Community Influence Maximization Based on Flexible Budget in Social Networks

Mengdi Xiao[1,2], Peng Li[1,2(✉)], Weiyi Huang[1,2], Junlei Xiao[1,2], and Lei Nie[1,2]

[1] College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei, China
lipeng@wust.edu.cn

[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-Time Industrial System, Wuhan, Hubei, China

Abstract. The influence maximization (IM) problem is a vital issue in social networks. In community structure, community influence maximization (CIM) chooses the seed nodes based on the characteristics of the community structure instead of blindly selecting seed nodes from the entire network. However, it depends on the community size, which results in high influence nodes not being selected due to a lack of budget. In this paper, we propose a budget allocation strategy for the CIM problem. To solve the problem of less influence spread in sparse community structure, we propose a community influence maximization algorithm based on a flexible budget and adopt the reverse influence sampling (RIS) approach to sample the network structure, which reduces the time complexity of the greedy algorithm. Then, we consider the imbalance of influence expansion between communities, and we propose a balanced community influence maximization algorithm, which maintains the relative balance of the influence spread ratio between communities. In addition, we analyze the time complexity of our proposed algorithms and give a theoretical guarantee. Finally, we conduct extensive experiments on three real datasets. Compared with other baseline algorithms, the results show that the proposed algorithms have a good performance in terms of influence maximization and community influence balance.

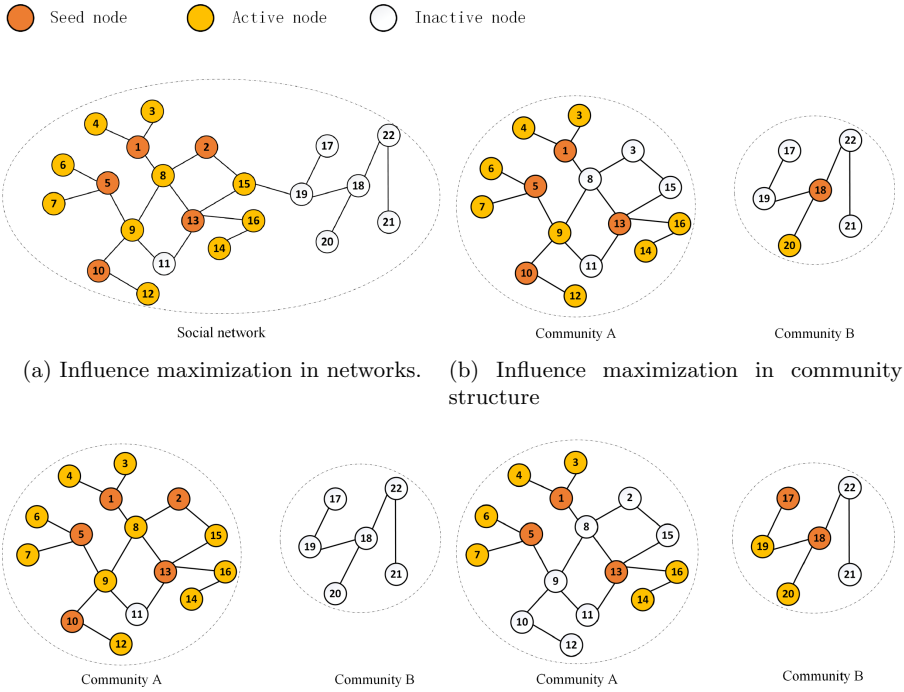
Keywords: Influence maximization · Reverse influence sampling · Social network

1 Introduction

The problem of influence maximization in social networks has attracted wide attention in recent years, and it has many applications in various aspects, such as rumor control and viral marketing. However, some influence maximization applications (such as public service advertising, safety education) are unbiased,

P. Li—This work is partially supported by the NSF of China (No. 61802286).

H. Gao and X. Wang (Eds.): CollaborateCom 2021, LNICST 406, pp. 534–553, 2021.
https://doi.org/10.1007/978-3-030-92635-9_31



(a) Influence maximization in networks. (b) Influence maximization in community structure

(c) Flexible budget based community influence maximization (d) Flexible budget based balanced community influence maximization

Fig. 1. The distribution of influence under the four conditions.

the balance of influence has become an issue that needs consideration. How to maximize the influence and ensure the balance of influence diffusion has become a critical problem.

Influence maximization aims to select seed nodes S with the maximum influence spread in the social networks. In the traditional methods [8,9], they calculate the marginal influence gain of each node by traversing the entire network, rank the nodes with the influence spread gain in each iteration, and get the maximum influence coverage with seed set S after k iterations. However, the above process may differ somewhat based on its structural characteristics in a community structure. The relationships within communities are tight, and the relationships between communities are spares. So the influence maximization method works within communities, but it does not work well between communities. Therefore, we consider influence maximization in community structure, and we use the budget allocation method between communities.

In community structure, the existing solution of the influence maximization problem is to maximize the influence spread with the predefined budget without considering the network balance. For example, as shown in Fig. 1, there are 22 nodes in the network, which can be divided into two communities A and community B with 16 nodes and 6 nodes, respectively. The result of the traditional

influence maximization algorithm is shown in Fig. 1a. While the given budget of total seed nodes k is 5, the seed set is $\{1, 2, 5, 10, 13\}$ and the number of activated nodes is 15. Figure 1b is the result based on budget allocation in the community structure. The community budget allocation is $[5 * 16/22] : [5 * 6/22] = 4 : 1$. Then, the seed set size of community A is 4, and that of community B is 1. The seed set is $\{1, 5, 10, 13, 18\}$ and the activated nodes number of community A is 12, and that of community B is 2. The total number of activated nodes in Fig. 1b is less than in Fig. 1a. Therefore, we propose a flexible budget solution as shown in Fig. 1c, where the budget fluctuation range is 1. The goal of this solution is to maximize the total influence spread, and the activated nodes number in Fig. 1c is equal to Fig. 1a. However, there is a big difference in the number ratio of active nodes between communities A and B. It shows that the community influence spread of A and B is not balanced. Therefore, we further propose a flexible budget-based balanced community influence maximization algorithm, which is shown in Fig. 1d. It considers not only the influence maximization but also considers community influence balance. Where the seed set is $\{1, 5, 13, 17, 18\}$ and the real influence of the community is 14, and the influence ratio is $(9/16):(4/6) = 27:32$, which is close to the balance.

How to solve the problem of influence maximization and community balance is what we are considering now. Some existing solutions mainly traverse the entire network to find the maximum marginal influence, and its time complexity is very high. To improve the algorithm's efficiency, we sample the community structure to obtain approximate influence spread estimates. Although our algorithm reduces the time complexity, it causes new problems. First, the fairness of the budget allocation method needs to be guaranteed. Second, we still need to maximize the whole network influence. Finally, community influence balance and influence maximization need to be weighed.

To solve the above problems, we propose a community influence maximization solution based on a flexible budget. First of all, to ensure the algorithm's fairness, we propose a simple budget allocation strategy based on community size. Second, we propose a flexible budget-based community influence maximization algorithm based on the reverse influence sampling (RIS) method. The RIS algorithm based on community structure estimates the community influence diffusion with ensuring computational efficiency. It selects the seed nodes iteratively with a strategy based on a flexible budget to obtain the maximum influence. Then, we propose a community balance influence algorithm based on RIS to solve the imbalance of community influence caused by a flexible budget. It selects seed nodes iteratively to get a trade-off between influence maximization and influence balance. We evaluate the performance of our proposed solution based on three real-world datasets. Extensive experiments show that the performance of our proposed algorithm is very close to that of traditional influence maximization, and the balance of our proposed algorithm is better than that of the existing algorithms.

In summary, we have made the following contributions:

- We propose an influence maximization algorithm for a flexible budget based on reverse influence sampling to maximum influence within a certain budget fluctuation range.
- Considering the balance of community influence distribution, we improve the above algorithm and propose a balanced influence maximization algorithm with a flexible budget. We analyze the time complexity of our proposed algorithm and give a theoretical guarantee.
- We evaluate our proposed algorithm in three real-world social datasets, and the results show our proposed algorithms have outstanding performance in both influence maximization and influence balancing, respectively, compared with other algorithms.

2 Related Work

Domingos and Richardson *et al.* [6] are the first to propose the problem of maximizing the influence in social networks. They transform the problem of viral marketing into the problem of maximizing the influence in a formulaic form. Then Kempe *et al.* [8] study the nature of the influence maximization problem. They propose that the problem is NP-hard, and they find that the greedy algorithm can get a $1 - 1/e$ approximate ratio. The greedy algorithm has very high time complexity due to its dependence on Monte Carlo simulation and the scale of the network. To reduce the time cost of the greedy algorithm, Leskovec *et al.* [9] propose a lazy-forward method (CELF), which is based on the diminishing marginal influence gain property in each round of node selection process. Talukder A. *et al.* [17] propose a knapsack-based reverse influence maximization model, where a linear threshold model is used in reverse order to estimate the minimum node cost required for nodes to be activated. Tang Jing *et al.* [18] propose a new online influence maximization algorithm, where a pause mechanism for the lack of interactivity and flexibility is provided due to the long-running time of the algorithm. The mode of information propagation in an undirected graph is studied by Schoenebeck Grant *et al.* [14]. Sun, Lichao and Huang *et al.* [16] study the multi-round influence maximization problem of adaptive and non-adaptive. Becker and R. *et al.* [3] study the influence distribution and seed node selection in attribute, community, and preference networks, and they weigh the relationship between influence maximization and balance.

Now there are some studies on the influence maximization caused by node attributes or relationships under different scenarios. Chen *et al.* [5] studies the selection of influence seed nodes when variety of marketing strategies are mixed, Banerjee *et al.* [1,2] study the complementary and competitive characteristics between projects and propose utility-driven influence maximization model. In the above papers, the marketing scenarios and the relationship between nodes in the network are considered. They all combine the maximization of influence with the possible scenarios in the network but do not consider the structural characteristics of the network. Stoica and Ana-Andreea *et al.* [15] propose a

method to solve the fairness problem in the maximization of social influence. Yang, Yu and Mao *et al.* [19] model the problem of maximizing the sustainable influence, and they propose the coordinate descent framework to consider the project discount within the budget scope. Guo *et al.* studied the influence maximization method in the community structure in [7], and proposed a simplified pipage rounding method based on guaranteeing the approximate ratio, which reduces the complexity of IMCB-Framework further. Lin *et al.* [11] proposes a RIS-based attribute sampling algorithm to tradeoff the maximization of influence and the balance of attributes. Inspired by it, we study the relationship between influence maximization and influence balance on non-attributed networks. In a non-attribute network, community structure can be easily discovered based on the tightness of the relationships between nodes.

Unlike the point of focus in the above papers, this paper focuses on the tradeoff between maximization of influence and balance based on the flexibility of community budget in community structure. In other words, we study the distribution of an unbiased advertisement across communities within the community structure. We propose a community influence maximization algorithm based on upper budget bounds and a balancing algorithm based on upper and lower budget bounds to solve the above problems.

3 System Model and Problem Formulation

In this section, we model the problem and propose a formulaic solution. We summarize the frequently used notations in Table 1.

Table 1. Frequently used notations.

Symbol	Description
$G = (V, E)$	Network G, nodes set V, edges set E
n, m	n is the number of nodes in V, m is the number of edges in E
$N(v_i)$	The neighbor nodes set of node v_i, $v_i in V$
$p_{i,j}$	The probability that node v_i active node $v_j \in N(v_i)$
C, C_i	Community partition in G, C_i is the i-th community in C
k	k is the size of seed nodes
$\sigma(S)$	The number of actived nodes by seed set S
γ	Weighting parameter
$R(v_i)$	The RR set of node v_i

3.1 System Model

We formalize the social network as graph $G = (V, E)$. $V = \{v_1, v_2, ..., v_n\}$ represents the nodes set in the network, where v_i denotes the i-th user node. E represents the edges set in the network, where $e_{i,j} \in E$ denotes the relationship

between node v_i and its neighbor node v_j. We define social network nodes size as $n = |V|$, edges size $m = |E|$. In any two nodes v_i, v_j in the network, if there is an edge $e_{i,j}$ between them, it means that node v_i can active its neighbor node v_j with probability $p_{i,j}$. The state of node v_j in graph can be defined as a_i, if node v_i is actived, then $a_i = 1$, otherwise $a_i = 0$. Actived nodes have the potential to activate its neighbor nodes in the next iteration.

In social networks, information propagation follows the influence propagation model. The main propagation process of the influence propagation model is as follows: (1) Given the seed node S and the network graph G, the state of all nodes are unactived ($a_i = 0, \forall i$) in the network. (2) In the initial time $t = 0$, the nodes v_i of the seed set S are changed to $a_i = 1$. (3) For $\forall t > 0$, all activated nodes in V try to activate their neighbor nodes in a certain probability. Node v_i in time $t - 1$ was not activated and it is activated in t, then in time $t + 1$ node v_i will try to activate the v_j where $v_j \in N(v_i)$ with probability $p_{i,j}$. (4) Iterative spread until there is no active node in the network.

There are two classical influence propagation models in social networks: the independent cascading (IC) and linear threshold (LT) models. Every node has only one chance to activate its neighbor nodes in the IC model, and every node's activation behavior is independent and unrelated. Different from the IC model, the activation behavior of the nodes in the LT model is related. The node v_i activates its neighbor node v_j if and only if $p_{i,j} \geq \delta_j$ in the IC model, where δ_j represents the activation threshold of node v_j. However, in LT model, the node v_j will be activated while $\sum_{v_i \in N(v_j)} p_{i,j} \geq \delta_j$. If the node v_j is not jointly activated by its neighbors at time t, v_j will try to be activated again while any other neighbor node is activated at any time $T > t$.

In this paper, we adopt the trigger model as our influence propagation model, which is based on the independent cascade model and the linear threshold model. Given a directed network graph $G = (V, E)$, any node $v_i \in V$ has a trigger distribution $D(v_i)$, which is the probability distribution based on the subsets of incoming neighbors $N(v_i)$, the trigger set $T(v_i)$ is obtained by randomly sampling $D(v_i)$. Trigger model can be described as: for the trigger set of node v_i, delete any incoming edge from the node not in the trigger set $T(v_i)$ to get the trigger sub-graph G', and activate any node belonging to the seed set S to get the activated nodes set $\sigma(S)$.

Classic influence maximization problem is to find a seed set S from the nodes set V, where $k = |S|$ represents the size of S, the influence of seed set S is the number of nodes activated by S, which is expressed as $\sigma(S)$. The influence of seed set S can be formulated as $\sigma(S) = \sum_{v_i \in V} a_i$, then influence maximization can be formulated as $S^* = \arg\max \sigma(S)$.

3.2 Problem Formulation

In the social networks, there is usually a community structure in Graph G. The node connections within the community are relatively close, while the node connections between communities are relatively sparse. Therefore, we use the classical community partitioning algorithm to divide the entire social network G into t disjoint community subgraphs, which is defined as $C = \{C_1, C_2, ..., C_t\}$.

We split the problem of influence maximization in social networks into two subproblems, the problem of community influence maximization and the problem of budget allocation. The community budget is defined as $B = \{b_1, b_2, ..., b_t\}$ in the community structure, and the total budget is $k = \sum_{b_i \in B} b_i$. Without loss of generality, we define the cost of each seed as the same and as a unit cost, then the budget $b_i = |S_i|$ represents the seed set size of the i-th community C_i. Based on the budget b_i, we should select the seed set with the maximum influence spread for each community. The influence spread of seed set S_i in community c_i can be expressed as $\sigma_{C_i}(S_i) = \sum_{v_i \in C_i} a_i$. Therefore, the community influence maximization problem can be formulated as:

$$S^* = \arg\max \sum_{\forall C_i \in C} \sigma_{C_i}(S_i) \tag{1}$$

In this paper, we study the problem of community influence maximization based on the flexible budget.

Definition 1 ***(Flexible Budget based Community Influence maximization, FBCIM Problem).*** *Given community subgraph $C = \{C_1, C_2, ..., C_t\}$ and budget allocation B. The FBCIM problem is to find a flexible seed budget allocation strategy, where the number of joint active nodes is maximum.*

In Fig. 1, we can see that we solve the FBCIM problem may lead to the unbalanced distribution of influence spread in the community. Therefore, we define the flexible budget based balanced community influence maximization (FBBCIM) problem as follow.

Definition 2 ***(Flexible Budget based Balanced Community Influence maximization, FBBCIM Problem).*** *Given a graph G and community distribution subgraph $C = \{C_1, C_2, ..., C_t\}$ and budget allocation B. The FBBCIM problem is to find a flexible seed budget allocation strategy, where the influence maximization and community balance are considered together.*

We use $Q(S)$ represent the set of the ratio of influence and community size in each community, where $Q(S) = \{\sigma_{C_1}(S_1)/n_1, \sigma_{C_2}(S_2)/n_2, ..., \sigma_{C_t}(S_t)/n_t\}$, n_i represents the size of community C_i. We adopt the variance $Var(Q(S))$ of influence proportion as the scale of balance evaluation. At the same time, some high-diffusion communities may allocate few seed nodes, while low-diffusion communities may allocate too many seed nodes. To avoid excessive balanced influence, we have a tradeoff between maximizing influence and balancing influence. According to the above definition, the FBBCIM problem can be expressed as a dual objective optimization problem of maximizing influence and minimizing variance. Therefore, FBBCIM problem can be formulated as:

$$S^* = \arg\max \sum_{\forall C_i \in C} \sigma_{C_i}(S_i) - \gamma n Var(Q(S)) \tag{2}$$

Where γ is the parameter, which describes the importance of variance in the influence function. n is the size of the graph, and we use it to scale up the variance value appropriately.

4 Our Solutions

4.1 General Solution

First, we need to study the rationality of budget allocation. We adopt a simple rounding method to allocate the seed budget based on the ratio of the community size and the size of the seed set. For t communities $C = \{C_1, C_2, ..., C_t\}$, the size of each community is n_i, the total budget is k, and the budget of each community is floating around kn_i/n. We use the community budget rounding method to allocate surplus budget. The steps of the community budget rounding method are as follows:

First, we need to study the rationality of budget allocation. We adopt a simple rounding method to allocate the seed budget based on the ratio of the community size and the size of the seed set. For t communities $C = \{C_1, C_2, ..., C_t\}$, the size of each community is n_i, the total budget is k, and the budget of each community is floating around kn_i/n. We use the community budget rounding method to allocate surplus budget. The steps of the community budget rounding method are as follows:

(1) Let b_i be the integer part of $k * n_i/n$ and let f_i be the fractional part of $k * nc_i/n$. We use $F = \{f_1, f_2, ..., f_t\}$, and $B = \{b_1, b_2, ..., b_t\}$
(2) We sort each element in F from largest to smallest, and the new ordered set denoted by F'.
(3) Let $q = k - \sum_{b_i \in B} b_i$, which represents the number of unallocated seed nodes.
(4) We increment the integer part by 1, that is $b_i + 1$ if f_i is in the top q of F'.

To solve the FBCIM problem, we propose a greedy algorithm with diminishing returns of internal community influence as given in Algorithm 1. We select the seed nodes with the maximum community influence for t communities with a budget allocation of B. In each iteration of each community, we select the node with the maximum influence to be added into the seed set (lines 3–5). To avoid the excessive crossover problem of influence diffusion cover between seed nodes, we sort the marginal influence gain rather than the influence of nodes (line 4), which can be expressed as $\sigma_{C_i}(v|S_i) = \sigma_{C_i}(S_i \cup \{v\}) - \sigma_{C_i}(S_i)$.

Algorithm 1. Algorithm for community influence maximization

Input: social network $G(V, E)$, community $C = C_1, C_2, ..., C_t$, seed budget allocation $B = \{b_1, b_2, ..., b_t\}$
Output: seed set S

1: Initialize: $S \leftarrow \emptyset$
2: **for** $i = 1, \ldots, t$ **do**
3: **for** $j = 1, 2, ..., b_i$ **do**
4: $v = \arg\max \sigma_{c_i}(v|S_i)$
5: $S_i \leftarrow S_i \cup \{v\}$
6: Let $S \leftarrow S \cup S_i$
7: return S

We analyze the time complexity of the Algorithm 1. In Algorithm 1, seed nodes are selected for t communities according to their budget settings, and b_i seeding rounds are selected in each community. The marginal influence of nodes in each iteration is calculated. In the greedy algorithm, we most often use the Monte Carlo simulation method to estimate the influence spread of the seed set. In Monte Carlo simulation, the selected seed nodes simulate the information propagation process from the edge relationship according to the corresponding propagation model. The process of a simulation is traversing the edge of the network once, and the time complexity is $O(n_i m_i)$. Assuming we simulate H propagations for each community, the time complexity of Algorithm 1 can be expressed as $O(H \sum_{C_i \in C} b_i n_i m_i) = O(Htknm)$.

4.2 FBCIM Algorithm

It is well known that Monte Carlo greedy algorithm is not scalable and has high computational costs. In this paper, we propose a scalable algorithm based on Reverse Influence Sampling (RIS). It can give an approximate estimate of the influence and has a theoretical guarantee. We use the critical concept Reverse Reachable (RR) set to describe RIS.

Definition 3. ***Reverse Reachable (RR) set****: Given a graph $G = (V, E)$, the RR set $R(v)$ is obtained by sampling the trigger set $T(v)$ with a certain probability.*

According to Definition 3, any node in the RR set $R(v)$ can propagate influence to node v. Any node in the RR set $R(v)$ is activated, then v is activated. Therefore, the reverse reachable (RR) set represents the set of any node when the node is not selected.

Reverse Influence Sampling (RIS) algorithm can be described as follows: First, we randomly sample the trigger set $T(v)$ of any node v to obtain its reverse reachable set $R(v)$, where $R(v)$ represents the nodes set that can activate node v. Then, we sample enough reverse reachable sets to obtain the approximate influence results. When we sample more RR sets, the seed set S will cover more RR sets, the number of activated nodes will be more. Finally, we use a greedy algorithm to calculate the influence spread and get the seeds set with the maximum influence gain.

Lemma 1. *Given disjoint community partitioning $C = \{C_1, C_2, ..., C_t\}$, a RR set $R(v)$, the number of nodes n_i in community C_i, the influence of S can be estimated as: $\mathbb{E}[\sigma(S)] = \sum_{C_i \in C, S_i \in S} n_i \cdot Pr[S_i \cap R(v) \neq \emptyset | v \in C_i]$.*

Proof. For the seed set S_i of community C_i, the proportion that S_i covers the community C_i is equal to the probability that S_i activates random v. The influence of S_i can be estimated as: $\mathbb{E}[\sigma_{C_i}(S_i)] = n_i \cdot Pr[S_i \cap R(v) \neq \emptyset]$, where $Pr[S_i \cap R(v) \neq \emptyset]$ represents the probability that the intersection of the seed set S_i and the RR set $R(v)$ of a random node $v \in C_i$ is not empty. Therefore, the influence of S can be estimated as: $\mathbb{E}[\sigma(S)] = \sum_{C_i \in C, S_i \in S} n_i \cdot Pr[S_i \cap R(v) \neq \emptyset | v \in C_i]$.

For the CIM problem, we propose the Flexible Budget-based Community Influence Maximization (FBCIM) algorithm based on RIS. Our method includes two processes: RR set sampling and seed selection.

RR Set Sampling. First, we generate the RR sets in all community structures. In each community, we sample θ_i nodes and generate θ_i RR sets. Theoretically, the more RR sets are sampled, the higher the approximation of influence results. However, when the number of nodes in the network is relatively large, the number of RR sets is also significant, it will lead to the more calculation cost. So we need to calculate a lower bound on the number of RR sets, that is, to obtain the nearest similar influence result with fewer RR sets.

Unlike traditional influence propagation, node v is activated if and only if the intersection of $R(v)$ and S is not an empty set. In RIS sampling, the activation probability of a node is the probability that the seed set S contains the trigger set of any node. Let $\mathbb{E}$ represent expectation, $\sigma(S_i)$ represent the number of active nodes in the reverse influence sampling model, θ_i and represent the number of RR sets sampled by the community C_i. Based on Lemma 1, the influence of seed set S can be expressed as follows:

$$\sigma(S^*) = \sum_{\forall C_i \in C} \frac{n_i}{\theta_i} \mathbb{E}[\sigma_{C_i \in C}(S_i)] \tag{3}$$

Lemma 2. *The expected influence $\sigma(S)$ is a monotone and submodular function under RIS model.*

Proof. From Lemma 1, we know the influence of S can be expressed as $\sigma(S) = \sum_{C_i \in C, S_i \in S} \sigma_{C_i}(S_i)$. Let $f(S) = \sigma(S)$, $g(S_i) = \sigma_{C_i}(S_i)/n_i = Pr[S_i \cap R(v) \neq \emptyset]$. In the RIS algorithm, we select seed nodes according to the maximum marginal revenue of S covering R, that is, the revenue of nodes selected in each round is diminishing compared with that of the last round. However, with the increase in the number of seed nodes, the overall revenue is increasing. So, $g(S_i)$ is nonnegative, monotonic, and submodular. Because $f(S) = \sum_{S_i \in S} n_i \cdot g(S_i)$, the function $f(S)$ is nonnegative, monotonic, and submodular.

Lemma 3 ***(Chernoff Bound).*** *Let $X_i \in [0, 1]$ be θ i.i.d random variables with a mean μ. For any $\delta > 0$, $Pr[|\sum X_i - \theta\mu| \geq \delta \cdot \theta\mu] \leq 2\exp(-\frac{\delta^2}{2+\delta} \cdot \theta\mu)$.*

Lemma 3 is the classical Chernoff Bound theoretical analysis. We use it to get the lower bound of θ_i in every community. The low bound means the minimum number of RR sets what the influence estimation and approximation assurance needs.

Lemma 4. *Given community subgraph $C = \{C_1, C_2, ..., C_t\}$, the number of community nodes is n_i, if θ_i satisfies:*

$$\theta_i \geq \frac{n_i(\epsilon + 2)l \log 2n_i}{OPT_i \cdot \epsilon^2} \tag{4}$$

Then, for any set $S_i \in C_i$*, the following inequality* $|n_i Pr[S_i \cap R \neq \emptyset] - E[\sigma(S_i)]| < \epsilon \cdot OPT_i$ *holds with at least* $1 - 1/{n_i}^l$ *probability.*

Proof. In community C_i, let x_i represents $Pr[S_i \cap R \neq \emptyset]$, which is the probability that the intersection of the seed set with an random RR set is not empty. $E[\sigma(S_i)]$ is the expected influence spread of seed set S_i, μ_i represents the expected activation probability of the node, it can be denoted as $\mu_i = E[\sigma(S_i)]/n_i$. According to Lemma 3 (Chernoff bound), we have:

$$Pr[|n_i x_i - E[\sigma(S_i)]| \geq \epsilon \cdot OPT_i]$$

$$= Pr[|\theta_i x_i - \theta_i \mu_i| \geq \tfrac{\epsilon \cdot OPT_i}{n_i \mu_i} \cdot \theta_i \mu_i]$$

$$\leq 2\exp(-\tfrac{\epsilon^2 \cdot {OPT_i}^2}{n_i(2 n_i \mu_i + \epsilon \cdot OPT_i)} \cdot \theta_i)$$

$$\leq 2\exp(-\tfrac{\epsilon^2 \cdot OPT_i}{n_i(2+\epsilon)} \cdot \theta_i) \leq 1/{n_i}^l$$

Therefore, when $\theta_i \geq \frac{n_i(\epsilon+2) l \log 2n_i}{OPT_i \cdot \epsilon^2}$, the inequality $|n_i x_i - E[\sigma(S_i)]| \geq \epsilon \cdot OPT_i$ holds with probability $1/{n_i}^l$. Lemma 4 is proved.

Seed Selection. We further optimize Algorithm 1 and propose a FBCIM algorithm based on a flexible budget. Given each community budget and a float parameter, the algorithm allows the actual budget to fluctuate within this parameter. In FBCIM algorithm, the flexible budget of community C_i is defined as $b_i^{'} \leq b_i + \lambda$, where b_i is the budget allocated based on community size, and λ is a flexible parameter, which stipulates the scope of budget fluctuations. As shown in Algorithm 2, under the constraint of total budget k, the node with the maximum influence is iteratively selected to seed set, and the iteration is carried out until the size of seed set within each community is equal to the maximum budget fluctuation upper bound. This seed selection method ensures that the seed nodes selected in each round are optimal. That is, the seed node selection process is a monotonic submodule.

Algorithm 2. FBCIM

Input: social network $G(V, E)$, community subgraph $C = \{C_1, C_2, ..., C_t\}$, seed budget allocation $B = \{b_1, b_2, ..., b_t\}$, θ_i RR sets in random community C_i

Output: seed set S

Initialize: $S = \{S_1, S_2, ..., S_t\} \leftarrow \emptyset$
for $j = 1, 2, ..., k$ **do**
  for $i = 1, \ldots, t$ **do**
    if $|S_i| \leq b_i + \lambda$ **then**
      $v^i = \arg\max \sigma_{c_i}(v|S_i)$
  Let $\sigma_{v^z}(v^z|S_z)$ be the max in $\{\sigma_{v^1}(v^1|S_1), \sigma_{v^2}(v^2|S_2), ..., \sigma_{v^t}(v^t|S_t)\}$
  $S_z \leftarrow S_z \cup \{v^z\}$
return S

In Algorithm 2, we find the nodes with the most significant influence within each community (Lines 4–5), rank the influence of these nodes, select the nodes with the most influence, and allocate the budget to the community (Line 6). To ensure the algorithm's fairness, we set the size of the community seed set according to the upper bound of the given community budget (Line 4). The above process iterates k rounds and assigns the node with the most significant influence to the community as the seed node (Lines 2–7).

4.3 FBBCIM Algorithm

To alleviate the influence imbalance caused by influence maximization in Algorithm 2, we propose a Flexible Budget-based Balanced Community Influence Maximization (FBBCIM) algorithm as shown in Algorithm 3. Different from Algorithm 2, we restrict the upper and lower bounds of the community budget. The lower bound is set to avoid ignoring the fairness of the community budget for obtaining a more balanced influence. In addition, because the proportion of community influence changes during the seed selection process, we propose redistributing the community budget according to the change of seed size and corresponding influence. The algorithm process is as follows:

Algorithm 3. FBBCIM

Input: social network $G(V, E)$, community subgraph $C = \{C_1, C_2, ..., C_t\}$, seed budget allocation $B = \{b_1, b_2, ..., b_t\}$, θ_i RR sets in random community C_i
Output: seed set S

1: Initialize: $S = \{S_1, S_2, ..., S_t\} \leftarrow \emptyset$
2: **for** $i = 1, \ldots, t$ **do**
3: **for** $j = 1, 2, ..., b_i + \lambda$ **do**
4: $v = \arg\max \sigma_{C_i}(v|S_i)$
5: $S_i \leftarrow S_i \cup \{v\}$
6: $f(C_i, S_i) = \sigma_{C_i}(S_i)$
7: **if** $|S_i| \geq (b_i - \lambda)$ & $|S_i| \leq (b_i + \lambda)$ **then**
8: append $(j, f(C_i, S_i))$ to $list_i$
9: **if** $\sum_{C_i \in C} |S_i| = k$ **then**
10: $S^* = \arg\max \sum_{\forall c_i \in C} \sigma_{C_i}(S_i) - \gamma nVar(Q(S))$
11: return S^*

First, we find the node with the most significant influence within the upper budget bounds of each community (Lines 3–5) and store the influence change list of upper and lower budget bounds (Lines 7–8). The generated community influence list is selected, and the nodes with the most significant total community influence and the minor total variance are sorted as the seed set (Lines 9–10). Different from Algorithm 2, Algorithm 3 restricts the lower bound of the community budget to avoid excessive transfer of influence.

The time complexity of Algorithm 2 and Algorithm 3 are analyzed as follows. The running process can be divided into two parts in these two algorithms: RR set sampling and seed nodes selection. In Lemma 4, we have got the number of RR sets per community, which is θ_i. Then, we construct the sample subgraph and θ inverse reachable set, the time complexity is $O(\sum \theta_i m_i) = O(\sum_{C_i \in C} b_i m_i \log n_i/\epsilon^2)$. In the node selection stage, we calculate the number of occurrences of each node in the RR set, and each community iterates b_i times, then the time complexity can be expressed as $O(\sum b_i n_i)$. The total time complexity of the algorithm is $O(\sum_{C_i \in C} b_i(m_i + n_i) \log n_i/\epsilon^2) = O(tk(m+n) \log n/\epsilon^2)$, which is much less than the time complexity of Algorithm 1.

5 Performance Evaluation

In this section, we run our proposed algorithms on several real networks and study their distribution in the networks. By comparing our proposed algorithm with some other algorithms, we demonstrate the effectiveness and efficiency of the proposed algorithms.

5.1 Datasets and Parameters Setting

We run our experiment on the following three real-world networks with statistics summarized in Table 2. The Facebook [10] dataset is collected from survey participants using this Facebook app, which is from Stanford Network Analysis Project. LastFM [13] dataset is a music communication platform. The Government [12] dataset is one category of Facebook pages.

Table 2. Dataset statistics.

Dataset	n	m	Direction	Density
Facebook	4039	88234	Undirected	0.001
LastFM	7624	27806	Undirected	0.001
Government	7057	89455	Undirected	0.0036

In the given network datasets above, we use a community discovery algorithm to mine the community structure. In this paper, we use Louvain [4] algorithm to find the community structure of the networks. This algorithm calculates the benefit of modularity from nodes clustering and obtains the community structure by maximizing the network's modularity. We obtained 12, 15, and 19 communities through the community partitioning algorithm in the Facebook, LastFM, and Government datasets, respectively.

In our experiment, we set the propagation probability of the independent cascading model as 0.1. To ensure the accuracy of the influence calculation, we set the Monte Carlo simulation number to 1000, which is used to estimate the

influence of the greedy algorithm. In RIS sampling, we set the node sampling probability as 0.1 and set $\varepsilon = 0.1$. In addition, we set the parameter γ in algorithm FBCIM and algorithm FBBCIM to 1. In all experiments, we set the total budget k from 20 to 100, where the unit range is 20. For each experiment, we run five times under the same settings to get the influence estimate of the algorithm.

5.2 Comparison of Algorithms and Metrics

To evaluate our proposed algorithms, we test our proposed algorithms and the baseline algorithms as following:

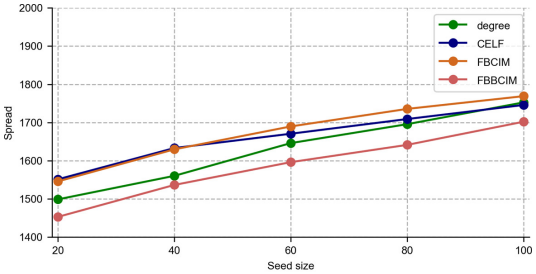
(1) CELF: CELF is an improved algorithm of the CELF algorithm, which reduces certain time complexity by simplifying the marginal revenue comparison process.
(2) degree: For the heuristic algorithm of degree centrality, the nodes with high degree centrality are selected as seed nodes. In this experiment, we use the DegreeHeuristic algorithm in combination with Algorithm 1 to obtain the comparative data.
(3) FBCIM: Our proposed Flexible Budget-based Community Influence Maximization algorithm considers reasonable budget allocation based on maximizing influence.
(4) FBBCIM: Our proposed Flexible Budget-based Balanced Community Influence Maximization algorithm takes balance into account based on our FBCIM algorithm.

To measure the performance of the algorithm, we use the four criteria:

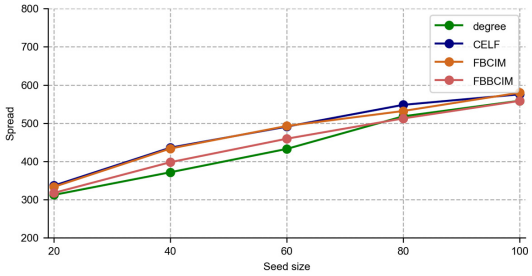
(1) Expected influence spread: the total number of activated nodes in a graph with seed set S.
(2) Variance of the influence spread ratio between communities: it is denoted by $\gamma nVar(S)$, where n represents the size of the network. This parameter is added to enlarge the variance value. This evaluation metric represents the balance of community influence spread.
(3) The trade-off between maximizing influence spread and community balance: It is denoted by $\sum_{c_i \in C} \sigma_{c_i}(S_i) - \gamma nVar(S)$, where $\sum_{c_i \in C} \sigma_{c_i}(S_i)$ is the influence spread, and $\gamma nVar(Q(S))$ is the community balance.
(4) Running time: The time it takes to run the algorithm. The unit of time is second.

5.3 Evaluation Results

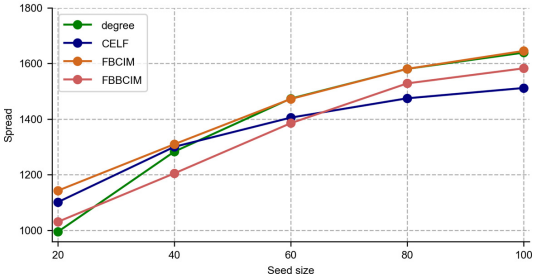
Figure 2 shows the performance of our proposed algorithms on the expected influence spread compared with two other algorithms. Expected influence spread is the most direct way to demonstrate the algorithmic effectiveness in influence maximization. The larger the result of the algorithm, the higher the quality of



(a) Facebook



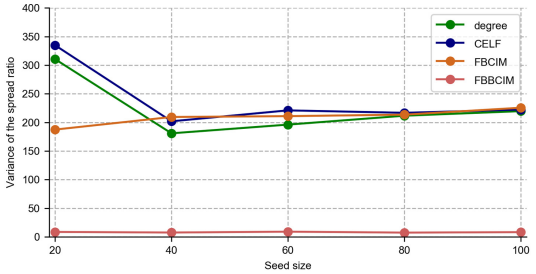
(b) LastFM



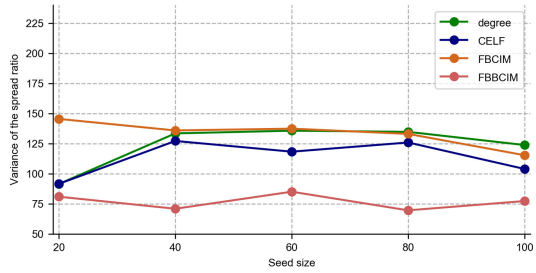
(c) Government

Fig. 2. Expected influence spread in the three different networks.

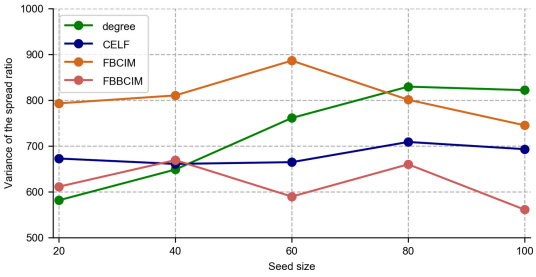
the algorithm. In all three subgraphs, FBCIM consistently outperforms all other algorithms. The reason for this is the flexible budget strategy of the FBBCIM algorithm. The performance of the FBBCIM algorithm is similar or slightly less than that of the other algorithms. This is because the FBBCIM will lose some influence to balance the community influence ratio, while the other three algorithms all seek to maximize influence spread. Therefore, it can be concluded from these results that our proposed FBCIM algorithm has outstanding performance in influence maximization, and the FBBCIM algorithm also performs well.



(a) Facebook



(b) LastFM

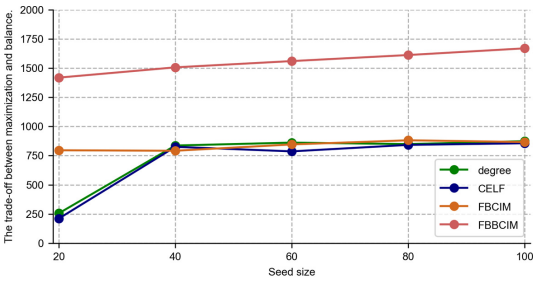


(c) Government

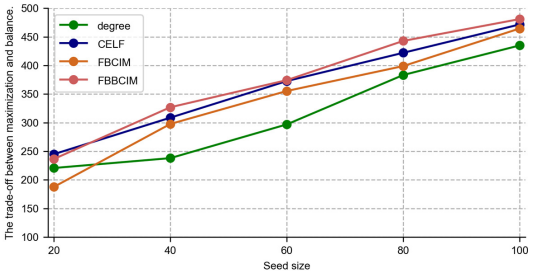
Fig. 3. Variance of the influence spread ratio between communities in the three different networks.

Figure 3 compares the variance of the influence ratio generated by the four algorithms in the three networks. To show more directly the difference in the variance of influence proportions, we multiply each result by the network size and scaling parameters γ. In this experiment, we set the parameter $\gamma = 10$. In the three subfigures, the variance of community influence ratio in FBCIM, CELF, and Degree algorithms is similar, and higher than that of FBBCIM. The reason is that seed nodes differ in their ability to disseminate information within the community, even if there is a relatively equitable budget allocation strategy. The FBBCIM algorithm that we propose reduces the difference in the proportion of community influence to minimize the difference in the propagation ability of

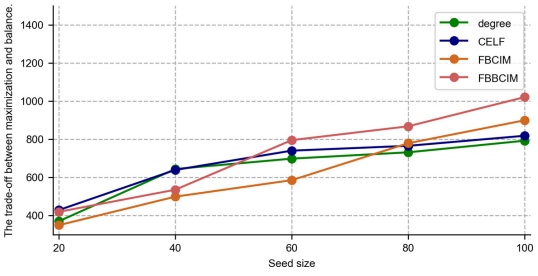
nodes. As depicted of that, the variance of the influence spread ratio in FBBCIM is the smallest among these four algorithms, so its performance is the best.



(a) Facebook



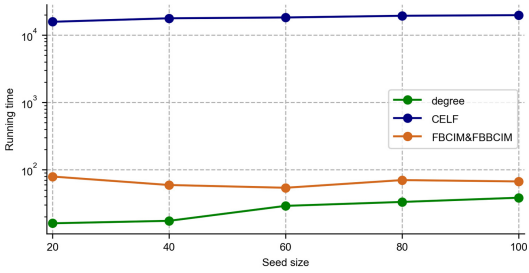
(b) LastFM



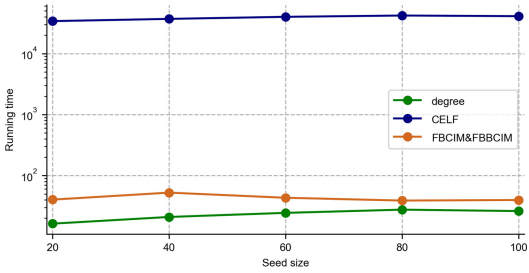
(c) Government

Fig. 4. The trade-off between community influence maximization and community balance in the three different networks.

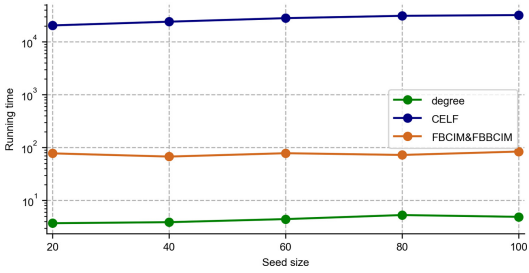
Figure 4 draws performance comparison of the trade-off between community influence maximization and community balance. Here, both the algorithm's influence scale and community balance are taken into account. As can be seen from the figure, the performance of FBBCIM proposed by us is better than the other three algorithms. Moreover, the FBCIM algorithm is not the worst among the three networks. Therefore, when community balance is valued, the FBBCIM algorithm is our best choice.



(a) Facebook



(b) LastFM



(c) Government

Fig. 5. Running time in the three different networks.

Figure 5 shows the running times of the four algorithms in three datasets. In the earlier algorithm analysis, we have known that the time complexity of FBCIM and FBBCIM is the equivalent, so their running time is coincident. Therefore, in this figure, we describe both FBCIM and FBBCIM with one line. From the three subfigures, the degree algorithm has the shortest running time in the three subfigures because the degree algorithm is heuristic. Moreover, the degree algorithm has no solid theoretical guarantee. The CELF algorithm has consistently had the most extensive running time, several orders of magnitude more than the FBCIM and Degree algorithms. It is because the CELF algorithm is based on the greedy algorithm. Although there is a theoretical guarantee in CELF, the running time is too great. The FBCIM and FBBCIM algorithms that

we propose are based on sampling, significantly reducing the calculation cost. In addition, these two algorithms have theoretical guarantees.

Summary: (1) Our algorithm FBCIM achieves a higher influence spread than degree and CELF. (2) FBBCIM algorithm is the most balanced algorithm compared to FBCIM, CELF and degree, and it performs best when both influence maximization and community balance are considered. (3) The time cost of FBCIM, FBBCIM, and Degree algorithm is much less than that of CELF, while the degree algorithm is a heuristic algorithm without a solid theoretical guarantee. The time cost of FBCIM and FBBCIM is the same and can guarantee the efficiency requirements in large algorithms.

6 Conclusion

In this paper, we study the influence maximization problem based on budget in community structure. We design a simple budget allocation strategy and propose two scalable algorithms: FBCIM and FBBCIM, which solve community influence maximization and balanced community influence maximization, respectively. The RIS method is adopted in CIM to reduce the calculation cost-effectively on guaranteeing the algorithm approximation. FBBCIM is based on FBCIM. We add variance parameters to reduce the imbalance between communities. Finally, we test the proposed algorithms on three real networks and show their effectiveness.

References

1. Banerjee, P., Chen, W., Lakshmanan, L.V.S.: Maximizing social welfare in a competitive diffusion model. Proc. VLDB Endow. **14**(4), 613–625 (2020)
2. Banerjee, P., Chen, W., Lakshmanan, L.V.S.: Maximizing welfare in social networks under a utility driven influence diffusion model. In: Proceedings of the 2019 International Conference on Management of Data (2019)
3. Becker, R., Corò, F., Angelo, G., Gilbert, H.: Balancing spreads of influence in a social network. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3–10 (2020)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**, P10008:1–12 (2008)
5. Chen, W., Wang, C., Wang, Y.: 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 (2010)
6. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2001)
7. Guo, J., Wu, W.: Influence maximization: seeding based on community structure. ACM Trans. Knowl. Discov. Data **14**(6), 1–22 (2020)
8. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)

9. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2007)
10. Leskovec, J., Sosič, R.: SNAP: a general-purpose network analysis and graph-mining library. ACM Trans. Intell. Syst. Technol. (TIST) **8**(1), 1–20 (2016)
11. Lin, M., Li, W., Lu, S.: Balanced influence maximization in attributed social network based on sampling. In: Proceedings of the 13th International Conference on Web Search and Data Mining (2020)
12. Rozemberczki, B., Davies, R., Sarkar, R., Sutton, C.: GEMSEC: graph embedding with self clustering. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019 (2019)
13. Rozemberczki, B., Sarkar, R.: Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020) (2020)
14. Schoenebeck, G., Tao, B.: Influence maximization on undirected graphs: toward closing the (1–1/e) gap. ACM Trans. Econ. Comput. **8**(4), 1–36 (2020)
15. Stoica, A.A., Chaintreau, A.: Fairness in social influence maximization. In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 569–574 (2019)
16. Sun, L., Huang, W., Yu, P.S., Chen, W.: Multi-round influence maximization. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018)
17. Talukder, A., Alam, M., Tran, N.H., Niyato, D., Hong, C.S.: Knapsack-based reverse influence maximization for target marketing in social networks. IEEE Access **7**, 44182–44198 (2019)
18. Tang, J., Tang, X., Xiao, X., Yuan, J.: Online processing algorithms for influence maximization. In: Proceedings of the 2018 International Conference on Management of Data, pp. 991–1005 (2018)
19. Yang, Y., Mao, X., Pei, J., He, X.: Continuous influence maximization. ACM Trans. Knowl. Discov. Data **14**(3), 1–38 (2020)