Article

A method to improve influence maximization in social networks based on community detection

Mansoureh Abolghasemi[1], **Esmaeil Bagheri**[2]
[1]Department of Computer, Khorasgan Branch, Islamic Azad University, Isfahan, Iran
[2]Department of Computer, Dehaghan Branch, Islamic Azad University, Isfahan, Iran
E-mail: Bagheri471@gmail.com

Received 11 July 2022; *Accepted 20 August 2022*; *Published 2 September 2022*; *Published 1 December 2022*

Abstract
With the emergence of social networks, human relationships on the internet have become a new form. Social networks are not only a communication tool for users, but can also be a basis for marketing and advertising products of different companies. Studying the impact of maximum penetration has attracted many researchers in recent years due to the benefits of viral marketing. Given a social network, the goal is to find a subset of K individuals as influential nodes that can generate maximum cascading influence through the network under a predefined diffusion model. The first research in this field did not work for large networks. After this effort, different methods were presented to maximize influence, among them, methods based on communities were proposed. Algorithms for maximizing community influence often use the influence of a node in its own community to approximate its influence in the entire network, so they can perform better. One of these community-based algorithms is the COFIM algorithm. In this paper, the efficiency of the COFIM algorithm, which is a community-based influence maximization method, is improved by distributing seed nodes through the community structure. The results of the proposed algorithm have been tested on six different data sets and then compared with the basic methods. The test results show the efficiency of the proposed method.

Keywords social networks; influence maximization; community detection.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Analysis of structures and behavior of social networks became one of the fundamental needs of commercial companies. Analysis of social networks is applicable in many utilizations including social networks management, analysis of market trend, identifying influential individuals and supporters, promoting efficiency of descriptive systems. In recent years, due to commercial needs, a lot of attention has been paid to social networks analysis in academic dimension. Today, not only information technology experts use this powerful tool, but also scholars of other majors like educational sciences, biology, communication sciences, economy, etc, use social network analysis as a key technique.

One of the most important subjects of research in the field of social networks, is maximizing efficiency in

social networks; that had attracted the attention of research community of data-mining significantly. This issue that in fact is provoked by viral marketing, has been performed by spreading innovations, new products, ideas, news, etc, in networks through recommendations and advertisements that are performed by network members and also according to effects that they have on each other (Bagheri et al., 2016).

Briefly, in viral marketing, through limited awareness about existence of a new product for several people and according to relations that people have together, we want to spread this news or product in word-of-mouth way. By modeling this concept in social networks, maximizing efficiency is the problem of selecting a group with definite numbers of most influential nodes in diffusion network, so that if the process of spreading these nodes starts, the amount of spread and diffusion of information, expected influence under a specific diffusion model, reach the maximum level and at the end, the most "effective nodes" will exist in network (Song et al., 2015).

The second section of this paper reviews researches on maximizing influence. The third section considers preliminaries of problem. In the fourth section, related works will be reviewed. In the fifth section, structure of proposed model will be considered, and evaluation of results is also performed and finally in the sixth section, the article concludes.

2 Background

The influence maximization problem was firstly formulated by Kempe et al. (2003). Under the IC or LT model, we use S to denote the seed set, i.e., the set of active nodes at step t = 0, and S(t) as the set of active nodes at step t. It is easy to see that the propagation stops when S(t) = ø. Then the number of overall activated nodes after the propagation stopped can be represented by ∞ t=0 |S(t)|. Since the diffusion models are usually stochastic, we use σ(S) to denote the expected number of overall activated nodes. The influence maximization problem is defined as follows.

Given a network and a diffusion model, the influence maximization problem aims to find a subset S of k nodes (|S| = k), such that the expected number of overall activated nodes σ(S) is maximized (Kempe et al., 2003):

$$S^* = \arg_S \max \sigma (S)$$

Kempe et al. (2003) proved that under IC and LT models, the influence maximization problem defined in Definition 1 is NP-hard and the objective function σ(S) is submodular. Based on the sub modularity, Kempe et al. further proposed a “hill climbing” greedy algorithm and proved that the algorithm provided a factor of (1 − 1/e − ε) guarantee to the optimal solution. Theoretically, the traditional greedy algorithm provides a 63% guarantee to the optimal solution. In real experiments, the solution provided by the greedy algorithm is quite close to the optimal solution. However, to have a good approximation of the objection function given the seed set S, the greedy algorithm requires tens of thousands of Monte–Carlo simulations, which seriously limits its application on large-scale networks. To solve the time efficiency problem of traditional greedy algorithm, a spectral of algorithms were proposed by researchers in recent years. Some works make use of sub modularity, such as the CELF algorithm proposed by Leskovec et al. (2007). Some research works assume that the influence can spread on the network only through shortest paths so that the objective function can be exactly computed. Another way to reduce the time complexity is to simply select top k nodes based on some heuristic metrics, such as the degree centrally, betweenness centrality and so on. However, since the heuristic methods take no consideration of propagation models, they usually give poor solutions.

In the age of big data, network scale grows in millions, if not billions. Traditional influence maximization methods either cannot handle large-scale networks, or provide inaccurate solutions with low influence spread. Recently, some research works were proposed to tackle the influence maximization problem

using community information. Community structure is defined as the partition of network nodes into groups, within which nodes are densely connected while between which they are sparsely connected (Girvan and Newman, 2002; Zhang and Li, 2016; Zhang, 2018, 2021).

Networks are a powerful way to model relational information among objects from social, natural, and technological domains. Networks can be studied at various levels of resolution ranging from whole networks to individual nodes. One way to understand networks at the level of groups is to identify sets of nodes with similar connectivity patterns (Zhang and Li, 2016; Zhang, 2018). Traditional methods aim to find network communities, which are defined as groups of nodes with many connections among the group's members, but few to the rest of the network. However, dense communities are but one kind of group structure in networks, and there may be other structures that help us to understand networks better. Thus, different structures of society can increase the diffusion of influence on social networks.

Recent works show that community-based influence maximization algorithms are generally faster than traditional greedy algorithms. based on the fact that different communities are sparsely connected to each other, we may use the influence of a node within its own community to approximate its influence on the whole network. However, the influence of a node within its own community can be computed more efficiently. Moreover, since they usually make the assumption that different communities are isolated, these algorithms naturally support parallelization.

3 Problem Description

The influence maximization problem in social networks, is a problem of selecting a definite number of most influential nodes of a social network so that if the process of spreading information under a specific diffusion model starts from these nodes, the amount of information diffusion or efficiency will be maximized and finally the highest number of involved nodes will be in the network.

In the problem of maximizing influence, social network is considered as a directed or undirected graph G = (V, E), that V is the collection of nodes and E is the collection of edges. Edge (u,v) ∈ E, shows a relation between two nodes u and v. The direction of information diffusion on a social network under a specific model through observing that model's assumptions, considers a small collection of that network nodes (seed) as active nodes. At each moment, an active node tries to activate its own inactive neighbors. Ultimately, according to above assumptions, algorithm ends and a number of inactive nodes will be added to active nodes. Different algorithms try to select seeds that can activate more nodes of networks.

We define influence of an initial node collection S on other nodes of social network as σ(S) that shows the number of activated nodes at the end of influence diffusion process by initial seed nodes S. The number of nodes in S collection equals to k. At the start of finding seed nodes, S=∅. We want to provide an algorithm that finds nodes of collection S so that influence diffusion in network under collection S will be maximized. Based on below formula, algorithms of maximizing influence in each stage try to add a node to collection S which can maximize influence diffusion than other nodes of network:

$$\operatorname{argmax}_{v \in V} (\sigma (S \cup \{v\}) - \sigma(S)) \quad (1)$$

According to mentioned points, finding seed nodes in whole network is time-consuming. Therefore, objective is finding an algorithm that can select the best seed nodes in the lowest time. On the other hand, despite useful features of communities in social networks, there's a little attention to role of communities in maximizing influence. Meantime, there are different structures of communities in social networks that create relations among individuals in different ways. Information circulation has high speed in these groups. In addition, some communities play important role in social networks and are a center for spreading information. On the other hand, actual networks consist a lot of nodes and calculating influence for each node is very

expensive. Local calculation of influence for each node in the community that (node) belongs to it can be performed very fast and accordingly time of algorithm execution will be improved. Furthermore, apportioning seed can help to better identification of seed nodes.

4 Related Works

Since Kempe et al. (2003) formally formatted the influence maximization problem and proposed the greedy algorithm, a lot of research works have been published to tackle this problem.

Cao et al. (2010) proposed the first community-based influence maximization algorithm OASNET (Optimal Allocation in a Social NETwork). They assume different communities are independent of each other and influence cannot spread across different communities. The community structure was detected by the CNM (Clauset–Newman–Moore) (Clauset et al., 2004) algorithm. The selection of seed nodes contains two phrases. In the first phase, from each community the algorithm selects k nodes using traditional greedy algorithm, resulting in a total number of C · k candidate nodes. In the second phase the algorithm selects k nodes as the seed set S from the C · k candidates using dynamic programming.

Zhang et al. (2013) studied the problem of identifying the influential nodes on networks with community structure. The authors firstly constructed an information transfer probability matrix from the weighted network. Then they applied the k-medoid clustering algorithm to identify the k seed (influential) nodes.

Chen et al. (2014) studied the community-based influence maximization problem using the HD (heat diffusion) model and proposed the CIM (Community-based Influence Maximization) algorithm. The algorithm can be divided into three phases: (i) community detection, (ii) candidate nodes generation, and (iii) seed nodes generation.

Li et al. (2015) considered the node conformity and proposed the community-based influence maximization algorithm CINEMA (Conformity-aware Influence Maximization). Based on conformity, the C2 (Conformity-aware Cascade) diffusion model is defined. In the diffusion process, the probability that an active node u activates its inactive neighbor v is defined as: $puv = \Phi(u) \cdot \Omega(v)$. where $\Phi(u)$ is the influence index of u and $\Omega(v)$ is the conformity index of v.

Another approach proposed in the field of influence maximization community-based was proposed by Shang et al. (2016) as COFIM. In this framework the influence propagation process is divided into two phases: (i) seeds expansion, and (ii) intra-community propagation. Experimental results show that this algorithm can significantly outperform other state-of-the-art methods in terms of time and memory efficiency.

Shang et al. (2018) proposed a community-based algorithm to solve the problem of memory consumption while at the same time influence maximization problem the IMPC algorithm: an influence maximization framework based on multi neighbor potential in community networks.

Bagheri et al. (2018) have proposed FSIM algorithm based on community detection. Without loss of quality, FSIM reduces the number of nodes that must be examined for finding seeds. Their method first detects communities from the input network and creates a new network from detected communities. The new network has m nodes, where each node represents a community. Hence only a limited number of nodes are examined so it is fast. Within each important community, important nodes are selected. Final seeds are selected after testing initial seeds.

Ghanbari et al. (2020) have proposed C-K-shell algorithm based on K-shell decomposition and community detection. They use SLPA algorithm for community detection and to make a better result use optimizing the decision-making in exploration and extraction of communities. K-shell analysis and community detection are used to choose the more influential nodes, which are proportional to the graph of social networks. C-K-shell reduces number of nodes which should be investigated to find seeds without

losing quality. Therefore, only a limited number of nodes are investigated so that speed is increased.

Bagheri (2020) has proposed a new method for maximizing influence on social networks based on node membership in communities. He studied the main challenges of other studies and found the lack of scalability and low speed. He proved influential nodes must also have local influence and global influence throughout the network so that they can affect the entire network at an acceptable time. His paper considers the important role of influential nodes in each community for influence propagation in that community and consequently propagating the influence throughout social network. Therefore, it finds the nodes that have more membership strength to their community. His proposed algorithm is tested on several real and synthetic social networks. Experimental results show that his proposed method can effectively find appropriate seed nodes for influence maximization.

Bagheri et al. (2021) have proposed A fast and accurate influence maximization algorithm in social networks based on community structures called FAIMCS. It can quickly find influential nodes across large networks with high accuracy. FAIMCS reduces computational overhead considerably by eliminating major portions of the social network graph which have little influence and uses community detection algorithm to determine each community's quota of influential nodes based on the structure of that community and finds influential nodes from the candidate nodes. Their experiment results show it is faster than other algorithms and provides a high level of accuracy for large social networks.

Kumar et al. (2022) proposed a novel method to solve the problem of influence maximization named Communities based Spreader Ranking (CSR), which is based on the notions of communities and bridge nodes. Their method identifies bridge nodes as influential nodes based on three concepts: community diversity, community modularity, and community density. Community diversity is used to identify bridge nodes and the rest two are used to identify significant communities. Extensive experimentation validation on various datasets using popular information diffusion models demonstrates that the proposed method delivers proficient results compared to numerous previously known contemporary influence maximization methods.

Recent works show that community-based influence maximization algorithms are generally faster than traditional greedy algorithms. Moreover, since they usually make the assumption that different communities are isolated, these algorithms naturally support parallelization. One of the problems with these algorithms is that algorithms use the influence of a node in their community to approximate their impact on the entire network, the accuracy of the algorithm will rely heavily on the structure of the underlying community. So an important point in the performance of these algorithms is to correctly identify the communities and use the appropriate algorithm to discover the community.

5 Proposed Method

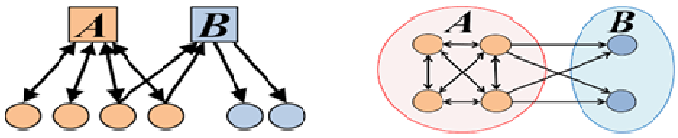
In this section, the framework of proposed method is described. Stages of suggested method are shown in the algorithm (Fig. 1).

Input graph file, number of seed node k.
Output: seed set s
Step 1: detecting community by CODA method
Step 2: determinate community structures
Step 3: computing seeds quotas for each community
Step 4: select seed S by CELF method

Fig. 1 Steps of proposed model.

5.1 Community detection

In first stage that communities should be detect, there is a graph of nodes and edges of data collection. To detect community, CODA algorithm has been used. This model starts with a bipartite affiliation graph, where nodes of the underlying network represent one 'layer' of the bipartite graph and communities represent the other. Edges between network-nodes and community-nodes in the affiliation graph represent memberships of nodes to communities. (Fig. 2(a)). Edges of the underlying network (Fig. 2(b)) then arise due to shared community affiliations of nodes. Consider for a moment an undirected network; when a node belongs to a community in such a network it typically means that the node has (undirected) edges to other members of the community. This type of community affiliation can be modeled using a bipartite graph of nodes and communities where undirected affiliations are formed between nodes and communities (Lattanzi and Sivakumar, 2009). In directed networks, however, we need a richer notion of community affiliation (Fig. 2(a)): a node may create edges to other members of a community, and it also receive edges from other members of the community, or both. Therefore, it can be assuming that nodes in directed networks can have two "types" of community affiliation: "Outgoing" affiliations from nodes to communities mean that in the network the node sends edges to other members of the community. And, "incoming" affiliations from communities to nodes mean that nodes receive edges from other community members. This article is referred to as "Incoming Membership" and "Outgoing Membership".



(a) (b)

Fig. 2 (a) Directed node community affiliation graph. Squares: communities, Circles: nodes of network G. Affiliations from nodes to communities indicate that nodes create edges to other members in those communities, while affiliations from communities to nodes indicate that nodes receive edges from others. Community A is cohesive, while B is a 2-mode community. (b) Network G corresponding to model in (a).

Formally, its denote a bipartite affiliation graph as B(V, C, M), where V is the set of nodes of the underlying network G, C the set of communities, and M a set of directed edges connecting nodes V and communities C. An outgoing membership edge of node u ∈ V to community c ∈ C is denoted as (u, c) ∈ M, and an incoming membership is denoted as (c, u) ∈ M. Now, given the affiliation graph B(V, C, M), it need to specify a process that generates the edges E of the underlying directed network G(V, E). To this end its consider a simple parameterization where this assign a single parameter pc to every community c ∈ C. The parameter pc models the probability of a directed edge forming from a member node u with an outgoing membership to community c to another member v of c with an incoming membership (Yang et al., 2014). Given B(V, C, M) and {pc}, the model generates a directed graph G(V, E) by creating a directed edge (u, v) from node u ∈ V to node v ∈ V with probability p(u, v):

$$P(u,v) = 1 - \prod_{k \epsilon C_{uv}} (1 - P_k) \qquad (2)$$

where $C_{uv} \subset C$ is a set of communities through which u has a 2- step directed path to v :

$$C_{uv} = \{C | (u,c), (c,v) \in M\} \qquad (3)$$

If $C_{uv} = \emptyset$ then:

$$P(u \, \mathcal{s} v) = \frac{1}{|V|} \qquad (4)$$

As noted earlier, we distinguish nodes' incoming memberships and outgoing memberships. In particular,

let M_{uc} indicate whether the node u belongs to community c with an outgoing membership, and L_{vc} indicate whether node v has an incoming membership for c. Now Eq. 2 can be represented as:

$$P(u,v) = 1 - \prod_C (1 - P_C)^{M_{uc}L_{vc}} \quad (5)$$

5.2 Community structure detection

By the estimated values for M_{uc} and L_{vc} can be determined the community affiliations of nodes. Its achieve this by thresholding M_{uc} and L_{vc} with a constant δ, i.e., Its regard u has an outgoing membership to community c if $M_{uc} \geq \delta$, and an incoming membership from c if $L_{vc} \geq \delta$. We choose the value of δ so that every pair of members in community c has edge probability higher than the background edge probability 1/|V | (Yang and Leskovec, 2015):

$$\frac{1}{|V|} \leq 1 - exp(-\sigma^2) \quad (6)$$

This inequality leads to:

$$\sigma = \sqrt{-log\left(1 - 1/|V|\right)} \quad (7)$$

It should be noted that various experiments performed in Yang et al. (2014) have shown that this choice works well for σ. After this step, two outputs are created, one of them represents community nodes with " incoming membership" and the other one represents community nodes with " outgoing membership". Once the network communities are identified, the community structure is detecting in next step. Thus Jaccard similarity has been used for this purpose (Yang et al., 2014):

$$J(c) = \frac{|O(c) \cap I(c)|}{|O(c) \cup i(c)|} \quad (8)$$

where O(c) is a set of nodes with " outgoing membership" and I (c) is a set of nodes with " incoming membership ".

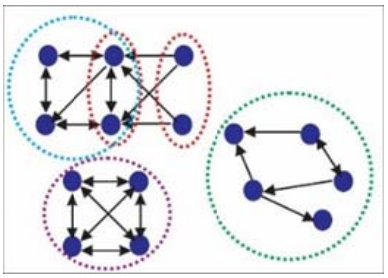


Fig. 3 Different structures of detected communities.

In a completely cohesive community, this Jaccard similarity is 1 because two sets of members are identical, whereas it is 0 in a completely 2-mode community. its regard a community c as 2-mode if J(c) is lower than some threshold γ or as cohesive otherwise (J(c) ≥ γ). Based on the tests performed in Ref.16, the threshold value is estimated as λ = 0.2. Therefore, based on the proposed model, four types of structures are considered for communities. Based on Jaccard similarity, if the jacquard similarity value is less than 0.2, the community is considered 2-mode community. If this value is between 0.2 and 0.5 the communities are semi-cohesive. And the greater the amount and to be closer to one is to have a cohesive community structure. This value for threshold is obtained by testing on the dataset, that the results of which are presented in the next section. The next structure is the overlapping structure of the communities that can be identified by searching for communities and finding common nodes in different communities of overlapping communities. Fig. 3 shows the types of structures in communities.

5.3 Seed quota computation

Third stage is determining seed share of each community. In fact, apportioning seed, determines how many influential nodes are needed in the community to have the highest number of active nodes. To apportion, following equation is used:

$$Q = K \times \frac{N_c}{N} \qquad (9)$$

where K is the number of power (seed) nodes, N_c is number of nodes in community and N is whole nodes. Now, according to identified structures, and through considering each community features, a specific share of seed will be assigned to that community. Reasoning of assigning seed here, is as follow: since two-status communities include two parts, a part of it only includes nodes that receive edges from other nodes (like famous people) and a part includes nodes that send some nodes to first part (like fans of famous people). Therefore, these communities don't receive a share of seed because they are likely member of other communities.

About systematic communities, it can be told that since their connections are dense, so they can take a lower share of seed because they may activate a lot of nodes with those low number of seed. This reasoning can also be used for overlapping communities, that means nodes which are in the common areas among communities, are members of other communities that can be activated and also activate their neighbor nodes. Due to this reasoning, low share of seed has also been assigned to overlapping communities.

But, there are structures that have none of these systems, i.e. neither they are two-status communities nor systematic communities ($0.2 \leq J(C) \leq 0.5$) and they may have no overlap with other communities. Therefore, such communities can activate their neighbor nodes with a low probability and so they take more share of seed for increasing expectable number of active nodes. So, if shares Q are given to overlapping communities, with a constant coefficient, more shares can be assigned to semi-systematic communities:

$$Q = \alpha \times K \times \frac{N_c}{N} \qquad (10)$$

Here, $\alpha = 2$ has been used, that this amount was suggested through test on several datasets. Number of seed is selected more than k to perform the best selection for seed nodes, on the other hand, there is a subset of these seeds so that the most number of nodes are activated.

5.4 Seed node selection

Now that each community's share of seed number has been determined, a method should be used to select power nodes. According to COFIM method, CELF algorithm is used. In this method, for influence diffusion, algorithm doesn't need to repetition, but a priority queue is used for descending sort of nodes. If the amount of influence diffusion on the top node of queue is more than other nodes, this node as an influential node, will be added to initial seed collection S.

Since this method wants to improve COFIM algorithm, therefore weighted cascade model (WC) is used. As it was explained before, in this model, the probability that each node can activate its own inactive neighbor depends on node degree. The likelihood of activation for each node is obtained from Eq.1.

5.5 Evaluation

In this section, the proposed model is evaluated by other methods on different datasets. By considering previous methods for investigating suggested method, different data collections in real world and also artificial networks for testing suggested method, are used.

In this research, it is tried to use model for different data collections with various measures to consider model's efficiency in large networks. In Table 1, information related to these data collections are shown.

Table 1 Real-World Networks datasets.

Dataset	Nethept	Epinions	Amazon	Patent	LiveJournal	Orkut
# Nodes	15K	76K	335K	3.8M	4M	**3.1M**
#Edges	31K	406K	926K	16.5M	35M	**117M**
Max Degree	64	3044	290	793	14724	**33313**
Avg. Degree	4.12	10.69	4.34	8.74	17.01	**76.17**
#Communities	2262	10307	12326	15563	116288	**5118**

In this study, in addition to real-world networks, we also use synthetic network. Since our framework is based on community structure, we choose to use the LFR algorithm (Lancichinetti et al., 2008) to generate synthetic benchmark networks. The LFR benchmark is provided by Lancichinetti et al. (2008), which produces synthetic graphs that are very similar to real-world graphs. In this benchmark, the size of each community and the degree of each node are derived from a distribution of power- law. The LFR model has several parameters. The most important of these is the mixing parameter μ, which controls the number of edges between communities. If $\mu = 0$ all edges are in the community, if $\mu = 1$ all edges are between nodes in different communities. The degree of nodes and sizes of community according to the power-law, distributed with varying power. The LFR algorithm accepts the following parameters: the number of nodes N, the average degree k, the maximum degree k_{max}, the power exponent of the degree distribution τ_1 and community size distribution τ_2. By experimenting with Synthetic networks LFR we can see how networks with different community structures affect the performance of our proposed algorithm.

5.6 Comparable algorithms and evaluation criteria

In this research, different algorithms of maximizing influence are used for comparison through suggested method. These algorithms are: influence algorithm of independent path (IPA) (Kim, 2013), two-stage algorithm for maximizing influence (TIM+) (Shang et al., 2014), algorithm of maximizing influence through betting (IMM) (Tang et al., 2015), degree algorithm (Li et al., 2015), IMPC algorithm (Shang, et al., 2018) and algorithm of influence maximization based on community (COFIM) (Shang, 2016). Also, for comparing and evaluating efficiency of suggested model through other methods, two measures are used that are: the amount of influence diffusion and execution time.

5.7 Experimental method

Since the suggested method is based on community structure, to create communities from data collection, community detection method CODA has been used (Yang et al., 2014). This method is used due to identification of systematic and two-status structures and also its capability in identifying interfered and overlapping structures to help improving the algorithm of maximizing influence based on community. Diffusion model in this algorithm, like COFIM algorithm uses weighted cascade diffusion method (WC).

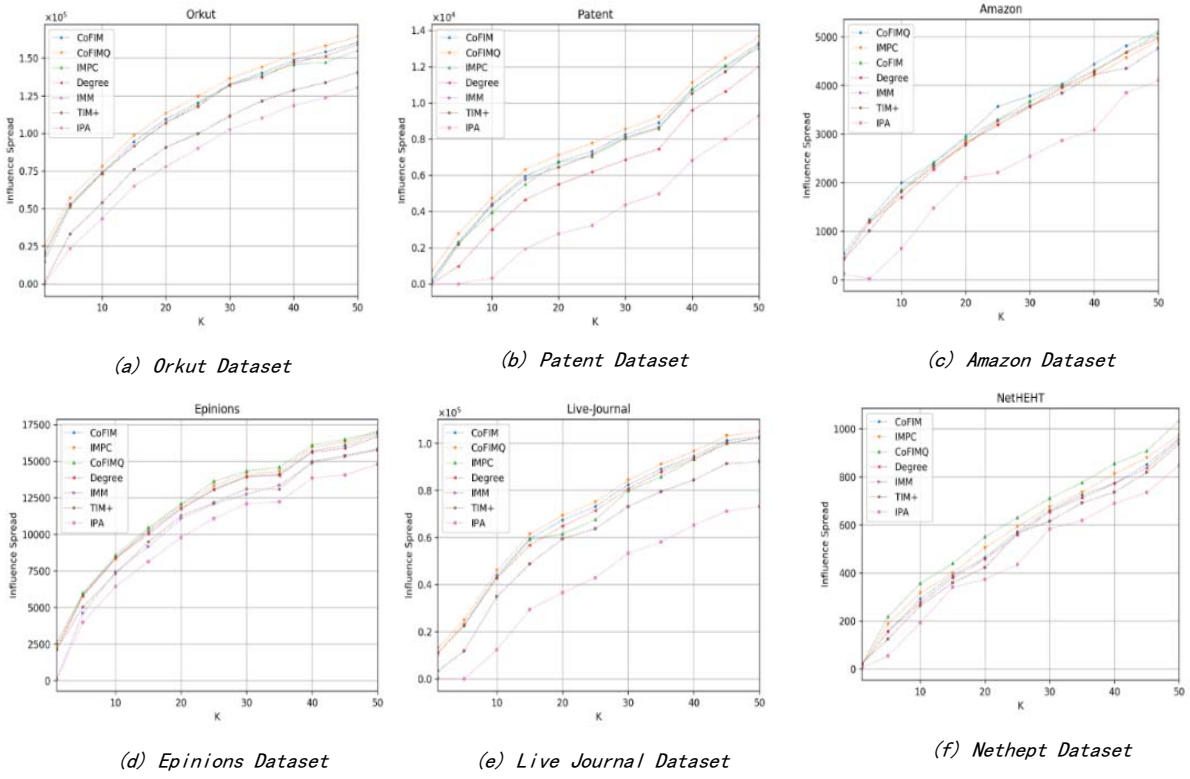
5.8 Experiment environment

The experiments are performed on a computer with Intel 3.5GHz Processor, 8 GB Ram and Linux operating system. Codes of this algorithm are written with C++ programming language, and programs are single-process and single-thread. Source codes of all basic algorithms are also written in C++ by other authors. In this section, we analyze obtained results of experiments from data collections of real world and also artificial networks based on comparison of expressed evaluation measures.

5.9 Real word networks results

First, we analyze influence diffusion. As it's obvious from Fig. 4, axis X shows number of power node k and vertical axis Y shows the amount of influence of different algorithm on data collections which are obtained

from Monte Carlo simulation. As it is observed, our suggested algorithm is generally among algorithms that have the most influence development. IPA algorithm always has lower performance than other algorithms. DEGREE algorithm that is an algorithm based on centrality of node, in some data collections has good performance in influence diffusion but it cannot be a guarantee for its efficiency in all data collections. As we can see in figure, this algorithm doesn't have good performance in PATENT data collection (Fig. 4 (b)). And has lower influence diffusion than other algorithm (except IPA). Of course, it has also relatively weak performance in NETHEPT (Fig. 4 (c)). Therefore, from this difference we can conclude that algorithms which use communities' information (IMC, COFIM, and suggested algorithm) has better performance in influence diffusion. Another obtained result is that since IMM algorithm is the developed form of TIM+ algorithm, so the amounts of their influence diffusion are approximately similar except in data collections LIVE JOURNAL and Orkut (Fig. 4 (a) and (e)). COFIM and IMPC algorithms have competitive performance with our suggested algorithm in some data collections but through change in the structure of community we could provide improved algorithm in terms of influence diffusion. In figures of experiment is obvious that our suggested algorithm has better performance than other methods.



(a) Orkut Dataset (b) Patent Dataset (c) Amazon Dataset

(d) Epinions Dataset (e) Live Journal Dataset (f) Nethept Dataset

Fig. 4 Influence spread of deferent algorithms in real world networks.

Since, time can be helpful in evaluation of algorithm efficiency and can be effective in algorithm improvement, therefore, in this section we consider execution time of different algorithms on some data collections, although considering time depends on hardware and programming language. In Fig. 5, results of these experiments are shown that we analyze them. Axis X shows the number of influence nodes (K) and axis Y shows time (second). Execution time in Fig. 5 has considered on data collections AMAZON, NETHEPT, EPINIONS and LIVE JOURNAL. In this section, as it can be seen in figures of execution time, suggested algorithm (COFIMQ) has more execution time than COFIM and IMPC, that is because of using CODA algorithm for detecting communities. Although time of our algorithm is slower than compared

community-based algorithms in this study, however, it has better execution time than other algorithms. Since IMM is developed model of TIM+, its execution time is lower than TIM+. Another compared algorithm is IPA that its execution time is linear and in terms of time, shows more time in output and figures than COFIMQ.

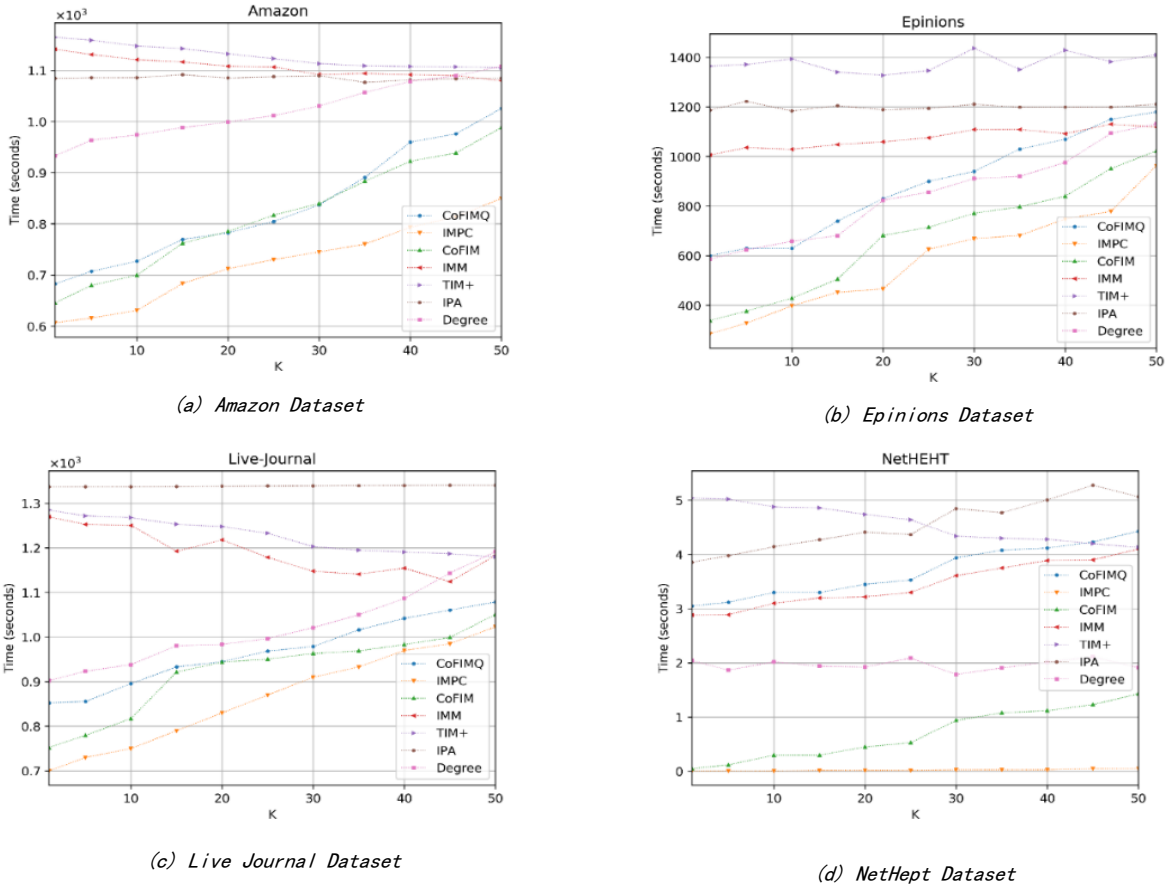
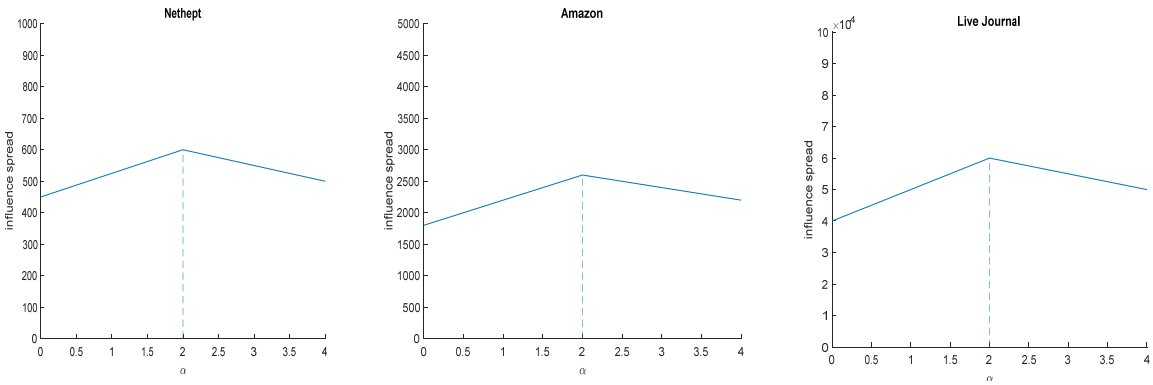


Fig. 5 Execution time of deferent algorithms in real world networks.

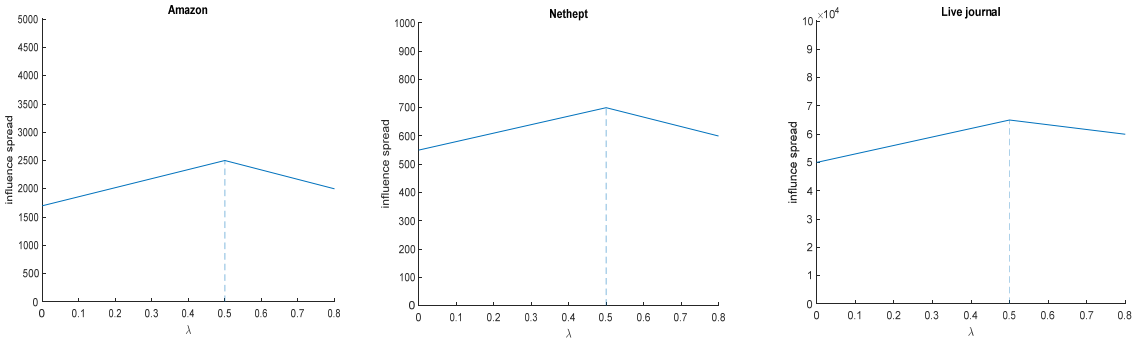
Parameter α is used for correct assignment of seed number to semi-dense structures to determine how much more seed share is needed for activating more nodes. This parameter has been tested on three data collection with three different size. The results of these experiments are shown in Fig. 6. As it's obvious, the amount of influence diffusion in different data collections has the highest amount in $\alpha = 2$.

In the suggested model, we use threshold amount for Jaccard similarity coefficient to detect different structures of communities. According to results for dense communities, threshold interval for Jaccard similarity coefficient is $\lambda \in [0.2, 0.9]$, this amount was obtained for highest influence diffusion through test on different data collections and its results are shown in Fig. 7. The most amount of influence diffusion is for $\lambda = 0.5$. Axis x in this figure shows different thresholds for community density and identification of its structure, and axis Y shows influence diffusion.



(a) NetHept Dataset *(b) Amazon Dataset* *(c) Live Journal Dataset*

Fig. 6 The effect of α parameter in influence spread.



(a) NetHept Dataset *(b) Amazon Dataset* *(c) Live Journal Dataset*

Fig. 7 The effect of λ parameter in influence spread.

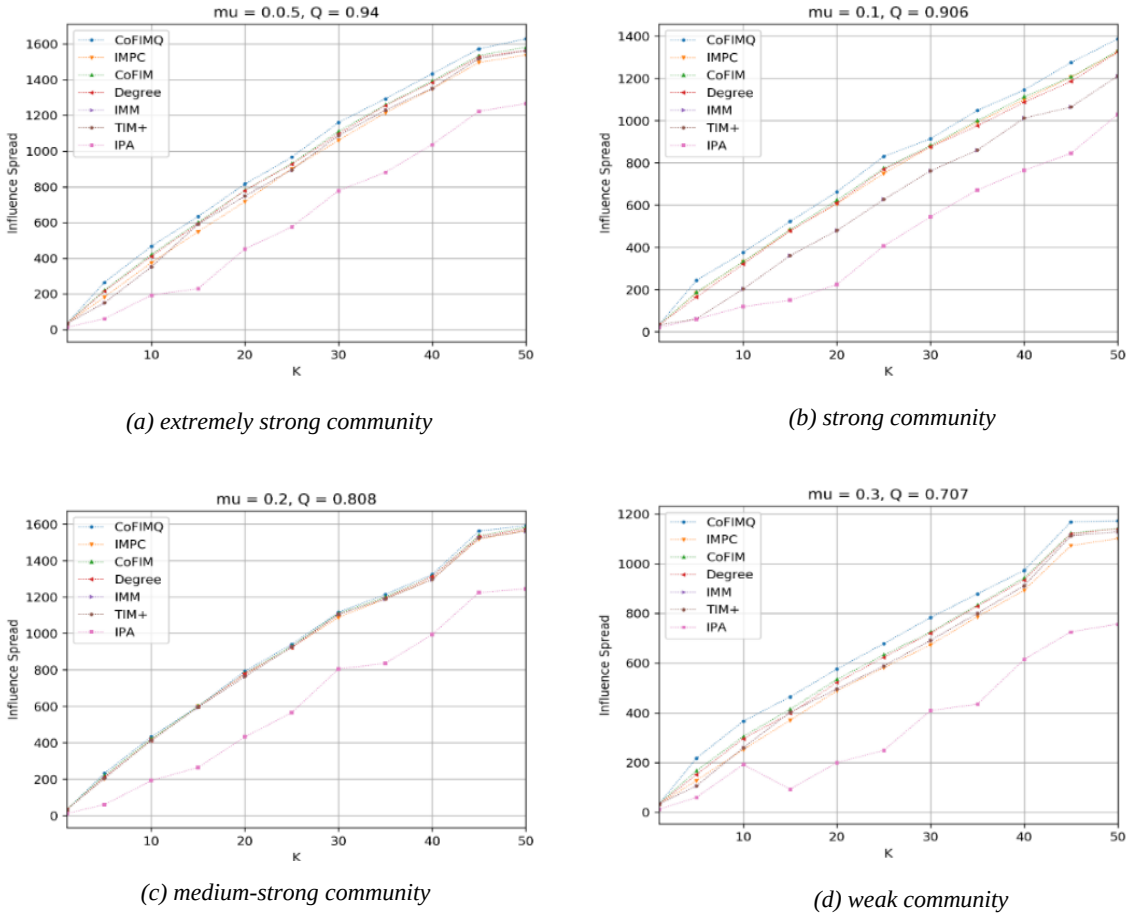
5.10 Synthetic networks results

In data collections of real world, results showed high scalability of our COFIMQ algorithm in terms of influence diffusion and relative efficiency. For more test of algorithm, we evaluated it in artificial networks that are created by using LFR algorithm (Lancichinetti et al., 2008). Since our algorithm is on the basis of community-based structure, we want to see how it influences on performance of COFIM algorithm. To do this experiment, LFR algorithm has been selected for creating artificial networks with field of adjustable social structure.

As it was mentioned in previous sections, LFR algorithm accepts these parameters: number of node N, average degree of node (k), highest degree k_{max}, index of degree distribution power τ_1 and index of community size distribution τ_2, maximum/minimum size of communities s_{max}/s_{min} and mixture parameter μ. Among them, four parameters (τ_2, s_{max}, s_{min}, μ) can be used to adjust the community structure of created networks. In our experiments, we just considered μ, because it determines community structure directly. Smaller μ means that relation in communities is more than relation between communities that means stronger structure of community. Fig. 8 presents results of four artificial LFR networks with different community

structures. Very strong community (μ = 0.5), strong community (μ = 0), average community (μ = 0) and weak community (μ=0). Other parameters of networks like N = 1000, K_{max}= 100, K = 15, S_{max} = 100, S_{min} =20, τ_1 = 2 and τ_2 = 1 are fixed.

From results of four LFR networks, it can be found that since proposed algorithm always obtains the best performance in finding effective seed nodes, again shows its improvement than other algorithms. In addition, it is obvious that community structure, can influence on algorithm efficiency significantly. As can be seen in Fig. 8, the level of identifying stronger communities can cause more improvement in algorithm and accordingly increase the amount of influence diffusion in network.



(a) extremely strong community

(b) strong community

(c) medium-strong community

(d) weak community

Fig. 8 Influence spread of Synthetic networks LFR.

6 Conclusion and Future Works

Maximizing influence is a classic diffusion optimization problem that is studied in the field of analyzing social networks and viral marketing. Although Greedy algorithm suggested by Kempe et al. (2003), provides a relative factor for optimized solution, but according to high number of Monte Carlo simulation, cannot be used for high-scale networks. Although, in terms of efficiency in execution time other sub-modularity-based or node-centrality-based exploratory algorithms, are limited but relatively good, but in terms of accuracy and precision, they cannot provide any warranty of performance. Other methods are also stated for this problem that are based on community. Experiments and researches show that due to the use of information and

characteristics of community, these methods can have more efficiency than other methods. One of the algorithms that has been known as a framework for maximizing methods of influence in community-based social networks, is COFIM algorithm that its efficiency and proficiency have been proven. However, like any other algorithm, it has disadvantageous that this method can be used through removing them. According to mentioned points in this study, we tried to improve this algorithm. To do this, since this method is based on community, through improving detected communities we tried to go forward in accordance with maximizing influence. In other words, by using different structures that obtained from communities, we could assign a different share of seed to each community and therefore improve maximizing influence and increase basic algorithm efficiency. In addition, by changing community algorithm, identification of overlapping communities has been considered.

Since diversity of virtual communications is increasing daily, therefore more researches and efforts should be performed in this field to improve these methods and create more efficient and better methods. Apportioning seed as proposed method, is performed based on communities' structure but other factors can be also used to assign seed to different communities. For example, apportioning seed can be used based on characteristics of node and social networks communities, in other words, characteristics like node special characteristics and characteristics of social, political, artistic and other communities can be used. Another suggestion is developing model in a way that supports dynamic social networks.

References

Bagheri E. 2020. A new method for maximizing influence on social networks based on node membership in communities. Network Biology, 10(4): 92-107

Bagheri E, Dastghaibyfard G, Hamzeh A. 2016. An efficient and fast influence maximization algorithm based on community detection. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). 1636-1641

Bagheri E, Dastghaibyfard G, Hamzeh A. 2018. FSIM: A fast and scalable influence maximization algorithm based on community detection. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 26(3): 53-65

Bagheri E, Dastghaibyfard G, Hamzeh A. 2021. FAIMCS: A fast and accurate influence maximization algorithm in social networks based on community structures. Computational Intelligence, 37(4): 1779-1802

Bozorgi AH. 2016. INCIM: A community-based algorithm for influence maximization problem under the linear threshold model. Information Processing and Management, 52(6): 1188-1199

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

Cao T, Wu X, Wang S, Hu X. 2010. OASNET: an optimal allocation approach to influence maximization in modular social networks. Proceedings of the 2010 ACM Symposium on Applied Computing. 1088-1094, ACM, USA

Chen W, Yuan, Y, Zhang L. 2010. Scalable influence maximization in social networks under the linear threshold model. The 2010 IEEE International Conference on Data Mining. Washington DC, USA

Chen YC, Zhu WY, Peng WC, Lee WC, Lee SY. 2014. CIM: Community-based influence maximization in social networks. ACM Transactions on Intelligent Systems and Technology, 5(2)

Clauset A, Newman ME, Moore C. 2004. Finding community structure in very large networks. Physical Review E, 70(6): 066111

Domingos P, Richardson M. 2001. Mining the network value of customers. The Seventh International Conference on Knowledge Discovery and Data Mining KDD 01. 57-66, ACM, New York, USA

Evans T, Lambiotte R. 2009. Line Graphs, Link Partitions and Overlapping Communities. arXiv, 0903.2181

Ghanbari A, Bagheri E. 2020. An influence maximization algorithm in social network using K-shell decomposition and community detection. Network Biology, 6(1): 163-168

Girvan M, Newman ME. 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of USA, 99(12): 7821-7826

Goyal A, Lu W, Lakshmanan LV. 2011. CELF++: Optimizing the greedy algorithm for influence maximization in social networks. The 20th International Conference Companion on World Wide Web. Hyderabad, India

Goyal A, Lu W, Lakshmanan LV. 2011. SIMPATH: An efficient algorithm for influence maximization under the linear threshold model. The 2011 IEEE 11th International Conference on Data Mining (ICDM). 211-220, Washington, USA

Guille A, Hacid H, Favre C, Zighed D. 2013. Information diffusion in online social networks: a survey. ACM SIGMOD Record, 42: 17-28

Jung K, Heo W, Chen W. 2012. Irie: Scalable and robust influence maximization in social networks. 2012 IEEE 12th International Conference on Data Mining (ICDM). 918-923, IEEE, USA

Kempe D, Kleinberg J, Tardos E. 2003. Maximizing the spread of influence through a social network. The ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 03. 137-146, ACM, Washington, USA

Kim J. 2013. Scalable and parallelizable processing of influence maximization for large-scale social networks. IEEE 29th International Conference on Data Engineering (ICDE). 266-277, IEEE, USA

Kumar S, Gupta A, Khatri I. 2022. CSR: A community based spreaders ranking algorithm for influence maximization in social networks. World Wide Web. https://doi.org/10.1007/s11280-021-00996-y

Lancichinetti A, Fortunato S, Kertesz J. 2009. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 11(3): 33-45

Lancichinetti A, Fortunato S, Radicchi F. 2008. Benchmark graphs for testing community detection algorithms. Physical Review, 4(78)

Lattanzi S, Sivakumar D. 2009. Affiliation networks. Proceedings of the forty-first annual ACM symposium on Theory of Computing. 427-434, ACM, USA

Leskovec J, Krause A, Guestrin, C, Faloutsos C, VanBriesen J, Glance N. 2007. Cost-effective outbreak detection in networks. the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA

Li H, Bhowmick S, Sun A, Cui J. 2015. Conformity-aware influence maximization in online social networks. The International Journal on Very Large Data Bases, 24(1): 117-141

Li Y, Fan J, Wang, Y., & Tan, K. L. 2018. Influence Maximization on Social Graphs: A Survey. IEEE Transactions on Knowledge and Data Engineering. IEEE, USA

Mohamadi-Baghmolaei R, Mozafari N, Hamzeh A. 2015. Trust based latency aware influence maximization in social networks. Engineering Applications of Artificial Intelligence, 41: 195-206

Nepal S, Bista S, Paris C. 2014. Behavior-Based Propagation of Trust in Social Networks with Restricted and Anonymous Participation. Computational Intelligence, 31(4): 642-668

Ohsaka N, Akiba, T, Yoshida, Y, Kawarabayashi KI. 2014. Fast and Accurate Influence Maximization on Large Networks with Pruned Monte-Carlo Simulations. 138-144, AAA, USA

Ok J, Jin Y, Shin J, Yi Y. 2014. On maximizing diffusion speed in social networks: impact of random seeding

and clustering. The 2014 ACM International Conference On Measurement and Modeling of Computer Systems (SIGMETRICS '14). 301-313, ACM, New York, USA

Palla G, Derényi I, Farkas I, Vicsek T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435: 814-818

Raghavan UN, Albert R, Kumara S. 2007. Near linear time algorithm to detect community structures in large-scale networks. Physics and Society, 76(3)

Rahimkhani K, Aleahmad A, Rahgozar M, Moeini A. 2015. A fast algorithm for finding most influential people based on the linear threshold model. Expert Systems with Applications, 42(3): 1353-1361

Richardson M, Agrawal R, Domingos P. 2003. Trust management for the semantic web. The 2nd International Semantic Web Conference (ISWC2003). Berlin Heidelberg, Germany

Shang JZ. 2016. CoFIM: A community-based framework for influence maximization on large-scale networks. Knowledge-Based Systems, 117: 88-100

Shang J, Liu L, Xie F, Wu C. 2014. How overlapping community structure affects epidemic spreading in complex networks. IEEE 38th International Computer Software and Applications Conference Workshops (COMPSACW). 240-245, IEEE, USA

Shang J, Wu H, Zhou S, Zhong J, Feng Y, Qiang B. 2018. IMPC: Influence maximization based on multi-neighbor potential in community networks. Physica A: Statistical Mechanics and its Applications, 512: 1085-1103

Song G, Zhou X, Wang Y, Xie K. 2015. Influence maximization on large-scale mobile social network: a divide-and-conquer method. IEEE Trans. Parallel and Distributed System, 26: 1379-1392

Sun H, Liu J, Huang J, Wang G, Jia X, Song Q. 2016. LinkLPA: A Link-Based Label Propagation Algorithm for Overlapping Community Detection in Networks. Computational Intelligence, 33(2): 308-331

Tang Y, Shi Y, Xiao X. 2015. Influence maximization in near-linear time: A martingale approach. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 1539-1554, ACM, USA

Wasserman S, Faust K. 1994. Social network analysis: Methods and applications. Cambridge University Press, USA

Xie J, Szymanski BK, Liu X. 2011. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. The 2011 IEEE 11th international conference on data mining workshops.

Yang J, Leskovec J. 2015. Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems, 42(1): 181-213

Yang J, McAuley J, Leskovec J. 2014. Detecting cohesive and 2-mode communities indirected and undirected networks. Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA

Zhang WJ. 2018. Fundamentals of Network Biology. World Scientific Europe, London, UK

Zhang WJ. 2021. Construction and analysis of the word network based on the Random Reading Frame (RRF) method. Network Biology, 11(3): 154-193

Zhang WJ, Li X. 2016. A cluster method for finding node sets / sub-networks based on between- node similarity in sets of adjacency nodes: with application in finding sub-networks in tumor pathways. Proceedings of the International Academy of Ecology and Environmental Sciences, 6(1): 13-23

Zhang X, Zhu J, Wang Q, Zhao H. 2013. Identifying influential nodes in complex networks with community structure. Knowledge-Based Systems, 42: 74-84