Identification of influential nodes in social networks with community structure based on label propagation

Yuxin Zhao[a,c,*], Shenghong Li[a,b], Feng Jin[c]

[a]*Department of Electronic Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China*
[b]*School of Information Security Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China*
[c]*IBM China Research Laboratory, 399 Ke Yuan Road, Shanghai 201203, China*

Abstract

Social network is an abstract presentation of social systems where ideas and information propagate through the interactions between individuals. It is an essential issue to find a set of most influential individuals in a social network so that they can spread influence to the largest range on the network. Traditional methods for identifying influential nodes in networks are based on greedy algorithm or specific centrality measures. Some recent researches have shown that community structure, which is a common and important topological property of social networks, has significant effect on the dynamics of networks. However, most influence maximization methods do not take into consideration the community structure in the network, which limits their applications on social networks with community structure. In this paper, we propose a new algorithm for identifying influential nodes in social networks with community structure based on label propagation. The proposed algorithm can find the core nodes of different communities in the network through the label propagation process. Moreover, our algorithm has low time complexity, which makes it applicable to large-scale networks. Extensive experiments on both synthetic and real-world networks under common diffusion models demonstrate the effectiveness and efficiency of our proposed algorithm.

Keywords: Social network, Influential node, Community structure, Label

*Yuxin Zhao
Email address: zhaoyuxin1@sjtu.edu.cn (Yuxin Zhao)

propagation
PACS: 89.75.Fb, 05.10.-a

1. Introduction

Social network is an abstract representation of real-world social systems consisting of large numbers of individuals and relationships between individuals [1, 2, 3]. The natures of social network are the interactions between different individuals, which leads to the spread of ideas, information and influences in the network [4]. With the increasing popularity of online social networks, such as Facebook, Twitter, MicroBlog and WeChat, it has become an effective and promising marketing strategy by conducting product promotions through social influences among individual cycles of friends and families [5]. A motivating application is the viral marketing, which aims to select a small number of influential users to adopt a product, and subsequently trigger a large cascade of further adoptions by utilizing the "Word-of-Mouth" effect in social networks.

Motivated by this background, an essential issue that has received considerable attention is the influence maximization problem, which aims to find a set of most influential individuals in a social network so that they can spread influence to the largest number of nodes in the network [7, 8, 9, 10]. Formally, influence maximization problem can be described as follows: given a positive integer k, identify a node set containing k nodes to maximize the influence effect under specific diffusion model, where the influence effect is quantitatively measured by the expected number of influenced nodes during the whole spreading process [7].

In recent years, a number of methods for finding the influential nodes in networks have been proposed to solve the influence maximization problem. These influence maximization methods can be roughly classified into two categories: centrality-based algorithms and greedy algorithms. Centrality-based algorithms evaluate the centrality or importance of the nodes according to some topological measures and identify the nodes with largest centrality as the influential nodes [11, 12]. Degree centrality [13], betweenness centrality [14] and closeness centrality [15] are the most basic centrality measures. Many other complicated centrality measures [16, 17, 18, 19] to identify core nodes in the network are also proposed from different perspectives. On the other hand, greedy algorithms formulate the influence maximization problem

as a discrete optimization problem and use greedy strategy to achieve the approximate optimal solution. Kempe et al. [7] first came up with a greedy algorithm to solve the influence maximization problem and used Monte-Carlo simulations to estimate the influence scope of initial node sets. However, this greedy algorithm is extremely time consuming and only applicable to small networks. To reduce the calculation complexity, Leskovec et al. [20] put forward the CELF("cost-effective lazy-forward") method to avoid redundant calculations of influence scope according to the submodularity property of influence spreading. Some other methods [21, 22, 23, 24, 25] estimate the influence scope using some heuristic strategies instead of Monte-Carlo simulations, which can effectively improve the time efficiency.

Community structure is a common and important property of real-world networks [26]. A community can be generally described as a group of nodes with dense internal connections and relatively sparse connections to the nodes in other groups [27, 28, 29]. Community structure is beneficial to understand the function and organization of social networks, since communities often correspond to real social associations and organizations. Some recent researches have shown that community structure has important effect on the spreading process in networks [30, 31, 32]. However, most existing influence maximization methods do not take into account the influence of community structure in the network, which limits their applications on social networks with community structure.

In this paper, we propose an algorithm to identify influential nodes in social networks with community structure based on label propagation. Our main contributions are summarized as follows:

1. We successfully introduce the label propagation process to identify the influential nodes in social networks with community structure.
2. Our proposed algorithm is parameter-free and requires no prior information about the community structure. It also exhibits very low time complexity, which makes it applicable to large-scale networks.
3. The tests on both synthetic and real-world networks under common diffusion models demonstrate the effectiveness and efficiency of our algorithm.

The rest of the paper is organized as follows. Section 2 makes an introduction of the related work of influence maximization problem. In Section 3, we introduce the basic concept of community structure in networks and the label propagation process. Our proposed method for identifying influential

nodes in social networks with community structure is described in detail in Section 4. The experimental results and discussions are reported in Section 5. Finally, Section 6 gives the conclusion of this paper.

2. Influence maximization

Influence maximization problem is how to identify a node set containing k nodes to maximize the influence effect where influence is propagated in the network under specific diffusion model [7]. Here, we first introduce the diffusion model adopted in our paper, and then give a brief review of the existing influence maximization methods.

2.1. Diffusion model

Diffusion model is the key to simulate the actual spreading process of ideas and information in social networks. In our study, we employ two widely-used diffusion models, namely independent cascade model (IC model) and linear threshold model (LT model). In both diffusion models, any node in the network can stay in one of two states: active state and inactive state. Active state indicates that the corresponding individual adopts the information, while inactive state contrarily means that the individual does not accept the information. A node can convert from inactive state to active state under the influence of its neighbor nodes in the network, but cannot convert in the opposite direction. The spreading process starts from an initial set of active nodes and unfolds in discrete time steps.

Let a social network be represented by a graph $G = (V, E)$, where V and $E = \{(u, v)|u, v \in V\}$ are respectively the set of nodes and edges in the network. Given the initial active node set $S \subset V$, the number of active nodes at the end of the spreading process is denoted by a variable $\varphi(S)$. The influence effect is measured by $\sigma(S)$, which is defined as the expected value of $\varphi(S)$. We call $\sigma(S)$ the influence scope of S. Since $\sigma(S)$ is very difficult to calculate precisely, Monte-Carlo simulations are used to estimate the influence scope $\sigma(S)$ in practical calculation.

2.1.1. Independent cascade model

In independent cascade model (IC model) [33, 34], every link (u, v) in the network is associated with a diffusion probability $p_{u,v}$, which indicates the probability that node u successfully actives node v. Any node in the network attempts to activate its inactive neighbors only at the time that it just gets

activated. When the initial active node set S is given, the spreading process under IC model unfolds according to the following rule. At each time step t, if node u is newly activated at time step $t-1$, it can make an attempt to activate every inactive neighbor node v with probability $p_{u,v}$. If the attempt successes, node v would become active at time step $t+1$; otherwise, it still stays inactive. No matter whether the attempt succeeds or not, node u can never try to activate other nodes at later time steps. When more than one newly activated node tries to activate a node, these activation attempts are independent from each other and can be proceed sequentially in an arbitrary order. The spreading process finally terminates if there are no more newly activated nodes in the network.

In our paper, due to the lack of information, the diffusion probabilities of different links in the IC model are set to a uniform value, $p_{uv} = p,\ \forall (u,v) \in E$.

2.1.2. Linear threshold model

In linear threshold model (LT model) [35, 36], every link (u,v) in the network is associated with a weight $w_{u,v}$, which indicates the influence that node u exerts on node v. The weights of links satisfies the constraint: $\sum_{v \in N(u)} w_{u,v} \leq 1, \forall u \in V$, where $N(u)$ is the set of neighbor nodes of node u in the network. Given an initial active node set S, the spreading process under LT model unfolds according to according to the following rule. First, each node in the network is assigned with a random threshold in the range [0, 1]. The threshold reflects the tendency of the node to convert to active state, so that it is harder to make the node with larger threshold get activated. Then, at each time step t, each inactive node v in the network is activated only if the influence sum of all its active neighbor nodes exceeds its corresponding threshold, i.e., $\sum_{u \in \Gamma(v)} w_{u,v} \geq \theta_v$, where $\Gamma(v)$ is the set of active neighbor nodes of node v and θ_v is the threshold for node v. This spreading process continues until no more activations are possible in the network.

2.2. Influence maximization methods

Most existing influence maximization methods can be roughly classified into two categories: centrality-based algorithms and greedy algorithms.

2.2.1. Centrality-based algorithm

The basic idea of centrality-based algorithm is to evaluate the importance of nodes in the network using some centrality measures and taking the nodes with large centrality as the initial active nodes.

Degree centrality is the simplest centrality measure. It is believed that, in social networks with broad degree distribution, the most connected people are the hubs for extensive influence spreading [13]. Although degree centrality is proven to be related with influence scope to some extend, but it has great bias since it dose not consider the whole network topology. Betweenness centrality and closeness centrality are two other typical centrality measures for influence maximization problem. Betweenness centrality [14] is defined as the number of shortest paths crossing through the node, which reflects the interpersonal influence a person put on others in social network theory [11]. Closeness centrality [15] is defined as the reciprocal of the sum of geodesic distances of one node to all the other nodes in the network, which is a measure of how long it takes to spread information from a given node to other reachable nodes in the network.

Chen et al. [16] proposed the local centrality which is an extension of degree centrality to local neighborhood. We et al. [17] identified influential nodes in the network based on the coritivity theory of complex network. Coritivity theory measures the importance of a node set by the number of connected components showing up after deleting the nodes and their incident edges from the network. Kitsak et al. [18] distinguished the nodes with different influential degree using the k-shell decomposition. The influential nodes identified by k-shell decomposition do not correspond to the nodes with high degree and betweenness, but are located in the center of network topology. Lu et al. [19] improved the well-known PageRank algorithm [37] to find leaders in social networks, which especially performs well on directed networks.

2.2.2. Greedy algorithm

Kempe et al. [7] made the first attempt to solve the influence maximization problem using a simple greedy algorithm. This greedy algorithm uses large numbers of Monte-Carlo simulations to estimate the influence scope and takes the influence scope as the optimization objective. A hill climbing strategy is adopted to pursue the optimal solution. At each iteration, the node with maximal marginal gain of influence scope is added to the initial active node set. It has been demonstrated that the greedy algorithm is $(1 - 1/e - \epsilon)$ optimal for the influence maximization problem, where e is the base of the natural logarithm and ϵ is a very small positive real number. Thus, the influence scope achieved by the greedy algorithm is at least better than 63% of actual global optimum. Extensive experiments have shown that

the greedy algorithm significantly outperforms the typical centrality-based algorithms in influence scope.

However, the main drawback of the greedy algorithm is the high computational complexity. Let n and m be respectively the number of nodes and edges in the network. Each simulation of spreading process takes $O(m)$ time, so that the calculation of influence scope for any initial active node set requires $O(Rm)$ time, where R is the number of repeated Monte-Carlo simulations. Therefore, the total time complexity of the greedy algorithm is $O(knRm)$.

In order to improve the calculation efficiency of greedy algorithm, Leskovec et al. [20] presented a CELF("cost-effective lazy-forward") method to avoid redundant calculations of influence scope according to the submodularity property of influence spread. In CELF greedy algorithm, the marginal gain of influence scope for a node does not need to be re-calculated if its value at previous iteration is already smaller than that of any other node at the current iteration. CELF greedy algorithm can achieve approximately same solutions as original greedy algorithm with much faster calculation speed.

Many other methods try to estimate the influence scope using some heuristic strategies instead of Monte-Carlo simulations. Kimura and Saito [21] exerted a strict constraint to the IC model that each node can be only activated at specific time steps and deduced the mathematical expression of influence scope. Chen et al. proposed the DiscountIC [22] algorithm which estimates the marginal gain of influence scope under the IC model using degree discount heuristics. Chen et al. [23] and Goyal et al. [24] estimated the influence scope under the LT model by exploring all the paths within a small neighborhood in the network. Kimura et al. [25] evaluated the marginal influence scope through a bond percolation process in social networks.

3. Community structure

3.1. Definition of Community structure

Community structure is a common and important property of real-world networks [26]. The communities in social networks have great practical significance since they correspond to the real associations or organizations. Therefore, the identification of community structure can provide an important insight into the function, organization and dynamics of social networks. Generally, a community can be defined as a group of nodes with dense internal connections and relatively sparse connections to the rest of the network

[27, 28, 29]. In order to make a more explicit description of community structure, we introduce the quantitative definitions of community structure which are widely adopted in the literature.

Given a network represented by a simple graph $G = (V, E)$, where V is the set of nodes and $E = \{(u, v)|u, v \in V\}$ is the set of edges between the nodes. The topology of the network is fully specified by the adjacency matrix A, where $A_{uv} = 1$ if node u and node v are directly connected and $A_{uv} = 0$ otherwise.

Radicci et al. [38] proposed two local definitions of community, respectively in a strong sense and a weak sense. Considering a subnetwork $C \subset G$, to which node i belongs, the total degree of node u is split into two contributions: $k_u = k_u^{in}(C) + k_u^{out}(C)$, where $k_u^{in}(C) = \sum_{v \in C} A_{uv}$ is the number of edges connecting node u to the nodes belonging to subnetwork C and $k_u^{out}(C) = \sum_{v \notin C} A_{uv}$ is clearly the number of edges connecting node u towards the rest of the network. The strong community is a subnetwork that satisfies the constraint:

$$k_u^{in}(C) > k_u^{out}(C), \quad \forall u \in C \tag{1}$$

In a strong community, each node has more connections within the community than with the rest of the network. Compared with strong community, weak community is under a relaxed constraint:

$$\sum_{u \in C} k_u^{in}(C) > \sum_{u \in C} k_u^{out}(C) \tag{2}$$

Weak community requires that the sum of node degrees within the community is larger than the sum of node degrees toward the rest of the network. According to the definitions, a strong community is also a weak community, whereas the converse is not true.

3.2. Label propagation

Label propagation [40] is an important dynamics process in networks, which aims at detecting community structure in the network. The procedure of label propagation is conceptually simple. Initially, each node in the network is assigned with a unique label. Then, at each time step, each node adopts the label shared by the maximum number of its neighbors in synchronous or asynchronous fashion. The label updating rule can be formulated as:

$$l_u^{'} = \arg\max_{l} \sum_{v \in \sigma(u)} \delta(l, l_v). \tag{3}$$

where, l'_u is the new label for node u, l_v is the current label for node v, $\sigma(u)$ is the set of the neighbor nodes of node u, $\arg\max_l$ returns label l for which the sum function attains the largest value, and $\delta()$ is the Kronecker delta function, i.e., $\delta(l, l_v) = 1$ when $l = l_v$; otherwise, $\delta(l, l_v) = 0$. If more than one maximal label exists, one of the maximal labels is chosen at random. As the time increases, the densely connected groups of nodes quickly reach a consensus on a unique label and expand outward to take over more nodes. The process finally converges when the label for each node is the maximal label among its neighbors. At the end of label propagation, one label would cover all the nodes in a community and different labels indicate different communities in the network.

The community structure obtained by label propagation approximately accords with the definition of strong community. While strong community requires each node to have strictly more connections within its community than outside the community, the label propagation guarantees that each node has at least as many connections within its community as it has with each of the other communities.

Label propagation is a simple dynamics process and exhibits near linear time complexity $O(Tm)$, where T is the number of time steps during the process and m is the number of the edges in the network. However, the label propagation process have the drawback of weak robustness due to the random nature. For the same community in the network, it may be covered by different labels with different runs of label propagation. Refer to [41] for detailed analysis of the label propagation process.

4. Our method

The spreading process in social networks with community structure has its unique pattern. Due to the structural compactness of the community, information can easily propagate within the community but has little chance to spread outside the community [39]. Therefore, it is necessary to identify the core nodes in different communities as the initial active nodes, so that they can spread influence to the largest range on the network.

Our proposed influence maximization algorithm based on label propagation (IM-LPA) especially aims at identifying influential nodes in social networks with community structure. We introduce the label propagation process to solve the influence maximization problem and propose a novel heuristic that the most influential node of a community could propagate its

label to all the nodes within the community during the label propagation process. The IM-LPA algorithm is composed of two phases: seeding phase and label propagation phase. In the seeding phase, some special nodes are extracted as the seeds of the communities in the network. Then, in the label propagation phase, the algorithm propagates the labels from the seed nodes and measures the centrality of these seed nodes for their communities according to the label propagation process. The details of the algorithm IM-LPA are presented below.

4.1. Seeding phase

Seeding techniques have been adopted in a variety of researches related with community structure in networks [42, 43]. Generally, a seed is a subnetwork which can be thought of as the potential core of a community in the network. The choice of seed is crucial for finding and revealing the community structure in the network. The nodes with high degree or clustering coefficient, fully-connected cliques and maximal cliques are usually taken as the seed of communities in the network [44].

In the seeding phase of the IM-LPA algorithm, we identify some specific nodes as the seeds of communities in the network according to the degree of the nodes in the network. Given a network $G = (V, E)$, where V is the set of the nodes and E is the set of the edges connecting the nodes. The procedure of the seeding phase is described below.

Step 1. Initialize the set of seed nodes $\Omega = \varnothing$, and the set of candidate nodes $W = V$.

Step 2. Calculate the degree k_v of each node v in the network.

Step 3. Find the node with the largest degree in the candidate node set, $v_{max} = \arg\max_{v \in W} k_v$.

Step 4. Add v_{max} to the seed node set, $\Omega = \Omega \cup \{v_{max}\}$.

Step 5. Remove v_{max} and its neighbor nodes $N(v_{max})$ from candidate node set, $W = W \setminus (\{v_{max}\} \cup N(v_{max}))$.

Step 6. Go back to Step 3 and repeat the process until the candidate node set is empty $W = \varnothing$.

Step 7. Output the seed node set Ω.

In social network analysis, the membership contribution of a node to its community has been proven to be highly related with its degree [45]. The node with larger degree is more likely to be the core of the community it

belongs to. The seeding phase of the IM-LPA algorithm can ensure that the chosen seed nodes have relatively large degree. Moreover, the seeding phase also guarantees that the seed nodes are independent from each other, which means that the geodesic distance of any two seed nodes is at least 2. The independence of seed nodes is crucial for the next label propagation phase, because it can eliminate the interference to the label propagation process.

4.2. Label propagation phase

After the seed nodes are extracted from the network, the algorithm IM-LPA expands these seeds to reveal the community structure based on label propagation. Considering that a node may belong to multiple communities in real social networks [46], we allow a node to have more than one label in the label propagation process. Specially, when a node is unlabeled, it has an empty label set.

The label propagation phase follows a simple procedure. First, each seed node is assigned with a unique label and the other nodes in the network are assigned with empty label sets. Then, at each time step t, every node in the network updates its label set synchronously according to the following label updating rule:

$$L_v^{'}(t) = \arg\max_{l} \sum_{u \in N(v)} \delta(l, L_u(t-1)). \tag{4}$$

where $L_v(t)$ is the label set for node v at time step t, $L_u(t-1)$ is the label set of node u at time step $t-1$, $N(v)$ is the set of the neighbor nodes of node v, $\arg\max_l$ returns the set of labels for which the sum function attains the largest value, and $\delta()$ is the extended Kronecker delta function, i.e., $\delta(l, L_u) = 1$ when $l \in L_u$; otherwise, $\delta(l, L_u) = 0$. Finally, the label propagation process terminates when the label set of any node in the network no longer changes.

The label updating rule of the IM-LPA algorithm is similar to that of the basic label propagation. The main difference lies on that when more than one maximal label exists, one of them is chosen at random in basic label propagation while all of them are retained in the new label set in the IM-LPA algorithm. For basic label propagation, the random choices of label happens very frequently at the first few iterations since each node initially has a unique label. This is the main reason that makes basic label propagation often achieve poor and unreasonable performances. Different from basic label propagation, the IM-LPA algorithm only initially assigns every

seed node with a unique label, which removes large numbers of misleading labels. Moreover, the improvement of label updating rule also eliminates the influence of the randomicity during the label propagation process. Thus, the label propagation in the IM-LPA algorithm is much more robust and stable than basic label propagation.

At the beginning of the label propagation process in the IM-LPA algorithm, the labels expand from the initial seed nodes and attempts to acquire more nodes in the neighborhood. When different labels reach on the same node, they start to compete for the occupation of the node, and the node only adopts the most frequent label or labels among its neighbors. We argue that the label derived from the most influential node of a community would finally defeat other labels and occupy all the nodes within the community. And the label derived from less important node would be taken over by the labels from most influential nodes and gradually disappear in the label propagation. At the end of the label propagation process, the nodes in the same community would be associated with the label from the most influential node of the community. In addition, the nodes with multiple labels indicate the overlaps between the communities in the network. Therefore, the whole label propagation process can reflect the centrality of different nodes in the network.

Let $l^{(v)}$ denote the label initially assigned to seed node v. According to the label propagation process, we define a process variable $N(v,t)$ denoting the number of nodes associated with label $l^{(v)}$ at time step t. Then, the centrality of each seed node can be measured as:

$$C(v) = \max_{0 \le t \le T} N(v,t) \tag{5}$$

where, $C(v)$ is the centrality of node v, T is the final time step of the label propagation process. The centrality reflects the importance and influence of a node for the community which it belongs to. The centrality of the most influential node for each community is equal to the number of nodes in the community. While, the centrality of less important node is determined by the compactness of the local structure near the node. After ranking these seed nodes according to the centrality, the top k nodes are identified as the most influential nodes for influence maximization.

4.3. Overview of the IM-LPA algorithm

By combining the seeding phase and label propagation phase, we can outline the IM-LPA algorithm as follows:

Algorithm IM-LPA

Input

- $A_{n\times n}$: the adjacency matrix of the network $G = (V, E)$, where V is the set of the nodes, E is the set of the edges connecting the nodes and $n = |V|$ is the number of nodes in the network. $A_{uv} = 1$ if node u and node v are directly connected; otherwise, $A_{uv} = 0$.
- k: the number of initial active nodes.

Output

The initial active nodes set S.

Method

Seeding Phase

Step 1: Initialize the set of seed nodes $\Omega = \varnothing$, and the set of candidate nodes $W = V$.

Step 2: Calculate the degree k_v of each node v in the network.

repeat

Step 3: Find the node with the largest degree in the candidate node set, $v_{max} = \arg\max_{v\in W} k_v$.

Step 4: Add v_{max} to the seed node set, $\Omega = \Omega \cup \{v_{max}\}$.

Step 5: Remove v_{max} and its neighbor nodes $N(v_{max})$ from candidate node set, $W = W \setminus (\{v_{max}\} \cup N(v_{max}))$.

until The candidate node set is empty $W = \varnothing$

repeat

Step 6: Assigned each seed node $v \in \Omega$ with a unique label $l^{(v)}$, and assign other nodes in the network with empty label sets.

Step 7: Update the label of each node in the network using Eq. (4) in synchronous fashion.

until The label set of any node in the network no longer changes.

Step 8: Calculate the centrality of each seed node $v \in \Omega$ using Eq. (5).

Step 9: Select the top k nodes with the largest centrality as the initial active nodes set S.

The proposed IM-LPA algorithm uses the label propagation process to identify the influential nodes in social networks with community structure. It can reveal the community structure in the network and measure the centrality of different nodes for their communities. The IM-LPA algorithm is

parameter-free and requires no prior information of the community structure in the network.

4.4. Time complexity analysis

We also make an analysis on the time complexity of the proposed IM-LPA algorithm. Let n and m respectively be the number of nodes and edges in the network, $\bar{d}$ be the average degree of the nodes which satisfies $O(m) = O(n) \cdot O(\bar{d})$.

1. In the seeding phase, it takes $O(m)$ time to calculate the degree of the nodes in the network. Ranking the nodes according to the degree also needs $O(m)$ time. Thus, the time complexity of the seeding phase is $O(m)$.
2. In the label propagation phase, each node has $O(\bar{d})$ neighbors and at most $O(\bar{d})$ labels in its label set. In the worst case, it costs $O(\bar{d}^2)$ time for one node to update its label set, since it needs to traverse $O(\bar{d}^2)$ labels among the neighbor nodes. Therefore, each time step of label propagation process requires $O(n\bar{d}^2)$ time and the total time of the label propagation phase is $O(Tn\bar{d}^2) = O(Tm\bar{d})$, where T is the number of time steps during the label propagation. According to our experiments, the value of T is generally very small compared with the size of the network.
3. The total time complexity of the IM-LPA algorithm is $O(m\bar{d})+O(Tm\bar{d}) = O(Tm\bar{d})$. For sparse networks, the time complexity is therefore $O(Tm)$, which is near linear with the number of edges in the network.

5. Experimental results and discussions

We have tested our proposed IM-LPA algorithm in comparison with other influence maximization methods on both synthetic benchmark networks and real-world social networks. The experiments adopt two basic diffusion models: independent cascade (IC) model and linear threshold (LT) model, as described in Section 2.1. For the IC model, the diffusion probability is uniformly set as $p = 0.025$, 0.05 and 0.1. The performances of different influence maximization methods is measured by the fraction of active nodes at the end of the spreading process, which is calculated over 10000 independent simulations. All the experiments are implemented by MATLAB 2009b running on a PC with a 2.7GHz processor and 3GB memory.

The influence maximization algorithms for comparison include the CELF greedy algorithm [20], degree centrality, betweenness centrality [14], closeness

centrality [15], local centrality [16], k-shell decomposition [18] and PageRank algorithm [37]. For the CELF greedy algorithm, the influence scope is estimated by 10000 simulations. For PageRank algorithm, the damping factor d is set to 0.85.

For all the test network, we also use a measure, which is called modularity [47], to evaluate the significance of the community structure in the network. Formally, modularity can be formulated as:

$$Q = \frac{1}{2m} \sum_{u,v \in V} (A_{uv} - \frac{k_u k_v}{2m}) \delta(u, v) \tag{6}$$

where, m is the number of edges in the network, A_{uv} is the element of the adjacency matrix for the network, k_u is the degree of node u, and $\delta(\cdot)$ is the extended Kronecker delta function, i.e., $\delta(u, v) = 1$ if node u and v are in the same community; otherwise, $\delta(u, v) = 0$. The term $k_u k_v / 2m$ indicates the expected number of edges connecting node u and node v in a random network of the same size and node degree distribution. If the number of edges within communities is greater than the expected number in a random network, the modularity value Q would be greater than 0. Larger value of modularity indicates more significant community structure in the network.

5.1. Synthetic networks

We first use the LFR benchmark model introduced by Lancichinetti et al. [48] to construct synthetic networks with community structure. A number of parameters are used to constrain the topological structure of the LFR benchmark networks. Both the node degree and the community size follow the power-law distribution, as commonly observed in real-world networks. The significance of the community structure is determined by a critical mixing parameter μ, which denotes the average fraction of the connections to other communities per node. The smaller value of mixing parameter μ leads to more significant communities. Fig. 1 shows the relationship between the mixing parameter μ and the modularity Q in LFR benchmark networks. We can see that the modularity Q decreases as the parameter μ increase from 0 to 1, namely the community structure in the network is gradually weakened with the increasing of the parameter μ.

Our experiments construct a series of LFR benchmark networks with obvious community structure. The network size is set to 1000 and the mixing parameter μ is set to 0.1. The power-law exponents of node degree and

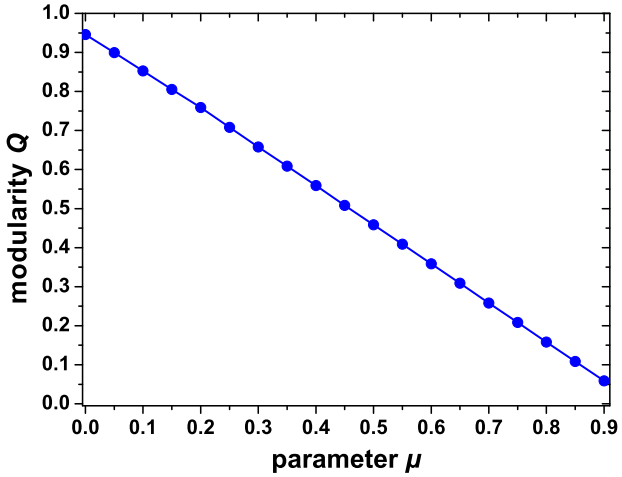


Figure 1: the relationship between the mixing parameter μ and the modularity Q in LFR benchmark networks.

community size are set to 2 and 1 respectively. The node degree is between [1, 50] and has an average value of 20. The first set of LFR benchmark networks contain small communities, whose sizes are in the range [20, 50]. In the first set of LFR benchmark networks, most influential nodes identified by the IM-LPA algorithm are the core nodes of different communities in the networks. Fig. 2 shows the performance of the IM-LPA algorithm on the first set of LFR benchmark networks in comparison with other algorithms, where each data point is an average over 10 different networks.

As is shown in Fig. 2, under IC model with $p = 0.025$ and $p = 0.1$, the greedy algorithm performs the best among all the algorithms. Our IM-LPA algorithm is only inferior to the greedy algorithm and shows significant advantages over the other algorithms in most situations. Under IC model with $p = 0.05$ and LT model, the IM-LPA algorithm performs the best among all the algorithms. It is slightly better than the greedy algorithm and greatly outperforms the other algorithms.

The second set of LFR benchmark networks contain large communities, whose sizes are in the range [100, 200]. In the second set of LFR benchmark networks, many influential nodes identified by the IM-LPA algorithm are in the same community since there exist only a few communities in the network. Fig. 3 shows the performance of the IM-LPA algorithm on the second set of LFR benchmark networks in comparison with other algorithms, where each

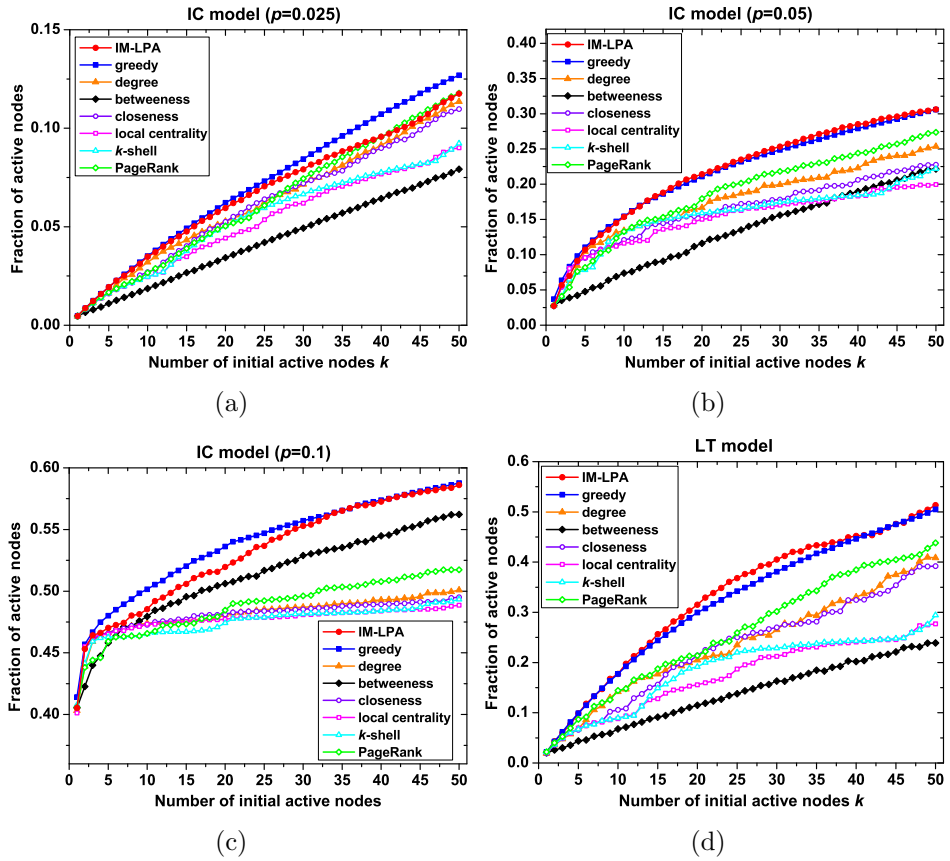


Figure 2: The influence spreading of different algorithms on LFR benchmark networks with small communities. (a) IC model, $p = 0.025$; (b) IC model, $p = 0.05$; (c) IC model, $p = 0.1$; (d) LT model.

data point is still an average over 10 different networks.

From Fig. 2, under IC model with $p = 0.025$ and $p = 0.05$, the performances of the IM-LPA algorithm, the greedy algorithm, degree centrality and PageRank algorithm are very close to each other. The IM-LPA algorithm is slightly superior to the other algorithms when the number of initial active nodes k is larger than 20. Under the LT model, the IM-LPA algorithm performs the best among all the algorithms and shows some advantages over the other algorithms. Only under the IC model with $p = 0.1$, our IM-LPA algorithm is inferior to the greedy algorithm and betweenness centrality, but it still performs better than the other algorithms.

From the above experimental results, we can see that the IM-LPA can

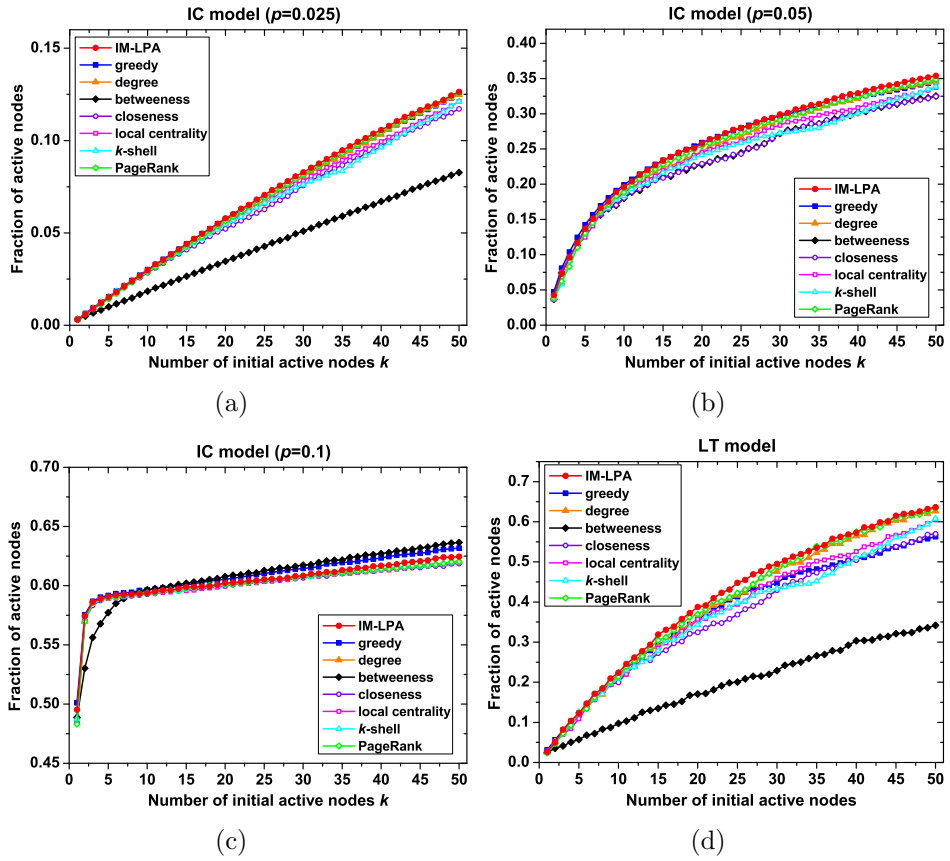


Figure 3: The influence spreading of different algorithms on LFR benchmark networks with large communities. (a) IC model, $p = 0.025$; (b) IC model, $p = 0.05$; (c) IC model, $p = 0.1$; (d) LT model.

effectively find the influential nodes in synthetic networks with community structure.

5.2. Real-world networks

We also compare the IM-LPA algorithm with other algorithms on several real-world social networks which are widely used in the literature. General information of these real-world social networks are shown in Table 1. It can be seen that Football Network, SFI network Facebook network and PGP network have significant community structure since the modularity is large. For Email network, the relatively small modularity indicates that the community structure is kind of indistinct.

Table 1: General information of the real-world social networks.

Network	Description	Node	Edge	Community	Modularity
Football	American College football union [49]	115	616	12	0.6010
SFI	Collaboration network of scientists at Santa Fe Institute [49]	118	200	8	0.7335
Email	E-mail interchanges between members of URV [50]	1133	5451	14	0.4876
Facebook	Friend relationship on Facebook Site[51]	4038	88234	8	0.7379
PGP	The network of users of PGP algorithm [52]	10680	24316	240	0.8432

The experiments on real-world social networks excludes the IC model with $p = 0.025$, due to the fact that influence can only spread to a tiny range on these networks when the diffusion probability of IC model is small. Fig. 4 shows the performances of the IM-LPA algorithm on the real-world social networks in comparison with other algorithms.

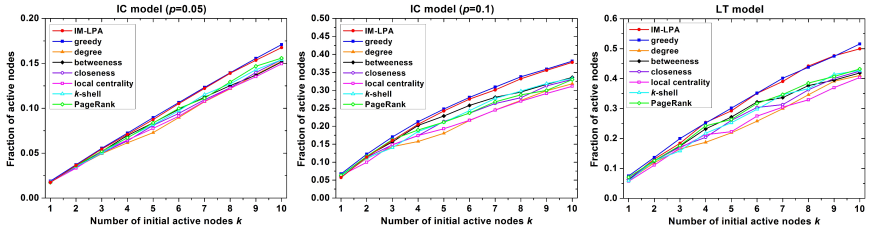
On Football Network, the performance of the IM-LPA algorithm is approximate to that of the greedy algorithm and significantly outperforms the other algorithms.

On SFI Network, when the number of initial active nodes $k < 4$, the IM-LPA algorithm performs very closely to the greedy algorithm and shows obvious superiority over the other algorithms. But as the number of initial active nodes k exceeds 4, the performance of the IM-LPA algorithms falls behind that of the greedy algorithm and gets close to that of degree centrality, k-shell decomposition and PageRank algorithm.

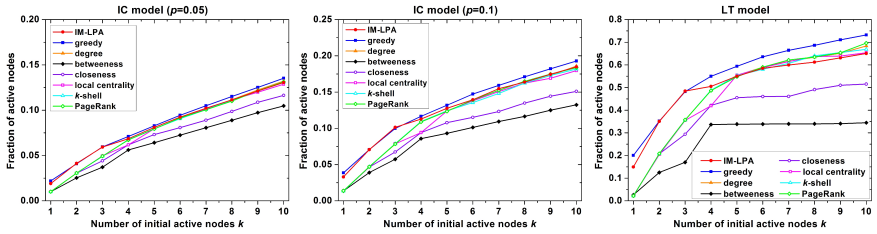
On Email Network, the greedy algorithm performs the best among all the algorithms, but there are little differences between the performances of most algorithms. Our IM-LPA algorithm does not perform well mainly due to the indistinct community structure of Email Network. But it still performs better than betweenness centrality, closeness centrality, local centrality and k-shell decomposition in most situations.

On Facebook network, the proposed IM-LPA algorithm gives excellent results and performs almost the same as the greedy algorithm under the IC model. Under the LT model, the greedy algorithm makes the best performances, and the IM-LPA algorithm still performs better than most of the other algorithms.

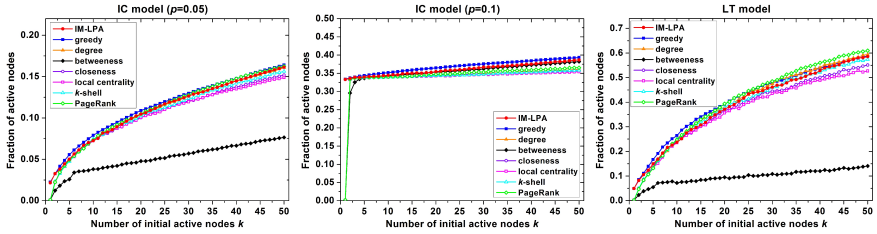
On PGP network, the greedy algorithm performs better than the other algorithms. The IM-LPA algorithm shows significant advantages over the other algorithms under the IC model. Under the LT model, the IM-LPA algorithm is inferior to PageRank algorithm and very close to degree centrality.



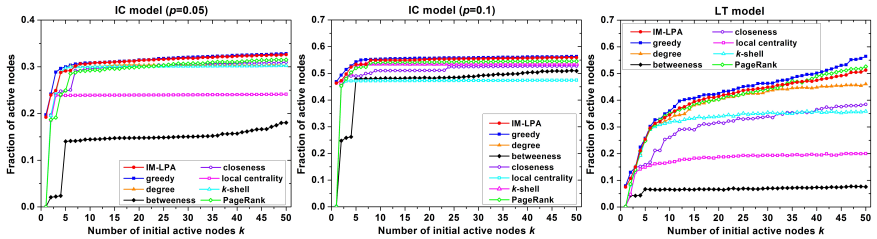
(a)



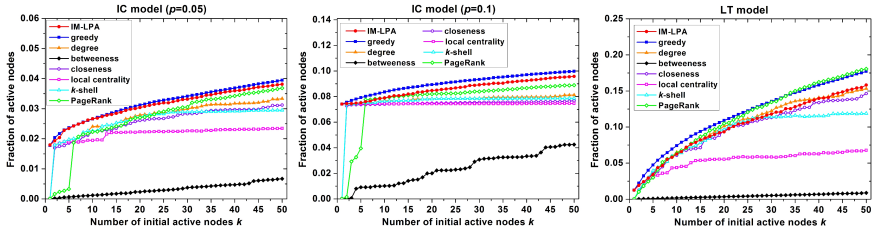
(b)



(c)



(d)



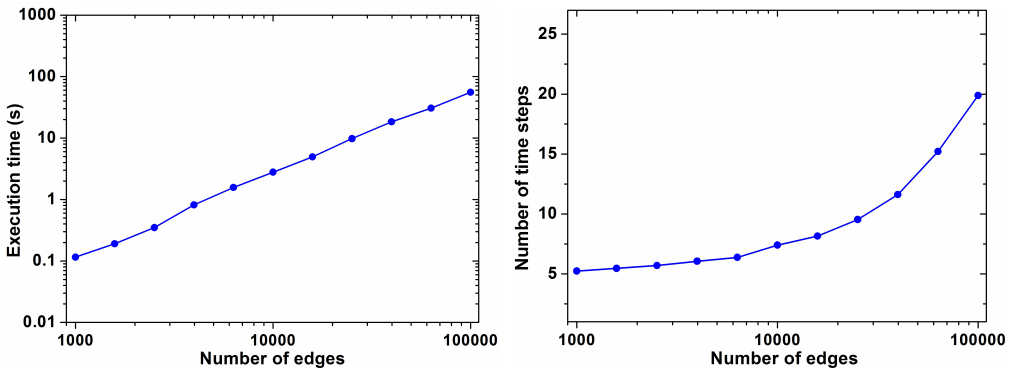
(e)

Figure 4: The influence spreading of different algorithms on real-world social networks. (a) Football Network; (b) SFI Network; (c) Email Network; (d) Facebook Network; (e) PGP Network.

The above experimental results demonstrate that the IM-LPA algorithm is promising and effective for identifying the influential nodes in real-world social networks with community structure.

5.3. Time complexity

Finally, we experimentally measure the time complexity of the IM-LPA algorithm. The LFR benchmark networks are still used in our experiments. The mixing parameter μ is set to 0.1, average node degree is set to 20 and the community size is in the range [20, 200]. The number of nodes in the network inceases from 100 to 10000, so that the number of edges varies from 1000 to 100000. The execution time and the number of time steps of the IM-LPA algorithm on benchmark networks are shown in Fig. 5, where each data point is an average over 10 networks.



(a) (b)

Figure 5: (a) Average execution time of the IM-LPA algorithm on benchmark networks with different sizes; (b) average number of time steps of the IM-LPA algorithm on benchmark networks with different sizes.

Seen from Fig. 5, the execution time of the IM-LPA algorithm increases a little rapidly than linearly to the number of edges in the network. This is mainly because that the number of time steps of the label propagation process in IM-LPA algorithms increases along with the number of edges. Therefore, the time complexity of the IM-LPA algorithm is accordant with the previous analysis in Section 4.3. The timing results demonstrate that our proposed IM-LPA algorithm is a fast influence maximization algorithm and applicable to large-scale networks.

6. Conclusion

In this paper, we propose the IM-LPA algorithm to solve the influence maximization problem in social networks with community structure. The label propagation process is introduced to identify the influential nodes in the network. Our proposed algorithm is based on a novel heuristic that the most influential node of a community could propagate its label to all the nodes within the community during the label propagation process. Hence, we make the labels propagate from some seed nodes and evaluate the centrality of these seed nodes according to the label propagation process.

The IM-LPA algorithm is parameter free and requires no prior information about the community structure in the network. Moreover, the IM-LPA algorithm has near linear time complexity, which makes it applicable to large-scale networks. We test the IM-LPA algorithm along with several other influence maximization methods on both synthetic and real-world networks for comparison. The experimental results demonstrate the effectiveness and efficiency of our proposed algorithm.

Acknowledgements

This research work is funded by the National Science Foundation of China (61271316), 973 Program of China(2013CB329605)the National Social Science Foundation of China (14ZDB167) Shanghai Key Laboratory of Integrated Administration Technologies for Information Security.

References

[1] C. Haythornthwaite, Social network analysis: An approach and technique for the study of information exchange, Libr. Inform. Sci. Res., 18 (1996) 323-342.

[2] S. Wasserman, Social network analysis: Methods and applications, Cambridge: Cambridge University Press, 1994.

[3] S.H. Strogatz, Exploring complex networks, Nature, 410 (2001) 268-276.

[4] M.J. Keeling, P. Rohani, Modeling Infectious Diseases: in Humans and Animals, Princeton: Princeton University Press, 2008.

[5] W. Chen, C. Wang, Y. Wang, Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks, In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, USA, 2010, pp. 1029-1038.

[6] V. Mahajan, E. Muller, F. M. Bass, New Product Diffusion Models in Marketing: A Review and Directions for Research, The Journal of Marketing, 54 (1990) 1-26.

[7] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, USA, 2003, pp. 137-146.

[8] E. Even-Dar, A. Shapira, A note on maximizing the spread of influence in social networks, Inform. Process. Lett., 111 (2011) 184-187.

[9] C. Gao, X. Lan, X. Zhang, Y. Deng, A bio-inspired methodology of identifying influential nodes in complex networks, PLoS One, 8 (2013) e66732.

[10] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, H. A. Makse, Identification of influential spreaders in complex networks, Nat. Phys., 6 (2010) 888-893.

[11] N. E. Friedkin, Theoretical foundations for centrality measures, Am. J. Sociol., 96 (1991) 1478-1504.

[12] T. Opsahl, F. Agneessens, J. Skvoretz, Node centrality in weighted networks: generalizing degree and shortest paths, Soc. Netw., 32 (2010) 245-251.

[13] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett., 86 (2001) 3200-3203.

[14] L. C. Freeman, A set of measures of centrality based on betweenness, Sociometry, 40 (1977) 35-41.

[15] G. Sabidussi, The centrality index of a graph, Psychometrika, 31 (1966) 581-603.

[16] D. Chen, L. Lu, M. S. Shang, Y. C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A, 391 (2012) 1777-1787.

[17] Y. Wu, Y. Yang, F. Jiang, S. Jin, J. Xu, Coritivity-based influence maximization in social networks, Physica A, 416 (2014) 467-480.

[18] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys., 6 (2010) 888-893.

[19] L. Lu, Y. C. Zhang, C. H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PLoS One, 6 (2011) e21202.

[20] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, USA, 2007, pp. 420-429.

[21] M. Kimura, K. Saito, Tractable models for information diffusion in social networks, In: Knowledge Discovery in Databases: PKDD 2006, 2006, pp. 259-271.

[22] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009, pp. 199-208

[23] W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, In: Proceedings of IEEE 10th International Conference on Data Mining (ICDM 2010), Sydney, Australia, 2010, pp. 88-97.

[24] A. Goyal, W. Lu, L. V. Lakshmanan, SIMPATH: an efficient algorithm for influence maximization under the linear threshold model, In: 2011 IEEE 11th International Conference on Data Mining (ICDM 2011), Vancouver, Canada, 2011, pp. 211-220.

[25] M. Kimura, K. Saito, R. Nakano, H. Motoda, Extracting influential nodes on a social network for information diffusion, Data Min. Knowl. Disc., 20 (2010) 70-97.

[26] M. E. J. Newman, The structure and function of complex networks, SIAM Rev., 45 (2003) 167-256.

[27] M. E. J. Newman, Detecting community structure in networks, Eur. Phys. J. B, 38 (2004) 321-30.

[28] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E, 69 (2004) 026113.

[29] S. Fortunato, Community detection in graphs, Phys. Rep., 486 (2010) 75-174.

[30] X. Wu, Z. Liu, How community structure influences epidemic spread in social networks, Physica A, 387 (2008) 623-630.

[31] W. Huang, C. Li, Epidemic spreading in scale-free networks with community structure, J. Stat. Mech., 2007 (2007) P01014.

[32] X. Chu, J. Guan, Z. Zhang, S. Zhou, Epidemic spreading in weighted scale-free networks with community structure, J. Stat. Mech., 2009 (2009) P07043.

[33] J. Goldenberg, B. Libai, E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Mark. Lett., 12 (2001) 211-223.

[34] J. Goldenberg, B. Libai, E. Muller, Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata, Acad. Mark. Sci. Rev., 9 (2001) 1-18.

[35] M. Granovetter, Threshold models of collective behavior. Am. J. Sociol., 1978 (1978) 1420-1443.

[36] D.J. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. USA, 99 (2002) 5766-5771.

[37] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web, Stanford: Stanford InfoLab Publication, 1999.

[38] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA, 101 (2004) 2658-2663.

[39] X. Chu, J. Guan, Z. Zhang, S. Zhou, Epidemic spreading in weighted scale-free networks with community structure. J. Stat. Mech., 2009 (2009) P07043.

[40] U. N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E, 76 (2007) 036106.

[41] M. J. Barber, J. W. Clark, Detecting network communities by propagating labels under constraints, Phys. Rev. E, 80 (2009) 026129.

[42] A. Lancichinetti, S. Fortunato, J. Kertesz, Detecting the overlapping and hierarchical community structures in complex networks, New J. Phys., 11 (2009) 033015.

[43] C. Lee, F. Reid, A. McDaid, N. Hurley, Detecting highly overlapping community structure by greedy clique expansion, In: Proceedings of International Workshop on Social Network Mining and Analysis (SNAKDD 2010), Washington DC, USA, 2010, pp. 33-42.

[44] C. Lee, F. Reid, A. McDaid, N. Hurley, Seeding for pervasively overlapping communities, Phys. Rev. E, 83 (2011) 066107.

[45] J. Scripps, P. N. Tan, A.-H. Esfahanian, Exploration of link structure and community-based node roles in network analysis, In: Proceedings of 7th IEEE International Conference on Data Mining (ICDM 2007), 2007, pp. 649-654.

[46] Y. Zhao, S. Li, S. Wang, Agglomerative clustering based on label propagation for detecting overlapping and hierarchical communities in complex networks, Adv. Complex Syst., 17 (2014) 1450021.

[47] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E, 69 (2004) 026113.

[48] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E, 78 (2008) 046110.

[49] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA, 99 (2002) 7821-7826.

[50] M. R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Phys. Rev. E, 68 (2003) 065103.

[51] J. Leskovec, J. J. Mcauley, Learning to discover social circles in ego networks, In: Advances in Neural Information Processing Systems, 2012, pp. 539-547.

[52] M. Boguña, R. Pastor-Satorras, A. Díaz-Guilera, A. Arenas, Models of social networks based on social distance attachment, Phys. Rev. E, 70 (2004) 056122.