorithm(SIR2) and 3-hop heuristic algorithm(RSPN3).

III. RELATED TECHNOLOGY

A. SOCIAL NETWORK STRUCTURE

1) NODE DEGREE

In graph theory, the degree is the indicator used to measure the importance of the node in the network. The degree of the node refers to the number of nodes directly connected to node i. In particular, in a directed graph, the node degree is divided into the out degree and the in degree.

As FIGURE 1 shows, there are four edges pointing to node 7, and two edges starting from node 7, so we can get $d_{in} = 4, d_{out} = 2, d = 6$.

2) NETWORK CLUSTERING COEFFICIENT

The clustering coefficient describes the characteristics of a graph (or network). In graph theory, the clustering coefficient is a coefficient used to describe the degree of clustering between the vertices of a graph. Specifically, it is the level to which the adjacent points of a point are connected to each other, for example, the degree to which your friends know each other on social networks. There is evidence that in various network structures reflecting the real world, especially in social network structures, nodes tend to form relatively dense network groups. In other words, real-world networks have higher clustering coefficients than networks that are randomly connected between two nodes.

In real social life and virtual social networks, the phenomenon of friend networks is very obvious, that is, in a person's social network, two of their friends may know each other. In social network research, the clustering coefficient can be used to measure the possibility that two friends are also friends of each other. The larger the clustering coefficient is, the closer the network of friends.

Denote a node in an undirected social network as u, and the degree d_u means that there are d_u nodes connected to node u in this network; therefore, there will be at most $\frac{d_u(d_u-1)}{2}$ edges between the d_u nodes. If the actual existence edge of the d_u nodes is A_u in the social network, the clustering coefficient C_u of node u can be obtained as follows:

$$C_u = \frac{A_u}{\frac{1}{2}d_u(d_u-1)} = \frac{2A_u}{d_u(d_u-1)} \tag{1}$$

B. INDEPENDENT CASCADE MODEL

Considering operational models for the spread of an idea or innovation through a social network G, we define two states of each individual node as being *active* or *inactive*, and the active node can affect the inactive neighbors to switch to the active state. It turns out that this assumption can easily be lifted later. Thus the process will look roughly as follows from the perspective of an initially inactive node v: as time unfolds, more and more of v's neighbors become active; at some point, this may cause v to become active, and v's decision may in turn trigger further decisions by nodes to which v is connected.

Independent Cascade Model is a kind of probability diffusion model, we start with an initial set of active nodes A_0, and the process unfolds in discrete steps according to the following randomized rule. When node u first becomes active in step t, it is given a single chance to activate each currently inactive neighbor v, it succeeds with a probability $p_{u,v}$ independently of the history thus far. if v has mutiple newly activated neighbors, their attempts are sequenced in an arbitrary order. If u succeed, then v will become active in step $t + 1$; but whether or not v succeeds, it cannot make any further attempts to activate v in subsequent rounds. Again, the process runs until no more activations are possible.

Here we give an example of independent cascade model. To simplify the process, we define the probability of influencing possibility as 1. The green node represents the active node, the red node represents the node that can active it's neighbors, and the orange node represents the node being activated during the current round. As FIGURE 2 shows: we start with active node 3, in phase (a), nodes 7,4 are affected and become active; In phase (b), nodes 7,4 try to influence their neighbors, and nodes 8,5,9 become active; In the same way, node 2 becomes active in phase (c); we can find in phase (d), there are no more activations are possible, so the diffusion process stops. And we get active nodes 2,3,4,5,7,8,9 in the end.

C. GREEDY ALGORITHM

Research on greedy algorithms is based on the hill climbing greedy algorithm. In each node selection, the node that can provide the maximum impact value is selected, and the global optimal is approximated by the local optimal solution. In greedy algorithm, there are two important concepts, namely, the marginal benefit and the submodule function.

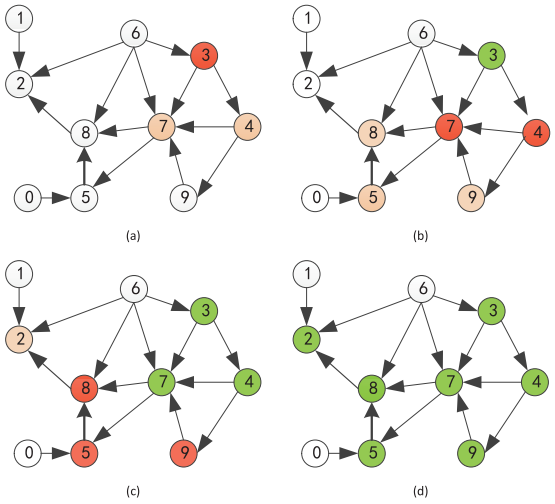

FIGURE 2. Indenpendent cascade model.

1) SUBMODULE FUNCTION

For an arbitrary function $f(\cdot)$ that maps a subset of the finite set U to a non-negative real number, if the function $f(\cdot)$ presents diminishing returns, the function $f(\cdot)$ is a submodule function. Diminishing returns means that the marginal benefit of adding any element v_i to the set S is not less than the marginal benefit of increasing the superset TâŁĞS of the element v_i to S, which can be described as follows:

$$\sigma_{v_i}(S) = \sigma(S \cup v_i) - \sigma(S) \tag{2}$$

2) MARGIN BENEFIT

In the problem of influence maximization, the marginal benefit of the influence value function $f(\cdot)$ refers to the increase in the final influence value that can be obtained by adding an additional node v_i as the initial active node based on the current active node S set, which can be described as follows:

$$\sigma_{v_i}(S) = \sigma(S \cup v_i) - \sigma(S) \tag{3}$$

D. COMMUNITY DETECTION ALGORITHM

1) LOUVAIN ALGORITHM

Considering that in real life and in virtual social networks, the user group will have a group structure, and by rationally utilizing the characteristics of the group structure, the computational complexity of the algorithm can be reduced.

We select the Louvain algorithm [28] as the community discovery algorithm for clustering to obtain a better community structure. The Louvain algorithm mainly uses the modularity Q as a measure of the community. Modularity:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{i,j} - \frac{k_i k_j}{2m}]\delta(c_i, c_j)$$

$$\delta(c_i, c_j) = \begin{cases} 0 & c_i, c_j \text{ belong to the same community} \\ 1 & c_i, c_j \text{ belong to different communities} \end{cases} \tag{4}$$

where m is the total number of edges in the graph, and k_i represents the sum of all the edge weights pointing to node I, and k_j is the same,$A_{i,j}$ represents the weight of the edge between node i and node j in the adjacency matrix of the network. When the network is a non-weighted graph, it can be regarded as the number of times that node i interacts with node j or directly as 1.

Module change:

$$\triangle Q = k_{i,in} - \frac{\Sigma_{tot} * k_i}{m} \tag{5}$$

where $k_{i,in}$ represents the incident of node i. The sum of the weights in the community is c, Σ_{tot} represents the total weight of the community to the community c, and k_i represents the weight of the membership to node i.

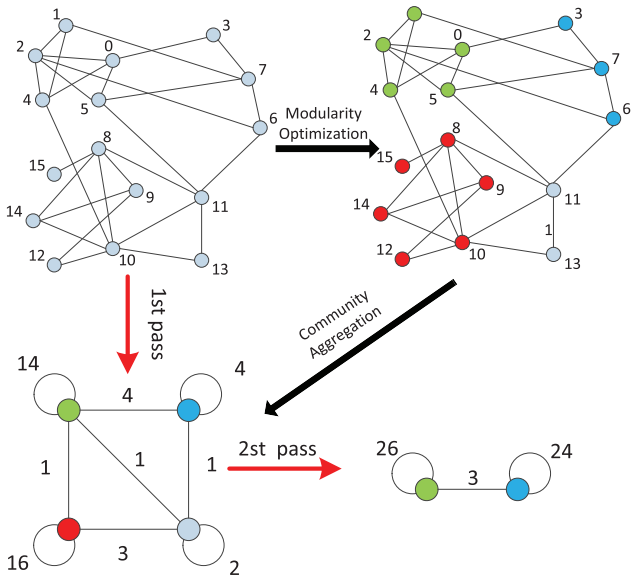

FIGURE 3. Louvain algorithm diagram.

The flow diagram of Louvain algorithm is shown in FIGURE 3. By using the Louvain algorithm, we can divide the social network into disjoint user communities.

Louvain algorithm is an algorithm based on multi-level optimized Modularity, which is fast and accurate. Louvain is considered one of the best community discovery algorithms for social networks of all sizes. The actual graph (700,000 points, 2 million edges) was tested and the time required for the iteration to complete the convergence was 1.77 seconds. So we choose Louvain algorithm to implement community detection.

IV. METHOD OF MAXIMIZING INFLUENCE BASED ON COMMUNITY DISCOVERY AND INDEPENDENT CASCADE MODEL

In this section, we introduce the method for influence maximization based on the community structure, as well as the reverse reachable index method based on random walk and the improved IC model.

A. REVERSE REACHABLE INDEXING METHOD

Inspired by Borgs *et al*. [10], we utilize the RR method to select potential high-influence nodes. Imagine the following

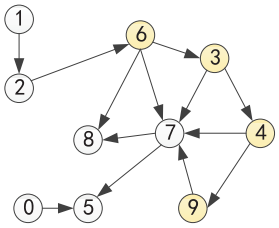

FIGURE 4. Reverse reachable method diagram.

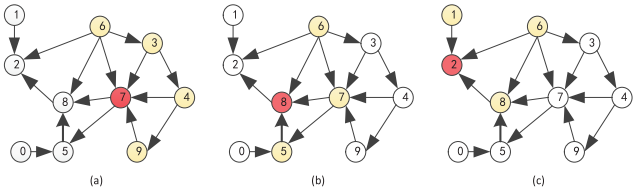

FIGURE 5. Reverse reachable method based on random walk diagram.

sampling strategy: randomly select some different nodes; then, according to the principle of reverse reach, for each node v, we select a node set. The nodes in the node set can affect node v and activate node v. According to this principle, if we repeat this process multiple times, we can obtain a sequence of the occurrences of the process nodes. In this process, the node with the highest occurrence rate is likely to be a high-influence node in the community.

As shown in Figure 4, if we randomly select node 7 as node v in a random selection, we can easily see that nodes 6, 3, 4, and 9 may affect node 7 and make its state change; thus, we record the four nodes, repeat the above process, and select the node with higher frequency as the candidate seed node at the end.

B. REVERSE REACHABLE INDEX METHOD BASED ON RANDOM WALK

First, we randomly select a node; then, we randomly select the next node according to the state of the current node and continuously iterate d times, where d is recorded as the maximum diameter of the current community. Each record in the process can affect the node whose node status changes; when the process ends, we select a new non-repeating node, repeat the above process, record the number of times that all nodes appear in the process in the community, and then select the node with the higher ranking, subsequently being added to the alternate seed node set.

As shown in FIGURE 5, in phase (a), node 7 is randomly selected; it can be seen that the nodes that can affect the state of node 7 are 3, 4, 6, and 9. Then, we randomly walk to node 8 and transition to phase (b). We easily find that the nodes can change the state of node 8 are 5, 6, and 7 and then we move to node 2 in phase (c), as can be observed. Those nodes are 1, 6, and 8. We observe that the number of occurrences of node 6 is the highest, that is, that node can affect the most nodes in the community and make their states change. If a high-impact node is selected in this community, we should give priority to node 6.

C. DIFFERENTIATION PROBABILITY

In most influence propagation models, when node u attempts to influence node v, the probability $p_{u,v}$ usually uses the reciprocal of the node u out-degree:

$$p_{u,v} = \frac{1}{d_u} \tag{6}$$

where d_u is the out-degree of node u.

Such influence probability only considers the degree of centrality of the node in the graph and uses the same influence probability for all neighbor nodes of the node; it does not consider the influence of the number of interactions between the nodes on the influence probability. Therefore, we improve the independent cascade model to realize the differentiation of the influence probability between any two nodes u and v to make it more in line with the actual situation.

1) ESTIMATION OF NODE INFLUENCE

In real social networks, some nodes have lower degrees of centrality but relatively high influence. Only considering the number of neighboring nodes of the node itself cannot fully represent the importance of the node. Therefore, based on the degree of the node, we introduce the network aggregation coefficient and integrate the importance of the node itself in the local area, the relative position of the node, and the node distribution of the node neighbor subgraph to estimate the potential impact of the node in the network.

As introduced in Section III, the parameter network aggregation coefficient C_i, which characterizes the attribute of the degree of interconnection between all adjacent points of a node in the graph in social networks. We can use C_i to describe the user's degree of mutual understanding between friends. For influence maximization, we can use it to describe the importance of the node in the process of influence propagation; when C_i is large, the connections of neighboring nodes of this node are relatively dense, that is, if there are nodes adjacent to the node that can replace its role in the propagation process, then the role of the node in the process of influence propagation is less important; when C_i is small, then the links between the adjacent points of the node are relatively sparse, indicating that the role played by the node propagation is relatively important. Then, the estimation formula of the influence potential of the node is defined as follows:

$$P_i = d_i + \sum_{j \in N_i} d_j(1 - C_j) \tag{7}$$

where P_i is the potential of the node, d_i is the degree of the node, N_i is the neighbor of node i, d_j is the degree of the node, and C_j is the aggregation coefficient of the node, that is, the degree of the node itself and the neighboring point of the node.

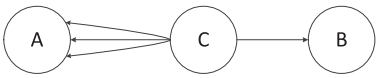

FIGURE 6. Diagram of multiple interactions.

2) THE EFFECT OF USER INTERACTION

Aggarwal *et al*. [12] proposed that the influence probability between nodes should gradually increase with the number of interactions between two nodes. In a social network, we have a similar situation: if node i interacts with node j frequently, then when node i transmits information to node j, node j has a relatively high probability of receiving the information. Based on this, we define the impact of the number of interactions between nodes on the propagation probability as follows:

$$f_{i,j}(\delta) = 1 - e^{-\lambda_{i,j}\delta_t} \tag{8}$$

where δ_t represents the number of interactions between nodes i and j and $\lambda_{i,j}$ controls the transmission rate between nodes.

As shown in FIGURE 6, it can be seen from the above figure that it is assumed that the probability of node C affecting node A and node B is the same, node C interacts with the node A three times, and interacts with the node B once. Here, we will set $\lambda_{i,j}$ to 0.5 uniformly.

$$f_{C,A}(\delta) = 1 - e^{-\lambda_{C,A}\delta_t} = 1 - e^{-0.5*3} = 0.77687$$
$$f_{C,B}(\delta) = 1 - e^{-\lambda_{C,B}\delta_t} = 1 - e^{-0.5*1} = 0.39347$$

As can be seen, the node C is more likely to affect node A than node B, this is more in line with the actual propagation law.

3) DIFFERENTIAL INFLUENCE PROBABILITY

Based on the above two sections, we comprehensively consider the influence of the node influence potential and the number of interactions between nodes on the propagation probability, thereby improving the independent cascade model. The influence probability p of node u on node v is modified from the traditional probability to

$$p_{u,v} = \frac{P_u}{\sum_{j \in N_u} P_j} * (1 - e^{-\lambda_{u,v}\delta_t}) \tag{9}$$

where $p_{u,v}$ is the influence possibility, P_v P_j represents the influence estimation of the node v and node j, N_u represents the neighboring node of the nodes, $\lambda_(u, v)$ represents the transmission rate control parameter between nodes u and v, in our paper, we set $\lambda u, v$ as 0.5 uniformly, and δ_t represents the number of interactions between node u and v.

As shown in FIGURE 7, the degrees of node 1 and node 3 are 8 and 6, respectively. If the node importance is determined by the node degree, it is obvious that node 1 is more important than node 3. However, from the perspective of influence propagation, if node 1 is selected as the seed node for influence propagation, it is not necessarily able to propagate the influence to the entire network. If node 3 with a relatively low degree is selected, a higher probability will spread the influence to the entire social network.

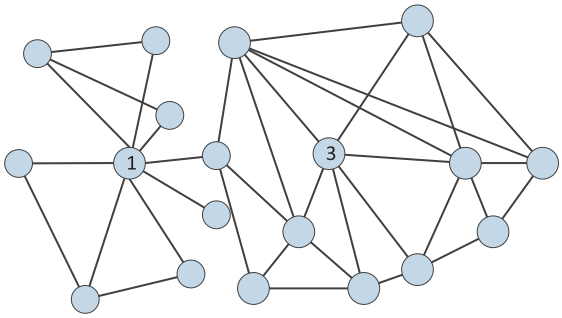

FIGURE 7. Example of improved independent cascade model.

Algorithm 1: Estimate the Node Influence

Input: Social network $G(V, E)$, in-degree matrix D, clustering coefficient matrix C
Output: Node estimate influence matrix I

Initial influence matrix P;
foreach *node* $i \in V$ **do**
  record the clustering coefficient C_i ;
  record the in-degree d_i;
end
foreach *node* $i \in V$ **do**
  find u's neighbors $neighbors_u$;
  foreach *node* $j \in neighbors_u$ **do**
    $N_u \leftarrow N_u + d_j * (1 - C_j)$
  end
  $I_u \leftarrow d_u + N_u$
end
Output the influence matrix I

According to the influence estimation used in this paper, the estimated influence of node 1 is 20.33, and the estimated influence of node 3 is 25.78, which is consistent with our calculation and also shows the validity of the influence estimation.

4) IMPROVEMENTS TO TRADITIONAL IC MODEL

Our improvements to the traditional IC model can be divide into two phases. First we calculate the estimate influence, the process is shown as Algorithm 1. Then we can get the differential influence possibility for two nodes, shown as Algorithm 2. Through the process, we extend the influence probability in the traditional independent cascade model method to the UADIC model (User Attribute Dependent independent cascade model) by combining the user influence and user interaction times.

D. SCREENING OF CANDIDATE SEED SETS

Through the above steps, we obtain the candidate seed set. Then, based on the improved independent cascade model and

Algorithm 2: Get Node Influence Possibility

Input: Social network $G(V, E)$, Influence matrix I, Interaction matrix R, node u,v

Output: influence possibility $p_{u,v}$

find u's neighbors $neighbors_u$;
foreach *node* $j \in neighbors_u$ **do**
  $NI_u \leftarrow NI_u + I_u$
end
foreach *node* $u, v \in neighbors_u$ **do**
  $p_{u,v} \leftarrow \frac{I_u}{NI_u} * (1 - e^{-0.5 * R_{u,v}})$
end
Output the influece possibility $p_{u,v}$

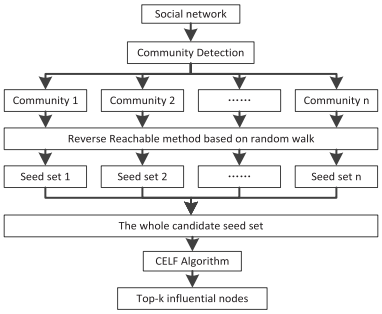

FIGURE 8. Algorithm overall flowchart.

classical greedy algorithm, the seed set is selected from the candidate seed set as output.

The overall flow chart of our method is shown in FIGURE 8.

E. TIME COMPLEXITY ANALYSIS

In this part, we analysis the time complexity of CBIM-RSRW we proposed. From FIGURE 8, we can see that, the algorithm mainly divided into three part:(i) the community detection part; (ii)select the high influential nodes in community by RSRW; (iii) use celf algorithm to select top-k nodes from candidate seed set.

In part(i), louvain algorithm mainly contains two-phase, the time Complexity of each iteration phase 1 is $O(n)$, and the second phase is $O(M + n)$, we can find the complexity is smaller than part(iii) $O(KcRm)$. In part(ii), the time complexity denpens on the nodes and edges in community, and it is much smaller than the number of nodes and edges of social network. So the time complexity of the whole process is $O(KcRm)$. wher m, n represent the number of nodes and edges in socail network, R represnet the Monte carlo simulation number, M represent the number of nodes in this iteration, c represent the all candidate seeds set and c is far less than n.

In summary, we believe the time complexity of CBIM-RSRW $O(KcRm)$ is far less than CELF $O(KnRm)$, and we can realize the goal of reducing the time taken to find the high influential nodes while ensuing that the selected seed nodes has relative high influence.

TABLE 1. Dataset description.

Dataset	Digg	Email	Facebook	Wiki
Nodes	8193	36692	4039	7115
Edges	56440	183831	88234	103689
Average degree	13.77	20.04	43.69	29.15
Average clustering	0.12	0.50	0.61	0.14

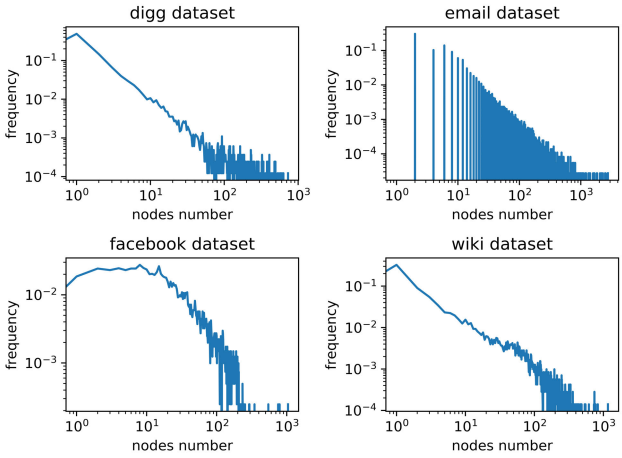

FIGURE 9. Dataset degree distribution histogram.

V. EXPERIMENTS

A. DATASET

We use four publicly available real dataset for experiments: Digg,Email,Facebook,Wikipedia. Digg dataset [29], [30] is a subset of data scrapped from Digg by Munmun De Choudnury during January 2009. Email dataset [31] covers all the email communication within a dataset of around half million emails, which posted by the Federal Energy Regulatory Commission during its investigation. Facebook dataset consists of 'circles' (or 'friends lists') from Facebook. Facebook dataset [32] was collected from survey participants using this Facebook app. The dataset includes node features (profiles), circles, and ego networks. Wikipedia dataset [33] use the latest complete dump of Wikipedia page edit history (from January 3 2008), and extracte all administrator elections and vote history data. The detail of the four datasets is shown as TABLE 1.

FIGURE 9 shows the node degree distribution of the four datasets, we can see that the four data sets of low-grade node density is higher, and the height of the node is relatively small, including Digg data set and the Wikipedia data set of low-grade nodes is more, Email the distribution of the data set is even, Facebook data set relative height of node more some, four types of data sets, respectively, represents the different data distribution, embodies the comprehensive nature of the selected data set.

B. BASELINE METHOD

We compare the CBIM-RSRW method with the following methods:

Random: Randomly select K nodes as the seed node set. This is one of the commonly used method to select the seed node set for comparing the general influence maximization problem.

PageRank: Google's left-hand ranking, also known as page ranking, is a technique that search engines calculate based on hyperlinks between web pages and is one of the elements of page rankings that are used to influence maximization issues.

CGA: Community-based Greedy algorithm for mining top-K influential nodes, proposed by Wang *et al*. [34], which providing an algorithm for detecting communities in a social network by taking into account information diffusion and a dynamic programming algorithm for selecting communities to find influential nodes.

CELF: The algorithm proposed by Leskovec *et al*. [35], which optimizing the greedy algorithm, greatly reduces the calculation amount and improves the calculation efficiency by 700 times.

CoFIM: *CoFIM* is a community-based framework to address the influence maximization. The overall influence is calculated by $g(S) = N(S)+\gamma|NC(S)|$, where γ a parameter, $N(S)$ is the neighbor set of S, and $NC(S)$ is the neighbor community set of S.

C. EVALUATION CRITERION

In this section, we use the methods described in Section IV to apply the node discovery algorithm to the initial dataset and community-discovered communities for finding potential high-impact nodes and making experimental comparisons at the end of this section. We compare two criterion: influence spread and running time.

In the influence maximization problem, when we compare algorithms, the commonly used evaluation indicators are the influence spread and time complexity. The evaluation indicators involved are defined as follows:

Influence spread: The range of propagation, the number of user nodes activated by the seed node simulated by the propagation model. The larger this value is, the better the algorithm.

Running time: The algorithm execution time. The smaller the value is, the higher the efficiency of the algorithm.

D. EXPERIMENTS

In this section, we set up two sets of experiments, one for testing the proportion of nodes randomly selected in the community and the other for comparing the performances of the algorithms against the proposed method. In order to get more accurate results, we set the number of Monte Carlo simulations to 10,000 in our experiments.

1) INITIAL NODE RATIO EXPERIMENT

To choose the number of initial nodes selected in the community, we decide which percentage of the initial nodes should be selected. We selected 5, 10, 20, 30, 40, and 50 seed nodes as tests and 5%, 10%, 20%, 30%, 40%, and 50% initial ratios of the community nodes for the random walks.

In this experiment, the step size of the random walks was set to 200 to ensure that each random walk could walk to as many nodes as possible to reduce the impact of randomness on the experimental results. Here the measure of time is seconds(s).

The algorithm is outlined in Algorithm 3. It first detects communities(line 1),we set the threshold(line 2),then the algorithm iteratively operate the community(line 4-20). Initial the candidate set(line 4), check the nodes number of the community to filter out small communities, when the number of community nodes is less than 1% of the network nodes, the community is skipped(line 4) and seed nodes were screened according to the proportion of community nodes in the network nodes (lines 5). In lines 10-16, the algorithm use the method RSRW we proposed to select potential high impact nodes. In line 19, add the high impact nodes into the candidate seeds set. In lines 20, we simulate the influence propagation process, statistic the influence spread, running time and output the result.

Algorithm 3: Initial Node Ratio Experiment

Input: social network $G(V, E)$, seed node number sequence S, Initial node sequence I, threshold ratio γ
Output: sees, influence spread and running time

```
1  Use the Community detection algorithm obtain the
   divided community Coms;
2  Determine threshold: θ = N * γ;
3  foreach I_m ∈ I do
4      Initial candidate seeds set seeds;
5      foreach community com_i ∈ Coms do
6          if N_i < θ then
7              continue;
8          end
9          Determine the node number to select:
           num_i = (N_i / N) * 2 * S_i;
10         foreach node i ∈ com_i do
11             Random select 10% nodes as initial for
               j = 1 to step do
12                 Record neighbors nodes can active node
                   i;
13                 Random walk to the next node;
14                 j ← j + 1
15             end
16         end
17         Choose the top num_i nodes with the most most
           occurrences join to seeds;
18     end
19     Use CELF select top-k nodes from seeds with the
       improved IC model;
20     Output seeds, influence spread and running time;
21 end
```

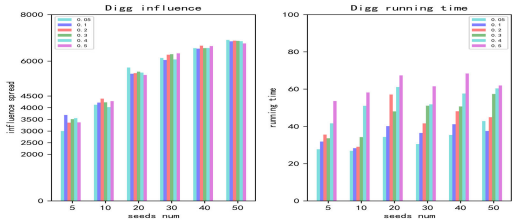

FIGURE 10. Digg dataset experimemt.

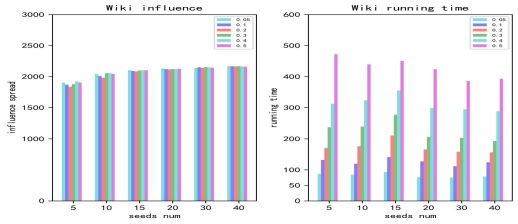

FIGURE 11. Wikipedia dataset experiment.

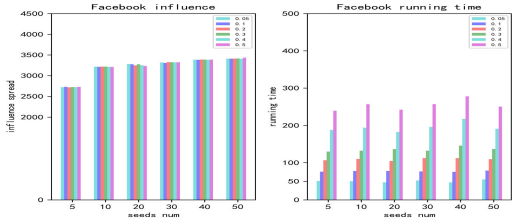

FIGURE 12. Facebook dataset experiment.

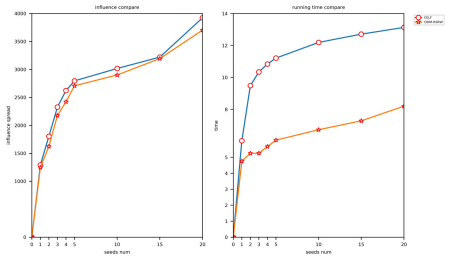

FIGURE 13. Compare with CELF algorithm.

The experimental results for the initial node selection ratio of the community are shown in FIGURE 10, FIGURE 11 and FIGURE 12. It can be easily found that the larger the initial ratio, the bigger the running time, and when we choose 10%, we can usually get better performance.

Therefore, considering the impact of the number of nodes ultimately affected and the time spent in the experiment, in subsequent experiments, we randomly selected 10% of the community nodes as the initial number of random walks.

2) COMPARATIVE EXPERIMENTS

Due to the *CELF* algorithm belong to the greedy algorithm which takes a lot of time to calculate, we choose Digg Dataset to demonstrate the efficiency of the algorithm, and we choose seed nodes from 1 to 20 to prove it, as FIGURE 13 shows.

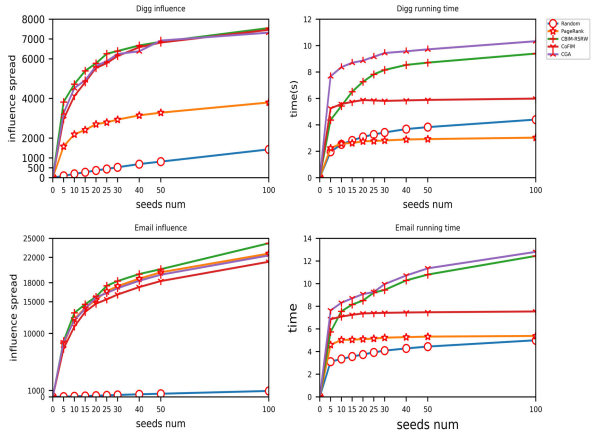

FIGURE 14. Digg dataset and Email dataset.

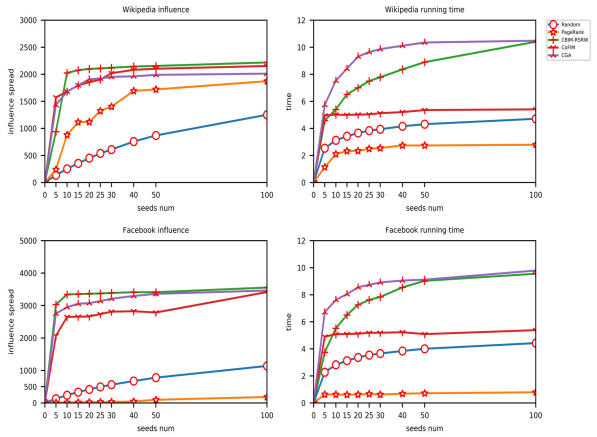

FIGURE 15. Wikipedia dataset and facebook dataset.

Compared with the other two algorithms, we selected 5, 10, 15, 20, 25, 30, 40, 50, and 100 seed nodes for the influence diffusion experiments. Here, because the time difference between the algorithms is large, we use the logarithmic value to represent the ordinate in the time-consumption comparison experiment.

As FIGURE 15 shows, compared to *CELF* algorithm, our method is relatively low in final impact, but our algorithm is faster than *CELF* algorithm. FIGURE 13, FIGURE 14 show the experiment results on four data sets: Digg, Email, Wikipedia, Facebook. We can see that, the method *CBIM* − *RSRW* has a good performance in four dataset. It can get relatively high influence spread, and affect a relatively large number of user nodes. Although compared to the two heuristic algorithm: *Random* and *Pagerank*, and the algorithm *CoFIM* it takes more time to calculate the influential nodes, compared to algorithm *CELF* and *CGA*, *CBIM* − *RSRW* can improve the running speed of the algorithm on the premise of ensuring the accuracy of the algorithm.

In summary, the *CBIM* − *RSRW* influence maximization method proposed in this paper can select nodes with high influence in relatively short time, and obtain a relatively

higher influence compared to the traditional heuristic algorithms. And compared to the greedy algorithm, the time taken to find the seed node is reduced while also ensuing that the selected node has a high influence.

VI. CONCLUSION

In this paper, we study the influence maximization problem in social network. We propose a method called the reverse reachable index method based on random walk(*RSRW*) to select the potential high influence node. By utilizing the community structure, we accelerate the speed of the influence maximization algorithm. At the same time, we improve the traditional independent cascade model to realize the goal of different weights for different users, and the differentiation of influence probability. The experimental result on four datasets shows *CBIM − RSRW* can reduce the time taken to find the high influential seed node while also ensuing that the selected node has a high influence.

For future works, We consider applying the *RSRW* method to the rapidly changing dynamic social network, so that we can obtain nodes with high influence in a short time. Meantime, we will take into account our method to the actual application, such as opinion formation and recommendation system.

APPENDIX

We provide our code in this part, Github's address is https://github.com/BaoACheng/CBIM-RSRW.

REFERENCES

[1] P. M. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2001, pp. 57–66.
[2] E. Tardos, D. Kempe, and J. Kleinberg, "Maximizing the spread of influence in a social network," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 137–146.
[3] L. Cui, H. Hu, S. Yu, Q. Yan, Z. Ming, Z. Wen, and N. Lu, "DDSE: A novel evolutionary algorithm based on degree-descending search strategy for influence maximization in social networks," *J. Netw. Comput. Appl.*, vol. 103, Feb. 2018, Art. no. 119130.
[4] Y.-Y. Ko, K.-J. Cho, and S.-W. Kim, "Efficient and effective influence maximization in social networks: A hybrid-approach," *Inf. Sci.*, vol. 465, pp. 144–161, Oct. 2018.
[5] Y. Jie and L. Jing, "Influence maximization-cost minimization in social networks based on a multiobjective discrete particle swarm optimization algorithm," *IEEE Access*, vol. 6, pp. 2320–2329, 2017.
[6] K. Ali, C.-Y. Wang, and Y.-S. Chen, "A novel nested q-learning method to tackle time-constrained competitive influence maximization," *IEEE Access*, vol. 7, pp. 6337–6352, 2019.
[7] R. Angell and G. Schoenebeck, "Don't be greedy: Leveraging community structure to find high quality seed sets for influence maximization," in *Proc. Int. Conf. Web Internet Econ.*, 2017, pp. 16–29.
[8] A. Bozorgi, S. Samet, J. Kwisthout, and T. Wareham, "Community-based influence maximization in social networks under a competitive linear threshold model," *Knowl.-Based Syst.*, vol. 134, pp. 149–158, Oct. 2017.
[9] F. Ye, J. Liu, C. Chen, G. Ling, Z. Zheng, and Y. Zhou, "Identifying influential individuals on large-scale social networks: A community based approach," *IEEE Access*, vol. 6, pp. 47240–47257, 2018.
[10] C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," 2012, *arXiv:1212.0884*. [Online]. Available: https://arxiv.org/abs/1212.0884
[11] Y. Yang, Y. Xu, E. Wang, K. Lou, and D. Luan, "Exploring influence maximization in online and offline double-layer propagation scheme," *Inf. Sci.*, vol. 450, pp. 182–199, Jun. 2018.
[12] C. C. Aggarwal, S. Lin, and P. S. Yu, "On influential node discovery in dynamic social networks," in *Proc. SIAM Int. Conf. Data Mining*, 2012, pp. 636–647.
[13] Q. Liqing, Y. Jinfeng, F. Xin, J. Wei, and G. Wenwen, "Analysis of influence maximization in temporal social networks," *IEEE Access*, vol. 7, pp. 42052–42062, 2019.
[14] B. Liu, G. Cong, Y. Zeng, D. Xu, and Y. M. Chee, "Influence spreading path and its application to the time constrained social influence maximization problem and beyond," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1904–1917, Aug. 2014.
[15] Y. Wang, Q. Fan, Y. Li, and K.-L. Tan, "Real-time influence maximization on dynamic social streams," *Proc. VLDB Endowment*, vol. 10, no. 7, pp. 805–816, 2017.
[16] G. Tong, W. Wu, S. Tang, and D. Du, "Adaptive influence maximization in dynamic social networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 112–125, Feb. 2017.
[17] M. Azaouzi and L. B. Romdhane, "An efficient two-phase model for computing influential nodes in social networks using social actions," *J. Comput. Sci. Technol.*, vol. 33, no. 2, pp. 286–304, Mar. 2018.
[18] M. Jaouadi and L. B. Romdhane, "DIN: An efficient algorithm for detecting influential nodes in social graphs using network structure and attributes," in *Proc. EEE/ACS 13th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2016, pp. 1–8.
[19] S. Kim, D. Kim, J. Oh, J.-H. Hwang, W.-S. Han, W. Chen, and H. Yu, "Scalable and parallelizable influence maximization with random walk ranking and rank merge pruning," *Inf. Sci.*, vols. 415–416, pp. 171–189, Nov. 2017.
[20] S. Jendoubi, A. Martin, L. Liétard, H. B. Hadji, and B. B. Yaghlane, "Two evidential data based models for influence maximization in Twitter," *Knowl. Based Syst.*, vol. 121, pp. 58–70, Apr. 2017.
[21] W.-X. Lu, C. Zhou, and J. Wu, "Big social network influence maximization via recursively estimating influence spread," *Knowl.-Based Syst.*, vol. 113, pp. 143–154, Dec. 2016.
[22] J. Shang, S. Zhou, L. Xin, L. Liu, and H. Wu, "CoFIM: A community-based framework for influence maximization on large-scale networks," *Knowl.-Based Syst.*, vol. 117, pp. 88–100, Feb. 2017.
[23] M. Hosseini-Pozveh, K. Zamanifar, and A. Naghsh-Nilchi, "Assessing information diffusion models for influence maximization in signed social networks," *Expert Syst. Appl.*, vol. 119, pp. 476–490, Apr. 2019.
[24] Q. He, X. Wang, Z. Lei, M. Huang, Y. Cai, and L. Ma, "TIFIM: A two-stage iterative framework for influence maximization in social networks," *Appl. Math. Comput.*, vol. 354, pp. 338–352, Aug. 2019.
[25] Q. He, X. Wang, M. Huang, J. Lv, and L. Ma, "Heuristics-based influence maximization for opinion formation in social networks," *Appl. Soft Comput.*, vol. 66, pp. 360–369, May 2018.
[26] Q. He, X. Wang, B. Yi, F. Mao, Y. Cai, and M. Huang, "Opinion maximization through unknown influence power in social networks under weighted voter model," *IEEE Syst. J.*, to be published.
[27] Q. He, X. Wang, C. Zhang, M. Huang, and Y. Zhao, "Iimof: An iterative framework to settle influence maximization for opinion formation in social networks," *IEEE Access*, vol. 6, pp. 49654–49663, 2018.
[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, no. 10, pp. 155–168, 2008.
[29] T. Jie, J. Sun, W. Chi, and Y. Zi, "Social influence analysis in large-scale networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2009, pp. 807–816.
[30] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2010, pp. 199–208.
[31] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Math.*, vol. 6, no. 1, pp. 29–123, 2009.
[32] J. Leskovec and J. Mcauley, "Learning to discover social circles in ego networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 539–547.
[33] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. 19th Int. Conf. World Wide Web*, Apr. 2010, pp. 641–650.
[34] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-K influential nodes in mobile social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 1039–1048.

[35] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2007, pp. 420–429.

FENG CAI is currently pursuing the master's degree with the Information Engineering Department, Minzu University of China. His research interests include data mining and big data analysis.

LIRONG QIU received the Ph.D. degree in computer sciences from the Chinese Academy of Science, in 2007. She is currently an Associate Professor of computer sciences with the Information Engineering Department, Minzu University of China. Her current research interests include different aspects of natural language processing, artificial intelligence, and distributed systems.

XINKAI KUAI is currently a Senior Engineer with the School of Information Engineering, Minzu University of China. His current research interests include image processing, natural language processing, and other related fields.

HONGSHUAI ZHAO received the bachelor's degree from the Minzu University of China, in 1998. He is currently with the Information Engineering Department, Minzu University of China. His research interests include artificial intelligence and big data processing. He awarded Senior Experimenter, in 2009.