


RESEARCH

Open Access



# Non-submodular model for group profit maximization problem in social networks

Jianming Zhu<sup>1\*</sup> , Smita Ghosh<sup>2</sup>, Weili Wu<sup>2</sup> and Chuangen Gao<sup>3</sup>

\*Correspondence:

jmzhu@ucas.ac.cn

<sup>1</sup> School of Engineering Science, University of Chinese Academy of Sciences, 19A Yuquan Rd., Beijing 100049, China

Full list of author information is available at the end of the article

## Abstract

In social networks, there exist many kinds of groups in which people may have the same interests, hobbies, or political orientation. Sometimes, group decisions are made by simply majority, which means that most of the users in this group reach an agreement, such as US Presidential Elections. A group is called *activated* if  $\beta$  percent of users are influenced in the group. Enterprise will gain income from all influenced groups. Simultaneously, to propagate influence, enterprise needs pay advertisement diffusion cost. *Group profit maximization* (GPM) problem aims to pick  $k$  seeds to maximize the expected profit that considers the benefit of influenced groups with the diffusion cost. GPM is proved to be NP-hard and the objective function is proved to be neither submodular nor supermodular. An upper bound and a lower bound which are difference of two submodular functions are designed. We propose a submodular–modular algorithm (SMA) to solve the difference of two submodular functions and SMA is shown to converge to a local optimal. We present an randomized algorithm based on weighted group coverage maximization for GPM and apply sandwich framework to get theoretical results. Our experiments verify the efficiency of our methods.

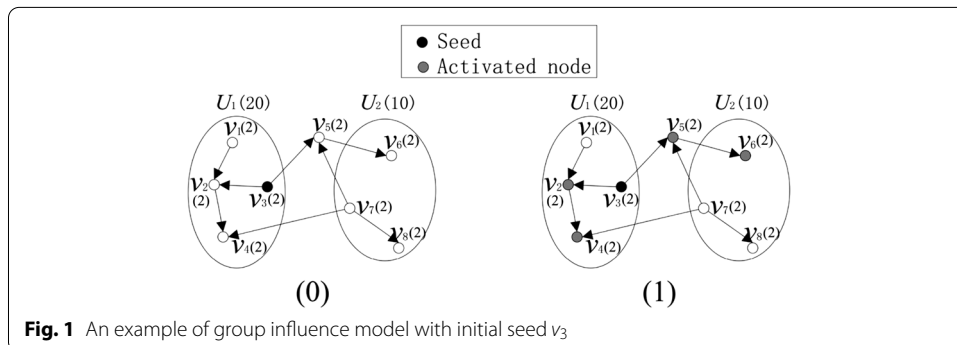
**Keywords:** Group profit maximization, Non-submodular, Submodular–modular algorithm, Social networks

## Introduction

In social society, no one is isolated and he must belong to some communities. Understanding his behavior needs to understand his groups [1]. People's behavior is influenced by group behavior and many of the world's decisions are done by groups or teams. Various types of groups exist not only in real-world society but also online social networks (OSN). The size of some groups may be smaller with only several members like family, while some groups may consist with hundreds of people, for example a school, even a whole country. With the rapid development and rising population of OSN such as Facebook with about 2.2 B users, WeChat with more than 1.0 B users, and Twitter with over 0.34 B users, etc. [2], hundreds of millions of users are able to be friends and exchange information with each other. Users with the same interests or hobbies may formulate group to talk over the common topics. In Wechat platform, WeChat group and circle of friends are very popular functions for each Wechat user. US Presidential Elections is another example. Presidential candidate will

gain all votes in one state if he gets the majority of tickets in the state. For simplification, the benefits from all activated groups are assumed to be calculated as economic indicator. Then, in this paper, both the cost and benefit are considered as monetary.

Since group holds an important role not only in real-world society but also online social networks, the enterprise (such as company), or government attempts to activate group. For example, when we consider family’s decision, usually, only one decision is done to buy some brand product according to the advertise of different brands. Similarly, a company need to purchase computers for their employees, while this company may use majority method to decide which brand of computer. The employee may be influenced by different brand of computers, while only one brand is purchased. A group is called to be activated if a certain percent of members are influenced. Enterprise producers often draw support from the OSN providers to diffuse their advertisements, so that all possible potential groups could be influenced. Zhu et al. [3] have presented the group influence maximization problem in social networks. In which, a group is called to be *activated* if  $\beta$  percent of members in this group are activated. Enterprises will gain income from all activated groups. Simultaneously, to propagate influence, enterprise needs pay advertisement diffusion cost to the OSN provider, while the cost is usually up to total hits on these advertisements. In this paper, we aim to pick  $k$  seed users to maximize the expected profit that maximize the expected profit that equals the benefit of influenced groups minus the diffusion cost. This optimization problem is called *group profit maximization* (GPM) problem. Given a social network  $G = (V, E, P)$ ,  $P$  is the influence probability for each directed edge  $(u, v)$  that means  $u$  could activate  $v$  with probability  $P$  after  $u$  becomes activated.  $\beta$  is called group activated threshold. For each activated group  $U$ , the benefit is  $b(U) \geq 0$ . Meanwhile, the diffusion cost  $c(v) \geq 0$  is required if  $v$  is activated. An example is shown in Fig. 1. There are 8 nodes in this graph and the influence probability equals 1. There are two groups.  $U = \{U_1 = \{v_1, v_2, v_3, v_4\}, U_2 = \{v_6, v_7, v_8\}\}$ . Benefit of group  $U_1$  is 20 and  $U_2$  is 10, while diffusion cost of each node is 2. Assume the activation threshold  $\beta = 0.5$  which means a group will be activated if at least half of nodes are activated. Figure 1(1) chooses  $v_3$  as the seed, and then,  $\{v_2, v_3, v_4, v_5, v_6\}$  will be activated and only group  $U_1$  is activated under the activation threshold 0.5. Then, the total profit is  $b(U_1) = 20 - 5 \times 2 = 10$ .



### Related works

Kempe et al. [4] first presented influence maximization (IM) problem. They showed that IM problem was NP-hard under independent cascade (IC) model. And the objective function of IM was submodular. Following Kempe's work, [5–12] have studied different types of IM problems. Realizing the exist of crowd influence, IM problem with considering crowd influence was studied by Zhu [13–15].

Optimizing the profit return in viral marketing has proved much more difficult than only maximizing the influence propagation [16], since the number of seeds picked yields a trade-off between the benefit and cost of viral marketing. Several recent publications studied profit maximization problems from the advertiser's point [16–18]. These works considered the cost of seed selection which is modular and implies that their profit metric is still submodular. Ref. [19] proposed a profit maximization problem which took into account the cost of information propagation, whose profit function could be decomposed into the difference of two submodular functions.

However, most of the existing methods are either too slow for billion-scale networks such as Facebook, Twitter, and World Wide Web or fail to retain the  $(1 - 1/e - \epsilon)$ -approximation guarantees. The sampling method is the bottleneck of solving IM. Borgs et al. [20] proposed a novel sampling method named reverse influence set (RIS) which can reduce the sampling complexity. Two-phase influence maximization (TIM)/TIM+ [21] and Influence Maximization via Martingales (IMM) [22] were introduced for solving IM problem. Nguyen et al. [23] made a breakthrough and proposed Dynamic-Stop-and-Stare Algorithm (D-SSA) which was much faster while guarantee the same approximation ratio. Zhu [13] presented weighted RIS sampling method.

Since the objective function of GPM is non-submodular which will be shown in the following section, the existing social IM methods can not be applied to solve the GPM. Schoenebeck [24] presented the 2-quasi-submodular function optimization problem whose objective was non-submodular. Narasimhan and Bilmes [25] presented an approximation method for solving submodular + supermodular function which substituted the supermodular function by a modular function. Bach [26] proved that any non-submodular function could decompose as a difference of two submodular function. Another approach named sandwich approximation strategy was presented by [27], which approximates the objective function by formulating its lower bound and upper bound. More recent results can be found in [28, 29].

### Contributions

We summarize our contributions as follows:

1. Motivated by the group structure in social network, group profit maximization (GPM) problem is presented which select  $k$  seeds, such that the expected profit is maximum.
2. We evaluate the challenges of the GPM by analyzing computational complexity. First, GPM is proved to be NP-hard under IC model. Second, the objective function of GPM is shown neither submodular nor supermodular.

3. To obtain approximate solution, we propose a lower and upper bound for the objective function. We show that maximizing the lower bound and upper bound are still NP-hard. Meanwhile, both lower bound and upper bound can be decomposed to the difference of submodular functions. We also present a submodular–modular algorithm to solve the difference of submodular functions.
4. Then, we propose a weighted group coverage maximization algorithm for solving GPM. Second, we formulate a sandwich approximation framework, which preserves a theoretical analysis result. We verify our algorithm on real-world data sets.

This paper is organized as follows: first, we present the group profit maximization (GPM) problem; then, the proof of NP-hardness and properties of objective function will be given; third, we propose lower bound and upper bound, and present our algorithms; experiments are presented in the following section; and finally, the paper is concluded. Table 1 summarizes the symbols and their meaning.

### Problem formulation

Independent cascade (IC) model is an information propagation model with widely application. IC model will be introduced first, and then, the group profit maximization (GPM) problem is presented.

#### Independent cascade model [4]

Given an social network  $G = (V, E, P)$ , where  $V$  is a set of users and  $E$  is a set of directed edges. For each edge  $e = (u, v)$ ,  $P_e$  is the weight on  $e$ , representing the information activation probability ( $0 \leq P_e \leq 1$ ). Specifically,  $u$  will attempt to activate  $v$  with activation probability  $P_e$  after  $u$  is activated.

Assume  $S \subseteq V$  is the initial seed users. Let  $S_t$  be the nodes which are activated in step  $t$  ( $t = 0, 1, \dots$ ). At the beginning,  $S_0 = S$ . The propagation process is as follows step by step. At step  $t$ , for each activated node in  $u \in S_t$ ,  $u$  will try to activate each inactivated neighbor

**Table 1** Frequently used notation

Notation	Description
$G = (V, E, P)$	A social network with user set $V$ and edge set $E$ . $P$ is the influence probability. $P_e$ represents influence probability on edge $e$ where $0 \leq P_e \leq 1$
$G = (V, C, E, P, f)$	A candidate seed set $C \subseteq V$ . Each user has a weight $f$
$\mathcal{U}$	The set of groups, $b(U)$ is the benefit when $U$ is activated for $U \in \mathcal{U}$
$c(v) \geq 0$	$c(v)$ is the cost to activate $v$
$n =  V $	The number of users
$m =  E $	The number of edges
$l =  \mathcal{U} $	The number of groups
$\beta$	The threshold of a group being activated, $0 < \beta \leq 1$
$k$	The number of seeds
$\beta(S)$	The expected benefit of all activated groups with seed set $S$ .
$\gamma(S)$	The expected diffusion cost of all activated users with seed set $S$
$\rho(S) = \beta(S) - \gamma(S)$	The expected profit with seed set $S$

$v$  with the activation probability of  $P_{(u,v)}$ . IC model assumes that  $u$  has only one chance to activate its inactivated neighbor  $v$ .

**Group profit maximization**

Given an instance of GPM with directed graph  $G = (V, E, P)$ , a **group**  $U$  is a subset of  $V$ . Let  $\mathcal{U}$  be a collection of groups. The number of total groups is  $l$ .  $0 < \beta \leq 1$  is the activation threshold. When  $\beta$  percent of users in a group are activated, this group is said to be **activated**. For each activated group  $U$ , there is a benefit  $b(U) \geq 0$ . Simultaneously, there is a diffusion cost  $c(v) \geq 0$  for each activated user.

Now, a realization of random graph will be introduced which can help us to understand the IC model.  $G = (V, E, P)$  is a random directed graph, a realization  $g$  is a subgraph of  $G$ , where  $V(g) = V(G)$  and  $E(g) \subseteq E(G)$ . The influence probability of each edge in  $E(g)$  is 1. The generation process is: (1) for each edge  $e \in E(G)$ , uniformly generate a random number  $r$  between 0 and 1; (2) this edge  $e$  is kept in  $g$  if and only if  $r \leq P_e$ .  $\mathcal{G}$  represents the set of any realizations of  $G$ . Obviously, there are  $2^{|E(G)|}$  sample graphs in  $\mathcal{G}$ .  $g$  is generated with probability  $P[g]$ . Then, we have:

$$P[g] = \prod_{e \in E(g)} P_e \prod_{e \in E(G) \setminus E(g)} (1 - P_e).$$

Let  $\mathcal{U}_g(S)$  represent the set of groups activated by the initial seed set  $S$ .  $V_g(S)$  is the set of nodes activated by the initial seed set  $S$ . Now, the benefit of activated groups is:

$$\beta(S) = \sum_{g \in \mathcal{G}} P[g] \sum_{U \in \mathcal{U}_g(S)} b(U),$$

and the cost of activated nodes is:

$$\gamma(S) = \sum_{g \in \mathcal{G}} P[g] \sum_{v \in V_g(S)} c(v).$$

We define the profit as  $\rho(S) = \beta(S) - \gamma(S)$ . Then, Group profit maximization (GPM) considers information propagation in social network. The objective aims to select  $k$  seed users to maximize the profit  $\rho(S)$ :

$$\max \rho(S) \tag{1}$$

$$\text{s.t. } |S| \leq k. \tag{2}$$

Figure 1 shows an example to explain the information diffusion process of GPM, where there exists 8 nodes and the influence probability on each edge is 1. Let  $\beta = 0.5$ . At the beginning,  $v_3$  is the seed. At the first time step,  $v_2, v_5$  are activated by  $v_3$ , as shown in Fig. 1(1). At the second time step,  $v_4$  is activated by  $v_2$  and  $v_6$  is activated by  $v_5$ , as shown in Fig. 1(1). Finally, activated node set is  $\{v_2, v_3, v_4, v_5, v_6\}$ . Since activation threshold  $\beta = 0.5$ , group  $U_1$  is *activated* and  $U_2$  is inactivated.

### Properties of GPM

In this section, GPM will be proved to be NP-hard. The properties of the objective function  $\rho(\cdot)$  will be discussed.

#### Hardness results

It is known that any generalization of an NP-hard problem is also NP-hard. Kempe et al. have proved that the influence maximization (IM) problem is NP-hard [4], which is a special case of GPM. Each node is considered as a group and benefit of each group is 1. There does not exist cost on each node. Let  $\beta = 1$ . Obviously, the GPM is NP-hard.,

**Theorem 3.1** *The group profit maximization problem is NP-hard.*

For any instance of GPM, it is difficult to compute the objective  $\rho(S)$  even for fixed seed set  $S$ . To estimate  $\rho(S)$ , Monte Carlo method is always used to estimate  $\rho(S)$ . First, a large number of sample graphs of  $G$  are generated, and then computer  $\rho(S)$  on each sample graph. Finally, the average of  $\rho(S)$  is the estimation value. Kempe et al. have proved that computing the objective of IM was #P-hard [4], and then, the following result is true.

**Theorem 3.2** *Given a seed node set  $S$ , computing  $\rho(S)$  is #P-hard under the IC model.*

#### Modularity of objective function

A set function  $f : 2^V \leftarrow \mathbb{R}$  is called *submodular* [30] if it holds that  $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$  for any subsets  $A \subset B \subseteq V$  and  $v \in V \setminus B$ . On the other hand, if, for any subsets  $A \subset B \subseteq V$  and  $v \in V \setminus B$ , it satisfies that  $f(A \cup \{v\}) - f(A) \leq f(B \cup \{v\}) - f(B)$ ,  $f$  is *supermodular*. A set function  $f : 2^V \leftarrow \mathbb{R}$  is called *monotone nondecreasing* if it satisfies  $f(A) \leq f(B)$  for any  $A \subseteq B \subseteq V$ .  $f$  is said to be a *polymatroid function* if it is monotone nondecreasing, submodular and  $f(\emptyset) = 0$ .

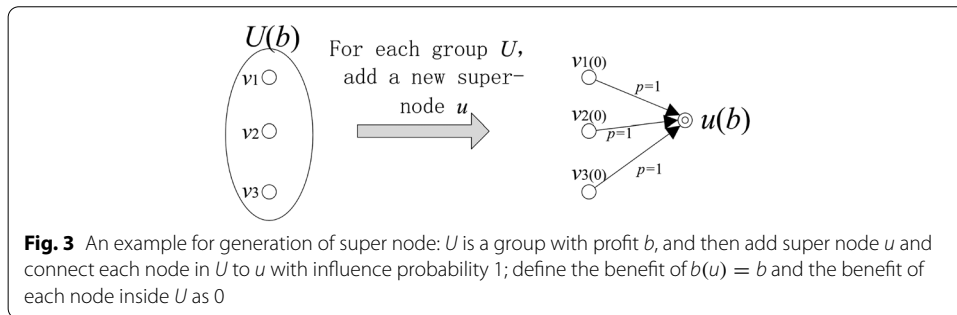
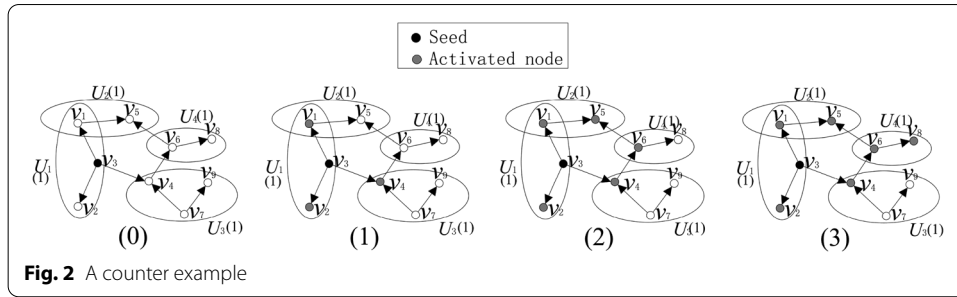
Greedy algorithm guarantees  $(1 - 1/e)$ -approximation for polymatroid maximization problem with cardinality constraints [31]. Also, we have the following result for  $\gamma$ .

**Theorem 3.3**  *$\gamma(\cdot)$  is monotone nondecreasing, submodular, and  $\gamma(\emptyset) = 0$ .*

Meanwhile,  $\beta(\cdot)$  is neither submodular nor supermodular, although  $\beta(\emptyset) = 0$  and  $\beta(\cdot)$  is monotone nondecreasing.

**Theorem 3.4**  *$\beta(\cdot)$  is neither submodular nor supermodular under IC model even when  $b(U) = 1$  for any  $U \in \mathcal{U}$ .*

*Proof* We prove by a counter example. When  $b(U) = 1$  for any  $U \in \mathcal{U}$ ,  $\beta(S)$  is the expected number of eventually activated groups for



initial seed set  $S$ . Consider an instance of GPM problem, as shown in Fig. 2 where there are 9 nodes and the influence probability of each edge is 1. There exist 4 groups  $U = \{U_1 = \{v_1, v_2, v_3\}, U_2 = \{v_1, v_5\}, U_3 = \{v_4, v_7, v_9\}, U_4 = \{v_6, v_8\}\}$  and  $b(U_1) = 1, b(U_2) = 1, b(U_3) = 1, b(U_4) = 1$ . Assume the activation threshold  $\beta = 0.5$ .  $\square$

First, we will prove that  $\beta(\cdot)$  is not submodular. Let  $A = \emptyset, B = \{v_3\}$ , and  $v_9 \in V \setminus B$ . We have  $\beta(A) = 0, \beta(B) = 3$ . Putting  $v_9$  into  $A$  and  $B$ , we have  $\beta(A \cup \{v_9\}) = 0$ , since  $v_9$  can not activate any group.  $\beta(B \cup \{v_9\}) = 4$ , since all groups are eventually activated. Thus,  $\beta(A \cup \{v_9\}) - \beta(A) = 0$  and  $\beta(B \cup \{v_9\}) - \beta(B) = 4 - 3 = 1$ . Therefore,  $\beta(A \cup \{v_9\}) - \beta(A) < \beta(B \cup \{v_9\}) - \beta(B)$  means  $\beta(\cdot)$  is not submodular.

On the other hand,  $\beta(\cdot)$  is not supermodular. Let  $A = \emptyset, B = \{v_3\}$ , and  $v_7 \in V \setminus B$ . We have  $\beta(A) = 0, \beta(B) = 3$ . Putting  $v_7$  into  $A$  and  $B$ , we have  $\beta(A \cup \{v_7\}) = 3$  since  $v_7$  can activate  $\{v_4, v_5, v_6, v_7, v_8, v_9\}$ .  $\beta(B \cup \{v_7\}) = 4$ , since all nodes are eventually activated. Thus,  $\beta(A \cup \{v_7\}) - \beta(A) = 3$  and  $\beta(B \cup \{v_7\}) - \beta(B) = 4 - 3 = 1$ . Therefore,  $\beta(A \cup \{v_7\}) - \beta(A) > \beta(B \cup \{v_7\}) - \beta(B)$  means that  $\beta(\cdot)$  is not supermodular.

We also have the following corollary.

**Corollary 3.1**  $\rho(\cdot)$  is neither submodular nor supermodular under IC model.

**Lower bound and upper bound**

To optimize a non-submodular function is very hard. Lu et al. presented a sandwich approximation framework (SAF) [27]. SAF attempts to find a lower bound and upper bound for the original objective function. Now, we will design lower bound and upper bound for  $\rho(\cdot)$ . Simultaneously, the properties of these two bounds will be analyzed.

**The upper bound**

A new set function  $\bar{\beta}(\cdot)$  is defined which satisfies  $\beta(S) \leq \bar{\beta}(S)$ . In this paper, we formulate the upper bound in two steps. First, a relaxed GPM (r-GPM) problem is generated by modifying group activation rules. For r-GPM, a group is said to be activation if at least 1 activated node is activated in this group. Second, we add a super node for each group. The benefit  $b(u)$  of this super node is defined as the benefit  $b(U)$  of the corresponding group. Then, connect every node in this group to this super node and set influence probability 1. An example is shown in Fig. 3.

$W$  represents the super node set and  $E'$  represents the edge set for nodes in  $V$  to super nodes in  $W$ . Next, a general weighted influence maximization (WIM) is defined as follows.  $V \cup W$  is node set and  $E \cup E'$  is edge set.  $C \subseteq V$  is the set of candidates of seed users. Node weight function  $f$  satisfies:

$$f(v) = \begin{cases} b(v), & v \in W \\ 0, & v \in V \end{cases}$$

$\bar{\beta}(S) = \sum_{v \text{ is activated}} f(v)$  is the expected weight of activated nodes for seed set  $S$ . Let  $G = (V, C, E, P, f)$  be an instance of general Weighted IM problem, where  $C$  is the candidate seed set. We can prove  $\bar{\beta}(\cdot)$  is monotone, submodular, and  $\beta(S) \leq \bar{\beta}(S)$ .

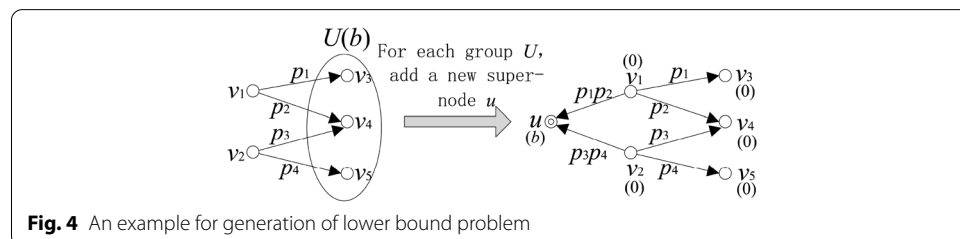
**Theorem 4.1** *Let  $G = (V, E, P)$  be an instance of GPM, and then, we have  $\bar{\beta}(\cdot)$  is an upper bound of  $\beta(\cdot)$ .*

Define  $\bar{\rho}(\cdot) = \bar{\beta}(\cdot) - \gamma(\cdot)$ , and then, we have:

**Theorem 4.2** *Let  $G = (V, E, P)$  be an instance of GPM, and then, we have  $\bar{\rho}(\cdot)$  is an upper bound of  $\rho(\cdot)$ . Simultaneously,  $\bar{\rho}(\cdot)$  can be represented as the difference of two sub-modular functions.*

**The lower bound**

In this subsection, a lower bound will be formulated. The idea is to keep some groups and delete some groups. If at least  $\beta$  percent of nodes in a group can be activated simultaneously, this group will be kept. It means that there must exist 1 node that connects to  $\beta$  percent of nodes in this group. An example is shown in Fig. 4. The activation threshold is  $\beta = 0.5$ . Since  $v_1$  and  $v_2$  connect to 2 nodes in group  $U$ , group  $U$  will be kept. A super node  $u$  related to group  $U$  will be generated, and new directed edges



**Fig. 4** An example for generation of lower bound problem



$(v_1, u), (v_2, u)$  will be added with influence probability  $p_{(v_1, u)} = p_1 p_2, p_{(v_2, u)} = p_3 p_4$ . The benefit of  $u$  is set  $b$  and the other nodes are 0.

The following process is the detail. Let  $G = (V, E, P)$  be an instance of GPM. For a group  $U_i$  with benefit  $b_i$ , assume  $H_i = \{v \in V | v \text{ links to at least } \beta \text{ percent of nodes in } U_i\}$ . If  $H_i \neq \emptyset$ , a super node  $u_i$  is generated and directed edges  $\{(v, u_i) | v \in H_i\}$  are added. For each  $v \in H_i$ , let  $U'_i$  be the set of nodes in  $U_i$  which  $v$  links to. Then,  $p_{(v, u_i)} = \prod_{v' \in U'_i} p_{(v, v')}$ ,  $b(u_i) = b_i$ , and benefits of all other nodes are 0. Next, a general weighted influence maximization (WIM) can be generated. The node set is  $V \cup W$ , and the edge set is  $E \cup E'$ .  $E'$  is the set of all new added edges. The candidate seed set  $C \subseteq V$ . The weight function of node  $f$  satisfies:

$$f(v) = \begin{cases} b(v), & v \in W \\ 0, & v \in V \end{cases}$$

$\underline{\beta}(S) = \sum_{v \text{ is activated}} f(v)$  is the expected weight of activated nodes for seed set  $S$ . Let  $G = (V, C, E, P, f)$  be the instance of general WIM problem.  $\underline{\beta}(\cdot)$  is monotone, submodular, and  $\beta(S) \geq \underline{\beta}(S)$ .

**Theorem 4.3** *Given an instance GPM  $G = (V, E, P)$ ,  $\underline{\beta}(\cdot)$  is an lower bound of  $\beta(\cdot)$ .*

Certainly, let  $\underline{\rho}(\cdot) = \underline{\beta}(\cdot) - \gamma(\cdot)$ , then we have the following result:

**Theorem 4.4** *Let  $G = (V, E, P)$  be an instance of GPM, and then, we have  $\underline{\rho}(\cdot)$  is an lower bound of  $\rho(\cdot)$ . Simultaneously,  $\underline{\rho}(\cdot)$  can be represented as the difference of two submodular functions.*

### Algorithm

Since computing the objective function of GPM is #P-hard, the reverse influence set (RIS) sampling method will be extended to estimate  $\bar{\rho}(\cdot)$  and  $\underline{\rho}(\cdot)$ . Next, an submodular-modular algorithm will be proposed for solving the lower bound and upper bound problems. Then, we will propose an randomized algorithm which is base on weighted group coverage maximization strategy. Finally, a sandwich approximation framework will be presented with theoretical analysis.

We will apply  $(\epsilon, \delta)$ -approximation method [32] to analyze our algorithm. The absolute error is  $\epsilon$  and the confidence is  $(1 - \delta)$ . Let  $\Upsilon = 4(e - 2) \ln(2/\delta)/\epsilon^2$  and  $\Upsilon_1 = 1 + (1 + \epsilon)\Upsilon$ , and then, the Stopping Rule Algorithm [32] has been shown  $(\epsilon, \delta)$  approximation.

### Extended reverse influence set (RIS) sampling

In this section, we will present an extended version of the RIS sampling method. Given a weighted directed graph  $G = (V, C, E, P, f)$ , which represents a general weighted influence maximization problem and  $C$  is the candidate. The influence probability is  $P$  and  $f$  is the node weight function. Assume  $S$  is the seed set.  $\rho'(S) = \sum_{v \text{ is activated}} f(v)$  is the expected weighted number of activated nodes. Looking for  $k$  seed users in  $C$  to maximize  $\rho'(S)$ . Obviously,  $\phi(S)$  is submodular and monotone. Extended RIS generates a set

$\mathcal{R}$  of random *weighted reverse reachable (WRR) sets*. Let  $R_j$  be a WRR set which can be formulated as follows,

**Definition 5.1** (*Weighted reverse reachable (WRR) set*) [13]. Given  $G = (V, C, E, P, f)$ , a random WRR set  $R_j$  is generated from  $G$  by (1) selecting a random node  $v \in V$ ; (2) generating a sample graph  $g$  from  $G$ ; (3) returning  $R_j$  as the set of nodes that can reach  $v$  in  $g$ ; and (4)  $w(R_j) = f(v)$ .

$S$  is the seed set. Let  $\text{Cov}_{\mathcal{R}}(S) = \sum_{R_j \in \mathcal{R}} \min\{|S \cap R_j|, 1\}$  be the coverage number of set  $S$  and  $W\text{Cov}_{\mathcal{R}}(S) = \sum_{R_j \in \mathcal{R}} w(R_j) \cdot \min\{|S \cap R_j|, 1\}$  be the coverage weight. This weighted coverage of set  $S$  might be used to estimate  $\rho'(S)$ .

**Lemma 5.1** [13]. Given  $G = (V, C, E, P, f)$ , a random WRR set  $R_j$  generated from  $G$ . For each seed set  $S \subseteq C$ , where  $C \subseteq V$  is candidate seed set:

$$\phi(S) = \sum_{v \in V} f(v) \Pr[S \text{ covers } R_j].$$

The estimation procedure for computing  $\phi(S)$  will be proposed as Algorithm 1, which also preserves the following theoretical result.

---

**Algorithm 1** Estimation Procedure (EP)

---

**Input:**  $G = (V, C, E, P, f)$  is an instance of WIM,  $0 \leq \epsilon, \delta \leq 1$ , seed set  $S$ .

**Output:**  $\hat{\phi}(S)$  such that  $\hat{\phi}(S) \leq (1 + \epsilon)\phi(S)$  with at least  $(1 - \delta)$ -probability.

- 1:  $\Upsilon = 1 + 4(1 + \epsilon)(e - 2) \ln(2/\delta)/\epsilon^2$
  - 2:  $\Upsilon_1 = 1 + (1 + \epsilon)\Upsilon$
  - 3:  $\mathcal{R} \leftarrow$  generate  $\Upsilon$  random WRR sets
  - 4:  $L = \text{Cov}_{\mathcal{R}}(S)$
  - 5: **while**  $L < \Upsilon_1$  **do**
  - 6:  $R' \leftarrow$  generate a new WRR set
  - 7: Add  $R'$  to  $\mathcal{R}$
  - 8:  $L = \text{Cov}_{\mathcal{R}}(S)$
  - 9: **end while**
  - 10:  $\hat{\phi}(S) \leftarrow \sum_{v \in V} f(v) \cdot W\text{Cov}_{\mathcal{R}}(S) / \sum_{j=1}^{|\mathcal{R}|} w(R_j)$
  - 11: **return**  $\hat{\phi}(S)$ .
- 

**Theorem 5.1** Algorithm 1 outputs an estimation  $\hat{\phi}(S)$  of  $\phi(S)$  which satisfies:

$$\Pr[(1 - \epsilon)\phi(S) \leq \hat{\phi}(S) \leq (1 + \epsilon)\phi(S)] \geq 1 - \delta.$$

**Submodular–modular algorithm**

Theorems 4.2 and 4.1 perform  $\bar{\rho}(\cdot)$  and  $\underline{\rho}(\cdot)$  are the difference of two submodular functions. Furthermore,  $\bar{\beta}(\cdot)$  and  $\underline{\beta}(\cdot)$  are objective functions of WIM problems. Then, we will propose a submodular–modular algorithm for solving such a function  $\phi(S) - \gamma(S)$  which satisfies  $\phi(\cdot)$  and  $\gamma(\cdot)$  are submodular functions.

At first, a modular upper bound and lower bound will be presented for  $\gamma(\cdot)$  according to [33]. The following formulas are two tight modular upper bounds which are tight at the given set  $X$ :

$$m_X^1(S) \triangleq \gamma(S) - \sum_{j \in X \setminus S} \gamma(j|X \setminus \{j\}) + \sum_{j \in S \setminus X} \gamma(j|\emptyset) \tag{3}$$

$$m_X^2(S) \triangleq \gamma(S) - \sum_{j \in X \setminus S} \gamma(j|V \setminus \{j\}) + \sum_{j \in V \setminus X} \gamma(j|X). \tag{4}$$

For brevity, we use  $m_X$  to refer either one. A modular lower bound  $h_X$  which is tight at a given set  $X$  will be formulated as follows. Assume  $\pi$  is any permutation of  $V$  and place all the nodes in  $X$  at the front. Let  $S_i^\pi = \{\pi(1), \pi(2), \dots, \pi(i)\}$  be a chain constructed by this permutation, where  $S_0^\pi = \emptyset$  and  $S_{|X|}^\pi = X$ . Define:

$$h_X^\pi(\pi(i)) = \gamma(S_i^\pi) - \gamma(S_{i-1}^\pi). \tag{5}$$

$h_X^\pi(S) = \sum_{v \in S} h_X^\pi(v)$  will be a lower bound for  $\gamma(S)$ , and it is tight at  $X$ . Then,  $h_X^\pi(S) \leq \gamma(S)$  holds for any  $S \subseteq V$  and specially  $h_X^\pi(X) = \gamma(X)$ . The following results can be proved.

**Theorem 5.2**  $\phi(S) - m_X(S) \leq \phi(S) - \gamma(S) \leq \phi(S) - h_X^\pi(S)$  and these two bounds are difference of submodular and modular functions.

Using  $\phi(S) - m_X(S) \leq \phi(S) - \gamma(S)$ , we can propose the submodular–modular algorithm. In each iteration, run maximization procedures for these two modular upper bounds and select the better one. Algorithm 2 can be proved convergency to a local maximal solution.

---

**Algorithm 2** Submodular-Modular Algorithm (SMA)

---

```

1:  $X^0 = \emptyset; t \leftarrow 0$ 
2: while not converged (i.e.,  $(X^{t+1} \neq X^t)$ ) do
3:   Randomly choose a permutation  $\pi^t$  whose chain contains the set  $X^t$ 
4:    $X^{t+1} := \arg \max_X \phi(X) - m_{X^t}(X)$ 
5:    $t \leftarrow t + 1$ 
6: end while
7: return  $X^t$ .

```

---

**Theorem 5.3** *Algorithm 2 monotonically increasing. Furthermore, assuming a local maxima  $\phi(X) - m_{X^t}(X)$  is returned from the submodular maximization procedure, then Algorithm 2 outputs a local optima solution.*

*Proof* For either modular upper bound, we have:

$$\begin{aligned} \phi(X^{t+1}) - \gamma(X^{t+1}) &\geq \phi(X^{t+1}) - m_{X^t}(X^{t+1}) \\ &\geq \phi(X^t) - m_{X^t}(X^t) = \phi(X^t) - \gamma(X^t). \end{aligned}$$

To show that this algorithm converges to a local maxima, we assume the submodular maximization procedure converges to a local maxima. Then, if the objective value does not increase in an iteration under both upper bounds, it implies that  $\phi(X^t) - m_{X^t}(X^t)$  is already a local optimum in that (for both upper bounds), we have  $\phi(X^t \cup \{j\}) - m_{X^t}(X^t \cup \{j\}) \leq \phi(X^t) - m_{X^t}(X^t), \forall j \notin X^t$  and  $\phi(X^t \setminus \{j\}) - m_{X^t}(X^t \setminus \{j\}) \leq \phi(X^t) - m_{X^t}(X^t), \forall j \in X^t$ . Note that

$m_{X^t}^1(X^t \setminus \{j\}) = \gamma(X^t) - \gamma(j|X^t \setminus \{j\}) = \gamma(X^t \setminus \{j\})$  and  $m_{X^t \cup \{j\}}^2 = \gamma(X^t) + \gamma(j|X^t) = \gamma(X^t \cup \{j\})$ , and hence, if both modular upper bounds are at a local optima, it implies

$$\phi(X^t) - \gamma(X^t) = \phi(X^t) - m_{X^t}^1(X^t) \geq \phi(X^t \setminus \{j\}) - m_{X^t}^1(X^t \setminus \{j\}) = \phi(X^t \setminus \{j\}) - \gamma(X^t \setminus \{j\})$$

Similarly,  $\phi(X^t) - \gamma(X^t) = \phi(X^t) - m_{X^t}^2(X^t) \geq \phi(X^t \cup \{j\}) - m_{X^t}^1(X^t \cup \{j\}) = \phi(X^t \cup \{j\}) - \gamma(X^t \cup \{j\})$ . Hence,  $X^t$  is a local optima.  $\square$

### Group coverage maximization algorithm

In this section, we will propose weighted group coverage maximization algorithm for solving GPM. Let  $\mathcal{U}$  be the set of groups.  $\mathcal{U}(S)$  represents the set of groups which includes at least one node in  $S$ , i.e.,  $\mathcal{U}(S) = \{U \in \mathcal{U} | U \cap S \neq \emptyset\}$ . Then,  $b(\mathcal{U}(S)) = \sum_{U \in \mathcal{U}(S)} b(U)$ . Algorithm 3 is shown below by selecting the maximum marginal gain at each step and at most  $O(knl)$  time complexity. Greedy algorithm may give better solution, but the running time is  $O(kn\Gamma(nm + nl))$ . We will compare several different strategies by experiments.

---

#### Algorithm 3 Weighted Group Coverage Maximization Algorithm (WGCMA)

---

**Input:** An instance of GPM  $G = (V, E, P)$ , the number of seeds  $k$ .

**Output:** a set of seed nodes,  $S_k$ .

- 1:  $S_k = \emptyset$
  - 2: **for**  $i = 1$  to  $k$  **do**
  - 3:    $v^* \leftarrow \arg \max_{v \in V} (b(\mathcal{U}(S \cup \{v\})) - b(\mathcal{U}(S)))$
  - 4:   Add  $v^*$  to  $S_k$
  - 5: **end for**
  - 6: **return**  $S_k$ .
- 

### Sandwich approximation framework

For GPM, we have formulated the lower bound and upper bound for  $\rho(\cdot)$ . Algorithm 4 gives the sandwich approximation framework.

---

#### Algorithm 4 Sandwich Approximation Framework

---

**Input:** Given an instance of CPM  $G = (V, E, P)$ ,  $0 \leq \epsilon, \delta \leq 1$  and  $k$ .

**Output:** a set of seed nodes,  $S$ .

- 1: Let  $S_L$  be the output seed set of solving the lowerbound  $\underline{\rho}$  by Submodular-Modular algorithm (Algorithm 2)
  - 2: Let  $S_Z$  be the output seed set of solving the upperbound  $\bar{\rho}$  by Submodular-Modular algorithm (Algorithm 2)
  - 3: Let  $S_A$  be the output seed set of solving  $G = (V, E, P)$  by Algorithm 3.
  - 4:  $S = \arg \max_{S_0 \in \{S_L, S_Z, S_A\}} \text{EP}(G, \epsilon, \delta, S_0)$  (by Algorithm 1)
  - 5: **return**  $S$
- 

For sandwich approximation framework, we can prove the following theoretical result.

*Theorem 5.4* Let  $S$  be the seed set returned by Algorithm 4, and then, we have:

$$\rho(S) \geq \max \left\{ \frac{\rho(S_Z)}{\bar{\rho}(S_Z)}, \frac{\underline{\rho}(S_L^*)}{\rho(S^*)} \right\} \frac{1 - \epsilon}{1 + \epsilon} \alpha \rho(S^*), \tag{6}$$

where  $S_L^*$  is the optimal solution to maximize the lower bound problem,  $S^*$  is the optimal solution of GPM, and  $\alpha$  is the approximation ratio of Algorithm 2.

*Proof* Let  $S_Z^*$  be the optimal solution to maximize the upper bound problem. Then, we have:

$$\begin{aligned} \rho(S_Z) &= \frac{\rho(S_Z)}{\bar{\rho}(S_Z)} \bar{\rho}(S_Z) \geq \frac{\rho(S_Z)}{\bar{\rho}(S_Z)} \alpha \bar{\rho}(S_Z^*) \\ &\geq \frac{\rho(S_Z)}{\bar{\rho}(S_Z)} \alpha \bar{\rho}(S^*) \geq \frac{\rho(S_Z)}{\bar{\rho}(S_Z)} \alpha \rho(S^*) \end{aligned}$$

and

$$\rho(S_L) \geq \underline{\rho}(S_L) \geq \alpha \underline{\rho}(S_L^*) \geq \frac{\rho(S_L^*)}{\rho(S^*)} \alpha \rho(S^*).$$

Let  $S_{\max} = \arg \max_{S_0 \in \{S_L, S_Z, S_A\}} \rho(S_0)$ , and then:

$$\rho(S_{\max}) \geq \max \left\{ \frac{\rho(S_Z)}{\bar{\rho}(S_Z)}, \frac{\rho(S_L^*)}{\rho(S^*)} \right\} \alpha \rho(S^*).$$

Since  $\forall S_0 \in \{S_L, S_Z, S_A\}$ ,  $(1 - \epsilon)\rho(S_0) \leq \hat{\rho}(S_0) \leq (1 + \epsilon)\rho(S_0)$ , we have:

$$(1 + \epsilon)\rho(S) \geq \hat{\rho}(S) \geq \hat{\rho}(S_{\max}) \geq (1 - \epsilon)\rho(S_{\max}).$$

It follows that:

$$\rho(S) \geq \frac{1 - \epsilon}{1 + \epsilon} \rho(S_{\max}) \geq \max \left\{ \frac{\rho(S_Z)}{\bar{\rho}(S_Z)}, \frac{\rho(S_L^*)}{\rho(S^*)} \right\} \frac{1 - \epsilon}{1 + \epsilon} \alpha \rho(S^*).$$

□

Sadly, the performance of sandwich framework depends on  $\alpha$ . Although we have proved the convergence of Algorithm 2 to a local optimal, the ratio  $\alpha$  is still an open problem. According to Theorem 5.4, the difference between  $\rho(S^*)$  and  $\underline{\rho}(S_L^*)$  has great influence on the performance of Algorithm 4. Iyer and Bilmes [33] studied the minimization problem of the difference between submodular function. While the difference between  $\rho(S^*)$  and  $\underline{\rho}(S_L^*)$  may be bounded, we have the following result.

**Theorem 5.5** *Let  $S_L^*$  be the optimal solution to maximize the lower bound problem and  $S^*$  is the optimal solution of GPM, and then, we have:*

$$\rho(S^*) - \underline{\rho}(S_L^*) \leq \max_{S, |S|=k} (\bar{\rho}(S) - \underline{\rho}(S)). \tag{7}$$

### Comparison with different heuristic strategies

We will compare Sandwich Approximation Framework (SAF) with Greedy Strategy (GS) proposed by Kempe [4] and Maximum Outdegree (MO) method by choosing the first  $k$  largest outdegree nodes. Algorithm 3 is called Weighted Group Coverage Maximization Algorithm, which represents as MC for simplification.

## Experiments

To evaluate our algorithms, we will test on two datasets coming from [34, 35]. Facebook-like Forum Network is the first dataset which was collected from the online community of Facebook. Users' activities in this forum are recorded in this dataset, in which there are one-mode and two-mode data. There are 899 users and the relationship between users is stored in the one-mode data. Beside one-mode data, there are 522 topics and the two-mode data contain the interesting network of 899 users and 522 topics. Users related to a topic are represented as a group. Newman's scientific collaboration network is the second dataset, which represents the co-authorship network. These data are based on preprints published to Condensed Matter section of arXiv E-Print Archive from 1995 to 1999. The one-mode data indicate the relationship among the co-authors. The relation between an author and the paper is shown in the two-mode data. The authors related to the same paper are considered as a group. Table 2 shows the details of these two datasets.

## Procedure

The instances are formulated from the above datasets. The basic graph is constructed by the one-mode dataset. The set of groups come from two-mode dataset. The benefit of a group is derived from the size of group. In this paper, by multiplying the size of the group by a factor of 10 is defined as the benefit. The cost of each activated node is generated as a random number from 0 to 1. We use Python 3.6 to write all programs and run on a Linux server with 16 CPUs and 256 GB RAM.

## Experimental results

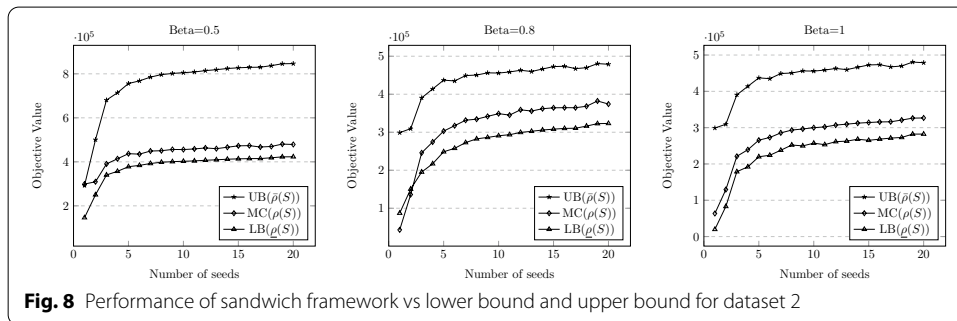
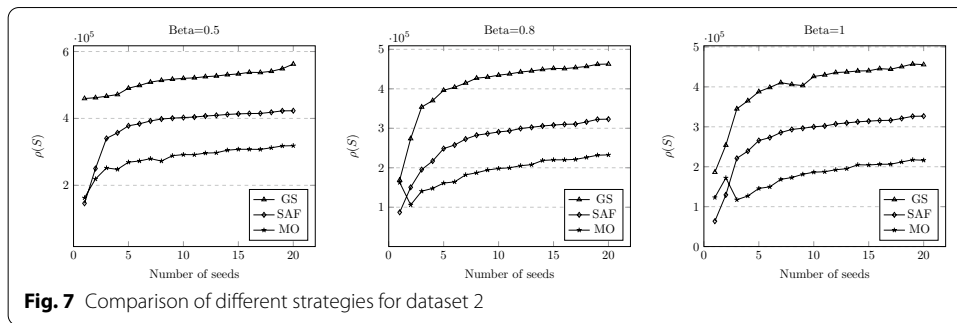
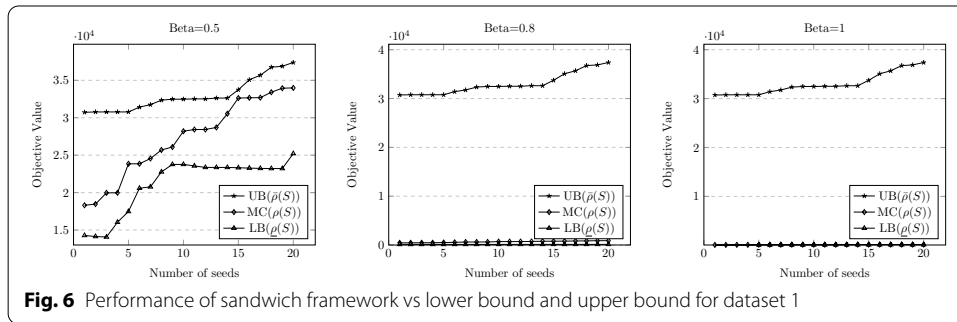
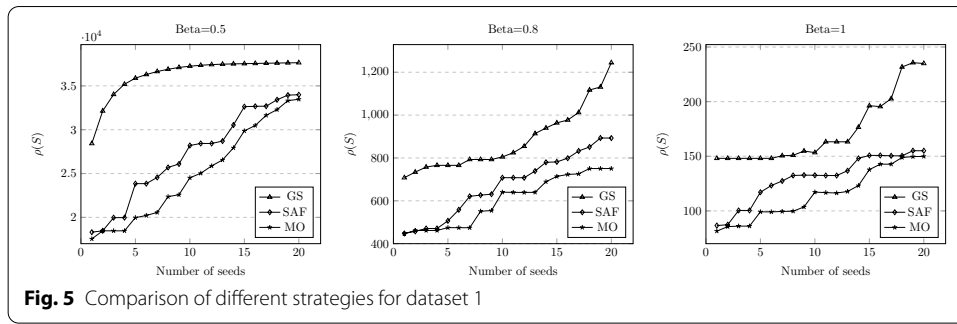
From the comparison of three different seed selection strategies, Greedy Strategy (GS) returns a comparatively higher benefit than SAF and MO methods. The MO strategy initially gives higher profit than the MC. The SAF outperforms MO as the number of seed nodes increasing. Figures 5 and 6 show the experimental results. Figures 7 and 8 show performance of SAF for dataset 1 and 2, respectively. The main results are as follows:

### *Profit increases with increase of seed number for fixed $\beta$*

The experiments are carried out with three values for beta values 0.5, 0.8, and 1. From the graphs, it can be observed that, for a given beta value, the profit increases with the increase in the number of seeds in a set. Initially, a seed set of lesser number of seeds is able to activate fewer groups, thus resulting in a lesser profit being generated. However, as the size of seed set increases, it is more likely for larger number of groups to be activated, thus increasing the profit with an increase in the number of seeds.

**Table 2** Data statistics

	Nodes	Edges	Groups	Average group size
Dataset 1	899	142,760	522	14.6
Dataset 2	16,726	95,188	22,015	3.7



**Profit decreases with increase in  $\beta$**

The experiments are carried out with three values for beta values 0.5, 0.8, and 1. As the beta increases, it is observed that the number of groups activated decreases for a given seed set, which, in turn, results in the profit decreasing. As beta is the determining factor

for activation of the group, as beta becomes larger and larger, lesser groups get activated. As a result, the profit generated by a lower beta value is much higher as compared to the profit generated by a higher beta value. The seed set activates more nodes, but the activation of number of group decreases.

#### **Gap of upper bound and lower bound**

It is observed from the graphs of dataset 1 that with an increase in the beta value, the gap between upper bound and lower bound increases. The reason behind this result is because of the formulation of upper bound in our problem and the size of each group in dataset 1. In our experiments, the upper bound is fixed even as the beta varies. As beta increases, the profit decreases, and as the upper bound is fixed, the gap between upper bound and lower bound becomes large. However for dataset 2, the gap remains almost the same even as beta increases as the group size are smaller having an average group size of 3.7 as compared to group size in dataset 1 having average 14.6 as the group size.

#### **Conclusion**

This paper studied profit maximization problem of information propagation in online social networks. Group activation was considered in this novel IM model. Each activated group would give a benefit, while information diffusion cost was needed for every activated users. Then, our group profit maximization (GPM) problem attempted to look for  $k$  seed users to propagate information, such that the expected profit was maximum. The profit combined benefit of activated groups and the cost on each activated users. GPM was proved to be NP-hard and the objective set function was shown neither submodular nor supermodular. We proposed a weighted version of group coverage maximization strategy for solving GPM. Simultaneously, a sandwich approximation framework was presented with theoretical analysis. Finally, the experiment results shown that our proposed algorithms were effectiveness and the efficiency. For future research, novel efficient methods for solving non-submodular optimization are eager for paying attention.

#### **Acknowledgements**

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant no. 72074203, the US National Science Foundation (NSF) under Award no. 1747818.

#### **Further information**

Preliminary version of this paper appeared in: Ref. [36].

#### **Authors' contributions**

All authors have contributed to the analysis of the results and to writing the paper. All authors read and approved the final manuscript.

#### **Availability of data and materials**

Not applicable.

#### **Competing interests**

The authors declare that they have no competing interests.

#### **Author details**

<sup>1</sup> School of Engineering Science, University of Chinese Academy of Sciences, 19A Yuquan Rd., Beijing 100049, China.

<sup>2</sup> Department of Computer Science, University of Texas at Dallas, Richardson, USA. <sup>3</sup> Shandong University, Jinan, China.

Received: 13 January 2020 Accepted: 22 December 2020

Published online: 07 January 2021



## References

- Forsyth DR. Group dynamics. 2018. p. 1.
- Meeker M. Internet trends 2018—code conference. *Glokalde*. 2018;1(3).
- Zhu J, Ghosh S, Wu W. Group influence maximization problem in social networks. *IEEE Trans Comput Soc Syst*. 2019. <https://doi.org/10.1109/TCSS.2019.2938575>.
- Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, ACM. 2003. pp. 137–46.
- 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, ACM. 2007. pp. 420–9.
- 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, ACM. 2010. pp. 1029–38.
- Goyal A, Lu W, Lakshmanan LV. Simpath: an efficient algorithm for influence maximization under the linear threshold model. In: Data Mining (ICDM), 2011 IEEE 11th international conference on, IEEE. 2011. pp. 211–20.
- Cohen E, Delling D, Pajor T, Werneck RF. Sketch-based influence maximization and computation: scaling up with guarantees. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, ACM. 2014. pp. 629–38.
- Ohsaka N, Akiba T, Yoshida Y, Kawarabayashi K-i. Fast and accurate influence maximization on large networks with pruned monte-carlo simulations. In: AAAI. 2014. pp. 138–44.
- Du N, Liang Y, Balcan M-F, Gomez-Rodriguez M, Zha H, Song L. Scalable influence maximization for multiple products in continuous-time diffusion networks. *J Mach Learn Res*. 2017;18(2):1–45.
- Yang Y, Lu Z, Li VO, Xu K. Noncooperative information diffusion in online social networks under the independent cascade model. *IEEE Trans Comput Soc Syst*. 2017;4(3):150–62.
- Aslay C, Lakshmanan LV, Lu W, Xiao X. Influence maximization in online social networks. In: Proceedings of the eleventh ACM international conference on web search and data mining, ACM. 2018. pp. 775–6.
- Zhu J, Zhu J, Ghosh S, Wu W, Yuan J. Social influence maximization in hypergraph in social networks. *IEEE Trans Netw Sci Eng*. 2018. <https://doi.org/10.1109/TNSE.2018.2873759>.
- Zhu J, Ghosh S, Zhu J, Wu W. Near-optimal convergent approach for composed influence maximization problem in social networks. *IEEE Access*. 2019;7:142488–97.
- Zhu J, Ghosh S, Wu W, Gao C. Profit maximization under group influence model in social networks. In: International conference on computational data and social networks. Springer. 2019. pp. 108–19.
- Tang J, Tang X, Yuan J. Profit maximization for viral marketing in online social networks: algorithms and analysis. *IEEE Trans Knowl Data Eng*. 2018;30(6):1095–108.
- Lu W, Lakshmanan LV. Profit maximization over social networks. In: Data mining (ICDM), 2012 IEEE 12th international conference on, IEEE. 2012. pp. 479–88.
- Zhu Y, Lu Z, Bi Y, Wu W, Jiang Y, Li D. Influence and profit: two sides of the coin. In: Data mining (ICDM), 2013 IEEE 13th international conference on, IEEE. 2013. pp. 1301–6.
- Tang J, Tang X, Yuan J. Towards profit maximization for online social network providers. *arXiv preprint arXiv:1712.08963*. 2017.
- Borgs C, Brautbar M, Chayes J, Lucier B. Maximizing social influence in nearly optimal time. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms. SIAM. 2014. pp. 946–57.
- Tang Y, Xiao X, Shi Y. Influence maximization: near-optimal time complexity meets practical efficiency. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data, ACM. 2014. pp. 75–86.
- Tang Y, Shi Y, Xiao X. Influence maximization in near-linear time: a martingale approach. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, ACM. 2015. pp. 1539–54.
- Nguyen HT, Thai MT, Dinh TN. Stop-and-stare: optimal sampling algorithms for viral marketing in billion-scale networks. In: Proceedings of the 2016 international conference on management of data, ACM. 2016. pp. 695–710.
- Schoenebeck G, Tao B. Beyond worst-case (in)approximability of nonsubmodular influence maximization. In: International conference on web and internet economics. 2017.
- Narasimhan M, Bilmes JA. A submodular-supermodular procedure with applications to discriminative structure learning. *arXiv preprint arXiv:1207.1404*. 2012.
- Bach F, et al. Learning with submodular functions: a convex optimization perspective. *Found Trends<sup>®</sup> Mach Learn*. 2013;6(2–3):145–373.
- Lu W, Chen W, Lakshmanan LV. From competition to complementarity: comparative influence diffusion and maximization. *Proc VLDB Endow*. 2015;9(2):60–71.
- Wu WL, Zhang Z, Du DZ. Set function optimization. *J Oper Res Soc China*. 2018;3:1–11.
- Zhu J, Ghosh S, Wu W. Robust rumor blocking problem with uncertain rumor sources in social networks. *World Wide Web*. 2020. <https://doi.org/10.1007/s11280-020-00841-8>.
- Fujishige S. Submodular functions and optimization. In: *Of Annals Of discrete mathematics*, vol. 47. 2008.
- Nemhauser GL, Wolsey LA, Fisher ML. An analysis of approximations for maximizing submodular set functions. *Math Program*. 1978;14(1):265–94.
- Dagum P, Karp R, Luby M, Ross S. An optimal algorithm for monte carlo estimation. *SIAM J Comput*. 2000;29(5):1484–96.
- Iyer R, Bilmes J. Algorithms for approximate minimization of the difference between submodular functions, with applications. *arXiv preprint arXiv:1207.0560*. 2012.
- Opsahl T. Triadic closure in two-mode networks: redefining the global and local clustering coefficients. *Soc Netw*. 2013;35(2):159–67.
- Newman ME. The structure of scientific collaboration networks. *Proc Nat Acad Sci*. 2001;98(2):404–9.

36. Tagarelli A., Tong H, editors. Computational data and social networks. CSoNet 2019. Lecture notes in computer science, vol. 11917. Berlin: Springer. pp. 108–19.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---