CrossMark

Team formation with influence maximization for influential event organization on social networks

Cheng-Te Li[1] · Mei-Yuan Huang[2] · Rui Yan[3]

© Springer Science+Business Media, LLC 2017

Abstract Online event-based social services allow users to organize social events by specifying the themes, and invite friends to participate social events. While the event information can be spread over the social network, it is expected that by certain communication between event hosts, users interested in the event themes can be as more as possible. In this paper, by combining the ideas of team formation and influence maximization, we formulate a novel research problem, *Influential Team Formation* (ITF), to facilitate the organization of social events. Given a set L of required labels to describe the event topics, a social network, and the size k of the host team, ITF is to find a k-node set S that satisfying L and maximizing the *Influence-Cost Ratio* (i.e., the influence spread per communication cost between team members). Since ITF is proved to be NP-hard, we develop two greedy algorithms and one heuristic method to solve it. Extensive experiments conducted on Facebook and Google+ datasets exhibit the effectiveness and efficiency of the proposed methods. In addition, by employing the real event participation data in Meetup, we show that ITF with the proposed solutions is able to predict organizers of influential events.

✉ Cheng-Te Li
chengte@mail.ncku.edu.tw

Mei-Yuan Huang
lisagood1217@gmail.com

Rui Yan
ruiyan@pku.edu.cn

[1] National Cheng Kung University, No. 1, University Road, Tainan City, 701 Taiwan,
Republic of China

[2] Technische Universität München, ArcisstraBe 21, D-80333 Munich, Germany

[3] Peking University, No. 5, Yiheyuan Road., Haidian, Beijing, China

🌐 Springer

1 Introduction

Team formation [9] in a social network is to find a set of experts such that not only a set of required labels is covered but also team members have lower communication cost with one another (i.e., well-connected in the underlying network). It is apparent that team formation can be applied to many real applications, such as searching for a group of employers to execute a project in a company, and composing an activity group for a cocktail party with particular themes. However, team formation techniques [1, 7, 12] are not applicable for organizing *Influential Events* in event-based social services (e.g. Meetup,[1] Plancast,[2] and Facebook Events). Here we consider influential event organization is to find a set of persons who are interested in the themes of an event, have better social interactions (i.e., lower communication cost), and can attract a large number of people to participate in the event. It is common and realistic to organize influential events. The real-world scenarios on the demand of influential events may include organizing technical conferences, fund raising for earthquake victims, and initiating anti-nuclear campaign. In such scenarios, people attempt to maximize the number of participants since more participants mean a success of the events. One may think *Social Influence Maximization* [8], which aims at finding a set of seeding users such that the number of influenced users can be maximized, seems to be a solution. However, influence maximization techniques [3, 4, 19] are not applicable for influential event organization because they consider neither the set of required labels, nor the communication between the selected seed nodes.

This paper proposes a novel problem, *Influential Team Formation* (ITF), in a social network. Given a set of required labels and the size k of the team, the ITF problem is to find a set S of nodes such that (a) the query label set is covered by the discovered k-node set S, and (b) the *influence-cost ratio* of nodes in S is maximized. We propose the Influence-Cost Ratio (ICR) to quantify the influence spread of the selected k nodes per communication cost. ICR of a node set S is defined as $ICR(S) = \frac{\sigma(S)}{c(S)}$, where influence spread $\sigma(S)$ is the expected number of nodes activated by S while the communication cost $c(S)$ is the sum of all-pair shortest path lengths between nodes in S. A team can derive a higher ICR if the team members can lead to higher influence spread and are well-connected. The ITF problem is challenging since maximizing influence spread contradicts minimizing communication cost. Influence maximization tends to select well-separated nodes because their activated nodes can have less overlapping. But team formation prefers well-connected nodes since they can produce lower communication cost.

It is worthwhile to note that a team is a task-oriented group whose team members not only possess some skill labels to deal with the task, but also well collaborate with each other. Therefore, the team formation problem asks for a set of required skill labels as the

---

[1]www.meetup.com

[2]www.plancast.com

input, and expects that the discovered team members are equipped with some of the required skill labels and have good communication among them. Since we aim at forming influential "teams", the selected team members (i.e., seeds) need to rely on a required set of skill labels and be well-connected to have good communication. In addition, "influential" teams also require the team members to be influential, i.e., team members should lead to higher influence spread in the social network. Consequently, the proposed ITF problem is a combination of team formation and influence maximization.

We create an example, as shown in Figure 1, to exhibit the differences among team formation (TF), influence maximization (IM), and the proposed ITF. This example assumes the set of required labels is $\{a, b, c, d, e\}$ and $k = 3$. TF may select the set $S_{TF} = \{v_1, v_2, v_3\}$ since they cover more required labels and are well-connected. $ICR_{TF} = \frac{7}{3}$. IM will select the set $S_{IM} = \{v_1, v_4, v_6\}$ because they can lead to highest influence spread. $ICR_{IM} = \frac{11}{5}$. ITF will find the set $S_{ITF} = \{v_1, v_5, v_6\}$ that leads to the highest $ICR_{ITF} = \frac{10}{3}$. It is because not only the union of the activated sets of v_1, v_5, and v_6 leads to the largest activated set (i.e., $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_9, v_{14}, v_{15}\}$), but also v_1, v_5, and v_6 are inter-connected with a triangle structure in the network.

To this end, we formulate the ITF problem under the *Independent Cascade* (IC) model, and prove its NP-hardness. We propose a greedy algorithm with quality guarantee. While the greedy solution is effective but very inefficient, we further develop two greedy methods: *ICR Greedy* (ICR-Greedy) and *Mixed Influence-Cost Greedy* (M-Greedy), and one heuristic method: *Similar Influence Search* (SimIS). ICR-Greedy iteratively selects nodes with highest marginal gain of ICR scores. M-Greedy combines the NewGreedy IM method [3] with the original TF algorithm [9] in an interweaving manner. SimIS integrates Group-PageRank [15] with a best-first search in the social network. To validate the proposed methods, we have simulation-based and prediction-based experiments. The simulation-based experiments conducted on two real social network datasets, Facebook and Google+, and the results show both M-Greedy and SimIS can generate the highest ICR scores with satisfying time efficiency. The prediction-based experiments are conducted using the real event participation data of the event-based social service Meetup. The goal is to validate whether ITF
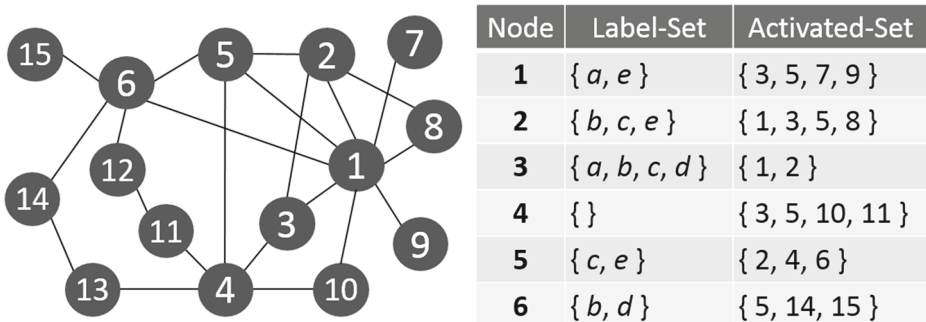

Node	Label-Set	Activated-Set
1	$\{a, e\}$	$\{3, 5, 7, 9\}$
2	$\{b, c, e\}$	$\{1, 3, 5, 8\}$
3	$\{a, b, c, d\}$	$\{1, 2\}$
4	$\{\}$	$\{3, 5, 10, 11\}$
5	$\{c, e\}$	$\{2, 4, 6\}$
6	$\{b, d\}$	$\{5, 14, 15\}$

Figure 1 A toy example of a social network (left), and a table (right) that describes the set of required labels possessed by each node and the set of nodes activated by each node. Note that a subset of nodes is shown in the table. Nodes except for v_1 to v_6 do not contain any required label

with the proposed solution can truly identify the organizers of influential events based on the required labels of the given event and the social network. The results exhibit satisfying accuracy.

The contributions of this work consist of four parts, as described in the following.

– We formulate the novel Influence Team Formation (ITF) problem to find a set of users as team organizers for initiating influential events on social network. By integrating Team Formation with Influence Maximization, we devise the Influence-Cost Ratio (ICR) to estimate the influence spread per communication cost among team members, and aim at maximizing ICR while covering the required labels in the ITF problem.

– We analyze the hardness of the ITF problem. And we develop three algorithms, ICR-Greedy, M-Greedy, and SimIS, to solve the ITF problem in either the greedy or heuristic manner.

– We conduct a comprehensive evaluation on our proposals with some baselines using real Google and Facebook datasets. Experimental results validates our idea and exhibit the promising effectiveness (in terms of ICR scores) and time efficiency of the proposed methods, especially for SimIS.

– In addition to the simulation-based evaluation, we further use real event participation data in Meetup to validate whether the proposed methods can accurately predict the organizers of influential events based on the required labels. Experimental results show satisfying accuracy, especially for SimIS.

2 Related work

The relevant studies consists of three parts: team formation, influence maximization, and social event organization. Team formation is first proposed to find a set of experts such that not only a set of required skills are covered, but also the communication cost among team members is minimized [9]. There are a series of follow-up extensions considering various real scenarios: jointly finding a team leaders and forming the team [7], simultaneously tackling multiple sets of required skills [1], specifying the number of experts for each skill [12], allowing geographical and team-size constraints [20], imposing swarm-based optimization [2], and recommending other individuals to replace some of existing team members [14]. *Community Search* [22] alternatively finds a densely-connected subgraph based on a set of given nodes, instead of required skills.

The Influence maximization problem is to find a set of seed nodes that can maximize the influence spread (i.e., the expected number of activated nodes) in a social network under either Independent Cascade or Linear Threshold model [8]. There are two mainstream solutions that attempt to strike a balance between efficiency and effectiveness. The first is to develop seed-selection heuristics, such as Degree Discount [3], Maximum Influence Arborescence [4], Group-PageRank [15], and IM-Rank [6]. The second is to speed up Monte-Carlo simulation-based greedy algorithm [8], such as Cost-Effect Lazy-Forward [10], NewGreedy [3], StaticGreedy [5], Pruned Monte-Carlo [19], and supervised Monte Carlo estimation [16]. In contrast to influence maximization, Zhang et al. [27] alternatively aims at finding a critical block of nodes to control the misinformation diffusion in a social network.

Social event organization aims at composing a group of persons that satisfying various kinds of event requirements. Socio-spatial Group Query [25] is to find a group of persons who are not only geographically close to each other but also acquainted with each other, and their extended work [21] maximizes the likelihood of friend making among group members. A follow-up work [11] recommends a group of users satisfying required labels and being acquainted with each other for an event host. SEO [13] further composes multiple event groups simultaneously. Marketing effect maximization [26] aims to find a set of nodes that are geographically close to the event location and attract more those users satisfying event themes. The bottleneck-aware social event arrangement (BSEA) [23] further considers social influence to recommend events for users.

3 Problem formulation

We first describe some preliminary notations. First, a social network is represented by a graph $G = (V, E)$, where V is the node set and E is the edge set. Each node u is associated with a set of labels L_u. Second, let $L(S)$ be the set of required labels covered by a node set S, i.e., $L(S) = L \cap (\cup_{u \in S} L_u)$, where L is the set of required labels. We define the *label coverage* π of a node set S, given by $\pi(S) = \frac{|L(S)|}{|L|}$. Third, we adopt the Independent Cascade model to propagate influence. In IC model, in time step t each active node u has a single chance to activate each of its inactive neighbors v with a pre-determined probability $p_{u,v}$. If u succeeds, v will become active in step $t+1$. Otherwise, u will not activate v again. The process ends when no more nodes can be activated. In this paper, $p_{u,v}$ is uniformly selected from the set {0.1, 0.2, ..., 0.9} based on the TRIVALENCY model [4]. Fourth, the *influence spread* of a node set S, denoted by $\sigma(S)$, is the expected number of activated nodes given S. Fifth, we define the *communication cost* of a node set S, denoted by $c(S)$, as the sum of all-pair shortest path lengths [7], i.e., $c(S) = \sum_{u,v \in S} len(u, v)$, where $len(u, v)$ is the length of the shortest path between u and v. To this end, we can define the *influence-cost ratio* of a node set S as $ICR(S) = \frac{\sigma(S)}{c(S)}$.

Influential team formation (ITF) Given a set of required labels L, a social network $G = (V, E)$, and the size k of the desired team, the ITF problem is to find a k-node set $S \subseteq V$ such that not only the label coverage $\pi(S) = 1$ but also the influence-cost ratio $ICR(S)$ is maximized.

Theorem 1 *For the IC influence propagation model, the influential team formation problem is NP-hard.*

Proof Sketch The ITF problem can be divided into two parts: team formation (TF) and influence maximization (IM). The TF part aims at finding a k-node set S that covers L and minimizes the communication cost $c(S)$. This part had been proved to be NP-hard [7] by a reduction from *3-Satisfaction* (3-SAT). The IM part is to find a k-node set that maximizes the influence spread $\sigma(S)$ under IC model, which had also been proved to be an NP-hard problem [8] by a reduction from *Set Cover* problem. Since maximizing $ICR(S)$ is composed by maximizing $\sigma(S)$ and minimizing $c(S)$ simultaneously, it is natural that the ITF problem is also NP-hard. □

4 Proposed methods

In this section, we propose three algorithms to solve the ITF problem. First, we describe a ICR-Greedy algorithm, which extends the "Lazy-Forward" influence maximization greedy method to deal with the set of required labels and ICR. Second, we develop a Mixed Influence-Cost Greedy algorithm, which integrates the team formation procedure with maximizing influence spread. Third, we present a heuristic method, Similar Influence Search, whose idea is that influential near-by nodes have similar distributions of influence spread over all the nodes in the network, and find team members by selecting nodes with similar influence distributions.

Algorithm 1 Lazy-forward ICR-greedy (ICR-greedy)

Input: A social network $G = (V, E)$, the team size k, and the set L of required labels.
Output: A k-node set S (team).

1 Initialize $S = \emptyset$;
2 $V^L = \{v | v \in V \text{ and } L_v \cap L \neq \emptyset\}$;
3 **for** *each node* $v \in V^L$ **do**
4       $\Delta_v = \sigma(\{v\})$;
5       $stamp_v = 0$;
6 **while** $|S| < k$ **do**
7       $v = \arg\max_{v \in V^L \setminus S} \Delta_v$;
8   **if** $stamp_v == |S|$ **then**
9           $S = S \cup \{v\}$;
10   **else**
11          $\Delta_v = ICR(S \cup \{v\}) - ICR(S)$;
12          $stamp_v = |S|$;

4.1 Lazy-forward ICR-greedy method

Since the monotonicity of both $\sigma(\cdot)$ and $c(\cdot)$ is straightforward and $\sigma(\emptyset) = 0$, let $c(\emptyset) = 1$, we can have $ICR(\emptyset) = 0$. Based on the set function theory [17], a simple greedy can generate an approximated result with $1 - 1/e$ ($\approx 63\%$) to the optimal solution. Since the ITF problem is NP-hard, we can estimate the $ICR(\cdot)$ by run Monte-Carlo simulation up to sufficient times R (e.g. $R = 20,000$), which was originally adopted by influence maximization [8]. However, such Monte-Carlo greedy method has time complexity $O(knRm)$, where $n = |V|$ and $m = |E|$. Therefore, we resort to the *Lazy-Forward* strategy [10], which exploits the submodularity to speed up the computation by reducing the times of estimating $\sigma(\cdot)$. The detailed algorithm, *Lazy-Forward ICR-Greedy*, is provided in Algorithm 1. V^L is used to retrieve the set of nodes possessing at least one required label (line 2). We use δ_v to record the marginal ICR (line 11), and implement δ_v using Priority Queue to efficiently find the node with the highest marginal ICR. Note that the proposed methods, i.e., ICR-Greedy, M-Greedy, and SimIS, does not directly ensure the label coverage $\pi(S) = 1$. Nevertheless,

we assume that the number of required labels is usually not large (e.g. 5). Therefore, with a proper team size k, it is easy to make $\pi(S) = 1$.

Algorithm 2 Mixed influence-cost greedy (M-greedy)

    Input: A social network $G = (V, E)$, the team size k, and the set L of required labels.
    Output: A k-node set S (team).
1 Initialize $S = \emptyset$;
2 $V^L = \{v | v \in V \text{ and } L_v \cap L \neq \emptyset\}$;
3 **for** *each node* $v \in V^L$ **do**
4      $\Delta_v = \sigma(\{v\})$;
5      $stamp_v = 0$;
6 **while** $|S| < k$ **do**
7   **if** $|S|\%2 == 0$ **then**
8         $v = \arg\max_{v \in V^L \setminus S} \Delta_v$;
9   **else**
10        $v = \arg\min_{v \in V^L \setminus S} dist_v$;
11   **if** $stamp_v == |S|$ **then**
12        $S = S \cup \{v\}$;
13   **else**
14        $\Delta_v = \sigma(S \cup \{v\}) - \sigma(S)$;
15        $dist_v = c(S \cup \{v\}) - c(S)$;
16        $stamp_v = |S|$;

4.2 Mixed influence-cost greedy method

While the previous ICG-Greedy directly maximizes $ICR(\cdot)$, we devise another greedy algorithm that aims at separately maximizing $\sigma(\cdot)$ and minimizing $c(\cdot)$ in an *interweaving* manner. The rationale is that directly maximizing $ICR(\cdot)$ may destroy the connectivity of the team members. That said, the shortest paths between nodes in S could pass through irrelevant nodes that contains no required labels. Since the communication cost $c(S)$ is the denominator of $ICR(S)$, including unconcerned nodes will increase $c(S)$ and lead $ICR(S)$ to an unsatisfied local maximal. To alleviate such damage on $ICR(\cdot)$, we propose to balance the the trade-off between $\sigma(\cdot)$ and $c(\cdot)$ by mixing maximizing $\sigma(\cdot)$ and minimizing $c(\cdot)$. The algorithm, *Mixed Influence-Cost Greedy* (M-Greedy), is shown in Algorithm 2. We take advantage of the Lazy-Forward strategy on the marginal gain of $\sigma(\cdot)$ and $c(\cdot)$ to minimize the cost function $c(\cdot)$. The first If-Statement (line 7), $|S|\%2 == 0$, is created to interweave maximizing $\sigma(\cdot)$ (line 8) with minimizing $c(\cdot)$ (line 10).

4.3 Similar influence search heuristic

Although the previous two greedy methods can be effective for finding the team S with high $ICR(S)$, they are quite inefficient for real-world practical usages. It is because of the

computation of influence spread $\sigma(\cdot)$, even the Lazy-forward strategy is applied. Therefore, we aim at developing an effective and efficient heuristic method, Similar Influence Search (SimIS). The central idea of SimIS is two-fold. The first is exploiting the heuristic linear influence modeling [24] to efficiently estimate the influence spread: under Independent Cascade model [8], the influence reaching a node v is a linear combination of the influence from v's neighbors. Such idea enables us to linearly compute the influence spread of an arbitrary set of nodes in a closed form. The second is that a node set S with lower communication cost means its members tend to be close to each other in the social network, and thus they are supposed to generate similar influence toward other nodes in $V \setminus S$. If the formation of team members starts from the most influential node, and iteratively selects the next team member by finding the node that can generate the influence spread (to all of the remaining nodes) as close as possible to the influence spread generated by the currently selected team members, we might be able to approximately lower down the communication cost while maintaining the influence spread. To implement such two ideas, we first elaborate the linear influence modeling, then give the detailed algorithm of SimIS.

Based on the linear influence model [24], we can approximately have the seed set S's influence on a node $v \notin S$, given by:

$$\sigma(S \rightarrow v) = \rho \sum_{u \in N(v)} p_{v,u} \sigma(S \rightarrow u), \tag{1}$$

where $N(v)$ is the set of neighbors of node v, and $\rho \in (0, 1)$ is the damping factor of influence propagation (note that for $v \notin S$, $\sigma(S \rightarrow v) = 1$). This representation reflects that the influence from S to v is a linear combination of the influence from S to each of v's neighbors, and allows efficient linear computation of influence propagation in an iterative manner. Equipped with $\sigma(S \rightarrow v)$, we can further derive the influence from S to a node set T ($S \cap T = \emptyset$, $S \in V$ and $T \in V$), denoted by $\sigma(S \rightarrow T)$, by summing up the influence from S to each node $v \in T$: $\sigma(S \rightarrow T) = \sum_{v \in T} \sigma(S \rightarrow v)$. Since the influence maximization part of our ITF problem is targeting at all nodes in the network, here we use the entire node set V in the network to be T. The influence spread of S, i.e., $\sigma(S) = \sigma(S \rightarrow V)$, can be derived with time complexity $O(|E|)$ [24].

Algorithm 3 Similar influence search (SimIS) heuristic

    Input: A social network $G = (V, E)$, the team size k, and the set L of required labels.
    Output: A k-node set S (team).
1   Initialize $S = \emptyset$;
2   $V^L = \{v | v \in V \text{ and } L_v \cap L \neq \emptyset\}$;
3   Let $\sigma(S \rightarrow V) = \mathbf{0}$;
4   **while** $|S| < k$ **do**
5   **for** *each node* $v \in V^L \setminus S$ **do**
6         $\sigma(\{v\} \rightarrow V) = ComputeInfluence(\{v\})$;
7       $v^\star = \arg\min_{v \in V^L \setminus S} \|\sigma(\{v\} \rightarrow V) - \sigma(S \rightarrow V)\|_F^2$;
        $S = S \cup \{v^\star\}$;
8   **if** $|S| < k$ **then**
9         $\sigma(S \rightarrow V) = ComputeInfluence(S)$;

Algorithm 4 $ComputeInfluence(S)$

    Input: a node set $S \subset V$.
    Output: $\sigma(S \rightarrow V)$: the influence of S on the entire node set V in network G.
1 **for** *each node* $v \in V$ **do**
2   **if** $v \in S$ **then**
3        $\sigma(S \rightarrow v) = 1$;
4   **else**
5        $\sigma(S \rightarrow v) = 0$;
6 **while** $\alpha < MAX\_ITERATIONS$ **do**
7   **for** *each node* $v \in V \setminus S$ **do**
8        $\sigma(S \rightarrow v) = \rho \sum_{u \in N(v)} p_{v,u} \sigma(S \rightarrow u)$;
9   α++;
10 $\sigma(S \rightarrow V) = [\sigma(S \rightarrow v_1), \sigma(S \rightarrow v_2), ..., \sigma(S \rightarrow v_n)]$;

Finding a k-node set that maximizes $ICR(S)$ can be considered as maximizing $\sigma(S)$ and minimizing $c(S)$ simultaneously when iteratively adding nodes to S. We propose a Similar Influence Search (SimIS) heuristic to approximately fulfill such idea based on the linear influence modeling $\sigma(S \rightarrow T)$. SimIS begins with finding the first team member $s^{(1)}$ by selecting the node with the highest influence spread, i.e., $s^{(1)} = \arg\max_{v \in V} \sigma(v \rightarrow V)$, then adds $s^{(1)}$ into S. Here we use *Group-PageRank* [15] $GPR(v \rightarrow V)$ to efficiently estimate a node v's influence spread $\sigma(v \rightarrow V)$. $GPR(v \rightarrow V)$ had been shown to be the upper bound for $\sigma(v \rightarrow V)$ [15]. To find each of the next k-1 nodes, i.e, $s^{(i)}(1 < i \leq k)$, we propose to select the node $v \in V \setminus S$ whose influence on the entire node set V is as similar as the influence of currently selected S on V. Specifically, let $\sigma(S \rightarrow V)$ be the vector that represents the influence of S on all nodes V in the network, i.e., $\sigma(S \rightarrow V) = [\sigma(S \rightarrow v_1), \sigma(S \rightarrow v_2), ..., \sigma(S \rightarrow v_n)]'$, where $n = |V|$. We aim at finding the next team member $s^{(i)}(1 < i \leq k)$ by selecting the node v that can minimize the difference between $\sigma(S \rightarrow V)$ and $\sigma(\{v\} \rightarrow V)$. The selection of next k-1 nodes can be expressed by:

$$s^{(i)} = \underset{v \in V \setminus S}{\arg\min} \|\sigma(\{v\} \rightarrow V) - \sigma(S \rightarrow V)\|_F^2, \tag{2}$$

where $i = 2, 3, ..., k$, and $\|.\|_F^2$ denotes the Frobenius norm. Since nodes that are close to one another in the network have higher potential to have similar vectors of influence on V, (2) can approximately find the team member with lower communication cost. In addition, since the formation of the team starts from the node with highest influence spread, (2) can approximately select the next nodes with least loss of influence spread, i.e., maintaining the influence spread of S as high as possible. The detailed algorithm of SimIS is given in Algorithm 3, along with Algorithm 4, $ComputeInfluence(S)$: the computation of the influence of a node set S on all nodes V in the network.

In short, the basic idea, nodes close to each other have similar influences, enables use to devise the SimIS algorithm to maximize the influence spread and minimize the communication cost at the same time. The key is that the 1st selected node is to purely maximize the influence spread. Then the 2nd node is selected based on its influence similarity with the first one. If the influence of the 2nd node is similar to the 1st one, it tends to maximize the influence by mimicking the influence of the 1st node. To mimic the 1st node's influence, the 2nd node should be as close as possible to the 1st in the network so that their influence can be close. Such policy is applied to the next nodes to be selected.

5 Experiments via simulation

The evaluation consists of two parts, simulation-based and prediction-based, which are presented in this and next section respectively. In the simulation-based evaluation, we aim to answer the following questions. First, what is the effectiveness (in terms of ICR) of the proposed methods for solving the ITF problem, compared with team formation and influence maximization techniques as baselines? Second, to allow organizing influential teams for real-world usages, any solution is supposed to be able to instantly report the team members. How about the run time of the proposed methods? Third, an effective team requires the members to communicate with each other in an efficient and direct manner, instead of requiring the involvement of other *inter-mediator* persons. For example, in Figure 1, node v_5 is the inter-mediator in the team $S_{IM} = \{v_1, v_4, v_6\}$ since v_5 does not cover any label but is passed by the shortest paths between team members. Can the proposed methods lead to less involvement of inter-mediators for the communication between team members? Note that in this simulation-based experiment, since the goal is to demonstrate the performance in terms of ICR and time efficiency, no real event participants are involved. In the next section, we will present the prediction-based experiment using real event participation data.

5.1 Data and evaluation settings

We collect two sub-networks from Facebook and Google Plus friendship respectively for the experiments. The data statistics, including numbers of nodes and edges in separate social networks and the total numbers of labels, is shown in Table 1. We consider the attribute values as so-called "labels." For example, we have labels male and female from Sex attribute, UIUC and CMU from Education attribute, and 1983 and 2000 from Birth Year attribute. Any accessible attribute values from users' profiles are treated as the labels. Figure 2a and b exhibit the numbers of users who own a certain label. We can find few labels are very popular while most of labels are unpopular (i.e., less than 10 users possess such labels) for both datasets. Figure 2c and d show the cumulative distributions that the number of labels possessed by a user. It can be observed that most users have less than 20 labels while only less than 10% users have label numbers higher than 20%. Such distributions may result from the fact that most users in online social services do not want to provide complete information in their profiles.

We compare the proposed three methods, ICR-Greedy, M-Greedy, and Similar Influence Search (SimIS) Heuristic, with two baselines. The first is the *Enhanced Team Formation* (ETF) algorithm [9] while the second is the modified *Linear Influence Maximization* (LIM) algorithm [15]. ETF is the state-of-the-art method that finds team members by selecting k nodes that cover the required labels and possess high communication cost toward each other. But ETF cannot consider the influence spread of the team. LIM is the state-of-the-art method that selects a set of seed nodes maximizing influence spread and maintaining time efficiency. We modify LIM by imposing a requirement: only those the k influential nodes that collectively cover the required labels can be selected. Note that under such modification,

Table 1 Statistics of the used datasets

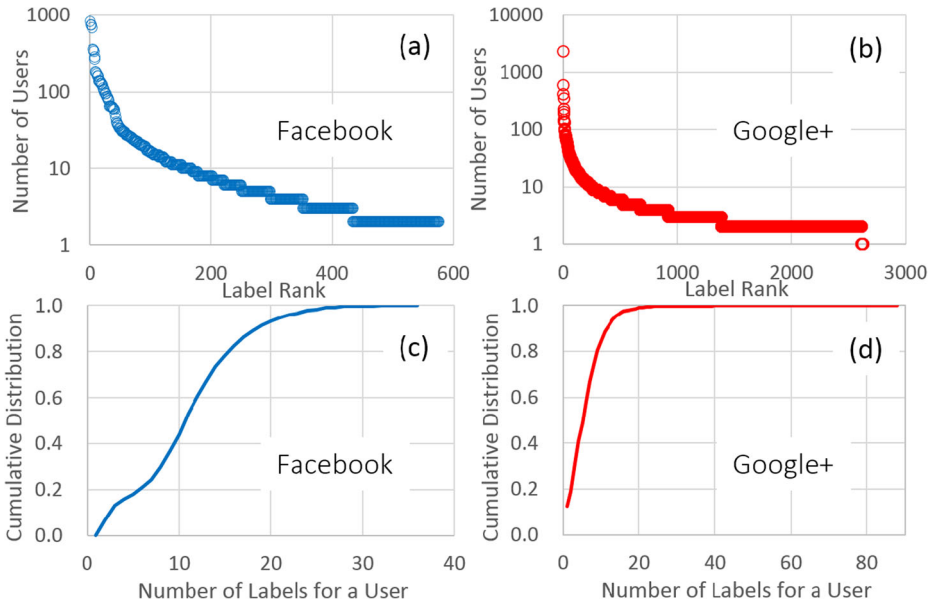|	#nodes	#edges	#labels
Facebook	1,045	53,498	576
Google+	3,478	435,569	2,627

Figure 2 Data distributions. **a** and **b** are the distributions of user numbers for labels. **c** and **d** are the cumulative distributions of the number of labels for a user

those nodes with largest marginal gain of influence spread could not be always selected since some of them could not cover any required labels. Nevertheless, LIM cannot minimize the communication cost between the selected seeds.

In the experiments, we mainly show how ICR scores and time efficiency (in second) change when varying the team size k. We randomly generate 20 sets of required labels for each dataset, in which each set contains 5 labels. Note that the size of required labels has also be changed and the results demonstrate similar trends. We use the IC model to estimate the influence spread, n which edge probabilities are uniformly selected from the set {0.1, 0.2, ..., 0.9} based on the TRIVALENCY model [4]. To further looking into the effect of different methods, we also show the results of influence spread and communication cost. In addition, since a good team does not need too much involvement of other inter-mediator users who help facilitate the communication between team members, we also report the *cardinality* of a team, which is defined as the number of nodes that are traversed via the shortest paths between team members. The cardinality of a good team is expected to be as close as to the team size. The final values, i.e., ICR, run time, cardinality, influence spread, and communication cost, are computed by averaging the results of 20 labels sets.

5.2 Experimental results

The main results of ICR and run time are shown in Figures 3 and 4 respectively. M-Greedy and SimIS are obviously able to generate higher ICR scores with satisfying time efficiency, comparing to baselines LIM and ETF. Though ICR-Greedy is close to M-Greedy and SimIS, it is very inefficient. In more details, M-Greedy has higher ICR scores for smaller teams while SimIS can be slightly better than M-Greedy for larger teams. Due to such results, we can suggest SimIS can be used for organizing larger events that may need more team
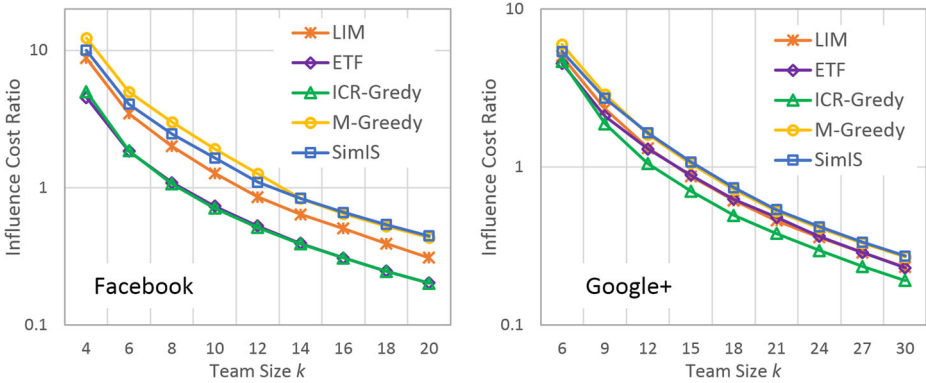
Figure 3 The performance of ICR by varying team sizes

members. The result that M-Greedy outperforms ICR-Greedy shows that maximizing ICR directly and greedily can lead to unsatisfying results while dealing with maximizing σ and minimizing c separately and greedily can strike a better balance between effectiveness and efficiency.

To understand why M-Greedy and SimIS lead to higher ICR scores but LIM and ETF perform worse, we show the influence spread σ and communication cost c in Figures 5 and 6 respectively. SimIS can generate higher σ than LIM on both datasets. We think the reason is that LIM is modified to cover required labels, so its σ is lowered down to some extent. Nevertheless, LIM's σ is quite close to M-Greedy and SimIS. On the other hand, we can also find Sim-IS's σ is higher than M-Greedy. It may results from half of M-Greedy algorithm is for minimizing c, so M-Greedy leads to lower c in Figure 6. Nevertheless, SimIS can also have lower c which is close to M-Greedy. It is natural to see ETF generates results with lowest c. In short, ETF and LIM lead to a trade-off between σ and c while M-Greedy and SimIS can strike a satisfying balance, and ICR-Greedy is an intuitive solution but generates worse results.

The last results for cardinality are shown in Figure 7. Both M-Greedy and SimIS need only few inter-mediator users to facilitate communication, i.e., requiring only additional
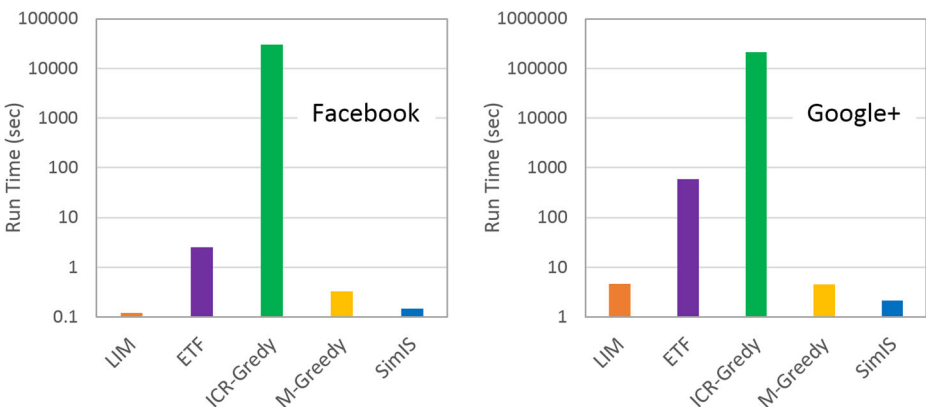

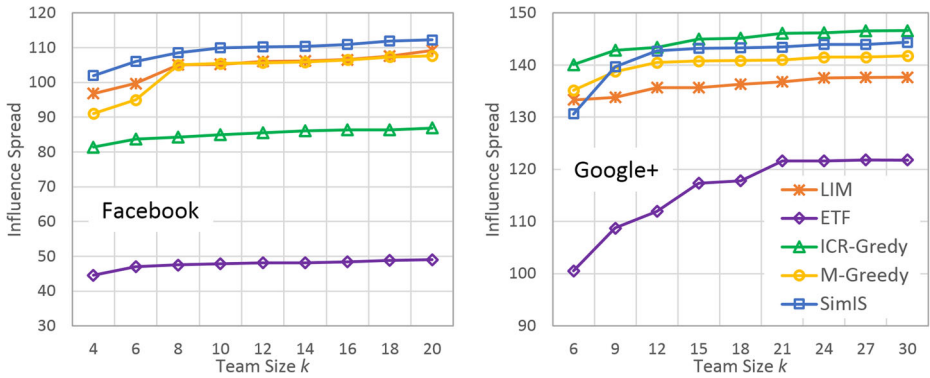**Figure 4** Run time of different algorithms

Figure 5 The influence spread σ for different methods

6 users for Facebook data and additional 12 users for Google+ on average. Such results are reasonable considering some efforts are created to reach near-by (but not directly-connected) influential users. Sometimes the cardinality values of LIM and ETF are low as well since they directly maximizing σ and minimizing c respectively. ICR-Greedy that leads to the worst cardinality exhibits that directly maximizing ICR will include too many inter-mediators.

Note that since our goal is to select a set of nodes that can maximize ICR, the ICR-Greedy is natural to be developed since it iteratively chooses the next node that can maximize the margin gain of ICR in a greedy manner. This ICR-Greedy is based on the idea of the greedy algorithm in solving the Influence Maximization problem. However, directly maximizing ICR will destroy the connectivity of the team members in the social network. That says, the communication of team members need irrelevant nodes as the intermediators. Since ICR consists of maximizing the influence spread and minimizing the communication cost among team members, an alternative greedy solution (i.e., M-Greedy) is to separately maximize the influence and minimize communication cost in an interweaving manner. Nevertheless, the experimental results show that the greedy algorithms are quite inefficient in running time, especially for the ICR-Greedy, as shown in Figure 4. The ICR-Greedy also works worse in ICR scores, as shown in Figure 3. Therefore, we further develop a heuristic algorithm,
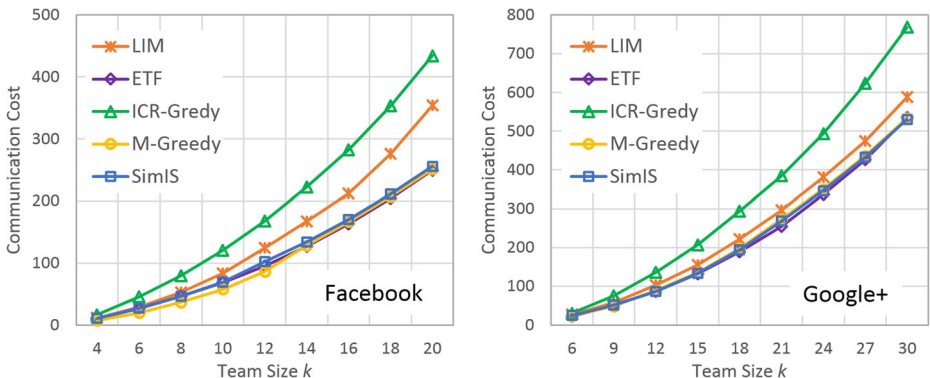

Figure 6 The communication cost c for different methods
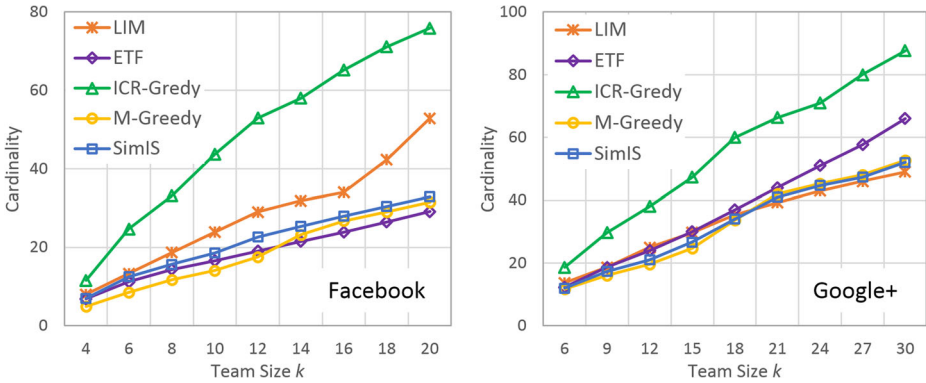
Figure 7 The cardinality for different methods

SimIS, to obtain the team members that can simultaneously possess the highest ICR scores and be derived with very high time efficiency. The results in Figures 3 and 4 truly validate the effectiveness and the efficiency of SimIS.

6 Evaluation with real event participation data

We further use real-world event-participation data to evaluate the performance of the proposed methods via prediction, in addition to examine the Influence-Cost Ratio in the previous section. The main question we aim to answer is whether or not the proposed methods can accurately predict the event organizers and participants, especially for the influential events. In other words, we expect those users who organize an event will be discovered based on the tags/labels of the event. Moreover, we also investigate whether the event participants can be identified by considering the members of the formed influential team to spread the event invitations.

6.1 Meetup data

We use the real event participation data in the social service Meetup for the evaluation. The Meetup data was collected during July 2013 to Oct 2013. Each user in Meetup can belong to multiple "online social groups". Users in Meetup is allowed to organize "offline social events" by specifying a set of tags that describe the topic of an event, and the event's geographical location (with latitude and longitude). Then the event organizers can launch events in online groups to invite and attract users in the same group or other groups to participate. In addition, users can also participate offline social events via the RSVP function (i.e., "yes", "no", or "maybe").

Since the Meetup service does not allow users to specify their social connections, we compile the social network based on online social groups that users join. Each user is considered as a node in the social network. We assume that two users have higher potential to have the social connection if they co-join more online social groups and their joined groups are smaller (since larger groups may lower down the possibility of acquaintance for two users). Let u and v be two users in the social network, and grp_i denote an online group with

$|grp_i|$ members. We take a similar approach as in [18] to measure the *connection strength* $cs(u, v)$, as formulated in the following:

$$cs(u, v) = \sum_{\forall grp_i, u \in grp_i \land v \in grp_i} \frac{1}{|grp_i|}. \tag{3}$$

We use a threshold parameter ρ to filter out those edge pairs with lower connection strength values. We consider $\rho = 0.06$ to construct the social network, and use such network to produce the experimental results. Note that the experiments are conducted using different ρ values, and the results exhibit similar trends. Here we report the result of $\rho = 0.06$.

We extract two data subsets in the Meetup data, which corresponds to two cities, Chicago (CHI) and San Francisco (SF). By considering the largest connected components as the final social networks and retrieving the corresponding offline events, the data statistics is provided in Table 2. Note that since we are concentrating on influential events, we consider only events whose number of participants is more than 100.

6.2 Evaluation settings

Our goal is to understand whether the proposed methods can truly find those users (i.e., event organizers) who host the influential events, and discover users who participates in the event. Therefore, the evaluation is designed to consist of two parts. The first is *Organizer Finding*, which aims at employing the proposed methods to find the influential team, and treating the team members as the predicted organizers, which will be examined using the *ground-truth event organizers*. The second is *Participant Forecasting* using the predicted event organizers to spread the event invitations, and investigating how many *ground-truth event participants* can be reported among the set of activated users. However, the Meetup data provides only which users are the participants of an event, and does not have the organizers of an event. To have the ground truth, therefore, we treat the earliest 10% participants of an event as the ground-truth organizers. Other 90% participants of an event are considered as ground-truth participants.

There are two evaluation metrics correspondingly. For Organizer Finding, by varying the team size $k = 10, 20, ..., 50$, we compute the *Organizer Hit Rate* (OHR) for the predicted organizer set S_k of size k. Let $\bar{S}$ be the set of ground-truth organizers of the corresponding event, OHR is defined as:

$$OHR(S_k) = \frac{hit(S_k, \bar{S})}{k}, \tag{4}$$

where $hit(S_k, \bar{S}) = |S_k \cap \bar{S}|$. OHR gets higher if more ground-truth organizers detected. In addition to OHR, we also wonder the social closeness between the predicted organizers and the ground-truth organizers. If a method can predict organizers very close to the ground truth, it can be considered as an effective method relatively. Therefore, we report the closeness scores between the predicted organizers S_k and the ground-truth organizers $\bar{S}$. Specifically, for each predicted organizer $s \in S_k$, we define its *Organizing Closeness*

Table 2 Statistics of the extracted dataset of Chicago (CHI) and San Francisco (SF)

	#nodes	#edges	#events	#average_tags
CHI	2,864	58,941	948	25.35
SF	3,302	96,552	955	23.29

(OClose) toward the nearest ground-truth organizer in the compiled social network, given by:

$$OClose(S_k) = \frac{\sum_{s \in S_k} dist(s, \bar{S})}{|S_k|}, \tag{5}$$

where $dist(s, \bar{S})$ is the length of the shortest path between the predicted organizer s and the set $\bar{S}$ of ground-truth organizers in the social network, in which edge weights are the connection strength values defined in (3). OClose gets lower if a method can predict organizers that are socially close to the ground truth.

For Participant Forecasting, we first need to obtain the set of forecasted participants based on a derived team S_k. We take advantage of the Independent Cascade (IC) model [8] to determine which users are considered as the forecasted participants. Specifically, by considering the predicted event organizers as the seeds, we run the IC model to simulate the propagation of event invitations starting from the seeds in the social network. Users successfully activated are regarded as those who accept the invitation to be the participants. Given the set S_k of predicted organizers of an event, the IC model starting from S_k will be performed up to 1,000 times, and nodes who are successfully actively up to 50 times will be considered as the forecasted participants of the event, denoted by $A(S_k)$. We measure the performance by computing the *Precision* and *Recall* of the forecasted participants. Let $\bar{A}$ be the set of ground-truth participants of the corresponding event, the measures of Prevision and Recall are defined as:

$$Precision(S_k) = \frac{hit(A(S_k), \bar{A})}{|A(S_k)|}, \tag{6}$$

$$Recall(S_k) = \frac{hit(A(S_k), \bar{A})}{|\bar{A}|}, \tag{7}$$

where $hit(A(S_k), \bar{A}) = |A(S_k) \cap \bar{A}|$, $|A(S_k)|$ is the number of activated users (i.e., forecasted participants), and $|\bar{A}|$ is the number of ground-truth participants. Note that in the execution of the IC model, we normalize the connection strength values $cs(u, v)$ between nodes to be within $[0, 1]$, and consider the normalized connection strength values as the propagation probabilities on edges. Furthermore, while Precision is to estimate the accuracy of forecasted participants in an overall perspective, sometimes the event organizers care more about who will participate in the events among their friends or follows. That says, in reality, the event organizers tend to invite their friends or followers, who are the immediate neighbors in the social network, to attend the events. Hence, to better facilitate event organization, a method should be able to successfully predict the participants among the immediate neighbors of the organizers. In order to obtain the accuracy of forecasted participants among immediate neighbors, we slightly modify Precision to obtain the *Neighboring Precision* ($nbr - Precision$):

$$nbr.Precision(S_k) = \frac{hit(A_1(S_k), \bar{A})}{|A_1(S_k)|}, \tag{8}$$

where $A_1(S_k) \subseteq A(S_k)$ is the subset of forecasted participants who are the immediate neighbors of organizers S_k in the social network. Higher $nbr.Precision(S_k)$ indicates better performance in forecasting immediate participants.

We aim to compare the performance of the methods SimIS, M-Greedy, ETF, and LIM, which can be treated as unsupervised learning approaches to predict event organizers and participants. The tags of the given event as the required labels to execute the methods. While SimIS derives the most promising results in the previous experiment, we would like to know
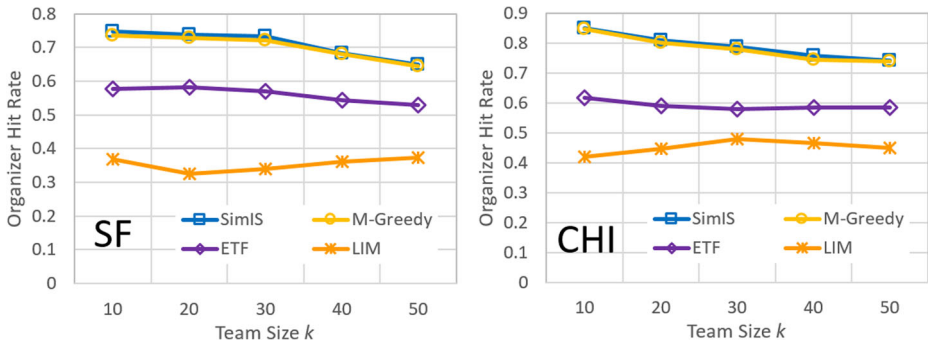
Figure 8 Organizer hit rate scores for organizer finding

whether it can also perform well on the evaluation using real event participation data. Note that the ICR-Greedy is not used for comparison here. The reason is that the time efficiency is very impractical, as indicated in Figure 4, to make the prediction.

6.3 Results and discussion

The performance for organizer finding is shown in Figures 8 and 9. We can find that the most effective methods are SimIS and M-Greedy, whose scores of Organizer Hit Rate are very close and significantly outperform those of ETF and LIM, and their OClose values are much lower than ETF and LIM. Such results lead to four implications. First, due to the satisfying hit rate scores (i.e., when team size $k \leq 30$, OHR is at least 0.7 for San Francisco and at least 0.8 for Chicago), the proposed Influential Team Formation is validated to be effective to capture the real-world event organization based on a set of specified labels that describe the topics of the event. Second, the proposed methods SimIS and M-Greedy are proven to possess sufficient predictability of real event organizers. Hence, the teams formed by SimIS and M-Greedy can be responsible for organizing influential team in practice. Third, owing to the worse performance of ETF and LIM, the discovered groups of users using
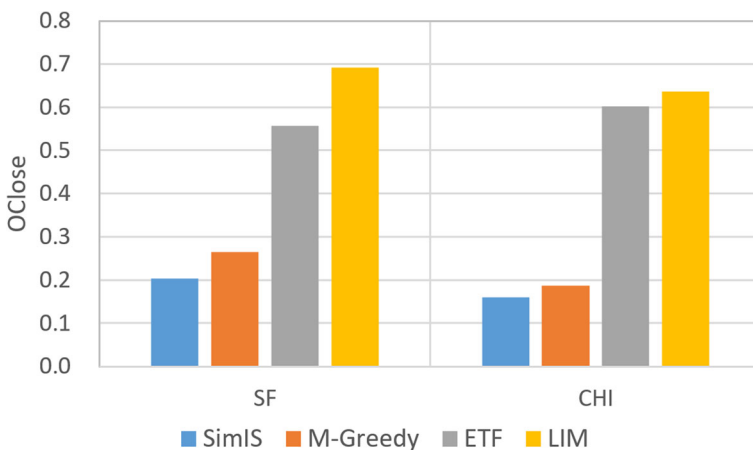

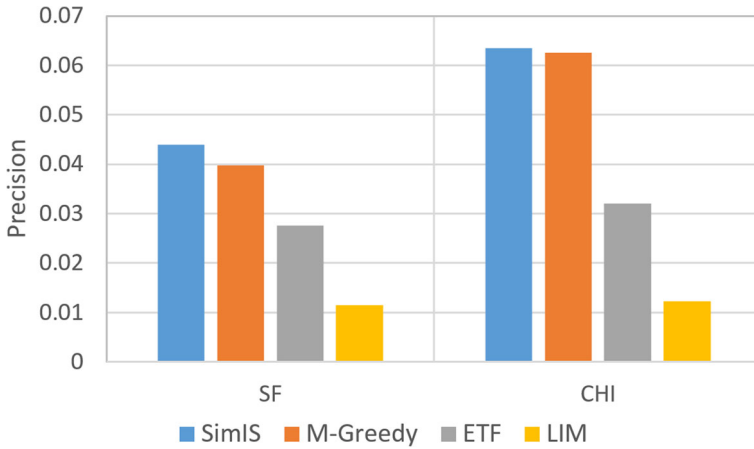**Figure 9** Organizing closeness scores for organizer finding

Figure 10 Precision scores for participant forecasting

either Influence Maximization [15] or Team Formation [9] fail to be considered as teams of organizers for influential events. Fourth, for those organizers cannot be found by these four methods, based on Figure 9, SimS tends to find users who are very close to the ground-truth organizers in the social network. This result indicates that even if the real organizers cannot be found (e.g. around 30% cannot be found according to Figure 8), organizers recommended by SimS can be also good-quality ones since they have strong relationships with the real organizers, and thus the messages of event invitation have higher potential to be spread through the real organizers.

We present the results for participant forecasting in Figures 10, 11, and 12, which correspond to Precision, nbr.Precision, and Recall respectively. It can be observed that SimIS generally outperforms the other three methods. Such results reflect that SimIS is able to not only effectively find influential event organizers, but also possess *relatively* high effectiveness on forecasting the event participants according to the found organizers. Nevertheless,
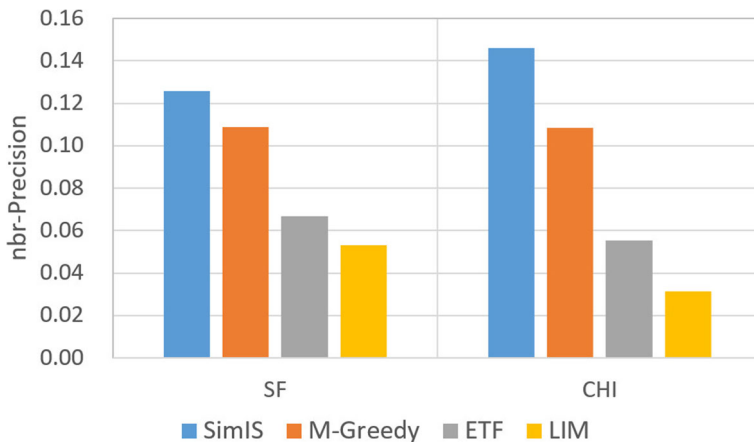

Figure 11 Neighboring (nbr) precision scores for participant forecasting
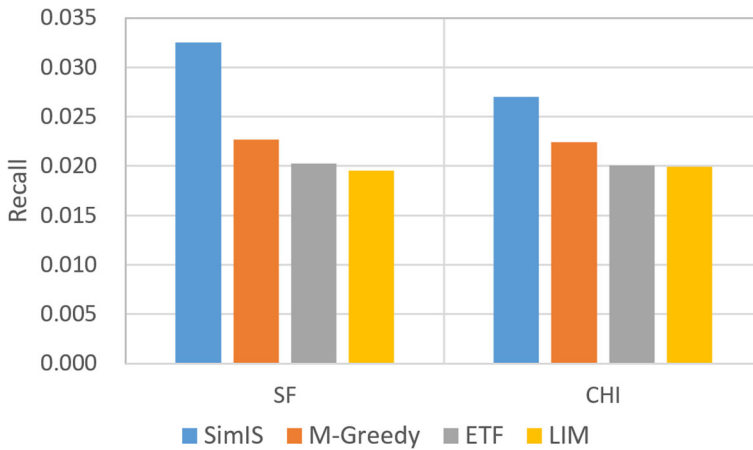
Figure 12 Recall scores for participant forecasting

it is apparent that both Precision and Recall scores are quite low, i.e., below 0.1 for Precision and below 0.05 for Recall. In fact, the low scores make the comparison insignificant and indicate very unsatisfying predictability. We think that such worse performance may result from the mechanism of spreading the influence/invitations starting from the event organizers. Recall that while developing an precise simulation model for the propagation of invitations is not the goal of this paper, we utilize the Independent Cascade model. Low precision and recall scores imply the IC model is unsuitable for the spreading of event invitations. Therefore, we can only conclude that the proposed SimIS is relatively useful under the IC model. We leave the development of an effective invitation propagation model and its evaluation for participant forecasting using SimIS in the future work. Nevertheless, we can also find that the immediate participants can be relatively better forecasted (i.e., the scores are higher than 0.12), as reported by nbr.Precision in Figure 11. Such results make sense because organizers usually invite their friends or followers to attend events, and people tend to accept the invitation of events hosted by their friends. Thus the proposed SimIS would be more applicable for predicting which friends or followers will participate in the events hosted by found organizers.

7 Conclusion and discussion

This paper proposes a novel Influential Team Formation (ITF) problem for organizing influential events on social networks. The objective of ITF is to find a team of users satisfying required labels and maximizing the influence spread per communication cost among team members. Two methods, M-Greedy and SimIS Heuristic, are developed and validated to be promising for solving the ITF problem in terms of effectiveness and efficiency. M-Greedy is suitable for organizing smaller events while SimIS is good for larger events. In addition, by considering SimIS and M-Greedy as two unsupervised learning approaches, we find they can lead to promising predictability on finding organizers for influential events using real event participation data in Meetup. Extensive analysis demonstrate the unsatisfying performance on forecasting participants, which encourages us to come up with an effective invitation propagation mechanism based on SimIS in the future.

Although the proposed techniques can effectively find the team members in both simulation-based and data-based experiments, there are still some limitations in our work. First, the influential team formation problem assumes that the label set of each individual is known in prior (e.g. being obtained from the user profiles). However, users in online social media tend to provide incomplete information, which might destroy the quality of the discovered team members. Second, the influence probabilities associated on edges of the social network is pre-determined. Unreal determination of influence probabilities could affect the real performance of the formed teams. How to automatically learn the influence probabilities from the action logs of information diffusion (e.g. retweets and replies) is a demanded issue. Third, we suppose the set of identified team members will perfectly join the event organization. But it is not the real case; that says, many of them may be unwilling or unavailable. A practical team formation should model the willingness of each individuals into the process of event organization. This can be a promising future direction.

By formulating and solving the proposed ITF problem with the effective algorithm SimIS, three potential extensions can be built. First, SimIS is able to not only find the team members, but also report the scores of both influence-cost ratio and influence spread. Hence, one can estimate the success of the formed team by the expected number of event partici-pants, or make a control on the event scale. Second, SimIS algorithm is easy to be extended based on the user-input team size. Therefore, we are allowed to find the alternative if any individual is not available or unwilling to join the team. Third, given that the real-world campaigns are opinion-aware and competitive, our ongoing work is to form the influential teams with the consideration of multi-party competitors, in which each party supports for a particular opinion.

References

1. Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A., Leonardi, S.: Online team formation in social networks. In: Proceedings of ACM International Conference on World Wide Web, WWW '12, pp. 839–848 (2012)
2. Basiri, J., Taghiyareh, F., Ghorbani, A.: Collaborative team formation using brain drain optimization: a practical and effective solution. World Wide Web Journal (2017)
3. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 199–208 (2009)
4. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, pp. 1029–1038 (2010)
5. Cheng, S., Shen, H., Huang, J., Zhang, G., Cheng, X.: Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In: Proceedings of ACM International Conference on Information & Knowledge Management, CIKM '13, pp. 509–518 (2013)
6. Cheng, S., Shen, H., Huang, J., Chen, W., Cheng, X.: Imrank: influence maximization via finding self-consistent ranking. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, SIGIR '14, pp. 475–484 (2014)
7. Kargar, M., An, A.: Discovering top-k teams of experts with/without a leader in social networks. In: Proceedings of ACM International Conference on Information and Knowledge Management, CIKM '11, pp. 985–994 (2011)
8. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, pp. 137–146 (2003)
9. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 467–476 (2009)

10. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective out-break detection in networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, pp. 420–429 (2007)

11. Li, C.-T., Shan, M.-K.: Composing activity groups in social networks. In: Proceedings of ACM International Conference on Information and Knowledge Management, CIKM '12, pp. 2375–2378 (2012)

12. Li, C.-T., Shan, M.-K., Lin, S.-D.: On team formation with expertise query in collaborative social networks. Knowl. Inf. Syst. **42**(2), 441–463 (2015)

13. Li, K., Lu, W., Bhagat, S., Lakshmanan, L.V., Yu, C.: On social event organization. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pp. 1206–1215 (2014)

14. Li, L., Tong, H., Cao, N., Ehrlich, K., Lin, Y.-R., Buchler, N.: Replacing the irreplaceable: fast algorithms for team member recommendation. In: Proceedings of ACM International Conference on World Wide Web, WWW '15, pp. 636–646 (2015)

15. Liu, Q., Xiang, B., Chen, E., Xiong, H., Tang, F., Yu, J.X.: Influence maximization over large-scale social networks: a bounded linear approach. In: Proceedings of ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, pp. 171–180 (2014)

16. Liu, X., Li, S., Liao, X., Peng, S., Wang, L., Kong, Z.: Know by a handful the whole sack: efficient sampling for top-k influential user identification in large graphs. World Wide Web Journal **17**(4), 627–647 (2014)

17. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximation for maximizing submodular set functions. Math. Program. **14**(1), 265–294 (1978)

18. Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys. Rev. E. **64**, 1–7 (2001)

19. Ohsaka, N., Akiba, T., Yoshida, Y., Kawarabayashi, K.-I.: Fast and accurate influence maximization on large networks with pruned Monte-Carlo simulations. In: Proceedings of AAAI Conference on Artificial Intelligence, AAAI '14, pp. 138–144 (2014)

20. Rangapuram, S.S., Bühler, T., Hein, M.: Towards realistic team formation in social networks based on densest subgraphs. In: Proceedings of ACM International Conference on World Wide Web, WWW '13, pp. 1077–1088 (2013)

21. Shen, C.-Y., Yang, D.-N., Lee, W.-C., Chen, M.-S.: Maximizing friend-making likelihood for social activity organization. In: Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD '15, pp. 3–15 (2015)

22. Sozio, M., Gionis, A.: The community-search problem and how to plan a successful cocktail party. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, pp. 939–948 (2010)

23. Tong, Y., She, J., Meng, R.: Bottleneck-aware arrangement over event-based social networks: the max-min approach. World Wide Web Journal **19**(6), 1151–1177 (2016)

24. Xiang, B., Liu, Q., Chen, E., Xiong, H., Zheng, Y., Yang, Y.: Pagerank with priors: an influence prop-agation perspective. In: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI '13, pp. 2740–2746 (2013)

25. Yang, D.-N., Shen, C.-Y., Lee, W.-C., Chen, M.-S.: On socio-spatial group query for location-based social networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, pp. 949–957 (2012)

26. Yu, Z., Zhang, D., Yu, Z., Yang, D.: Participant selection for offline event marketing leveraging location-based social networks. IEEE Trans. Syst. Man Cybern. Syst. **45**(6), 853–864 (2015)

27. Zhang, E., Wang, G., Gao, K., Yu, G.: Finding critical blocks of information diffusion in social networks. World Wide Web Journal **18**(3), 731–747 (2015)