IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 7, 2017, accepted September 25, 2017, date of publication October 2, 2017,
date of current version November 14, 2017.

*Digital Object Identifier 10.1109/ACCESS.2017.2758353*

# A New Structure-Hole-Based Algorithm For Influence Maximization in Large Online Social Networks

**JINGHUA ZHU, YONG LIU, AND XUMING YIN**
School of Computer Science and Technology, Heilongjiang University, Harbin, China

Corresponding author: Jinghua Zhu (zhujinghua@hlju.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61632010, Grant 61100048, Grant 61370222, and Grant 61602159, and in part by the Natural Science Foundation of Heilongjiang Province under Grant F2016034, and in part by the Education Department of Heilongjiang Province under Grant 12531498.

**ABSTRACT** The problem of influence maximization (IM) in a social network is to determine a set of nodes that could maximize the spread of influence. The IM problem has been vitally applied to marketing, advertising, and public opinion monitoring. Although recent studies have studied the IM problem, they are generally greedy or heuristic-based algorithms, which are time consuming for practical use in large-scale social networks. Based on the observation that structural hole nodes usually are much more influential than other nodes, in this paper, we develop a structure-hole-based influence maximization algorithm (SHIM) with an emphasis on time efficiency. The SHIM algorithm utilizes structure hole information to significantly decrease the number of candidates of seed nodes. To measure the structure importance of nodes, we propose an structure hole value calculate algorithm to calculate the structural hole value of nodes. We prove the SHIM is NP-hard and propose a structure-based greedy algorithm to select seeds with wide influence spread and high structural hole value. We conduct experiments on real data sets to verify our algorithm's time efficiency and accuracy, and the experimental results show that comparing with the existing algorithms, our algorithms are much more efficient and scalable.

**INDEX TERMS** Social network, influence maximization, structural hole, greedy algorithms.

## I. INTRODUCTION

The social network is a complicated structure composed of social individuals and relationships between them. Large scale online social networks like Sina Weibo, Tencent Wechat and Facebook have attracted millions of users recently [1], [2]. People would like to use social networks to communicate or diffuse information. For example, a company develops a new product, they want to advertise the product in a certain social network. The company has a limited budget so they can only give free sample products to a small number of users. They hope that the initial users could influence their friends to use the products, and their friends could influence their friends' friends. Through the word-of-mouth effect, a large number of users finally adopt the products.*Influence maximization* is a fundamental research problem in social networks [3], [4]. It selects a set of $k$ nodes as seeds in order to maximize the propagation of ideas, opinions and products *etc al*.

The problem of *Influence maximization* is *#P-hard*, the widely used baseline methods for computing influence spread are based on Monte Carlo simulation or heuristic algorithms. Most of the existing methods only take consider of the influence on nodes and propagation probability on edges, while ignoring the structure feature of nodes in social networks. In fact, structure positions act as bridge between individuals of different communities and have more control over information diffusion.

The absence of ties between two parts of social network is called *structural holes*. The notion forms the basis of theory of *structure holes*. Two parts can only make connections indirectly by the connection to the third individual. In this case, there is a hole between these two parts in terms of structure, which is called *structural hole* [5]. The third individual is called *structural hole spanner*. For example, in the social network as shown in Fig.1, node $a$ occupies a bridge position between two different groups $A$ and $B$.

However, nodes might not be selected as seed node by the traditional influence maximization algorithms if their influence or propagation probability are low in the traditional propagation model.

In order to improve the time efficiency and maximize the influence spread in large online social networks,

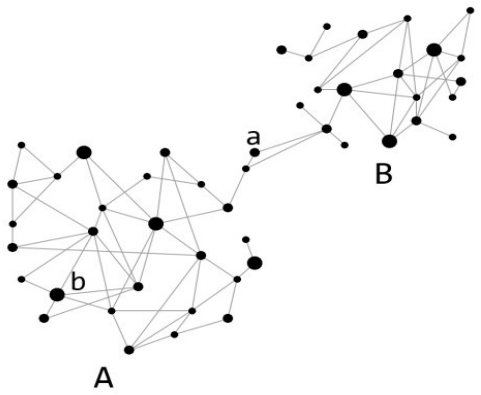



**FIGURE 1.** Example of structure holes between two groups.

we develop ***S**tructural **H**oles* based ***I**nfluence **M**aximization* algorithm (SHIM). The intuition behind SHIM is opinion leaders play a key role in spreading information within a community, while structure hole spanners are more important for spreading information between communities. In SHIM, we first identify structure hole spanners whose structure hole value above the given threshold. And then we compute the influence capability of each structure hole spanners. At last, we select the top-k seeds by combining the structure hole value and influence value. By this way, a large amount of non-structure hole spanners can be filtered out and the network scale can be shrank a lot. Furthermore, the spread of information can be improved by combination of structure hole value and influence value.

The aim of the existing influence maximization algorithms is to maximize the spread of influence and they mainly focus on the efficiency and accuracy of proposed algorithms. Few of them consider the privacy protection problem during the information diffusion process. In fact, due to the structural hole spanner nodes are much more influential than other nodes, they are more likely to be the information leakage nodes. Therefore, if we don't want to leak privacy information during the targeting advertising, we can revise our structure holes-based influence maximization algorithm by incorporating the information contents and the attitude (like or dislike) of user nodes.

The contributions of this paper are as follows:

- We propose SHIM: structural hole theory-based algorithm to solve the problem of influence maximization.
- We propose Algorithm1 to compute the structural hole value of nodes based on Spectral graph theory which is a useful tool in clustering and graph partition.
- We prove the problem of influence maximization is NP-hard and propose a greedy algorithm which is presented in Algorithm2 to solve it.
- We conduct experiments to verify the time efficiency and influence spread of our algorithm.

The rest of this paper is organized as following: Section 2 discusses the related work of influence maximization algorithms and the structure holes theory. Section 3 gives the definition of influence maximization problem and structure hole value. The structural holes based influence maximization algorithm and greedy algorithm will be presented in Section 4. The experimental results and analysis will be reported in Section 5. Finally, the conclusion will be drawn in Section 6 as well as the future work.

## II. RELATED WORKS

### A. INFLUENCE MAXIMIZATION (IM)

Domingos and Richardson [6] are the pioneers who study influence maximization problem in social networks. Kempe and David [7] modeled influence maximization problem as a discrete optimal problem and proved its NP-hardness. They proposed greedy algorithm with approximation ratio of (1-1/e).

Greedy algorithms need to perform several Monte Carlo Simulations in every epoch of seeds selecting iteratively, which leads to expensive costs and low efficiency when the number of nodes is very large. Leskovec *et al.* [8] presented CELF algorithm, an improved greedy algorithm. CELF was 70 0 times faster than basic greedy algorithm. Chen *et al.* [3] proposed NewGreedy algorithm to filter out those nodes with little contributions for information propagation. Several researches aim to design heuristic algorithm to improve computing efficiency [9], [10]. Although efficiency of heuristic algorithms is high, the approximation ratio is not guaranteed. Xiao-Dong [11] proposed a parallel influence maximization algorithm BUTA which has much shorter running time than greedy algorithms on cost of sacrifice of accuracy.

There are also many algorithms considering different types of IM problems by extending classic influence models. For example, topic-aware IM [12], [13] considers the influence diffusion under topic models; location-aware IM [14]–[16] focuses on maximizing the influence spread in certain spatial areas; and conformity-aware IM [17] considers users' conformity tendencies in the influence estimation; real-time IM [18] considers the stream influence maximization problem for dynamic social networks; Privacy ware IM [19]–[21] considers the Privacy-Preserving problem during the information diffusion in social network.

### B. STRUCTURAL HOLE

Structural holes theory is a sociological concept which is proposed by Burt [22]. Burt proved that the person who owns the structural hole has great advantages of control and information diffusion. Zhang *et al.* [23] solved the structural hole finding problem by Fiedler vector in Laplacian matrix and designed DGSH algorithm to detect structural holes. Xiao-Ping and Yu-Rong [24] used domain structural holes to detect most influential nodes and proposed N-Burt algorithm to accurately evaluate importance of nodes. However, they didn't evaluate the information propagation probability of nodes and thus the spread of selected nodes is not the most wide. Lou *et al.* [25] took advantage of information propagation probability to mine structural holes and designed HIS and MaxD model to find

structural holes owners. Their methods relied on cluster-based network.

In order to overcome the shortcomings of the above influence maximization algorithms, in this paper we exploited structural hole theory to the influence maximization problem and proposed SHIM algorithm to find $k$ seeds.

## III. PREMILINARIES

In this Section, we will introduce the model of structural hole based influence maximization problem, including the specific graph model, the information propagation model and the measurement of influence spread. We also prove the NP-hardness of the problem in this Section.

### A. GRAPH MODEL AND PROPAGATION MODEL

The social network can be treated as a directed graph $G(V, E, W, S)$, here $V$ stands for the set of vertices and $E$ is the set of edges. In the context of social network, $V$ can be treated as users and $E$ can be treated as relations between them, such as friendship or trust relation. $W$ is weights on edges representing influential probabilities among users. $S$ is set of structural hole values corresponding each node.

Independent Cascading (IC) Model and Threshold Model are two widely used propagation model in social network. In this paper, we use IC Model as our propagation model. In IC model, each node has two states: active or inactive. If node $u$ is activated in the $t$-th time step, it'll activate its neighbors in the next time step, which means each active node has only one chance to activate its neighbor nodes. P($u$, $v$) is the probability of node v be activated by node $u$.

Let $S_0$ be the seeds set, all nodes in $S_0$ are in active state at the first time step. The information diffusion process is as following in IC model: $S_{t-1}$ is the set of nodes which are active in the ($t$-1)-th step, and $S_t$ is the set of nodes which are active in the $t$-th step. In the($t+1$)-th step, each node $u$ in $S_t$ tries to activate its neighbor node $v$ with probability of P($u$, $v$). If such activation is successful, then $v$ changes state from inactive to active. Otherwise, node $u$ cannot activate $v$ anymore. Repeating the above process until no node in the network can be activated.

### B. PROBLEM STATEMENT

We use $\sigma$(S) to represent the influential spread of seed set S. The influential spread of seed nodes can be quantified by the number of nodes that will be activated by the seeds under the above propagation model. Given a social network graph $G(V, E, W, S)$, a positive integer $k$, positive real number $\alpha(0 \leq \alpha \leq 1)$, $\alpha(1-\alpha)$ represents user's preference, which shows the proportion of preference for the structure feature of nodes (resp. preference for the influence of the node). Then the structural hole theory based Influence maximization problem is to find a set of seed nodes under the IC model so that the nodes in the seed set are all structural holes and the seed node set has the most influence spread.

In the problem of influence maximization, Margin Gain (MG) of influential value function $\sigma(*)$ is influential gain by activating a node $v_i$ as initial active node based on the current active nodes set S. as described in the following formula:

$$\sigma_{v_i}(S) = \sigma(S \cup \{v_i\}) - \sigma(S) \tag{1}$$

*Theorem 1:* Structural Hole Theory based Influence Maximization problem (SHTIM) is NP-hard.

*Proof:* As defined above, SHIM can be regarded as Set Covering Problem. Set Covering Problem is defined as follow:

Set $A = \{a_1, a_2, a_3, \cdots, a_x\}$, Set $S = \{S_1, S_2, S_3 \cdots, S_y\}$ where $S_i$ is a subset of A, S contains all subsets of A.

The question is whether a set S' exists satisfying that the size of S' is k and S' covers all elements in A. Considering SHTIM, we construct a directed bipartite graph which contains A and S, as described in Fig.2, where A is the set of nodes which can be activated, S is the initial seeds set ($u_i \in$A, $v_i \in$ S').

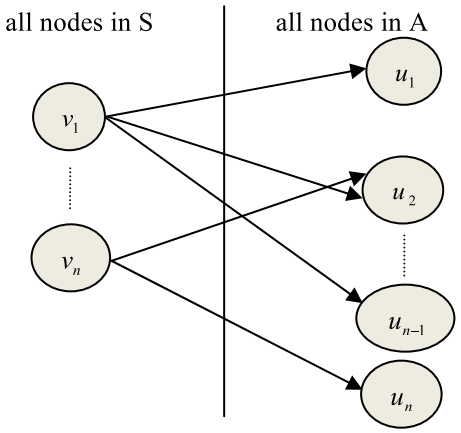



**FIGURE 2.** Directed bipartite graph.

Set covering problem is a problem of decision whether we can get a subset of size k, information can be propagated from these seed nodes in S to all other nodes in A.

### C. STRUCTURAL HOLES

Structural holes are nodes which function as bridge in social networks. For example, as shown in Fig.3, node 2 and node 3 are bridge nodes and if delete them, the graph will be divided into two unconnected parts. Such nodes are called structural hole nodes. they are connected by few neighbor nodes, but the network will be separated and the information propagation will be locked if these nodes are deleted. To decide whether a node is a structural hole node, we define a real number SH to evaluate the probability of a certain node be structural hole. In order to compute the SH, we will give introduce some notions first, then we will describe the solution in details.

#### 1) ADJACENT MATRIX

Given a network graph, the adjacent matrix is a widely used data structure to represent the network. The element $a_{ij}$ of the

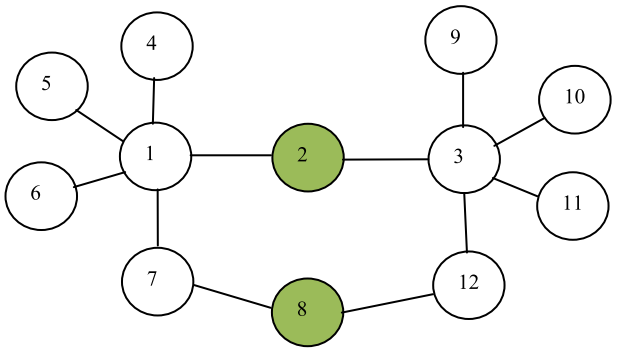



**FIGURE 3.** Structural holes in green circles.

adjacent matrix $A$ can be computed as the following way:

$$A = \begin{cases} a_{ij} = & if\ E(i,j) = 1 \\ 0 & otherwise \end{cases} \tag{2}$$

### 2) DEGREE MATRIX

The degree matrix $D$ is a diagonal matrix formed by the following way:

$$D = \begin{cases} d_i = p_i & if \quad i = j \\ 0 & otherwise \end{cases} \tag{3}$$

where $p_i = \sum_{j=1}^{n} a_{ij}$.

### 3) LAPLACIAN MATRIX

With the help of the *Adjacent Matrix* and the *Degree Matrix*, we define the *Laplacian matrix* as follows:

$$L = D - A \tag{4}$$

```
-0.18  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
 0.00  0.24  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
 0.00  0.00  0.47  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
 0.00  0.00  0.00  1.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
 0.00  0.00  0.00  0.00  1.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
 0.00  0.00  0.00  0.00  0.00  1.00  0.00  0.00  0.00  0.00  0.00  0.00
 0.00  0.00  0.00  0.00  0.00  0.00  1.00  0.00  0.00  0.00  0.00  0.00
 0.00  0.00  0.00  0.00  0.00  0.00  0.00  1.46  0.00  0.00  0.00  0.00
 0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  1.73  0.00  0.00  0.00
 0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  3.16  0.00  0.00
 0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  5.85  0.00
 0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  6.27
```

**FIGURE 4.** The Eigen value of the Laplacian matrix.

$L$ has many properties, we can compute the Eigen value and eigenvector of matrix $L$ as shown in Figure 4. The second smallest Eigen value and its eigenvector (also called Federal vector) as shown in Figure 5. We use the absolute values in Federal vector to map to nodes in Fig.6. The node which corresponds to the second smallest value is the best structural hole node and we defined the value as structural hole value *SH*.

$$SH = \min_{2}\{|F|\} \tag{5}$$

```
-0.30  0.15  0.21 -0.00  0.00 -0.00 -0.00 -0.05  0.12 -0.05 -0.69 -0.59
-0.22 -0.11  0.18 -0.00  0.00 -0.00 -0.00  0.05  0.89  0.15  0.02  0.29
-0.18 -0.35  0.06  0.00  0.00 -0.00 -0.00  0.08  0.12 -0.12  0.61 -0.66
-0.26  0.20  0.40  0.05  0.71  0.25  0.31  0.11 -0.17  0.02  0.14  0.11
-0.26  0.20  0.40  0.14 -0.09 -0.74 -0.31  0.11 -0.17  0.02  0.14  0.11
-0.26  0.20  0.40 -0.20 -0.63  0.49  0.00  0.11 -0.17  0.02  0.14  0.11
-0.58  0.24 -0.41 -0.00 -0.00 -0.00  0.00 -0.54  0.01 -0.32  0.14  0.13
-0.39  0.03 -0.43 -0.00 -0.00  0.00  0.00  0.30 -0.13  0.74  0.01 -0.07
-0.15 -0.46  0.11 -0.79  0.10 -0.19  0.05 -0.16 -0.16  0.06 -0.13  0.12
-0.15 -0.46  0.11  0.32  0.15  0.33 -0.66 -0.16 -0.16  0.06 -0.13  0.12
-0.15 -0.46  0.11  0.47 -0.25 -0.14  0.61 -0.16 -0.16  0.06 -0.13  0.12
-0.26 -0.18 -0.24  0.00 -0.00  0.00  0.00  0.70 -0.04 -0.54 -0.16  0.17
```

**FIGURE 5.** The Federal vector of the Laplacian matrix.

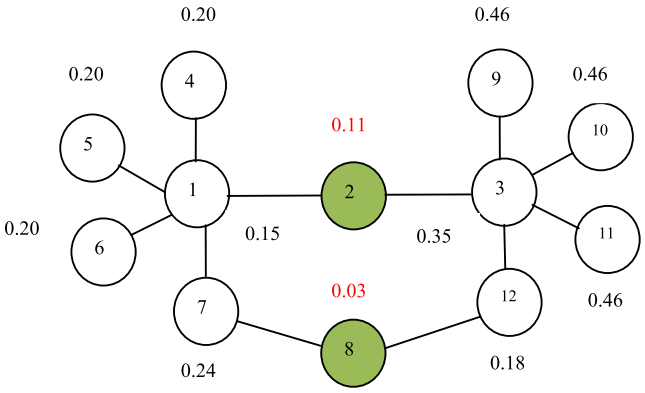



**FIGURE 6.** The Structural hole value of nodes in social network.

we can compute *SH* of a node based on the above formula, different structural hole node has different *SH* which can be seen from Fig.6.

## IV. STRUCTURAL HOLES BASED INFLUENCE MAXIMIZATION ALGORITHM

In this section, we first propose an algorithm called structure hole value calculate (SHVC) to compute the *SH* value for each node in network graph, we also analyze the complexity of this algorithm; then we present a Structure-based Greedy algorithm to solve Structural Hole Theory based Influence Maximization problem. Last, we analyze the time complexity of the greedy algorithm.

### A. COMPUTATION OF STRUCTURAL HOLES VALUE

Because the structure of node is needed to be considered when selecting the seed node, we need to know the Structure-Hole value(*SH*) of node. We design SHVC algorithm to calculate *SH* for each node. As shown in Algorithm1, given network graph $G$, we first set *SH* to 0 for all nodes (Line 1) and use the Federal vector to determine the optimal structure hole node in the current graph. Then we calculate and update the *SH* of this optimal structure hole node (Line 2-5). And then the value of the characteristic component is set to 1, and it is ignored in the next calculation (Line 6-7). And then back to Line 4 until calculating *SH* of all structural holes(Line 8-9). The pseudo-code is presented as follow:

Now we give analysis on the time complexity of SHVC algorithm. The cost of building matrix for a graph is $O(n^2)$. The cost of constructing degree matrix and adjacent matrix is $2O(n^2)$. The cost of constructing laplacian matrix is $O(1)$.

**Algorithm 1** SHVCalculate (SHVC)

Input: G(V, E,W,S), $\theta$
Output: $G_1$
1) u.SH=0, $\forall u \in V$;
2) Build the Laplacian Matrix LG;
3) Get the Federal vector F;
4) u$\leftarrow$min$_2${|F[i].value |} $1 \leq i \leq n$;
5) u.SH =min$_2${|F[i].value|};
6) F[i].value=1;
7) V=V-{u};
8) while(V$\neq \emptyset$)
9) GoTo Line 4
10) for $\forall u \in V$ do
11) $G_1 \leftarrow$G\{u|u.inf$< \theta$u.*SH*=0};
12) return $G_1$.

The line 4-5 to compute eigenvalue needs O(m+n) according to QR [26]. Therefore, the total time of finding the first structural hole node is $2O(n^2)+O(1)+O(m+n)=O(n^2)$. Then the algorithm will iteratively find the next structural hole node until the end. The cost of finding the first structural hole node is the most because the scale of networks in a new epoch is smaller than the previous one. We assume that the network is intense in which every node will be structural hole node, then finding all structural hole nodes will take $O(n^3)$ which is the upper bound of the algorithm.

### B. THE GREEDY INFLUENCE MAXIMIZATION ALGORITHM

In this paper, we call algorithm2 SG for short. Firstly, we use the SHVC algorithm to calculate the *SH* of each node and narrow down the candidate set. After getting the hole value of all nodes, we combine the structural holes attribute and nodes influence to remove those nodes which are less influential than the given threshold and non structural holes. We can select seeds on a smaller new graph. We define $\alpha$ ($0 < \alpha < 1$) as the weight of structure. The larger $\alpha$ means we pay more attention to the structural influence of a node. We combine structural hole value and the influence of nodes to select seeds. The pseudo-code is described as follows:

The time complexity of SHVC is $O(n^3)$. The time complexity of SG is $O(kRn^3)$, here R is the times of Monte-Carlo Simulation. The comparison of our algorithm with traditional greedy algorithm will be given in next section.

## V. EXPERIMENTS AND ANALYSIS

In this section, we verify the efficiency and correctness of our algorithms by conducting a set of experiments on two datasets.

### A. EXPERIMENT SETTING

#### 1) DATASET

We use Twitter [27] and RayLeague [28] as our datasets. As listed in Table I, we select 10 thousands nodes and 342732 edges from Twitter and 16 thousands nodes and 235440 edges from RayLeague.

**Algorithm 2** Structure-Based Greedy (SG)

Input: G(V, E,S,W), k, $\alpha$,$\theta$
Output: S ( |S| = k)
S=$\emptyset$
$G_1(V_1,E_1,W_1,S)\leftarrow$SHVCalculate(G,$\theta$);
for $i = 1$ to $k$ do
4) $CI_v = 0$; (v$\in$V)
5) for $j = 1$ to R do
6) for all v$\in$V do
7) $\sigma_v(S)= \sigma_v(S)\cup\{v\}-\sigma(S)$
8) $CI_v = CI_v+\alpha SH_v+(1-\alpha)\sigma_v(S)$
9) end for
10) end for
11) $v_{max}$=max$CI_v$/R
12) S=S$\cup${ $v_{max}$}
13) end for
14) return S.

**TABLE 1. Experiments Dataset**

| Dataset | Number of nodes | Number of edges |
|---|---|---|
| Twitter | 10000 | 342732 |
| RayLeague | 16000 | 235440 |

#### 2) EXPERIMENTS ENVIRONMENT

All experiments are conducted on Ubuntu-kylin16.04 with C++. The machine settings are: Intel(R) Core(TM) 2 Duo CPU, 2GB RAM.

**TABLE 2. Structural Holes Detection Algorithms**

| Algorithm | Source |
|---|---|
| DGSH | [23]EndeZhang. Generalized Structural Holes Finding Algorithm by Bisection in Social Communities. |
| HIS | [25]Tiancheng Lou.Mining structural hole spanners through information diffusion in social networks. |
| MAXD | The same as the above row |
| SHF | This paper. |

Firstly, we compare the structural hole finding algorithm presented in this paper called SHF with the existing structural holes detection algorithms (DGSH, HIS, MaxD) as shown in table II in terms of time complexity. Secondly, we analyze how the alpha and theta values will influence the algorithm performance like time and influence range. Lastly, we compare our Structure-based Greedy algorithm with CELF, NewGreedy, traditional Greedy and Degree Discount algorithm as shown in table III in terms of time complexity and influential range of seeds. In all experiments, the total rounds of Monte Carlo simulation is set to be 20000.

**TABLE 3.** Influence MAXIMIAZTION Algorithms

| Algorithm | Source |
|---|---|
| Greedy（G） | [6]Kempe D.Maximizing the spread of influence through a social network. |
| NewGreedy(NG) | [8]ChenW.Efficient influence maximization in social networks. |
| CELF | [7]Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks. |
| DegreeDiscount(DD) | [8]ChenW.Efficient influence maximization in social networks. |
| Strcture-based Greedy(SG) | This paper |

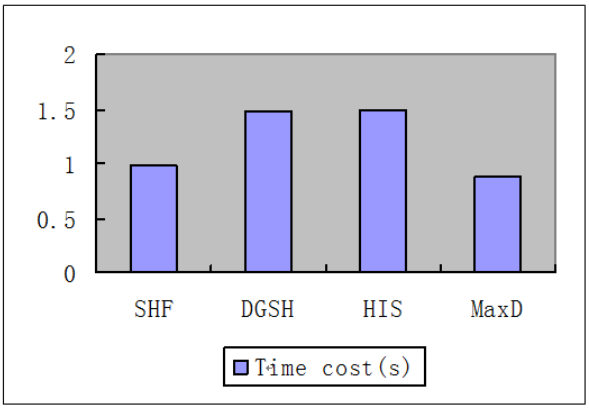



**FIGURE 7.** The time cost of structural hole detection algorithms.

### B. EXPERIMENT RESULTS

#### 1) STRUCTURAL HOLE DETECTION ALGORITHMS

We compare the performance of SHF, DGSH, HIS and MaxD in terms of time cost. We select 1000 nodes and the corresponding 12214 edges from Twitter. We select the first K = 100 detected structural hole nodes and compare them with nodes detected from existing algorithms. As shown in Fig.7, our SHF is the fastest one among all the structural hole detecting algorithms. This is because SHF builds the laplacian matrix and uses the Federal vector to determine the optimal structure hole node which can save a lot of time.

#### 2) EFFECTS OF ALPHA AND THETA

In this part, we evaluate influence range by varying the values of alpha and theta. We conduct experiments on RayLeague which contains 262111 nodes and 1234877 edges. We first select K=100 nodes with theta = 0.001 and comparing influential range and time cost with different alpha. The results are shown in Fig. 8(a). When alpha is closed to 0.5, the influence range reached to the maximum. However, alpha has no effect on time costs when selecting seeds.

In addition, we use the same dataset and set alpha to be 0.5, K = 100 to evaluate the effect of theta on performance. As shown in Fig.8(b),when theta is in range of 0.001 to 0.1, a better performance will be got in terms of time and influential range.

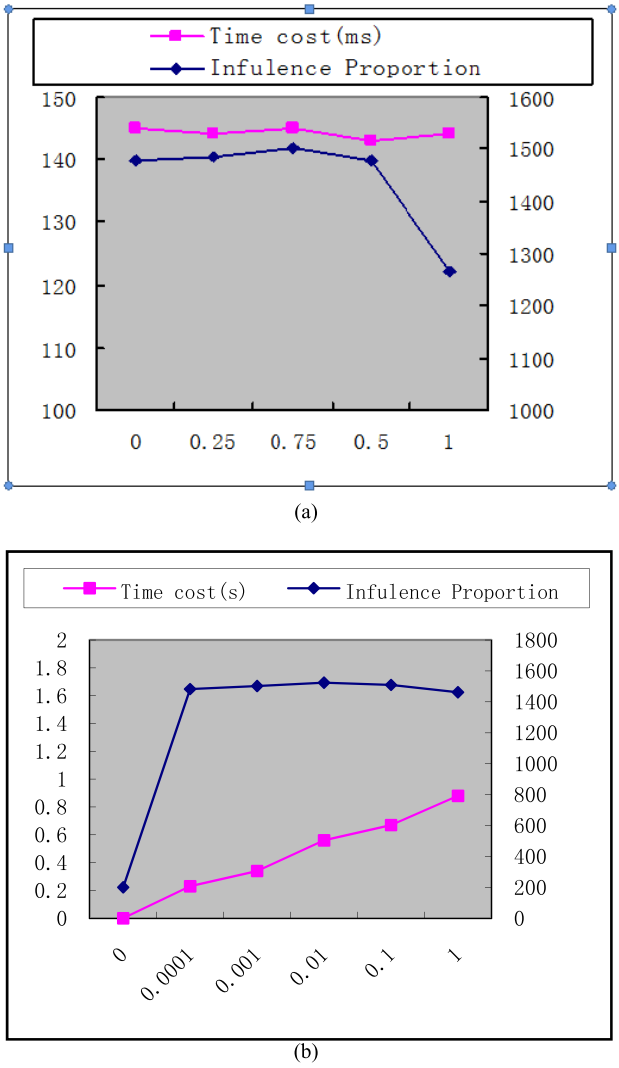
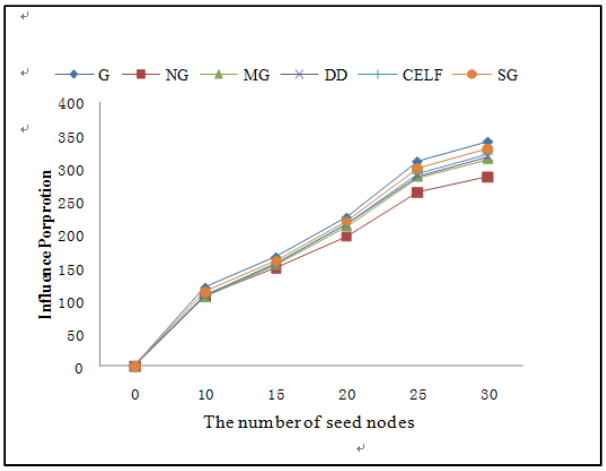


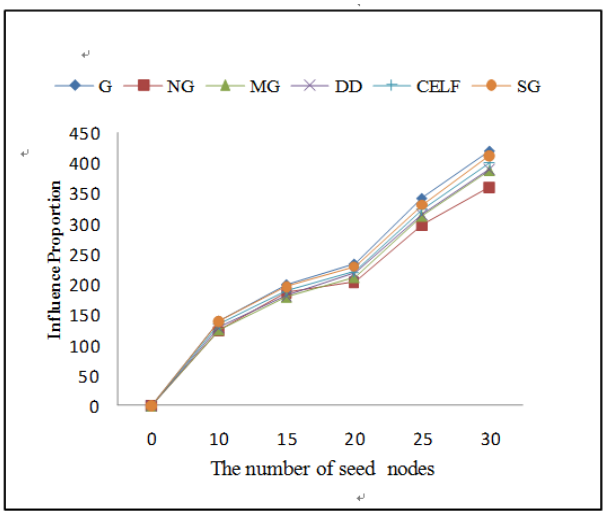

**FIGURE 8.** (a). The effect of alpha on algorithm performance. (b). The effect of theta on algorithm performance.

#### 3) COMPARE OF DIFFERENT INFLUENCE MAXIMIZATION ALGORITHMS

The first group of experiments aims to compare the traditional influence maximization algorithms with our proposed structural hole theory-based influence maximization algorithm in terms of time costs and influential range. We select 500 nodes and 1246 edges from Twitter and RayLeague respectively. We set theta as 0.001, alpha is 0.5. Experiments results are shown in Fig.9 to Fig.10, we compare the influential range of seeds from the above algorithms. As described in Fig.9(a) and Fig.9(b), our algorithm can influence more nodes because it considers nodes structure in network, nodes which occupy structural hole are better than non structural hole to propagate information.




(a)

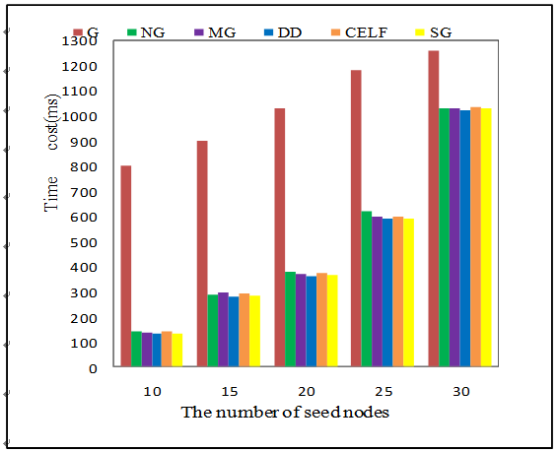



(b)

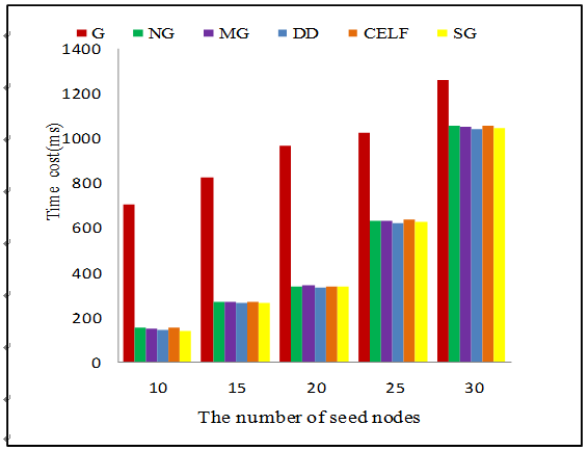
**FIGURE 9.** **(a).Results on Twitter Dataset. (b).Results on RayLeague dataset.**

The second group of experiments is designed to compare the time efficiency of various algorithm. MG algorithm is the fastest algorithm in which approximation ratio is guaranteed at present. From Fig.10(a) and Fig.10(b), we can see that our algorithm cost less time than MG. Therefore, according the above group of experiments, we can infer that our SG is better than MG in terms of time efficiency and results quality. As we can see, CELF has the perfect time efficiency while the seeds from CELF is not the best because it doesn't care about structural features.

To reduce time cost of iteratively computing the structural hole values of nodes SH, we adopt dynamic programming technique which can decrease the time complexity to O(n). Although SG costs more time than CELF, CELF is better than SG in selecting seeds. In fact, people hope to get the widest influential range by sacrificing some running time. The results show that our structure hole theory-based influence maximization algorithm has faster speed and larger influential range.




(a)




(b)

**FIGURE 10.** **(a). Time cost of different algorithm on Twitter Dataset. (b). Time cost of different algorithm on RayLeague Dataset.**

## VI. CONCLUSION AND FUTURE WORK

In this paper, we propose an novel algorithm to solve the problem of influence maximization based on structure hole theory. We defined the most influential nodes to be nodes that both have the strong ability of propagation and are structural hole nodes. We prove that structure hole theory-based influence maximization problem is NP-hard. To evaluate the structural importance of nodes in networks, we propose *SHVC* to compute SH of nodes. To solve the structure hole theory-based influence maximization problem, we propose Structure-based Greedy (*SG*) to select nodes which have strong ability of propagation and are with larger SH as seeds. We conduct experiments to verify the time efficiency and accuracy of our algorithm. Results on Twitter and RayLeague show that comparing with existing algorithms, our algorithm can solve influence maximization problem effectively and improve not only the influential range but time efficiency.

Social networks keep updating which makes the structural feature of nodes in networks keep changing. In the future, we'll extend our structure hole theory-based influence maximization algorithm to large scale dynamic networks and improve the scalability of our algorithm.

## REFERENCES

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *J. Comput.-Mediated Commun.*, vol. 38, no. 3, pp. 16–31, 2007.
[2] S. Xie *et al.*, "Research on topic-based local influence maximization algorithm in social network," *J. Front. Comput. Sci. Technol.*, vol. 10, no. 5, pp. 646–656, 2016.
[3] C. Wei, W. Yajun, and Y. Siyu, "Efficient topic-aware influence maximization using preprocessing," *CoRR*, vol. 9197, pp. 1–13, Mar. 2014.
[4] J. L. Z. Cai, M. Yan, and Y. Li, "Using crowdsourced data in location-based social networks to explore influence maximization," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFORM)*, Apr. 2016, pp. 1–9.
[5] L. Qian-Qian, "Research on mining structural hole spanners in weighted networks," Anhui University, Hefei, China, Tech. Rep., 2015, pp. 24–38.
[6] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 57–66.
[7] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 137–146.
[8] J. Leskovec *et al.*, "Cost-effective outbreak detection in networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 420–429.
[9] Z. He, Z. Cai, J. Yu, X. Wang, Y. Sun, and Y. Li, "Cost-efficient strategies for restraining rumor spreading in mobile social networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2789–2800, Mar. 2017.
[10] Z. He, Z. Cai, and X. Wang, "Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks," in *Proc. 35th IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2015, pp. 205–214.
[11] L. Xiao-Dong, "Research on efficient processing techniques of influence maximization in large-scale," Nat. Univ. Defense Technol., Changsha, China, Tech. Rep., 2013, p. 5.
[12] S. Chen, J. Fan, G. Li, J. Feng, K.-I. Tan, and J. Tang, "Online topic-aware influence maximization," *Proc. VLDB Endowment*, vol. 8, no. 6, pp. 666–677, 2015.
[13] W. Chen, T. Lin, and C. Yang, "Efficient topic-aware influence maximization using preprocessing," *CoRR*, vol. 9197, pp. 1–13, Mar. 2014.
[14] G. Li, S. Chen, J. Feng, K.-L. Tan, and W.-S. Li, "Efficient location-aware influence maximization," in *Proc. SIGMOD*, 2014, pp. 87–98.
[15] X. Wang, Y. Zhang, W. Zhang, and X. Lin, "Distance-aware influence maximization in geo-social network," in *Proc. ICDE*, 2016, pp. 1–12.
[16] X. Zheng, Z. Cai, J. Li, and H. Gao, "Location-privacy-aware review publication mechanism for local business service systems," in *Proc. 36th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2017, pp. 1–12.
[17] H. Li, S. S. Bhowmick, A. Sun, and J. Cui, "Conformity-aware influence maximization in online social networks," *VLDB J.*, vol. 24, no. 1, pp. 117–141, 2015.
[18] 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.
[19] Z. He, Z. Cai, Q. Han, W. Tong, L. Sun, and Y. Li. "An energy efficient privacy-preserving content sharing scheme in mobile social networks," *Pers. Ubiquitous Comput.*, vol. 20, no. 5, pp. 833–846, 2016.
[20] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Trans. Depend. Sec. Comput.*, to be published.
[21] X. Zheng, Z. Cai, J. Li, and H. Gao, "A study on application-aware scheduling in wireless networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 7, pp. 1787–1801, Jul. 2017.
[22] R. S. Burt, "Structural holes and good ideas," *Amer. J. Sociol.*, vol. 110, no. 2, pp. 349–399, 2010.
[23] E. Zhang, G. Wang, K. Gao, X. Zhao, and Y. Zhang, "Generalized structural holes finding algorithm by bisection in social communities," in *Proc. IEEE 6th Int. Conf. Genet. Evol. Comput.*, Aug. 2012, pp. 276–279.
[24] S. Xiao-Ping and S. Yu-Rong, "Leveraging neighborhood 'structural holes' to identifying key spreaders in social networks," *Acta Phys. Sinica*, vol. 64, no. 2, p. 020101, 2015, doi: 10.7498/aps.64.020101.
[25] T. Lou and J. Tang, "Mining structural hole spanners through information diffusion in social networks," in *Proc. Int. Conf. World Wide Web*, 2013, pp. 825–836.
[26] L. Fang and Z. Bi-Chun, "Two classical methods of taking characteristic value of matrix," *J. Yanbei Normal Univ.*, vol. 17, no. 6, pp. 11–12, 2001.
[27] *Twitter 2012 Facts and Figures (Infographic)*. Accessed: 2012. [Online]. Available: http://www.website_monitoring.com/blog/2012/11/07/twitter_2012_facts_and_figures_infograp-hic/
[28] (May 2016). *Rayleague Dataset*. [Online]. Available: http://www.rayleague.com/

**JINGHUA ZHU** received the B.S. degree in computer software and the M.S. degree in computer science in 1999 and 2002, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology in 2009. She has been a Professor with the School of Computer Science and Technology, Heilongjiang University, China, since 2016. She has published many high-quality conference and journal research papers. Her research interests include social networks, data mining, uncertain databases, and wireless sensor networks.

**YONG LIU** received the B.S. degree in computer software and the M.S. degree in computer science from Heilongjiang University, in 1998 and 2001, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology in 2010. He has been Associate Professor with the School of Computer Science and Technology, Heilongjiang University, China, since 2010. He has published many high-quality conference and journal research papers. His research interests include social networks and data mining.

**XUMING YIN** received the M.S. degree in computer science from Heilongjiang University in 2016. His research interests include social networks and data mining.