IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 26, 2019, accepted January 9, 2020, date of publication January 13, 2020, date of current version January 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2966056

TSIM: A Two-Stage Selection Algorithm for Influence Maximization in Social Networks

QIU LIQING, GU CHUNMEI, ZHANG SHUANG, TIAN XIANGBO, AND ZHANG MINGJV
Shandong Province Key Laboratory of Wisdom Mine Information Technology, College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding author: Qiu Liqing (liqingqiu2005@126.com)

This work was supported by the Nature Science Foundation of China under Grant 61502281 and Grant 71772107.

ABSTRACT The influence maximization problem is aimed at finding a small subset of nodes in a social./network to maximize the expected number of nodes influenced by these nodes. Influence maximization plays an important role in viral marketing and information diffusion. However, some existing algorithms for influence maximization in social networks perform badly in either efficiency or accuracy. In this paper, we put forward an efficient algorithm, called a two-stage selection for influence maximization in social networks (TSIM). Moreover, a discount-degree descending technology and lazy-forward technology are proposed, called DDLF, to select a certain number of influential nodes as candidate nodes. Firstly, we utilize the strategy to select a certain number of nodes as candidate nodes. Secondly, this paper proposes the maximum influence value function to estimate the marginal influence of each candidate node. Finally, we select seed nodes from candidate nodes according to their maximum influence value. The experimental results on six real-world social networks show that the proposed algorithm outperforms other contrast algorithms while considering accuracy and efficiency comprehensively.

INDEX TERMS Social networks, influence maximization, DDLF, heuristic method, TSIM.

I. INTRODUCTION

With the development and popularity of the Internet, billions of people are connected through online social networks, such as Facebook, Twitter and YouTube. Due to connecting billions of people, social networks generate tons of data every day. The generation of massive data promotes the research of social networks. Social networks are not only communication channels but also platforms for information propagation, public services and marketing [1]–[3].

In recent years, with the popularity of social networks, influence maximization problem has become one of the hot issues in this field [27]. The research of influence maximization problem stems from ‘‘viral marketing’’ [4]–[6]. Its initial purpose is to obtain the maximum commercial value and market return through ‘‘word-of-mouth’’ with the minimum marketing cost [7], [28], [29].

Driven by the wide applications in marketing, influence maximization was first regarded as an algorithmic problem by Domingos and Richardson [8]. Kempe *et al.* [9], [10] defined the influence maximization problem as a discrete optimization problem and proved that the problem is NP-hard. Moreover, they proposed a classical greedy algorithm to solve the problem and testified that the algorithm could guarantee $(1 - 1/e)$ to reach the optimal solution. However, the greedy algorithm has two obvious drawbacks as follows: (1) it needs to traverse all nodes in social networks; (2) it requires tens of thousands of Monte Carlo simulations to obtain an accurate result. Due to these limitations, the efficiency of the algorithm is very low, especially for the large social networks containing tons of nodes. In recent years, numerous studies have been done to optimize the efficiency of the naive greedy algorithm. Leskovec *et al.* [11] proposed an optimization method of the greedy algorithm, called Cost-effective Lazy Forward (CELF). The CELF algorithm utilizes the sub-modular property of the influence maximization objective function to greatly reduce the number of assessments on the influence spread of nodes. Therefore, the efficiency of CELF is 700 times faster than the naive greedy algorithm. Moreover, the algorithm needs to repeatedly calculate the marginal influence spread of each candidate node in the node selecting process. Therefore, in large social networks, the efficiency of CELF is also poor. Based on the greedy-based algorithms, aiming at solving the low efficiency

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

of the algorithm, the subsequent heuristic algorithms were presented. Chen *et al.* [12] proposed a representative heuristic algorithm, called Degree Discount algorithm. The main idea of the Degree Discount algorithm is that if a node is selected as a seed node, the degree of the node's neighbors should be discounted. The efficiency of the Degree Discount algorithm is better than the naive greedy algorithm. However, the algorithm only considers the relationship between the degree of nodes and neighbors, which ignores the structure of social networks. As a result, the accuracy of the algorithm is lower than other greedy algorithms.

In this paper, we propose an improved algorithm, called TSIM (a two-stage selection algorithm for influence maximization in social networks). The TSIM algorithm focuses on two aspects: Firstly, the TSIM algorithm utilizes the discount-degree descending technology and lazy-forward technology (called DDLF) search strategy to select $2k$ (where k represents the size of the seed set S) influential nodes as candidate nodes. After finishing the two-stage filtering of candidate nodes, k nodes can be selected as seed nodes from candidate nodes by the maximum influence value (MIV) function proposed in this paper. The double selection of candidate nodes ensures the influence of the final selection node.

Our contributions are summarized as follows:

(1) The DDLF strategy utilizes the discount-degree descending technology to select a part of candidate nodes and the lazy-forward technology to select another part of candidate nodes excluding the influence of the nodes.

(2) A new objective function is presented in this paper, called MIV, to select seed nodes, which improves the accuracy of the TSIM algorithm. The TSIM algorithm is based on the DDLF strategy for influence maximization in the social network. Moreover, the proposed algorithm first utilizes DDLF to select $2k$ candidate nodes. Then, the algorithm uses MIV to select k nodes as seed nodes from $2k$ candidate nodes.

(3) Extensive experimental results on six real-world social networks demonstrate that the proposed algorithm outperforms contrast algorithms when considering comprehensively efficiency and accuracy.

The structure of the paper is as follows. Section II recommends the related work for the influence maximization problem. Section III introduces the preliminaries about the diffusion model and problem definition. Section IV presents the main idea of the TSIM algorithm. Section V validates TSIM algorithm is more accurate than other algorithms through the experiment. Section VI concludes this paper.

II. RELATED WORKS

In 2001, Domingos and Richardson [8] investigated the influence maximization problem and defined it as an algorithmic problem. Since then, in 2003, Kempe *et al.* [9] studied the influence maximization problem based on two influence diffusion models, the Independent Cascade model [10], [13] and the Linear Threshold model [14], [15], which described how users of social networks spread their effects to their friends. Meanwhile, they proposed a greedy algorithm to solve the problem. Their experimental results show that the greedy algorithm is better than the degree heuristic algorithm and the centrality heuristic algorithm in terms of accuracy. However, the greedy algorithm needs to spend long time on the modern server machine, which cannot be widely used in large-scale social networks. In [11], [16], researchers presented lazy-forward optimization that significantly speeds up the greedy algorithm, but it still cannot scale to large networks with thousands of nodes and edges [17], [18]. In 2007, Leskovec *et al.* [11] proposed the CELF algorithm, which utilizes the sub-modularity to reduce the number of Monte Carlo simulations in node-selecting process. In 2009, Chen *et al.* [12] proposed a heuristic algorithm, called Degree Discount algorithm. The idea of Degree Discount algorithm is that if a neighbor node of a node is selected as the initial active node, the degree of the node needs to be discounted quantitatively. Then in 2010, the new heuristic algorithm, called PMIA [19], was presented by Chen et al. PMIA improves computation efficiency and result accuracy by utilizing maximum influence arborescence (MIA) model. PMIA is an algorithm based on the influence path, which adopts the MIA model to estimate the influence spread of nodes. In this process, it uses a threshold to prune unnecessary traveling, which speed up the node-selecting process. This algorithm needs to set the threshold factitiously. There is no uniform calculating method for the threshold. In different graphs, the value of the threshold should also be different. Hence, the method of setting the threshold factitiously may decrease the result accuracy. Moreover, in 2018, Cui *et al.* [20] posed the DDSE algorithm (degree-descending search evolution). DDSE is an evolutionary algorithm based on the degree descending search strategy, which is divided into four steps: initialization, mutation, crossover and selection. By simulating the biological evolution process, this algorithm can obtain the global optimal solution when the iteration times is enough. Although this algorithm can obtain the global optimal solution, it runs too slow due to the defect of the evolutionary algorithm. In addition, it may fall into the local optimal solution. And this algorithm utilizes expected diffusion value (EDV) as the evaluating standard, which saves running time, but sacrifices a lot of accuracy.

We gain the experience from the above algorithm and propose an improved algorithm, called TSIM, which combines the discount-degree technology with lazy-forward technology to optimize the problem.

III. PRELIMINARY

In this section, we first briefly introduce the notations used in this paper. The Independent Cascade (IC) model [10], [13] is selected as the influence diffusion model. Then, the definition of the influence maximization problem is formulated.

Given a graph $G = (V, E)$ and a number k, G represents a directed graph for a social network and k denotes the size of a seed set. In the graph G, V and E denote the nodes set and

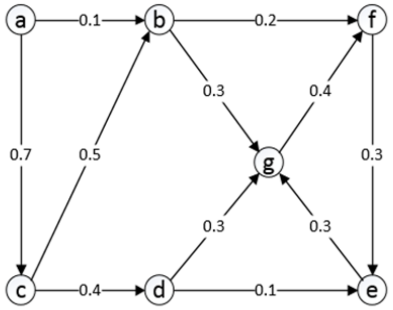


FIGURE 1. The diffusion process of the Independent Cascade model.

the edges set, respectively. Meanwhile, $p_{uv} \in [0, 1]$ expresses an active probability for each edge $(u, v) \in E$. A node in a social network G has two states: active or inactive. An active node u conveys an idea or innovation to its neighbor node v. If the node v accepts it, the node v is activated. Otherwise, the node v is inactive, and the node u can never attempt to activate the node v again. The process of interaction between nodes expresses the information propagation. To describe the interaction process and behavior pattern of nodes, we need a diffusion model.

A. THE INDEPENDENT CASCADE MODEL

In marketing research, Goldenberg *et al.* [21], [22] first put forward the Independent Cascade model as a diffusion model. The model was first proposed to solve the influence maximization problem by Kempe *et al.* [9]. In this paper, the Independent Cascade model is adopted to simulate the process of influence diffusion. Next, the propagation process of the Independent Cascade model is introduced as follows:

(1) Assume that the initial seed set S;
(2) At the time $t(t \geq 1)$, the active node u in the seed set S activates the inactive neighbor node v with probability p;
(3) If the node v is activated by u, at the time $t+1$, the node v becomes active and does the way as (1) to influence other neighbor inactive nodes of it. Otherwise, the condition of v will not change;
(4) The process of (1) and (2) will be repeated over and over again until there are no influential active nodes in the social network, the propagation process ends.

The diffusion process of the Independent Cascade model is shown in Fig. 1.

As shown in Fig. 1, the set of all nodes in this graph is $\{a, b, c, d, e, f, g\}$, and the weight of edges represents the node activation probability. At the time t, assuming that node a is the initial active node. At the time $t + 1$, node a tries to activate node b and c with the activation probability shown in Fig. 1. Supposing that node b is activated, node b activates its neighbor nodes f and g in the same way as node a. If node f is to be activated successfully, and then node f will attempt to activate its neighbor node e. If the activation of node f fails, node f has no chance to activate its neighbor nodes. The diffusion process will stop when no more new nodes are activated.

TABLE 1. Parameters of TSIM algorithm.

Parameters	Descriptions
k	number of seed nodes to be selected
n	number of nodes in G
m	number of edges in G
R	number of rounds of simulations in CELF
n'	number of nodes in G_{new}
m'	number of edges in G_{new}

B. PROBLEM STATEMENT

Influence maximization problem [23]: the purpose of the influence maximization problem is to find k nodes to maximize influence spread in the social network.

Given a social network $G = (V, E)$ and a number k, S is a seed set and $\sigma(S)$ denotes the influence spread of S in the Independent Cascade model in this paper.

Definition 1: $\sigma(S)$ is defined as the number of expected nodes activated by the initial active set S at the end of the information propagation [24], [26]. The formula for this is shown in (1):

$$\begin{cases} S = \arg\max \sigma(S) \\ S \subseteq V, |S| = k \end{cases} \tag{1}$$

The parameters of the TSIM algorithm are shown in TABLE 1.

IV. PROPOSED ALGORITHM

In this section, we introduce the framework of our improved algorithm and its details. As described in the previous section, some existing heuristic algorithms and greedy-based algorithms have some disadvantages in efficiency or accuracy. Hence, we utilize some strategies to optimize the Degree Discount algorithm and CELF algorithm. The Degree Discount algorithm selects seed nodes according to the degree of the node. However, the degree of a node cannot represent the influence spread of a node completely. The method of utilizing degree as the metric of the node-selecting process is suitable for the networks whose average degree is large but is not suitable for the networks whose average degree is small. Hence, the method of node-selecting sometimes neglects some influential nodes with a small degree. However, CELF calculates the influence of a node by tens of thousands of Monte Carlo simulations, which is accurate but time-consuming. In summary, the proposed algorithm is inspired by the advantage of the two algorithms. Therefore, a DDLF strategy combining the discount-degree descending technology with lazy-forward technology search strategies is proposed. Then, the proposed algorithm utilizes the strategy to obtain a large amount of efficiency by sacrificing acceptable accuracy.

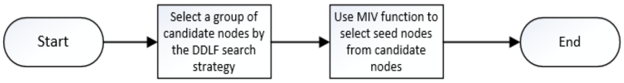

Start
Select a group of candidate nodes by the DDLF search strategy
Use MIV function to select seed nodes from candidate nodes
End

FIGURE 2. The framework of the TSIM algorithm.

A. FRAMEWORK OF TSIM ALGORITHM

The framework of our improved algorithm is given in Fig. 2.

From Fig. 2, we present the process of the TSIM algorithm.

(1) The TSIM algorithm sorts the nodes in the network G by the DDLF strategy. This strategy can be described as follows. First, the DDLF strategy sorts the nodes in the network G according to the degree of each node computed by the discount-degree descending technology and selects $top - k$ nodes as a part of candidate nodes. Second, the DDLF strategy constructs a subgraph G_{new} of the graph G by removing the descendant nodes of candidate nodes. Third, the DDLF strategy chooses $top - k$ nodes from the subgraph G_{new} as another part of candidate nodes by the lazy-forward technology. In the end, the DDLF strategy selects $2k$ candidate nodes.

(2) The algorithm estimates the influence spread of every candidate node by *MIV* and selects k nodes as seed nodes from candidate nodes. Next, we describe the TSIM algorithm by explaining the details of each step.

B. DDLF STRATEGY

In this paper, the discount-degree descending technology and lazy-forward technology search strategy, called DDLF, is proposed to select candidate nodes. The strategy has the following three steps: firstly, the DDLF sorts all nodes by discount degree of every node and selects $top - k$ nodes as a group of candidate nodes; secondly, construct a subgraph of initial graph G by excluding candidate nodes that have been selected; finally, the DDLF strategy selects another group of candidate node from subgraph by the lazy-forward technology. Hence, the TSIM algorithm not only improves the efficiency of greedy-based algorithms but also guarantees the accuracy of results.

1) DISCOUNT-DEGREE DESCENDING TECHNOLOGY

Discount-degree descending technology is a part of the DDLF strategy. The part gets experience from the Degree Discount algorithm. It utilizes the discount-degree as the influence of nodes, which saves plenty of time. Specifically, the part sorts all nodes in a network by their discount degree and selects $top - k$ nodes as a part of candidate nodes. Now, the Degree Discount algorithm is introduced as follows:

Definition 2 (Degree Discount Algorithm): [12] Considering the selection of node v as the new seed according to the degree of the node v, if the node u is selected as the seed node in the first round, the edge $\overrightarrow{vu}$ should not be calculated. Therefore, the Degree Discount algorithm discounts the degree of node v and makes the same discount on the degree of the node v for other neighbor nodes of the node v that already exist in the seed set.

Algorithm 1 Degree_Descend(G, k)

```
Input: Network $G = (V, E)$ and number $k$
Output: $S$ (a part of candidate sets)
1: $S = \emptyset$
2: for each node $v \in V$ do
3: compute the degree of $v$
4:    $dd_v = d_v$
5:    $t_v = 0$
6: end for
7: for $i = 1$ to $k$ do
8:    $u = \arg\max_v\{dd_v | v \in V \ S\}$
9:    $S = S \cup \{u\}$
10:    for each neighbor $v$ of $u$ and $v \in V \ S$
11:      $t_v = t_v + 1$
12:      $dd_v = d_v - 2\mathrm{t}_v - (d_v - t_v)t_v p$
13:    end for
14: end for
15: return $S$
```

Algorithm 2 *ConSubgraph(G,S,V')*

```
Input: Network $G = (V, E)$, $S$, and $V'$
Output:$G_{new}$
1: $E' = []$
2: for node $v$ in $S$ :
3:   for node $v'$ in $V'$ :
4:    if exist edge $(v, v')$ in $G$:
5:      $E' = E' \cup (v, v')$
6: end for
7: delete $E'$ and $V'$ from $G$ generate $G_{new}$
8:return $G_{new}$
```

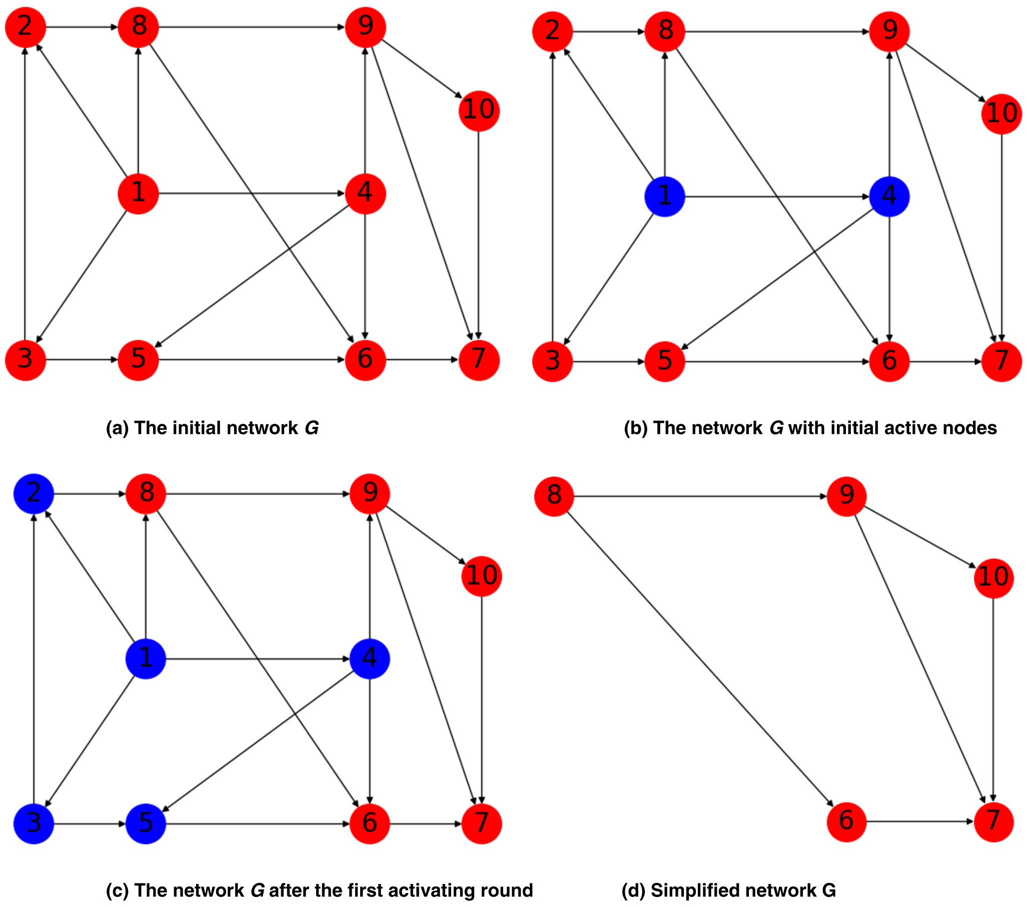
Next, this section introduces the Degree Discount algorithm in detail by pseudo-code.

As shown in Algorithm 1, firstly, the Degree Discount algorithm computes the degree of v in the set V(line 3-6). Secondly, it selects the largest degree nodes and picks them into the set S (line 7-8). According to the Degree Discount algorithm, if the neighbor v of the node u is already in the seed set, it does not need to consider the edge $\overrightarrow{uv}$. Therefore, it discounts the degree of the node u when selecting u as a seed node (line 10 $\sim$ 12).

The Degree Discount algorithm ignores the structure and actual operating effect, so that some nodes with the largest influence spread are ignored. Therefore, discount-degree descending technology is utilized to select the largest degree node in the graph G. Then, to find the most influential nodes ignored by discount-degree descending technology, we reduce the size of the network G. In the end, the lazy-forward strategy is applied to select nodes in the subgraph of G.

2) CONSTRUCTING A SUBGRAPH OF G

In this framework, to reduce the running time of DDLF, we make a new adjustment to the scale of the network graph



(a) The initial network *G*

(b) The network *G* with initial active nodes

(c) The network *G* after the first activating round

(d) Simplified network G

FIGURE 3. The process of constructing a subgraph.

$G = (V, E)$. As shown in Algorithm 2, S represents the subset of nodes selected by the discount-degree descending technology. The DDLF utilizes the Independent Cascade model to compute the influence spread of S in the network G, and put the activated nodes and nodes in S into a new node set V'. Meanwhile, the edges between S and V' are selected to form a new edge set E'(line 2 $\sim$ 5). Then, the activated nodes and related edges in G are deleted (line 7).Other nodes and edges in the network G to generate a new graph $G_{new} = (V_{new}, E_{new})$, where the node v in V_{new} belongs to the V, not belongs to V'. The edge e in E_{new} is a member of E instead of E'.

The operating steps are shown in Fig. 3, where blue nodes and red nodes represent active nodes and inactive node, respectively. We give an initial social network G whose node set is $\{1,2,3,4,5,6,7,8,9\}$. The simple network G is shown in Fig. 3(a). Assuming that the discount-degree descending technology selects the initial active set is $\{1,4\}$, as shown in Fig. 3(b), the blue node is the initial active node. Then, the initial active nodes in $\{1,4\}$ activate their neighbor nodes under the Independent Cascade model. Supposing that the nodes $\{2,3,5\}$ are activated, at the same time, diffusion propagation stops. The nodes $\{2,3,5\}$ and the relation edges from the network G is deleted to generate a new graph G, which is shown in Fig. 3(d). The process shows the method can reduce the scope of the G. The method can help to reduce the running time of the node-selecting process.

3) LAZY-FORWARD TECHNOLOGY

The final step in the DDLF strategy is to utilize lazy-forward technology to select another part of the candidate nodes. To ensure that the seed set has the maximum influence spread, after selecting a group of candidate nodes by degree-discount descending technology, the lazy-forward technology is applied to select another group of candidate nodes. Now, we introduce the lazy-forward strategy.

Definition 3 (Marginal Benefit) [25]: In the influence maximization problem, the marginal benefit of the influence value function $\sigma(\cdot)$ refers to the value that the $\sigma(\cdot)$ can increase for each additional node on the basis of the original node set S. The expression is as follows:

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

In (2), $\sigma(S)$ represents the influence value of initial seed set S, $\sigma(S \cup \{v_i\})$ expresses the influence value of node v_i after adding the seed set S. $\sigma_{v_i}(S)$ is the added

TABLE 2. Parameters of TSIM algorithm.

Node u	Marginal benefit $\sigma(S+u)-\sigma(S)$	Round t
u_A	18	0
u_B	11	0
u_C	10	0
u_D	8	0
u_E	7	0
u_F	5	0

TABLE 3. Influence spread increment after the first round.

Node u	Marginal benefit $\sigma(S+u)-\sigma(S)$	Round t
u_B	9	1
u_D	8	0
u_E	7	0
u_C	6	1
u_F	5	0

value of the node v_i. Next, we introduce the lazy-forward technology.

The process of lazy-forward technology is shown in TABLE 2 and TABLE 3. The first column, the second column and the third column represents the nodes, each node's marginal benefit and the node's round, respectively. At the round $t = 0$, the marginal benefit of all nodes needs to be computed (as shown in TABLE 2). Meanwhile, the node with the largest marginal benefit is inserted into the seed set S'. Then, the marginal benefit of the node u_B is the largest. At the round $t = 1$, the lazy-forward technology calculates the marginal benefit of u_B. Recalculate the node where the marginal return at the round $t = 0$ is larger than u_B at the round $t = 1$. If the value of the node is less than u_B, it is not necessary to calculate its value again. The node with the largest marginal benefit is selected from TABLE 3 and inserted into S' every time. The process will stop until the size of the seed set S' gets to k. The above lazy-forward part of the pseudo-code is shown as Algorithm 3.

C. OBJECTIVE FUNCTION

In this paper, the candidate nodes are selected by the discount-degree descending technology and the lazy-forward technology. To select the final seed set from the candidate set, this paper proposes a function called maximum influence value (MIV) to select seed nodes from candidate nodes. The MIV approximates the influence spread of a node in the seed set. The function is shown as follows:

$$MIV = w * Degree + (1-w) * Inf \tag{3}$$

The $Degree$ represents the degree of a node in a social network G. Inf represents the influence spread of single node under a specific diffusion model in a social network G, which leads to a fact that the influence overlapping of a node in a specific node set is neglected. Therefore, Inf in the MIV function is larger than the influence provides by this node in a specific node set. Considering above-mentioned problem, we discount the simulation results of the single node under the IC model to make sure that this value is close to the marginal benefit of this node. However, this treatment is not fair for some nodes, therefore, it is compensated according to the degree($Degree$) of each node. Hence, when utilizing the influence of a node as the metric of node-selecting, the MIV makes a discount for the influence of a node. And the margin benefit of a node is related to its degree. Hence, MIV utilizes them to make up a metric of the margin benefit, which avoids the repeated computation of the marginal influence spread of each candidate node to reduce the running time in the node-selecting process. And we conduct a lot

Algorithm 3 Lazy_Forward(G, k)

```
Input: Graph $G_{new} = (V_{new}, E_{new})$; a number $k$; the number of topic category, M; $Node(v, iis)$: the $\Delta iis$ of vertex $v$; $sis$: the influence of present seed; nQueue: the queue of $Node(v, iis)$
Output: another part of the candidate set $S'$
initialize $S' = \emptyset$, nQueue $= \emptyset$, $sis = 0$
**for** $v \in V_{new}$ do:
  $S_v = 0$
  **for** j = 1 to R do:
    $|S_v += |S'(\{v\})|$
  **end for**
pop the first node $v$ of the nQueue
$S' = \{node.v\}$ and $sis = node.iis$
**for** $i = 2$ to $k$ do:
  **while** true:
    pop the two element B, C
    **if** $Node(B, iis) >= C$:
      $S' = S' \cup \{node.v\}$
      $sis += node.iis$
      break
    **end if**
    $v = node.v$
    $S_v = 0$
    **for** $j = 1$ to R do:
      $S_v += |S'(\{v\})|$
    **end for**
    $S_v = S_v/R - sis$
    push $Node(v, S_v)$ into nQueue
  **end while**
**end for**
**return** $S'$
```

TABLE 4. The basic information of datasets.

Dataset	Wiki-Vote	ca-cond	soc-Epinions1	p2p-Gnutella31	web-Stanford	com-youtube
#Type	Directed	Directed	Directed	Directed	Directed	Directed
#Nodes	7115	23133	75879	62586	281903	1134890
#Edges	103689	186877	508837	147892	2312497	2987624
#Max.Degree	1345	558	11388	13126	38626	28754
#Avg.Degree	24.358	16.157	12.248	3.859	13.184	5.265

Algorithm 4 TSIM Algorithm

```
Input: network $G = (V, E)$,number $k$
Output: $S''$
 1. $S = \deg ree\_descend(G, k)$, $S'' = \emptyset$, $V' = \emptyset$, $E' = \emptyset$,
    $w = 0.01$
 2. calculate the influence range of the set $S$ in the graph $G$,
    add the relation of the nodes in $V'$, add the relation of
    the edges in $E'$
 3. $G_{new} = ConSubgraph(G, S, V')$
 4. $S' = Lazy\_forward(G_{new}, k)$
 5. $S_{new} = S \cup S'$
 6. for $v$ in $S_{new}$:
 7.     calculate the degree of $v$ and the $Inf$ of $v$
 8.     calculate $v's\ MIV = w * Degree + (1 - w) *$
        $Inf//MIV_{set}[v] = MIV$
 9. sort $MIV_{set}$ the value of $MIV$ in descending order
10. select $k$ nodes with the maximum  value to join the node
    set $S''$
11. return $S''$
```

of experiments to obtain the two discount values. In the *MIV* function, a large number of experimental data shows that the parameter w is set to 0.01, the final selected seed set has the best influence spread. We demonstrate the effect of this value 0.01 in the experimental section below. The value of *Inf* is calculated by the Independent Cascade model. The experimental results demonstrate that the function is effective.

D. TSIM ALGORITHM

The candidate nodes are selected through the introducing method above. In this section, candidate nodes are added into a set S'. The size of S' is $2k$. The TSIM algorithm utilizes the objective function as (3) to select k nodes with maximum influence spread in a social network G.

In this paper, the TSIM algorithm is shown in Algorithm 4. $G = (V, E)$ is the input graph and the number k is the size of the final seed set S''.

Lines 1 to 4 of Algorithm 4 represent our proposed strategy DDLF. Firstly, TSIM uses the discount-degree descending technology to select a part of candidate nodes in a graph G. And the nodes are put into S(line 1). Secondly, in line 2, the Independent Cascade model is used to calculate the influence range of the set S in the graph G. The activated nodes and associated edges are placed in a set V' and E', respectively. Thirdly, in line 3, we construct a new graph G_{new}. Fourthly, the lazy-forward technology is utilized to choose another part of candidate nodes from the graph G_{new}. The S' is another part of the candidate nodes set. Then, in line 5, the S_{new} is the candidate nodes set, consisting of S and S'. Next, in line 6-9, the *MIVs* of all the nodes in S_{new} need to be computed. Finally, we select k nodes with the largest *MIV* in line 10. Meanwhile, we get the final seed set S''.

E. COMPUTATION COMPLEXITY ANALYSIS

In this section, we analyze the time complexity of the proposed algorithm according to its process described by Algorithm 4. Firstly, we use the discount-degree technology to select a part of candidate nodes and the procedure needs $O(k \log n + m)$ [1] basic operations. Secondly, in Algorithm 3, lazy-forward technology has a time complexity of $O(kRm'n')$ [1] (line 4). Thirdly, the time complexity of the *MIV* computation is $O(k)$. Fourthly, the time complexity of sorting the $2k$ nodes is $O(2k \log 2k)$. To sum up, the overall time complexity of the algorithm is $O(k \log n + m) + O(kRm'n') + O(k) + O(2k \log 2k)$. Therefore, the total time complexity of TSIM is $O(k \log n + m) + O(kRm'n')$. $(n' < n, m' < m)$.

V. EXPERIMENT

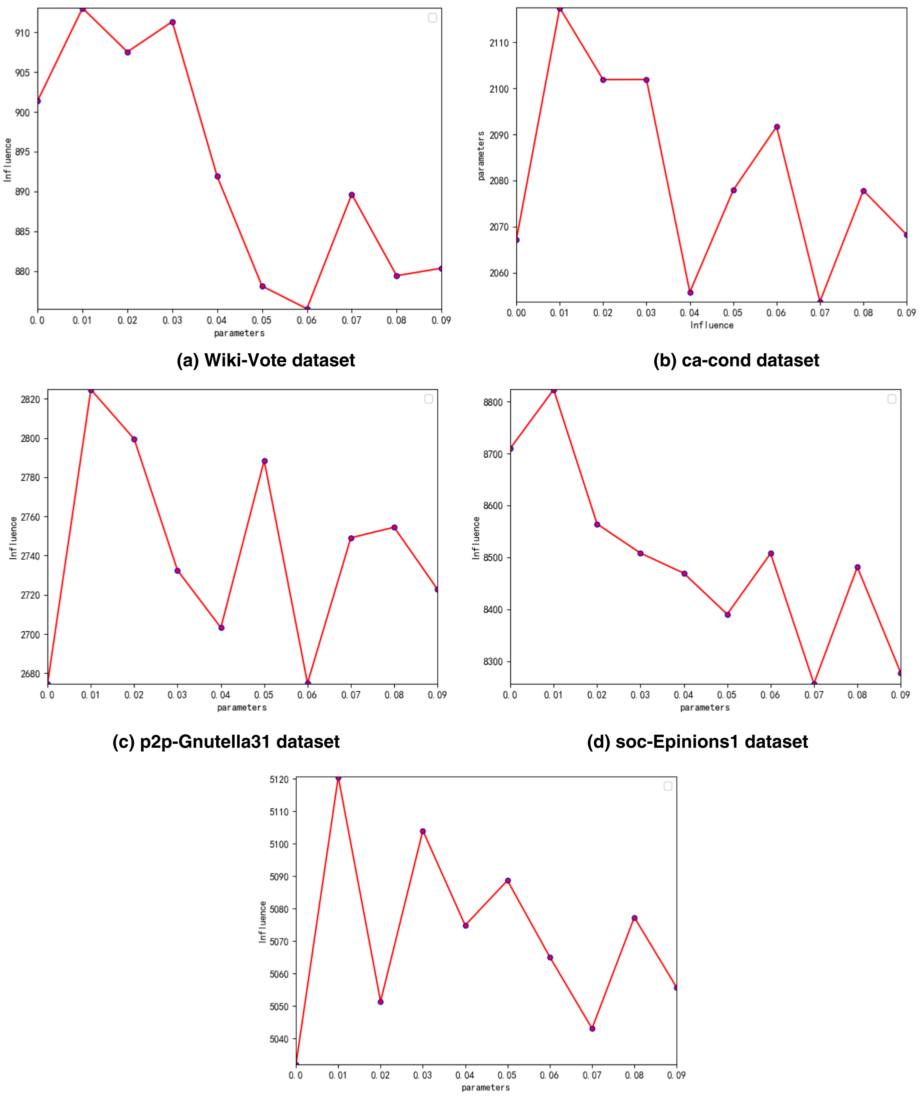
In this section, we evaluate our proposed algorithm on six real-world social networks and compare it with other approaches on the same social networks. Meanwhile, we compare the running time and the influence spread of the proposed algorithm with other approaches.

A. DATASETS AND EXPERIMENTAL SETTING

1) DATASETS

To ensure the authenticity of the experiment, we download six real-world datasets from SNAP. In TABLE 4, we introduce the characteristic of six social networks.

The first and second columns in the dataset represent nodes, and the two nodes in the same row represent edges between the two nodes. The third column in the dataset refers to activation probability p_{uv} that $p_{uv} = 1/N_{in(v)}$. Active probability p_{uv} is expressed as one in-degree of the node under the Independent Cascade model. And $N_{in(v)}$ denotes the incoming-degree of node v.



(a) Wiki-Vote dataset

(b) ca-cond dataset

(c) p2p-Gnutella31 dataset

(d) soc-Epinions1 dataset

(e) web-Stanford dataset

FIGURE 4. **The different parameters of *w* in different datasets.**

2) EXPERIMENTAL SETTING

To better demonstrate the advantages of the proposed algorithm, we compare the proposed algorithm with the other algorithms. As follows, we briefly introduce the Degree Discount algorithm, CELF, PMIA, and DDSE.

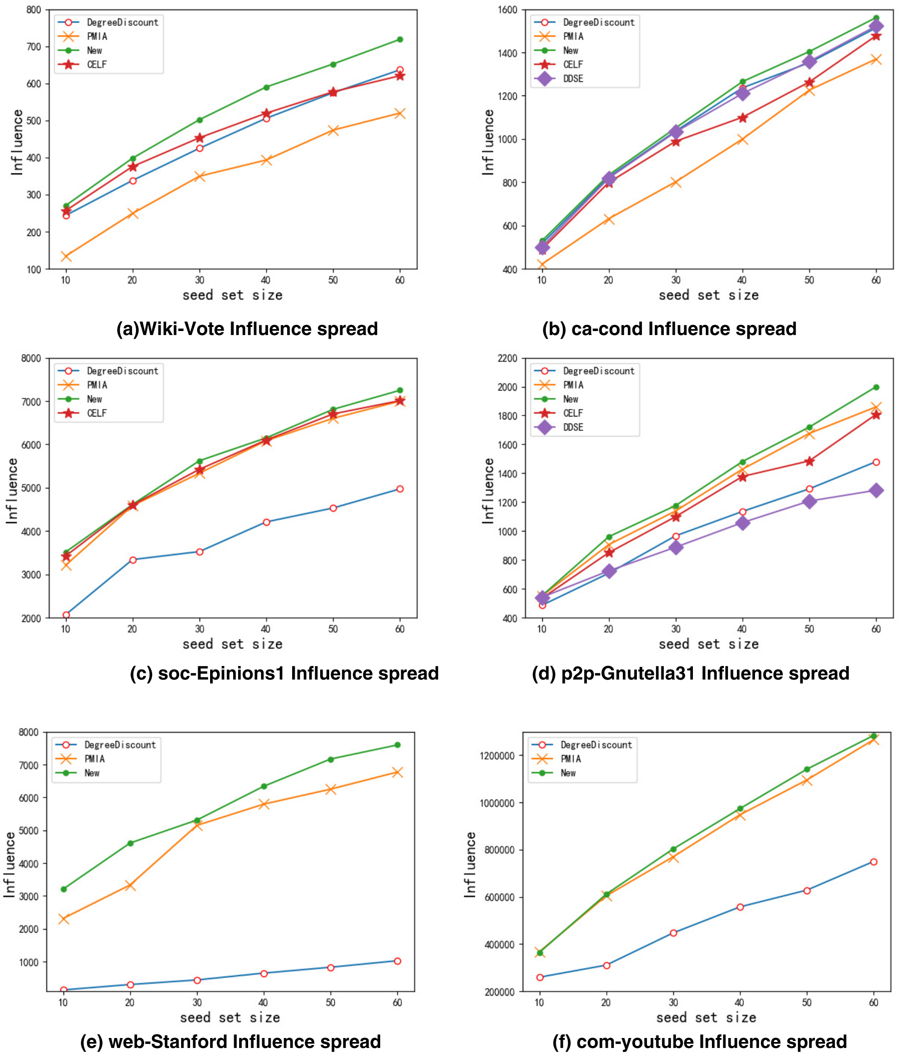
Degree Discount algorithm: the algorithm is introduced in section IV;

CELF algorithm: it has been introduced in section IV. In this algorithm, the number of Monte Carlo Simulation is set to 100.

PMIA: it uses the local influence subtree to simulate the influence of each node, so as to balance the accuracy and time efficiency of the greedy algorithm and heuristic algorithm. We set its influence threshold as $\theta = 1/10$;

DDSE: the algorithm bases on the strategy of degree descending search (DDS) and EDV. It overcomes the efficiency issue of greedy approaches by avoiding repeated simulations in the node selecting process. We set parameters: $g_{\max} = 200$, $diversity = 0.6, f = 0.1, cr = 0.4$.

All the algorithms in our paper are implemented by using python. And the python runs on Windows 8.1 64-bit operating system with a CPU of 1.80GHz and 8G of memory. The time of running DDSE and CELF algorithm on social networks with large average degree is unacceptale. Therefore, the DDSE algorithm is only run on the ca-cond network and p2p-Gnutella31 network. Moreover, we don't run the CELF algorithm on the web-Stanford network and com-youtube network.



(a)Wiki-Vote Influence spread

(b) ca-cond Influence spread

(c) soc-Epinions1 Influence spread

(d) p2p-Gnutella31 Influence spread

(e) web-Stanford Influence spread

(f) com-youtube Influence spread

FIGURE 5. **Influence spread.**

B. EXPERIMENTAL

1) EXPERIMENT FOR THE PARAMETERS OF MIV

Before the comparison experiments, we conduct some experiments for the parameters of *MIV* and find out the optimal values of these parameters.

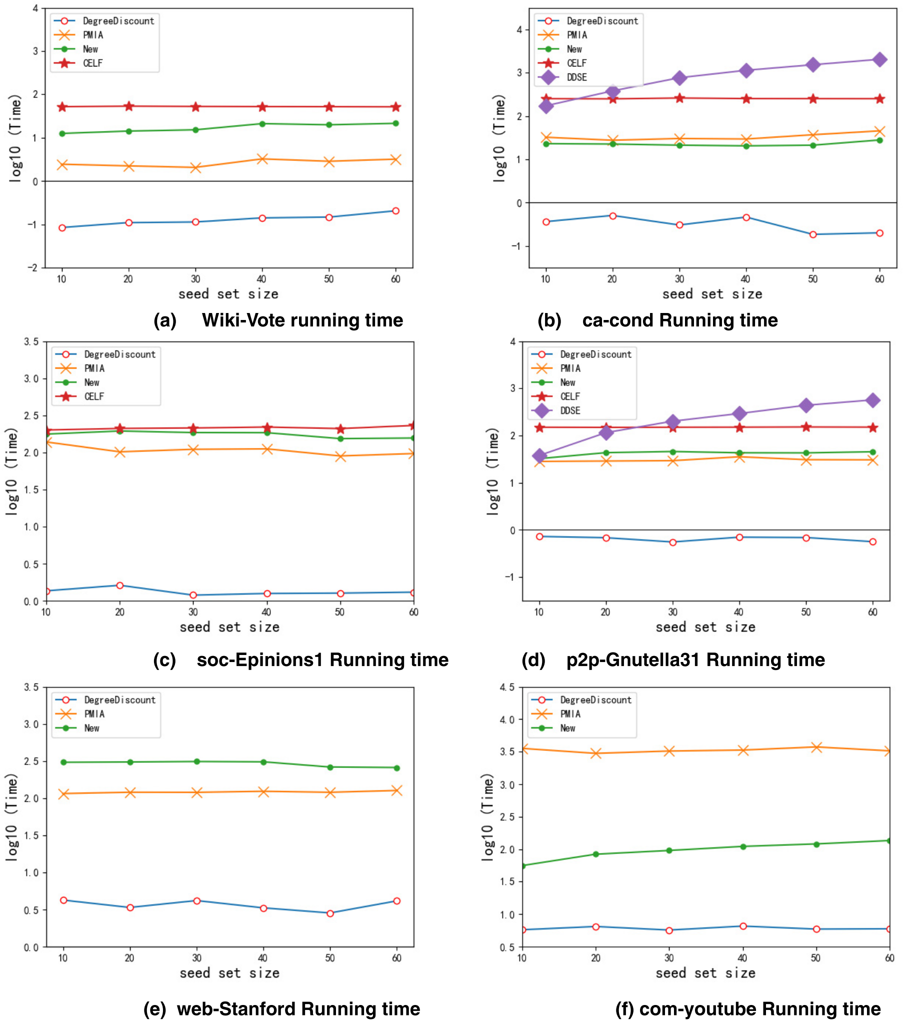
We utilize the above datasets to confirm the value of the parameter w of the *MIV* function. To ensure the accuracy of the experimental results, we set the size k of the final seed set as 100. First, the DDLF strategy selects $2k$ candidate nodes. Then, we set different values of w to select k seed nodes among the candidate nodes. As shown in Fig. 4, in the case of different w values, when $w = 0.01$, the influence of the seed node selected by *MIV* is greater than that of other nodes. It is easy to be known from this experiment that the proposed algorithm can get the best results when $w = 0.01$.

2) EXPERIMENTAL RESULTS

In this section, we compare our proposed algorithm with comparison algorithms in terms of running time and influence spread.

(1) The influence spread for the real-world datasets: the size of the seed set k is 10,20,30,40,50, and 60. Meanwhile, the accuracy of the algorithm in different seed sizes is compared. Fig. 5 (a)$\sim$(f) shows the influence spread results of five algorithms in six datasets.

As shown in Fig. 5, we can easily find the TSIM algorithm is better than others, which shows that we can obtain a better marketing effect. On the different datasets, the influence spread results of PMIA, CELF, Degree Discount, and DDSE algorithm perform differently. Compared with the above algorithms, the seed set selected by the TSIM algorithm has the largest influence spread. Fig. 5(a) shows the experimental results of the Wiki-Vote dataset, from which we can find the influence spread of TSIM is 15.8% better than the CELF algorithm in terms of accuracy. The TSIM algorithm is superior to the CELF algorithm because it not only considers the degree of nodes but also pays attention to the propagation of each node in the seed set. Fig. 5(b) shows that the influence spread of our proposed algorithm in the ca-cond dataset is 5.7% better than the CELF algorithm.



(a) Wiki-Vote running time

(b) ca-cond Running time

(c) soc-Epinions1 Running time

(d) p2p-Gnutella31 Running time

(e) web-Stanford Running time

(f) com-youtube Running time

FIGURE 6. **Running time.**

On the soc-Epinions1 and p2p-Gnutella31 datasets, as shown in Fig. 5(c)(d), TSIM is 3.6% and 7.5% better than PMIA in terms of result accuracy. The TSIM is an algorithm based on the degree and Monte Carlo simulation. Compared with PMIA, the TSIM algorithm has more stable accuracy, because PMIA needs to set threshold artificially. The experiment on the Web-Stanford dataset shows that TSIM is 12.2% better than PMIA. As shown in Fig. 5(f), our proposed algorithm is best, 1.42% and 71.21% better than PMIA and the Degree Discount algorithm, respectively. From the analysis, we can prove that the influence spread of TSIM is better than others. It can be explained as follows: TSIM utilizes the DDLF strategy to improve the low efficiency and the low accuracy of greedy-based algorithms. Specifically, discount-degree technology can save a large amount of running time. However, the approach of utilizing discount-degree as a metric of estimating the influence has obvious limitations. The discount-degree may neglect some nodes whose degree is small, but the influence is large. The lazy-forward technology makes up the drawback of discount-degree on social networks whose average degree is large in the node-selecting process.

(2) Running time on the six real-world datasets: under the condition of fixed seed set size, the comparison results of all algorithms in running time are given.

From Fig. 6(a) $\sim$ (f), the running time of the Degree Discount algorithm is faster than all the above algorithms. Because it is more possible to quickly select the largest degree node in graph G and specially tuned for the uniform IC model. Meanwhile, the running time of the TSIM algorithm is 58.82%, 88.84%, 32.17% and 92.09% faster than CELF on the first four datasets, respectively. In Fig. 6(a), (c), (d) and (e), the PMIA algorithm is superior in efficiency. PMIA utilizes a threshold to prune unnecessary traveling, but this pruning depends on the structure of social networks. The efficiency of the DDSE algorithm is lower than our proposed algorithm. Compared with DDSE, the TSIM algorithm can guarantee better result accuracy and computation efficiency. The local search of the DDSE algorithm takes a long time, and thus, the efficiency is very low. As shown in Fig. 6(f), the experimental results is clear. Degree Discount is best in terms of computation efficiency, and our proposed algorithm is only worse than Degree Discount algorithm. In detail, Degree Discount is 95.62% and 99.82% faster than TSIM and PMIA, respectively. The Degree Discount algorithm utilizes the discount-degree to estimate the influence spread of single node. The basic idea of this algorithm is that a node with a larger discount-degree has a larger influence spread. But this way of evaluating influence spread can not describe the influence spread of a node accurately. This method is suitable for the networks with a large average degree, but is not suitable for the social networks with a small average degree. In our experiments, the average degree of these social networks are relatively large, so the degree-discount algorithm performs well. Our proposed algorithm improve the result accuracy by sacrificing computation efficiency.

VI. CONCLUSION

In this paper, we tackle the influence maximization problem by proposing a new algorithm called TSIM. TSIM overcomes the low accuracy of the Degree Discount algorithm and the low efficiency of the CELF algorithm by combining the advantages of the two algorithms. The proposed algorithm combines the discount-degree descending technology with the lazy-forward technology, which utilizes the advantage of the CELF algorithm to make up the drawback of the Degree Discount algorithm and utilizes the advantage of the Degree Discount algorithm to improve the efficiency of the CELF algorithm. We conduct extensive experiments on real-world social networks. Experimental results demonstrate the proposed algorithm outperforms the other four comparison algorithms in efficiency and accuracy.

In the future, we will find a more accurate method to estimate parameter *w* of *MIV* and achieve the parallel processing in node-selecting process. By two methods, we will improve the computation efficiency and the result accuracy of our proposed algorithm.

REFERENCES

[1] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1852–1872, Oct. 2018.

[2] J. Wu, Z. Chen, and M. Zhao, "Information cache management and data transmission algorithm in opportunistic social networks," *Wireless Netw.*, vol. 25, no. 6, pp. 2977–2988, Aug. 2019.

[3] J. Wu, Z. Chen, and M. Zhao, "Weight distribution and community reconstitution based on communities communications in social opportunistic networks," *Peer-to-Peer Netw. Appl.*, vol. 12, no. 1, pp. 158–166, Jan. 2019.

[4] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, 2013.

[5] S. G. Kulkarni, A. Bilgi, D. Miskin, R. Oursang, P. Hiremath, and N. T. Pendari, "Spheres of influence for effective viral marketing," in *Proc. Int. Conf. Adv. Comput., Commun. Control Netw. (ICACCCN)*, Greater Noida, India, Oct. 2018, pp. 750–753.

[6] J. Zhang, S. Wang, Q. Zhan, and P. S. Yu, "Intertwined viral marketing in social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, San Francisco, CA, USA, Aug. 2016, pp. 239–246.

[7] 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.

[8] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining-KDD*, San Francisco, CA, USA, 2001, pp. 57–66.

[9] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining-KDD*, Washington, DC, USA, 2003, pp. 137–146.

[10] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," *Theory Comput.*, vol. 11, no. 4, pp. 105–147, 2015.

[11] 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-KDD*, 2007, pp. 420–429.

[12] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 199–208.

[13] Q. Wang, M. Gong, C. Song, and S. Wang, "Discrete particle swarm optimization based influence maximization in complex networks," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2017, pp. 488–494.

[14] D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *International Colloquium on Automata, Languages and Programming*. Lisbon, Portugal: Lecture Notes in Computer Science, 2005, pp. 1127–1138.

[15] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," *Knowledge Discovery in Databases: PKDD 2006* (Lecture Notes in Computer Science), vol. 4213. Berlin, Germany: Springer, 2006, pp. 259–271.

[16] A. Goyal, W. Lu, and L. V. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. 20th Int. Conf. Companion World Wide Web-WWW*, 2011, pp. 47–48.

[17] K. Jung, W. Heo, and W. Chen, "IRIE: Scalable and robust influence maximization in social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 918–923.

[18] L. Qiu, W. Jia, J. Yu, X. Fan, and W. Gao, "PHG: A three-phase algorithm for influence maximization based on community structure," *IEEE Access*, vol. 7, pp. 62511–62522, 2019.

[19] C. Wang, W. Chen, and Y. Wang, "Scalable influence maximization for independent cascade model in large-scale social networks," *Data Mining Knowl. Discovery*, vol. 25, no. 3, pp. 545–576, Nov. 2012.

[20] 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, pp. 119–130, Feb. 2018.

[21] J. Goldenberg, B. Libai, and E. Müller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Lett.*, vol. 12, no. 3, pp. 211–223, 2001.

[22] J. Goldenberg, B. Libai, and E. Müller, "Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata," *Acad. Marketing Sci. Rev.*, vol. 9, no. 3, pp. 1–18, 2001.

[23] J. Ge, L.-L. Shi, L. Liu, and X. Sun, "User topic preferences based influence maximization in overlapped networks," *IEEE Access*, vol. 7, pp. 161996–162007, 2019.

[24] J. Zhu, S. Ghosh, J. Zhu, and W. Wu, "Near-optimal convergent approach for composed influence maximization problem in social networks," *IEEE Access*, vol. 7, pp. 142488–142497, 2019.

[25] Q. Wang, Y. Jin, Z. Lin, S. Cheng, and T. Yang, "Influence maximization in social networks under an independent cascade-based model," *Phys. A, Stat. Mech. Appl.*, vol. 444, pp. 20–34, Feb. 2016.

[26] J. Yang and J. Liu, "Influence maximization-cost minimization in social networks based on a multiobjective discrete particle swarm optimization algorithm," *IEEE Access*, vol. 6, pp. 2320–2329, 2018.

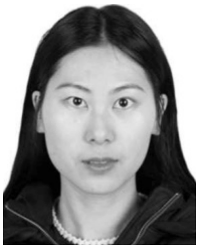
[27] J. Wu, G. Yu, and P. Guan, "Interest characteristic probability predicted method in social opportunistic networks," *IEEE Access*, vol. 7, pp. 59002–59012, 2019.

[28] P. Guan and J. Wu, "Effective data communication based on social community in social opportunistic networks," *IEEE Access*, vol. 7, pp. 12405–12414, 2019.

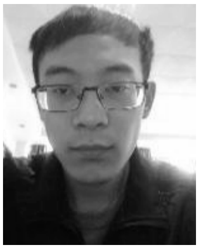
[29] J. Cheng, X. Wu, M. Zhou, S. Gao, Z. Huang, and C. Liu, "A novel method for detecting new overlapping community in complex evolving networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 9, pp. 1832–1844, Sep. 2019.

QIU LIQING was born in 1978. She received the Ph.D. degree in computer software and theory from Beihang University, Beijing, China. She is currently a Lecturer with the Shandong University of Science and Technology.

GU CHUNMEI was born in 1994. She received the B.S. degree from Taishan University, Taian, China, in 2018. She is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Shandong University of Science and Technology. Her current research interest includes social networks.



ZHANG SHUANG was born in 1994. She received the B.S. degree from the University of Jinan Quancheng College, Yantai, China, in 2018. She is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Shandong University of Science and Technology. Her current research interest includes social networks.



TIAN XIANGBO was born in 1995. He received the B.S. degree from the Shandong University of Science and Technology, Qingdao, China, in 2018, where he is currently pursuing the M.S. degree with the School of Computer Science and Engineering. His current research interest includes social networks.

ZHANG MINGJV was born in 1995. She received the B.S. degree from Shandong Women's University, Jinan, China, in 2018. She is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Shandong University of Science and Technology. Her current research interest includes social networks.