Group Influence Maximization Problem in Social Networks

Jianming Zhu, Smita Ghosh, and Weili Wu, *Member, IEEE*

***Abstract*— Group plays an important role in social society. Much of the world's decision or work is done by groups and teams. A group's decision should be made based on most of the members in the group that reach agreement on a concerned topic. If we want to spread a topic and maximize the total number of activated groups in a social network, which seed users should we choose. In this article, we will study a new influence maximization (IM) problem which focuses on the number of groups activated by some concerned topic or information. A group is said to be *activated* if β percent of users in this group are activated. Group IM (GIM) aims to select k seed users such that the number of eventually activated groups is maximized. We first analyze the complexity and approximability of GIM, which is NP-hard, and the objective function presented in this article is proven to be neither submodular nor supermodular. We develop an upper bound problem and a lower bound problem whose objective functions are submodular. Then, an algorithm based on group coverage will be proposed, and the Sandwich framework is formulated with theoretical analysis to solve GIM. Our experiments verify the effectiveness of our method, as well as the advantage of our method against the other heuristic methods.**

***Index Terms*— Group influence maximization (GIM), independent cascade (IC), nonsubmodular, sandwich framework, social networks.**

I. Introduction

UNDERSTANDING people's group is critical to understand people's personal behaviors, such as why they think and what they do [1]. Human behavior is mostly easily influenced by their group's behavior and most of the world's decisions or works are done by groups or teams. Sometimes, the group is small with only two or three members, such as family. While sometimes group may be large, such as a community, even a whole state or country.

Nowadays, several large-scale social networks emerge such as Facebook with 2.2B users, Twitter with 0.34B users, and WeChat with 1.0B users, etc. [2], and millions of people are more able to become friends and share information or topic with each other. People with the same character such as interest may create a group on the social network platform to discuss the concerned topics. Group plays an important role in these platforms. Another group decision example is U.S. Presidential Elections. Presidential candidate will win all votes in a state if he got the maximum number of tickets in this state.

Group can be formed traditionally offline. But with the deepening of Internet application, many offline group decision processes are switched to online. Especially in China, online communities formulate conveniently while members inside are stranger and not in the same place. These viral communities exist everywhere and various kinds of information diffuse inside the group. For example, the group purchase would ensure a relatively lower price and higher quality of various products for Chinese buyers, which means advertisement needs influence a certain number of members in this group where the group purchase can be carried out. A brief example is shown in Fig. 1(a). There are three different groups in this network with different sizes. Group U_1 contains four members, while groups U_2 and U_3 have three members, respectively. Note that member v_6 belongs to two different groups.

Either in real-world or online social network, the group plays an important role. In many cases, government or company tries to influence group rather to care more about personal influence in order to obtain a maximum benefit. A group will be activated if a certain number of members are activated. Given a random social network $G = (V, E, P)$, P represents the influence probability on each directed edge (u, v) which means u will try to activate v with probability P when u becomes activated. The traditional *influence maximization* (IM) problem aims to select k initial seed users inside the whole network to maximize the expected number of eventually activated users. While in this article, we consider there exists a set of groups in the social network. One group may contain several users, and this group will be activated if a certain number of users in this group are activated. Fig. 1 shows an example of group influence problem. There are nine nodes and the influence probability of each edge is 1, and there exist three groups. $\mathcal{U} = \{U_1 = \{v_1, v_2, v_3, v_4\}, U_2 = \{v_5, v_6, v_9\}, U_3 = \{v_6, v_7, v_8\}\}$. Assume the activation threshold $\beta = 0.5$ which means a group will be activated if at least half of nodes are activated. Fig. 1(a) chooses v_1 as the seed, then $\{v_1, v_2, v_3, v_4, v_5, v_7\}$ will be activated and only group U_1 is activated under the activation threshold 0.5. On the other hand, Fig. 1(b) chooses v_6 as the seed, then $\{v_3, v_4, v_5, v_6, v_7\}$ will be activated and U_2, U_3 are activated. Seed v_1 could activate six nodes and one group, while v_6 as seed could activate five nodes and two groups. That means the node has maximum influence may not always group IM (GIM). In this article, we will study the GIM problem

Manuscript received January 15, 2019; revised June 8, 2019; accepted August 27, 2019. Date of publication September 19, 2019; date of current version December 9, 2019. This work was supported in part by the U.S. National Science Foundation under Grant 1747818, in part by the National Natural Science Foundation of China under Grant 91324012, and in part by the University of Chinese Academy of Sciences through the Project of Promoting Scientific Research Ability of Excellent Young Teachers. *(Corresponding author: Jianming Zhu.)*

J. Zhu is with the School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jmzhu@ucas.ac.cn).

S. Ghosh and W. Wu are with the Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA.

Digital Object Identifier 10.1109/TCSS.2019.2938575

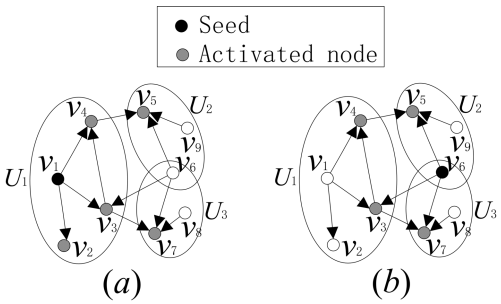

Fig. 1. Example of information diffusion process with initial seed v_3.

which aims to select k seeds such that the number of activated groups is maximized.

Here we present some other examples for group influence problem:

1) The group may be a family, a company, or a society in real-world social network. A family usually makes a decision of purchasing some product among different brands according to their advertisements. Another situation is that a company plans to purchase computers for each employee. If the company uses voting method to determine the brand of computer to purchase, each employee may influence by a different brand of computer and vote for that brand, while the eventually brand purchased is just one brand which gets the maximum votes.
2) For online social network, such as Wechat, people always formulate groups or join different kinds of groups.

For example, all students and teachers in one laboratory formulated a research group. They can easily share information and ideas anytime. In Wechat APP, each person may belong to tens to hundreds of groups. There are millions of types of groups, while each group must have its own characters. In some large groups, members inside the group may not know each other at the beginning of group formulation. Therefore, more and more information is propagating in a group platform, which becomes more valuable to study group characters and the group influence problem.

A. Related Works

IM problem was first presented in 2003 by Kempe *et al.* [3]. They formulated this problem as an optimization mathematic model and proved that the IM is NP-hard under independent cascade (IC) model, and the objective function is a submoduar function. Finally, they proposed a greedy algorithm which guarantees $(1-1/e-\epsilon)$-approximate for any $\epsilon > 0$. Motivated by IM problem, fruitful research works [4]–[10] have been developed. Aslay *et al.* [11] presented a novel model of incentivized social advertising problem. Li *et al.* [12] summarized the recent works on IM problem.

The most popular methods are two-phase influence maximization (TIM)/TIM+ [13] and influence maximization via Martingales (IMM) [14], which guarantee an $(1-1/e-\epsilon)$-approximation for IM under IC model. Another recent and more efficient technique is reverse influence set (RIS) sampling introduced by Tang *et al.* [13], [14] and Borgs *et al.* [15]. They all tried to generate a $(1-1/e-\epsilon)$-approximate solution with the numbers of samples as smaller as possible. Later, Nguyen *et al.* [16] proposed a new sampling algorithms named dynamic-stop-and-stare (D-SSA), which is faster than TIM+ and IMM with the same $(1-1/e-\epsilon)$-approximation. Zhu *et al.* [17] extended the RIS sampling technique to weighted version.

Few results [18] are proposed when the objective function violates the submodularity. Note that the objective function of group influences maximization problem is not submodular which will be proved in Section III, we cannot adapt existing social IM methods to solve the GIM directly. Narasimhan and Bilmes [19] have proposed an approximation method for submodular + supermodular function by substituting one of the two functions by a modular function. Bach *et al.* [20] proved that any nonsubmodular function could represent as a difference of two submodular functions. The latest method is based on the sandwich approximation strategy [21], [22], which analyzes the objective function by comparing with its lower bound and upper bound.

B. Contributions

1) Motivated by the group decision in social society, we propose the GIM problem that aims to maximize the number of eventually activated groups under IC model.
2) We assess the challenges of GIM problem by analyzing computational complexity and properties of the objective function. First, we show that GIM is NP-hard under IC model. Furthermore, the objective function of GIM is proven neither submodular nor supermodular.
3) To achieve a practical approximate solution, we develop a lower bound and upper bound of the objective function. We prove that the problems of maximizing these two bounds are still NP-hard under IC model. However, we also prove that both lower bound and upper bound are submodular.
4) For solving GIM, first we develop a group coverage maximization algorithm (GCMA). Second, we formulate a sandwich approximation framework, which preserves a theoretical analysis result. Our experiments on real-world data sets verify the effectiveness and the efficiency of the proposed algorithm.

The content of this article is as follows: first, we formulate the GIM problem; then, the statement of NP-hardness and properties of objective function will be given; afterward, we develop a lower bound and upper bound and present our algorithm; experiments are presented in Section VI; and finally, this article draws a conclusion. The symbols and their meaning used in this article are shown in Table I.

II. Problem Formulation

Based on the above concept, we will formulate a new IM problem called GIM problem in this section. The IC model is one of the most popular information diffusion models. Our GIM Problem is also based on the IC model. IC model

TABLE I
FREQUENTLY USED SYMBOLS AND THEIR MEANING

Notation	Description
$G = (V, E, P)$	A social network, where V represents the node set and E represent edge set. An influence probability P is associated with each edge in E.
$G = (V, C, E, P, f)$	A social network, where V represents the node set and E represent edge set. Each edge is associated with an influence probability P. $C \subseteq V$ represents the candidate seed set which means initial seed must be selected from C. f is the node weight function.
P_e	influence probability for any edge e, $0 \leq P_e \leq 1$
$\mathcal{U}$	the set of groups, each $U \in \mathcal{U}$ is a subset of V
$n = \|V\|$	number of nodes
$m = \|E\|$	number of edges
$l = \|\mathcal{U}\|$	the number of groups in G
β	the activation threshold, $0 < \beta \leq 1$
k	the number of initial seeds
$\rho(S)$	The expected number of activated groups with initial seed set S in G under given diffusion model.

specifies that each node u has only one chance to activate each of its neighbors after u is activated and all influence processes are independent.

A. Group Influence Maximization

Given a directed graph $G = (V, E, P)$, a *group* U is defined as a subset of V. Let $\mathcal{U}$ be the set of groups and l be the number of total groups. Given an activation threshold $0 < \beta \leq 1$, a group is said to be *activated* if β percent of nodes in this group are activated under IC model.

The GIM considers information propagation in social network under the IC model. The objective is to look for k initial seed users where the expected number of eventually activated groups is maximized

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

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

where S is the initial seed set and $\rho(S)$ is the expected number of groups eventually activated for given initial seed set S.

An example is shown in Fig. 2 to explain the diffusion process for GIM, where there are nine nodes and the influence probability of each edge is 1. And there exist four groups $\mathcal{U} = \{U_1 = \{v_1, v_2, v_3\}, U_2 = \{v_1, v_5\}, U_3 = \{v_4, v_7, v_9\}, and U_4 = \{v_6, v_8\}\}$. Assume the activation threshold $\beta = 0.5$. At the beginning, v_3 is selected as seed. At the first time step, $v_1, v_2, and v_4$ will be activated by v_3 as shown in Fig. 2(1). At the second time step, v_5 will be activated by v_1, and v_6 will be activated by v_4 as shown in Fig. 2(2). At the third time step, v_8 will be activated by v_6 as shown in Fig. 2(3). Then, the final activated nodes are $\{v_1, v_2, v_3, v_4, v_5, v_6, v_8\}$. According to the given activation threshold 0.5, groups U_1, U_2, U_4 are *activated*, while U_3 is inactivated. The number of eventually activated groups is $\rho(\{v_3\}) = 3$.

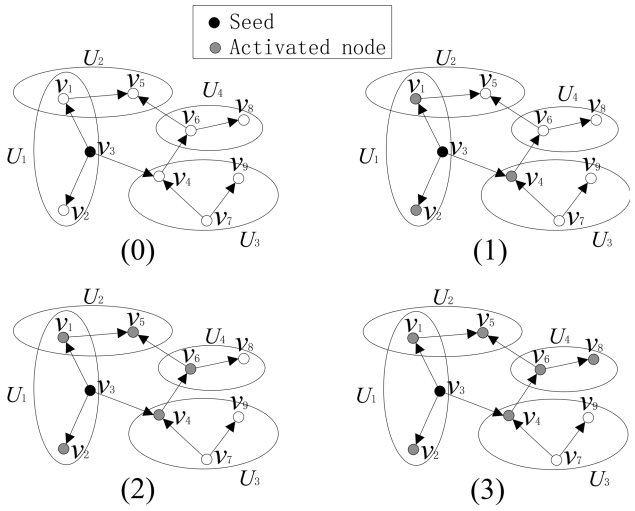

Fig. 2. Example of information diffusion process with initial seed v_3.

III. PROPERTIES OF GIM

In this section, we first present a statement of the hardness of the GIM, then discuss the properties of the objective function $\rho(\cdot)$.

A. Hardness Results

The IM problem has been proven to be NP-hard [3], which is a special case of our problem when each node is considered as a group and $\beta = 1$. Therefore, the GIM is obvious NP-hard.

Theorem 1: The GIM Problem is NP-hard.

Given an instance of GIM, it is difficult to compute the objective $\rho(S)$ even for a given seed set S since the activation process is not determinate but randomized according to the influence probability. Then, $\rho(S)$ is the expected value of the number of groups eventually activated. For such a problem, the Monte Carlo method is widely used to compute $\rho(S)$ by generating a large number of sample graphs of G and computing $\rho(S)$ on each sample graph. Finally, the output is the average value of all $\rho(S)$. The next section discussed on the number of graphs to be generated. Since computing the objective of IM had been proved #P-hard under the IC model [3], then we have the following result.

Theorem 2: Given a seed node set S, computing $\rho(S)$ is #P-hard under the IC model.

B. Realization of Random Graph

Given a general directed graph $G = (V, E, P)$, a realization g of G is a sample graph where g has the same node set with G and the edge set $E(g)$ of g is a subset of $E(G)$. While the influence probability on each edge in the sample graph is set 1, which means the influence process is determinate. The formulation process of a sample graph g is as follows: 1) for each edge $e \in E(G)$, randomly generate a number r between 0 and 1 uniformly and 2) e will be kept in g if and only if $P_e \geq r$. Now, g is a deterministic directed graph. Assume $\mathcal{G}$ contains all possible realizations of G. Not that there are

$2^{|E(G)|}$ graphs in $\mathcal{G}$. Let P_g be the probability that g can be generated. Then

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

Let $\rho^g(S)$ denote the number of groups activated by seed set S in g. Therefore, $\rho(S)$ can be expressed as

$$\rho(S) = \sum_{g \in \mathcal{G}} P_g \rho^g(S). \tag{3}$$

C. Modularity of Objective Function

Assume $f : 2^V \leftarrow \mathbb{R}$ is a set function. f is said to be *submodular* [23] if for any two subsets $V_1 \subset V_2 \subseteq V$ and $v \in V \setminus V_2$, $f(V_1 \cup \{v\}) - f(V_1) \geq f(V_2 \cup \{v\}) - f(V_2)$ holds. While if for any two subsets $V_1 \subset V_2 \subseteq V$ and $v \in V \setminus V_2$, it satisfies that $f(V_1 \cup \{v\}) - f(V_1) \leq f(V_2 \cup \{v\}) - f(V_2)$, f is *supermodular*. f is said to be *monotone* nondecreasing if for any $V_1 \subseteq V_2 \subseteq V$, it satisfies $f(V_1) \leq f(V_2)$. f is called a *polymatroid function* if it is submodular, monotone nondecreasing, and $f(\emptyset) = 0$, where $\emptyset$ denotes the empty set.

Polymatriod maximization problem with cardinality constraints has an $(1 - 1/e)$-approximation for greedy method [24]. Furthermore, this guarantee cannot be improved in general, since it cannot improve for Max k-cover(assuming $P \neq NP$) which is equivalent to polymatriod maximization problem, no polynomial algorithm could provide better approximation guarantees [25].

It is obvious that $\rho(\emptyset) = 0$ and $\rho(\cdot)$ is monotone nondecreasing. Unfortunately, $\rho(\cdot)$ is neither submodular nor supermodular under the IC model.

Theorem 3: $\rho(\cdot)$ is neither submodular nor supermodular in GIM under the IC model.

Proof: This theorem will be proved by formulating a counterexample. First, we will prove that $\rho(\cdot)$ is not supermodular. Consider an instance of GIM problem as shown in Fig. 2. Let $A = \emptyset, B = \{v_3\}$, and $v_9 \in V \setminus B$. We have $\rho(A) = 0, \rho(B) = 3$. Substituting v_9 into A and B, we have $\rho(A \cup \{v_9\}) = 0$ since v_9 cannot activate any group. $\rho(B \cup \{v_9\}) = 4$ since all groups are eventually activated. Thus, $\rho(A \cup \{v_9\}) - \rho(A) = 0$ and $\rho(B \cup \{v_9\}) - \rho(B) = 4 - 3 = 1$. Therefore, $\rho(A \cup \{v_9\}) - \rho(A) < \rho(B \cup \{v_9\}) - \rho(B)$ means $\rho(\cdot)$ is not submodular.

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

IV. Lower Bound and Upper Bound

A sandwich approximation strategy introduced by Lu *et al.* [21] for solving nonsubmodular optimization problem, which analyzes the objective function by comparing with its lower bound and upper bound. In this section, we will first design an upper bound for $\rho(\cdot)$, then a lower bound is presented for $\rho(\cdot)$.

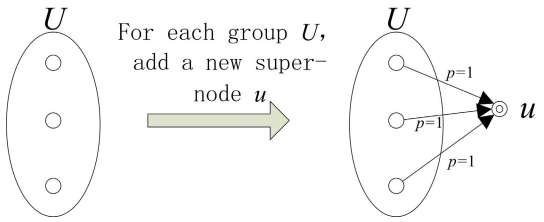

Fig. 3. Example for generating super node: assume U is a group, then add super node u and connect each node in U to u with influence probability 1.

A. Upper Bound

We will define a new set function $\overline{\rho}(\cdot)$ that satisfies $\rho(S) \leq \overline{\rho}(S)$ for any seed set $S \subseteq V$. The formulation process can be divided into two steps. Given an instance of GIM $G = (V, E, P)$, in the first step, we get a relaxed GIM (r-GIM) problem for given GIM by changing the group activation rules. For r-GIM problem, a group will be activated if there exists at least one activated node in this group. In the second step, for each group, we add a super node to the graph and connect each node in this group to the super node with influence probability 1. Fig. 3 shows an example.

Assume W is the super node set and E' is the edge set for node in V to node in W. Then, an instance of a general weighted IM (WIM) problem is defined as follows. $V \cup W$ is the node set and $E \cup E'$ is the edge set. $C \subseteq V$ is the candidate seed set, while all k seed nodes must be picked from C. For each node v, there is a weight $f(v)$ and f satisfies

$$f(v) = \begin{cases} 1, & v \in W \\ 0, & v \in V. \end{cases}$$

For all nodes that belong to super node set, the weight f is 1. For the other nodes, f is 0. Assume S is the initial seed set. Let $\overline{\rho}(S) = \sum_{v \text{ is activated}} f(v)$ be the expected weight of eventually influenced nodes. $\overline{\rho}(S)$ yields to count all activated super nodes only. Then, $G = (V, C, E, P, f)$ is called general WIM problem with candidate seed set $C \subseteq V$. $\overline{\rho}(\cdot)$ is monotone, submodular and $\rho(S) \leq \overline{\rho}(S)$ for any seed set $S \subseteq V$.

Theorem 4: Given an instance of GIM $G = (V, E, P)$, $\overline{\rho}(\cdot)$ is an upper bound of $\rho(\cdot)$.

B. Lower Bound

In this section, we will formulate a lower bound for GIM. The main idea is to delete some groups from G, and only keep such groups whose β percent of nodes can be activated at the same time. That means there exists at least one node connect to β percent nodes of this group in G. Fig. 4 shows an example with activation threshold $\beta = 0.5$. Group U will be kept since there exist v_1 and v_2 that connect to two nodes in group U. Then, generate super node u for group U and add new directed edges $(v_1, u), (v_2, u)$ with influence probability $p_{(v_1,u)} = p_1 p_2$, $p_{(v_2,u)} = p_3 p_4$.

The general construction process is as follows. Given an instance of GIM, for each group U_i, suppose $H_i = \{v \in V | v$ connects to at least β percent nodes of $U_i\}$, if $H_i \neq \emptyset$,

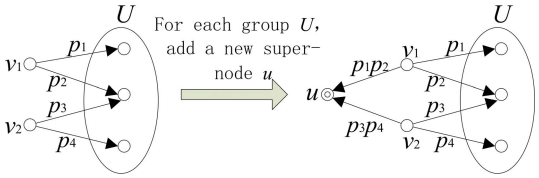

Fig. 4. Example for generation of lower bound problem.

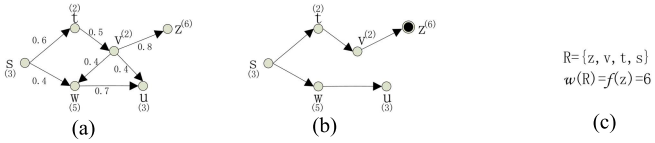

Fig. 5. Example of WRR set generation.

generate super node u and add directed edges $\{(v,u)|v \in H_i\}$. Assume W is the super node set. For each $v \in H_i$, assume U_i' contains all nodes in U_i which v connects to. Then, $p_{(v,u)} = \prod_{v' \in U_i'} p_{(v,v')}$. Finally, an instance of a general WIM problem could be formulated. $V \cup W$ is the node set, $E \cup E'$ is the edge set where E' contains all newly added edges, candidate seed set $C \subseteq V$ means k seed nodes must be selected from C. f is weight function of node which satisfy

$$f(v) = \begin{cases} 1, & v \in W \\ 0, & v \in V. \end{cases}$$

Suppose S is the initial seed set. Let $\underline{\rho}(S) = \sum_v$ is activated $f(v)$ be the expected weight of eventually influenced nodes. $\underline{\rho}(S)$ yields to count all activated super nodes only. Then, $\overline{G} = (V, C, E, P, f)$ is a general WIM problem with candidate seed set $C \subseteq V$. $\underline{\rho}(\cdot)$ is monotone, submodular and $\rho(S) \geq \underline{\rho}(S)$ for any seed set $S \subset V$.

Theorem 5: Given an instance GIM $G = (V, E, P)$, $\underline{\rho}(\cdot)$ is an lower bound of $\rho(\cdot)$.

V. Algorithm

Zhu *et al.* [17] have extended D-SSA [16] algorithm to solve a general WIM problem. Then, an estimation procedure for the objective function of GIM is proposed based on Monte Carlo method. Finally, a sandwich approximation framework will be presented for analyzing the performance of our algorithm.

A. (ϵ, δ)-Approximation

Sampling method needs to be used in solving random models since the expected number of influenced nodes $\rho(\cdot)$ is an expectation of a random process. To obtain an efficient estimation, we review the (ϵ, δ)-approximation in [26] which will be used in our algorithm.

Given ϵ as an absolute error of estimation and $(1-\delta)$ as the confidence, let X be a random variable and μ_X is a numerical characteristic of these random variables. (ϵ, δ)-approximation means a Monte Carlo estimator $\hat{\mu}_X$ of μ_X satisfies

$$\Pr[(1-\epsilon)\mu_X \leq \hat{\mu}_X \leq (1+\epsilon)\mu_X] \geq 1-\delta. \tag{4}$$

Define the sampling times as $\Upsilon = 4(e-2)\ln(2/\delta)/\epsilon^2$, then the Stopping Rule Algorithm given in [26] has been proven to be (ϵ, δ)-approximation. Algorithm 1 shows an estimation procedure for $\rho(\cdot)$ by applying (ϵ, δ)-approximation.

B. Extension for Reverse Influence Set Sampling

In this section, we will recall the extended version of the RIS sampling method for weighted cases. First, define a general WIM problem under the IC model as $G = (V, C, E, P, f)$. $C \subseteq V$ is the candidate seed set where seed must be selected from C, f is weight function of node and P is the influence probability. Assume S is the initial seed set. Let $\rho'(S) = \sum_v$ is activated $f(v)$ denote the expected weighted number of influenced users. The aim is to select k initial seed users set S in C to maximize $\rho'(S)$. Note that $\rho'(S)$ is submodular and monotone. Since G is a weighted random graph, the sampling method is necessary in order to estimate $\rho'(S)$. An extension of the RIS method is also based on RIS, which needs to generate a set $\mathcal{R}$ of *weighted reverse reachable (WRR) sets*. Each WRR set R is a subset of V, and an example of such a WRR set generation is shown in Fig. 5. Fig. 5(a) presents a weighted random graph with ten nodes. Each node is associated with a weight. Then the generation process first selects a random node z and a sample graph is generated according to the probability on each edge. Fig. 5(b) shows a sample graph. Based on the reverse reachable technique, $\{z, v, t, s\}$ can reach node z. Then, $R = \{z, v, t, s\}$ is a WRR set with weight $w(R) = f(z) = 6$ as shown in Fig. 5(c).

Reverse influence sampling method has been proven efficiently since it does not need to compute the objective function of a huge number of sample graphs. By generating a certain number of WRR sets and applying the greedy algorithm for weighted maximum coverage problem, RIS can estimate the objective function and output an efficient approximation solution. The details of this extended case can be found in [17], and they also proposed an extended D-SSA algorithm (Algorithm ED-SSA) for solving the WIM problem which also guaranteed the $(1-1/e)$-approximation.

C. Estimation of Objective Function $\rho(S)$

According to the definition of (ϵ, δ)-approximation, $\rho(S)$ could be estimated by Monte Carlo method for a given node set S. Given an instance of GIM $G = (V, E, P)$, for any realization g, $\rho^g(S)$ could be computed by applying breadth-first search (BFS) procedure for a given node set S. For any given absolute error ϵ and confidence degree $1-\delta$, Algorithm 1 shows the detailed estimation process and the running time is $O(\Gamma(nm+nl))$ where Γ is the number of sample graphs for given ϵ and δ, n and m are the number of nodes and edges, respectively, and l is the number of groups.

D. Group Coverage Maximization Algorithm

In this section, we will present an algorithm for solving GIM based on group coverage maximization method. Given

Algorithm 1 Estimation of Objective Function $\rho(S)$ (EOF)

Input: An instance of GIM $G = (V, E, P)$, $0 \leq \epsilon, \delta \leq 1$, seed set S.

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

1: $\Gamma \leftarrow 4(e-2)(1+\epsilon)^2 \ln(2/\delta)(1/\epsilon^2)$
2: $\mathcal{G} \leftarrow$ generate Γ realization of random graph $G = (V, E, P)$
3: $\hat{\rho}(S) = \frac{1}{\Gamma}\sum_{g\in\mathcal{G}} \rho^g(S)$
4: **return** $\hat{\rho}(S)$

Algorithm 2 GCMA

Input: An instance of GIM $G = (V, E, P)$, the number of initial seed users k.

Output: a set of seed nodes, S_k.

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

Algorithm 3 Sandwich Approximation Framework

Input: an instance of GIM $G = (V, E, P)$, the number of seeds k, ϵ, δ.

Output: a set of seed nodes, S.

1: Assume S_L is the seed set by solving the lower bound problem with Algorithm ED-SSA [17].
2: Assume S_X is the seed set by solving the upper bound problem with Algorithm ED-SSA [17].
3: Assume S_A is the seed set by solving $G = (V, E, P)$ with Algorithm 2.
4: $S = \arg\max_{S_0\in\{S_L,S_X,S_A\}} \text{EOF}(G, S_0, \epsilon, \delta)$
5: **return** S

$G = (V, E, P)$ and the set of groups $\mathcal{U}$, let $\mathcal{U}(S)$ be the set of groups that contains any one of nodes in S, i.e., $\mathcal{U}(S) = \{U \in \mathcal{U} | U \cap S \neq \emptyset\}$. Algorithm 2 is shown by selecting the maximum marginal gain at each step and at most $O(knl)$ time complexity. Greedy algorithm may give a better solution, but the running time is $O(kn\Gamma(nm+nl))$. We will compare several different strategies by experiments.

E. Sandwich Approximation Framework

For GIM, we have formulated a lower bound and an upper bound for $\rho(\cdot)$. Then, Algorithm 3 shows the whole sandwich approximation framework. The main idea is to solve the lower bound and upper bound problems in order to obtain a solution which can be bounded.

We can prove that there exists the following theoretical result for sandwich approximation framework.

Theorem 6: Assume S is the seed set obtained from Algorithm 3, then we have

$$\rho(S) \geq \max\left\{\frac{\rho(S_X)}{\overline{\rho}(S_X)}, \frac{\underline{\rho}(S_L^*)}{\rho(S^*)}\right\}\frac{1-\epsilon}{1+\epsilon}\left(1-\frac{1}{e}-\epsilon\right)\rho(S^*) \tag{5}$$

where S_L^* is the optimal solution for solving the lower bound IM problem and S^* is the optimal solution for the original GIM.

Proof: Let S_X^* be the optimal solution for solving the upper bound IM problem. Since Algorithm ED-SSA [17] guarantees a $(1-(1/e)-\epsilon)$-approximation, we have $\overline{\rho}(S_X) \geq (1-(1/e)-\epsilon)\overline{\rho}(S_X^*)$. Then, we have

$$\rho(S_X) = \frac{\rho(S_X)}{\overline{\rho}(S_X)}\overline{\rho}(S_X) \geq \frac{\rho(S_X)}{\overline{\rho}(S_X)}\left(1-\frac{1}{e}-\epsilon\right)\overline{\rho}(S_X^*)$$

where S_X^* is the optimal solution for upper bound, which yields $\overline{\rho}(S_X^*) \geq \overline{\rho}(S^*)$. Also, $\overline{\rho}(S^*) \geq \rho(S^*)$ because the objective function $\overline{\rho}(\cdot)$ is an upper bound of $\rho(\cdot)$ for any given seed set $S \subseteq V$. Then

$$\begin{aligned}\rho(S_X) &\geq \frac{\rho(S_X)}{\overline{\rho}(S_X)}\left(1-\frac{1}{e}-\epsilon\right)\overline{\rho}(S^*)\\ &\geq \frac{\rho(S_X)}{\overline{\rho}(S_X)}\left(1-\frac{1}{e}-\epsilon\right)\rho(S^*).\end{aligned}$$

On the other hand, Algorithm ED-SSA [17] also guarantees a $(1-(1/e)-\epsilon)$-approximation for solving the lower bound problem, and we have

$$\begin{aligned}\rho(S_L) \geq \underline{\rho}(S_L) &\geq \left(1-\frac{1}{e}-\epsilon\right)\underline{\rho}(S_L^*)\\ &\geq \frac{\underline{\rho}(S_L^*)}{\rho(S^*)}\left(1-\frac{1}{e}-\epsilon\right)\rho(S^*).\end{aligned}$$

Let $S_{\max} = \arg\max_{S_0\in\{S_L,S_X,S_A\}}\rho(S_0)$ which means $S_{\max}$ represents the set with the maximum expected objective value among $\{S_L, S_X, S_A\}$, then

$$\rho(S_{\max}) \geq \max\left\{\frac{\rho(S_X)}{\overline{\rho}(S_X)}, \frac{\underline{\rho}(S_L^*)}{\rho(S^*)}\right\}\left(1-\frac{1}{e}-\epsilon\right)\rho(S^*)$$

Since for any $S_0 \in \{S_L, S_X, S_A\}$, $(1-\epsilon)\rho(S_0) \leq \hat{\rho}(S_0) \leq (1+\epsilon)\rho(S_0)$ satisfies because the sampling method is (ϵ, δ)-approximation. In addition, $\hat{\rho}(S) = \max_{S_0\in\{S_L,S_X,S_A\}}\hat{\rho}(S_0)$ according to step 4. We have

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

It follows that

$$\begin{aligned}\rho(S) &\geq \frac{1-\epsilon}{1+\epsilon}\rho(S_{\max})\\ &\geq \max\left\{\frac{\rho(S_X)}{\overline{\rho}(S_X)}, \frac{\underline{\rho}(S_L^*)}{\rho(S^*)}\right\}\frac{1-\epsilon}{1+\epsilon}\left(1-\frac{1}{e}-\epsilon\right)\rho(S^*).\end{aligned}$$

From Theorem 6, we can find the difference between $\rho(S^*)$ and $\underline{\rho}(S_L^*)$ has a significant influence on the performance of Algorithm 3. Iyer and Bilmes [27] have studied the problem of minimizing the difference of two submodular functions and proved that the difference between $\rho(S^*)$ and $\underline{\rho}(S_L^*)$ could be

bounded. Then, the following result of the theoretical gap is true.

Theorem 7: Assume S_L^* is the optimal solution for the lower bound IM problem, and S^* is the optimal solution for the original GIM, then we have the following result:

$$\rho(S^*) - \underline{\rho}\left(S_L^*\right) \leq \max_{S,|S|=k} \left(\overline{\rho}(S) - \underline{\rho}(S)\right). \tag{6}$$

Proof: According to the definition of S_L^* and S^*, $\underline{\rho}(S_L^*) \geq \underline{\rho}(S^*)$ satisfies since S_L^* is the optimal solution of the lower bound problem. Then, we have

$$\rho(S^*) - \underline{\rho}\left(S_L^*\right) \leq \rho(S^*) - \underline{\rho}(S^*).$$

On other hand, $\overline{\rho}(\cdot)$ is an upper bound of $\rho(\cdot)$, which yields $\rho(S^*) \leq \overline{\rho}(S^*)$. In total, we have

$$\rho(S^*) - \underline{\rho}\left(S_L^*\right) \leq \overline{\rho}(S^*) - \underline{\rho}(S^*) \leq \max_{S,|S|=k} \overline{\rho}(S) - \underline{\rho}(S)).$$

According to Theorem 7, the performance of Algorithm 3 is up to the quality of lower bound and upper bound. A different definition of these two bounds will lead to different approximation ratios. We will analyze the proposed bounds by real-world data experiments in the next section.

VI. Experiments

In this section, we will verify our algorithms by testing on real-world social networks. First, two data sets are described. Then, three algorithms applied in the experiments will be explained in detail. Finally, experiment results are shown and our observations will be presented.

A. Data Description

In this article, we have used two data sets from [28] and [29]. The first data set is Facebook-like Forum Network which was collected from a similar online community like the Facebook online social network. It records users' activity in the forum. This data set contains one-mode and two-mode data. The two-mode data is an interesting network of 899 users liking a topic among 522 topics. The one-mode data represents the relationship among the 899 users. Each topic is assumed to be one group and the people who like the same topic are assumed to belong to a group. The second data set is the Newman's scientific collaboration network which represents the coauthorship network based on preprints posted to Condensed Matter section of arXiv E-Print Archive between 1995 and 1999. The two-mode data represents the relationship between an authors and the article that they have written, and the one-mode data represents the relationship among the coauthors. Each article is considered as a group, and the coauthors of an article are assumed to be in a group. The details about the data are mentioned in Table II.

B. Procedure

For both data sets, the one-mode data is used to build the graphs and all graphs are directed graphs. The influence probability was assigned based on the degree of node. Assume N is the number of indegree of a given node, then influence probability on its incoming edge is $1/N$. This method is widely used in the previous works of literature. Three approaches have been used to establish the input seed set. They are Greedy, Maximum Coverage, and Maximum Outdegree. All the programs are written in Python 3.6.3 and run on a Linux server with 16 CPUs and 256-GB RAM. For each strategy, we generated 10 000 sample graphs and iterated over all these sample graphs to obtain the results.

1) Greedy: In the greedy approach, the k seed set is chosen by iterating k times over the whole node set of the network and choosing the k nodes that activate the maximum groups. The output of this approach is comparatively the most optimal. However, the run time of the algorithm is really high as the algorithm iterates over the entire node set multiple times to determine the node that activates the maximum number of groups.

2) Maximum Coverage: In this approach, the k seed set is determined by the nodes that have the maximum coverage of groups. The seed set contains the top k nodes that are a part of a large number of groups in the network. The seed set in this approach is fixed and thus has a run time lower than the greedy approach. However, the result is comparatively lower than that of the greedy approach.

3) Maximum Outdegree: In this approach, the k seed set is fixed and is given by the first k nodes of the graph that have the maximum outdegree, i.e., the nodes that have the maximum number of edges going out of it. This algorithm has a run time lower than the greedy approach and almost similar to the maximum coverage approach but the result is comparatively lower than that of both the greedy approach and the maximum coverage approach.

TABLE II
Data Statistics Using in Our Experiments

	Number of nodes	Number of edges	Number of groups	Average group size
Dataset1	899	142760	522	14.6
Dataset2	16726	95188	22015	3.7

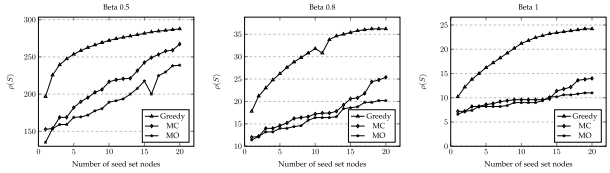

Fig. 6. Experimental results for data set 1.

C. Experimental Results

The observations of the experimental results are shown in Figs. 6 and 7 for data sets 1 and 2, respectively. From the graphs, the following results are obtained.

1) Greedy Algorithm Outperforms the Other Approaches: By comparing the three algorithms, it has been observed that the greedy approach gives a comparatively higher output than the maximum coverage and the maximum outdegree methods, i.e., given a network, greedy approach activates more number of groups compared to the others. The maximum outdegree

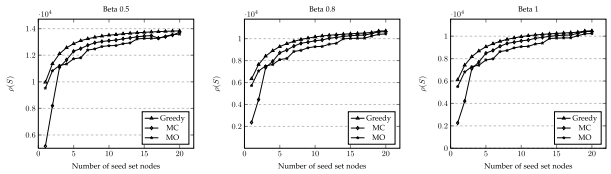

Fig. 7. Experimental results for data set 2.

TABLE III
DATA STATISTICS OF YOUTUBE DATA SET

	Number of nodes	Number of edges	Number of groups	Average group size
Youtube	1,134,890	2,987,624	8,385	13.50

initially gives higher group activate count than the maximum coverage but as the number of seed nodes increases, the maximum coverage approach outperforms the maximum outdegree approach but is still lower than the greedy approach. The run time of the greedy approach is however higher compared to that of the maximum coverage and the maximum outdegree approach as it iterates through the entire node set to get the k seed set.

2) With Increase in Beta, Groups Activated Decreases: The experiments are carried out with three values for beta 0.5, 0.8, and 1. As the beta increases, it is observed that the number of groups activated for a given node decreases. The seed set activates more groups but the activation by a single node decreases.

3) Group Size Effects the Output: For groups with different average sizes, the performance of the three algorithms is affected. Large size groups tend to show a bigger gap between the greedy approach and the other approaches, but small size groups do not have that much of gap between the three approaches.

D. Experiments on Large Data Set

In this experiment, we used the Youtube data set [30]. Youtube is a video-sharing web site that includes a social network. In the Youtube social network, users form friendship with each other and users can create groups where other users can join. The data statistics are provided in Table III.

The data set had two files, one representing the edges of the graph and the other was a list of the communities. We experimented with the top 5000 community. The graphs were sampled to give various subsets of the original graph and the experiments were run on all the sample graphs to calculate the count of groups activated. Finally, the average of the output of all results was taken to plot the graph. The edge probability was assigned based on the indegree of the node. For example, $p(u, v) = 1/d$, for an edge between u and v and d is the indegree of v. We also used the above three approaches for the experiments.

1) Experimental Results: We experimented for 100 seed set nodes. Fig. 8 shows the results.

1) The greedy algorithm outperforms the other approaches. When we analyzed the output of the three algorithms, it was observed that the greedy algorithm activated the maximum number of groups as compared to the other two methods. The maximum coverage algorithm initially gave higher results, but as the size of the seed set increased, the greedy algorithm gave better results. However, the run time of the greedy approach is comparatively higher.
2) Node set of maximum outdegree as seeds activated a very low number of groups. It was observed that for 100 seeds, the maximum outdegree nodes activated roughly 52 groups compared to the 100 nodes of maximum coverage (which activated around 300 groups). One possible reason for this happening could be the nodes that are activated by the maximum outdegree were not a part of any group. So even if the nodes with maximum outdegree did activate a large number of nodes, the resulting number of groups activated was less.

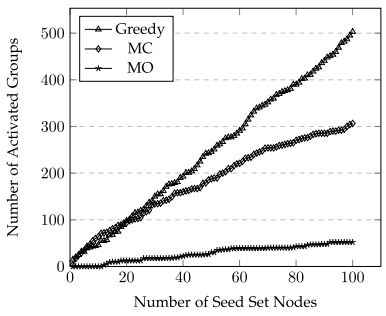

Fig. 8. Experimental results for Youtube with $\beta = 0.1$.

VII. CONCLUSION

In this article, we investigated the GIM problem in social networks. We proposed a new IM model with consideration of group activation. The GIM was formulated to select k nodes under IC model to maximize the number of influenced groups. We have shown that GIM was NP-hard and the objective function was neither submodular nor supermodular. For solving GIM, a group coverage maximization strategy was proposed. Finally, we formulate a sandwich approximation framework which guarantees a theoretical result, and the experiments show that the proposed algorithm is efficient and effective. For future research, novel efficient methods for solving nonsubmodular optimization problem are valuable to study.

REFERENCES

[1] D. R. Forsyth, *Group Dynamics*. Boston, MA, USA: Cengage Learning, 2018.
[2] M. Meeker and L. Wu, "Internet trends report 2018," *Kleiner Perkins*, vol. 5, p. 30, 2018.
[3] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 137–146.
[4] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Van Briesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data*, Aug. 2007, pp. 420–429.

[5] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 1029–1038.
[6] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "SIMPATH: An efficient algorithm for influence maximization under the linear threshold model," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 211–220.
[7] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation: Scaling up with guarantees," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, Nov. 2014, pp. 629–638.
[8] N. Ohsaka, T. Akiba, Y. Yoshida, and K.-I. Kawarabayashi, "Fast and accurate influence maximization on large networks with pruned Monte-Carlo simulations," in *Proc. AAAI*, Jun. 2014, pp. 138–144.
[9] Y. Yang, Z. Lu, V. O. Li, and K. Xu, "Noncooperative information diffusion in online social networks under the independent cascade model," *IEEE Trans. Comput. Social Syst.*, vol. 4, no. 3, pp. 150–162, Sep. 2017.
[10] C. Aslay, L. V. Lakshmanan, W. Lu, and X. Xiao, "Influence maximization in online social networks," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 775–776.
[11] C. Aslay, F. Bonchi, L. V. Lakshmanan, and W. Lu, "Revenue maximization in incentivized social advertising," *Proc. VLDB Endowment*, vol. 10, no. 11, pp. 1238–1249, Aug. 2017.
[12] Y. Li, J. Fan, Y. Wang, and K. L. Tan, "Influence maximization on social graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1852–1872, Oct. 2018.
[13] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2014, pp. 75–86.
[14] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2015, pp. 1539–1554.
[15] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. 25th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2014, pp. 946–957.
[16] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proc. Int. Conf. Manage. Data*, Jul. 2016, pp. 695–710.
[17] J. Zhu, J. Zhu, S. Ghosh, W. Wu, and J. Yuan, "Social influence maximization in hypergraph in social networks," *IEEE Trans. Netw. Sci. Eng.*, to be published.
[18] H.-J. Hung *et al.*, "When social influence meets item inference," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 915–924.
[19] M. Narasimhan and J. A. Bilmes, "A submodular-supermodular procedure with applications to discriminative structure learning," 2012, *arXiv:1207.1404*. [Online]. Available: https://arxiv.org/abs/1207.1404
[20] F. Bach, "Learning with submodular functions: A convex optimization perspective," *Found. Trends Mach. Learn.*, vol. 6, nos. 2–3, pp. 145–373, 2013.
[21] W. Lu, W. Chen, and L. V. Lakshmanan, "From competition to complementarity: Comparative influence diffusion and maximization," *VLDB Endowment*, vol. 9, no. 2, pp. 60–71, 2015.
[22] W. L. Wu, Z. Zhang, and D. Z. Du, "Set function optimization," *J. Oper. Res. Soc. China*, vol. 7, no. 2, pp. 183–193, 2019.
[23] S. Fujishige, *Submodular Functions and Optimization*, vol. 58. Amsterdam, The Netherlands: Elsevier, 2005.
[24] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions," *Math. Program.*, vol. 14, no. 1, pp. 265–294, Dec. 1978.
[25] U. Feige, "A threshold of ln n for approximating set cover," *J. ACM*, vol. 45, no. 4, pp. 634–652, 1998.
[26] P. Dagum, R. Karp, M. Luby, and S. Ross, "An optimal algorithm for Monte Carlo estimation," *SIAM J. Comput.*, vol. 29, no. 5, pp. 1484–1496, 2000.
[27] R. Iyer and J. Bilmes, "Algorithms for approximate minimization of the difference between submodular functions, with applications," 2012, *arXiv:1207.0560*. [Online]. Available:https://arxiv.org/abs/1207.0560
[28] T. Opsahl, "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients," *Social Netw.*, vol. 35, no. 2, pp. 159–167, 2013.
[29] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, 2001.
[30] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. 7th ACM/Usenix Internet Meas. Conf. (IMC)*, San Diego, CA, USA, Oct. 2007, pp. 29–42.

Jianming Zhu received the B.S. and M.S. degrees in mathematics from Shandong University, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree in operations research from the Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing, China.

He was a Visiting Scientist at the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA, from December 2017 to December 2018. He is currently an Associate Professor with the School of Engineering Science, University of Chinese Academy of Sciences, Beijing. His current research focuses on algorithm design and analysis for optimization problems in data science, wireless networks, and management science.

Smita Ghosh received the bachelor's degree in computer science and engineering from the West Bengal University of Technology, Kolkata, India, in 2015, and the master's degree in computer science from the University of Texas at Dallas, Richardson, TX, USA, in 2017, where she is currently pursuing the Ph.D. degree in computer science, under the supervision of Dr. W. Wu, with the Department of Computer Science.

Her current research interests include social networks and incorporating the big data analysis of social network with machine learning and deep learning.

Weili Wu (M'00) received the M.S. and Ph.D. degrees from the Department of Computer Science, University of Minnesota, Minneapolis, MN, USA, in 1998 and 2002, respectively.

She is currently a Full Professor with the Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA. Her current research interests include design and analysis of algorithms for optimization problems that occur in wireless networking environments and various database systems, especially data communication and data management.