Knowl Inf Syst
DOI 10.1007/s10115-013-0695-x

REGULAR PAPER

On team formation with expertise query in collaborative social networks

Cheng-Te Li · Man-Kwan Shan · Shou-De Lin

Received: 30 April 2012 / Revised: 18 September 2013 / Accepted: 25 September 2013

Abstract Given a collaborative social network and a task consisting of a set of required skills, the team formation problem aims at finding a team of experts who not only satisfies the requirements of the given task but also is able to communicate with one another in an effective manner. This paper extends the original team formation problem to a generalized version, in which the number of experts selected for each required skill is also specified. The constructed teams need to contain adequate number of experts for each required skill. We develop two approaches to compose teams for the proposed *generalized team formation tasks*. First, we consider the specific number of experts to devise the generalized Enhanced-Steiner algorithm. Second, we present a grouping-based method condensing the expertise information to a compact representation, *group graph*, based on the required skills. Group graph can not only reduce the search space but also eliminate redundant communication cost and filter out irrelevant individuals when compiling team members. To further improve the effectiveness of the composed teams, we propose a *density-based* measure and embed it into the developed methods. Experimental results on the DBLP network show that the teams composed by the proposed methods have better performance in both effectiveness and efficiency.

Keywords Team formation · Social network · Expertise query · Collaborative networks

C.-T. Li (✉) · S.-D. Lin
Graduate Institute of Networking and Multimedia,
National Taiwan University, Taipei City, Taiwan
e-mail: d98944005@csie.ntu.edu.tw

S.-D. Lin
e-mail: sdlin@csie.ntu.edu.tw

M.-K. Shan
Department of Computer Science, National Chengchi University,
Taipei City, Taiwan
e-mail: mkshan@nccu.edu.tw

Published online: 06 October 2013

1 Introduction

Team formation is an important research topic in the area of organization theory [4,5,9,10]. A successful project not only relies on the expertise of the participated members but also hinges on the effectiveness of communication and collaboration among them. In other words, to form an effective team of experts for a given task or project, it is critical to ensure the team members possess the professional skills that satisfy the required expertise and have excellent communication manners to effectively work together.

Given a *collaborative social network*, the formation of a team aims to find a crew of experts for a given *task* requiring a set of specific *skills*. A collaborative social network consists of a pool of *candidates*, in which each candidate is an expert possessing some skills, with their existing collaboration relationships. In addition, there is a weight assigned for each collaboration link to indicate the *communication cost* between the connected experts according to their previous collaborations. A lower weight indicates easier, smoother and more effective collaboration between such two experts. As a result, the *team formation problem* aims at finding a set of experts from these candidates to meet the requirement of a given task and to minimize the total communication cost between the team members.

For example, assume that a project leader aims to organize a team from a pool of six candidates, $P = \{1, 2, 3, 4, 5, 6\}$, for a given task requiring four distinct skills, $R = \{s_1, s_2, s_3, s_4\}$. Each candidate i is an expert of a set of skills X_i where $X_1 = \{s_4\}$, $X_2 = \{s_1\}$, $X_3 = \{s_1, s_3\}$, $X_4 = \{s_1, s_4\}$, $X_5 = \{s_2\}$, and $X_6 = \{s_3\}$. Also assume that there exists a social network connecting these experts, as shown in Fig. 1a. In Fig. 1a, bold lines indicate previous collaborations between the candidates while the edge weights stand for the communication costs between them. To form a team that meets the given task without considering the communication cost, we find four teams that satisfy the expertise requirement: $T_1 = \{1, 2, 3, 5\}$, $T_2 = \{1, 2, 5, 6\}$, $T_3 = \{3, 4, 5\}$, and $T_4 = \{4, 5, 6\}$. However, if the communication cost is considered, $T_4 = \{4, 5, 6\}$ becomes the best choice since the communication cost among its members is the smallest. The communication cost of T_4 is 0.3 and that of T_3 is 0.6, while both T_1 and T_2 are considered as invalid because the social network of the team members consists of disconnected components; meaning there is no previous collaboration between the team members.

The team formation problem is first proposed by Lappas et al. [14], who attempted to exploit the collaboration social network with the communication costs between the members to organize a team for a given task. However, they failed to meet the requirement of having a specific number of experts for each skill required in the task. In the rest of the paper, we

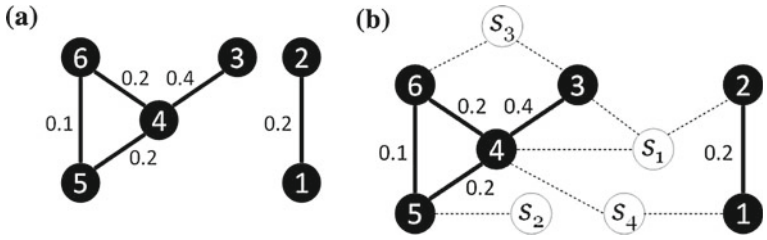


Fig. 1 **a** A collaborative social network. **b** An enhanced graph, where nodes with *black color* are experts and those with *white color* are skills

treat Lappas et al.'s setting as the *basic task*. In real-life situations, it is likely that more than one expert is demanded for some skills when composing a team for a specific task.

To meet the real-life requirement, we propose to generalize the team formation problem by allowing specification of the minimum required number of experts for each skill. Specifically, the team formed must satisfy the following conditions. (1) Its members possess all the required skills in the expertise query. (2) For each required skill, the team contains at least the specified number of experts. (3) The total communication cost between team members should be as low as possible. We regard such task as the *generalized team formation*. For example, in Fig. 1a, if a given task requires forming a team with **two** experts with skill s_1, one expert with skill s_3 and one expert with skill s_4, $T_5 = \{3, 4\}$ is the best team with the lowest communication cost of 0.4.

Lappas et al. [14] have proved that the team formation problem is NP-hard. By defining the communication cost as the sum of edge weights in the minimum spanning tree that connects the chosen experts, they propose two approximation methods, *Cover-Steiner* and *Enhanced-Steiner* algorithms, to solve the team formation problem for basic tasks.

We propose two novel methods to solve the generalized team formation tasks based on the existing Enhanced-Steiner algorithm. First, we modify the Enhanced-Steiner algorithm to deal with the generalized tasks. In addition, rather than selecting a seed node randomly as in the original Enhanced-Steiner algorithm, we devise a *density-based seed selection* strategy by considering the potential interactions between experts with the required skills and embed it into the Steiner algorithm. Second, we propose a novel grouping-based method to compose the team for generalized tasks. Our grouping-based method condenses experts in the collaborative social network to a *group graph* structure based on the required skills. To satisfy the required skills with specified numbers, we develop a *role composition* algorithm to extract the final collaborative subgraph of the team by connecting experts based on the group roles of individuals (i.e., within and between groups). The proposed methods are evaluated on five effectiveness measures and the time efficiency.

The remainder of this paper is organized as follows: In Sect. 2, we have a full review and summary about relevant literatures. The problem definition and notations are described in Sect. 3. Section 4 describes the generalized Enhanced-Steiner algorithm. Then, we propose the density-based measure to improve the generalized Enhanced-Steiner algorithm in Sect. 5. In Sect. 6, we propose the group-based team formation method. Section 7 exhibits experimental results, while Sect. 8 concludes this paper.

2 Related work

Existing works related to this paper can be divided into three categories: team formation, social group planning, and connection subgraph discovery.

2.1 Team formation

The team formation problem is extensively studied in the field of operations research. Wi et al. [26] solve the team formation problem by modeling it as an integer programming problem to find an optimal match between individuals and requirements. Fitzpatrick et al. [9] evaluate individuals' drive and temperament to assess the quality of a team. Chen and Lin [5] estimate the interpersonal attributes of experts from the psychological view when arranging a team. The above three studies aim to find a better match between the required skills and experts. However, the collaboration between experts is neglected. Gaston et al. [10] study

the potential correlation between the expertise network structures and the team performance. However, they neither take it as a computational problem nor propose a method to compose a team. Cheatham and Cleereman [4] simply collects the neighboring individuals surrounding each skill in a social-concept graph to form the team. Agustín-Blas et al. [1] aim to partition a staff-resource matrix such that the staff members in each team/group have maximum knowledge about the resources in the corresponding team/group. However, these works do not consider the social relationships and the communication cost between individuals.

Lappas et al. [14] is the first to solve the team formation problem by combining both the social network and the communication cost between experts. They propose two approximation methods, *Cover-Steiner* and *Enhanced-Steiner* algorithms, to solve the basic task team formation problem. Experimental results showed that Enhanced-Steiner outperforms Cover-Steiner. Lappas' work opens up the opportunity for data miners to investigate how to form effective teams of experts with a variety of user requirements. Anagnostopoulos et al. [2] design a fitness function to assign tasks to experts when forming the teams. In particular, they consider the balance of workload of experts to have the fair task assignment. Their follow-up work [3] extends such framework by integrating the social collaborations between experts and allows tackling multiple online tasks concurrently. Kargar and An [11] propose to find top-k teams of experts with the specification of team leaders. Majumder et al. [18] impose the capacity constraints which ensure that no experts are assigned tasks beyond his/her capacity values in the team formation problem. Sorkhi et al. [20] compose teams of experts by minimizing the communication cost with the consideration that different experts have diverse levels of skillfulness. Kargar et al. [12] assume each expert should be associated with a monetary weight to represent his/her personal cost for professional networking. They present the bi-objective team formation by minimizing both the communication cost and the personal cost of the project. We give a summary about social network-based team formation, as shown in Table 1. Some abbreviations of aspects are denoted: communication cost (CC), number of skilled experts (NE), capacity on experts (CE), balance of workload (BW), team leader (TL), skillfulness level of experts (SL), online multiple tasks (OT), and personal cost (PC). A marked cell indicates that the paper tackles the corresponding aspect.

Since one of our proposed methods extends from the *Enhanced-Steiner algorithm* [14], we describe more about its details below. The Enhanced-Steiner algorithm consists of two steps. The first step constructs an enhanced graph which enhances the collaborative social network by adding skill nodes and edges connecting each skill node to individuals who possess such skill. An example is shown in Fig. 1b. The second step aims to find a Steiner tree that densely

Table 1 Summarizing the differences between this paper and the recent advances about social network-based team formation in eight aspects

	CC	NE	CE	BW	TL	SL	OT	PC
Lappas et al. [14]	■							
Anagnostopoulos et al. [2]				■				
Kargar and An [11]	■				■			
Anagnostopoulos et al. [3]	■			■			■	
Majumder et al. [18]	■		■					
Sorkhi et al. [20]	■					■		
Kargar et al. [12]	■							■
This paper	■	■						

connects the required skills in the enhanced graph. Given a graph $G = (V, E)$, a required set of vertices $R \subseteq V$, a Steiner tree is a connected and acyclic subgraph of G which spans all vertices of R with the minimum cost. To find a Steiner tree, many algorithms are proposed. Lappas et al. [14] present a greedy heuristic algorithm shown in Algorithm 0. The algorithm starts by selecting a skill node randomly from the enhanced graph (line 2). Then, each round of the algorithm finds the skill node possessing the minimum distance to the set of nodes which have added to the solution (line 4). All the nodes along the shortest path from this skill node to the nodes in current solution are added to the new solution set as well (line 5 & 6).

Algorithm 0. Enhanced-Steiner for basic tasks.

Input: G=(V,E), V=$\{1,...,n\}$;
the skill sets $\{X_1, ..., X_n\}$ of individuals a task R=$\{s_1,...,s_q\}$.
Output: Team $V' \subseteq V$ and its induced subgraph $G[V']$.
1: H=(V_H, E_H) ← *EnhancedGraph*(G, R).
2: V' ← v, where v is a random node from R.
3: **while** $(R \setminus V') \neq \phi$ **do**
4: v^* ← argmin $_{u \in R \setminus V'}$ *dist*(u, V') in H.
5: **if** *Path* (v^*, V') $\neq \phi$ **then**
6: V' ← $V' \cup$ {*Path* (v^*, V')}.
7: V' ← $V' \setminus \{s_1, ..., s_q\}$.

2.2 Social group planning

Social group planning, whose goal is similar to team formation, aims at recommending a set of individuals who satisfy various requirements for a real-world activity or event. Sozio and Gionis [21] are the first to point out such problem in the context of social network mining. They define and solve a *Community-Search Problem*, which aims to find a group of individuals densely connected to a given set of persons. Assuming that each individual has a list of available time slots [27], propose *Social-Temporal Group Query* to find the most suitable activity time and the attendees with the minimum total social distance to the initiator. Considering that each individual is associated with a geospatial location [28], further propose *Socio-Spatial Group Query* to select a group of nearby attendees with tight social relation. Based on similar settings [17], propose *Circle of Friend Query* to find a set of friends who are close to each other in both spatial and social aspects. On the other hand, assuming that each person is associated with a set of attributed labels (e.g. name, interests, age, sex, school) [15], present *Context-based People Search* to identify who the users would like to find according to the given context labels. Li and Shan [16] further develop the *Activity Composter* system to facilitate users for initializing, inviting, and suggesting suitable friends to attend different kinds of social events or activities, such as cocktail party, study group, and group buying.

2.3 Connection subgraph discovery

Given a set of nodes, the *connection subgraph discovery* problem is to find a subgraph with the best connections between the query nodes. Its objective is similar to the team formation problem except for that each node is not associated with a set of skills. Faloutsos et al. [8] are the first to find the connection subgraph for a pair of nodes. The most well-known approach is the *Random Walk with Restart* (*RWR*) [22], which measures the proximity between nodes. Work well with a variety of input requirements, including allowing AND/OR constraints [23],

providing interactive feedback with users [24], querying a small graph describing the desired relationships between entity types [25], the RWR approach is considered a very effective approach to extract the diverse kinds of best connection subgraphs. More recently, a Steiner tree-based approximation algorithm, STAR [13], is proposed to find the connection subgraph in multi-relational graphs. Cheng et al. [7] consider the community structure with the *modularity* measure to discover the connection subgraphs. They also propose a correlation index to find the *groups* that the query nodes belong to, as well as the best connection structure among groups [6].

3 Problem definition

Definition 3.1 Let $A = \{a_1, \ldots, a_m\}$ be a universe of m skills, a *collaborative social network* is an undirected and weighted graph $G = (V, E)$ with each node i in $V = \{1, \ldots, n\}$ being an individual who possesses a set of skills $X_i \subseteq A$ and each edge (i, j) in E representing the collaboration relationship between two individuals. The weight assigned on each edge (i, j) stands for the communication cost between individuals i and j.

Note that edges with lower weight values represent better collaborations between two individuals and vice versa. For example, in the coauthorship network, if two researchers coauthor more papers together, the weight on the edge connecting the two authors would be lower assuming they will work more efficiently together than two authors who have never worked together before.

Definition 3.2 A *generalized task* $R = (S, K)$ consists of a set of required skills, $\{(s_i, k_i)|\forall i, 1 \leq i \leq q, s_i \in A, k_i \text{ is an integer}\}$, where k_i denotes the minimum required number of experts for skill s_i and q is the number of required skills. Note that if $k_i = 1(\forall\ i, 1 \leq i \leq q)$, the expertise query is reduced to the *basic task*.

Definition 3.3 Given a collaborative network $G = (V, E)$ and $V' \subseteq V$, the *communication cost* of V' is defined as the sum of edge weights in the *minimum spanning tree* of the induced subgraph $G[V']$, denoted by $CC(V')$.

Note that our definition of communication cost follows the approach of [14]. In fact, there are various ways to define the communication cost based on the pairwise distances between team members in the social graph. For example, the *diameter cost* [18] is defined by the largest shortest distance between any two nodes in the discovered subgraph. The *sum-of-distances* cost [11] is defined by the sum of the shortest distances between the experts possessing the pair of skills. In practice, the effect of using the pairwise distance-based measures on team formation is verified to exhibit similar tendency and results, as evidenced by Kargar and An [11]. What we choose (i.e., the Steiner cost) is the one that commonly used by all the team formation works [2,3,11,14,18]. Besides, it can be observed that according to such pairwise distances, assuming that the cost between node i and j is w_{ij} and the cost between j and k is w_{jk}, the cost between i and k equals to one of the following three cases: (a) $\text{cost}_{ik} = w_{ij} + w_{jk}$, if there is no direct link between i and k, (b) $\text{cost}_{ik} = w_{ik} < w_{ij} + w_{jk}$, if the direct link between i and k possesses the lowest cost, and (c) $\text{cost}_{ik} = w_{ij} + w_{jk} < w_{ik}$, if there exists a direct link between i and k but the corresponding weight w_{ik} is not the shortest distance between i and k. In case (c), it is reasonable because the collaboration or communication between i and k would be better through the coordination of node j.

Problem Definition Given a collaborative social network $G = (V, E)$ and a generalized task $R = (S, K)$, the team formation problem for the generalized task is to find a set of individuals $V' \subseteq V$ which forms an induced subgraph $G[V']$ such that

(1) $\forall(s_i, k_i) \in R, s_i \subseteq \bigcup_{j \in V'} X_j$,
(2) $\forall(s_i, k_i) \in R, k_i \leq \left|\{j | j \in V' \text{ and } s_i \in X_j\}\right|$, and
(3) The communication cost $CC(V')$ is minimized.

Theorem 1 *The generalized team formation problem is NP-hard.*

Proof We prove the theorem by a reduction from the *Group Steiner Tree* (GST) problem [19]. In the decision version of the GST problem, we are given an undirected graph $G = (V, E)$, a cost function $c : E \rightarrow \mathbf{R}$, a constant δ, and k subsets of nodes $\{g_1, \ldots, g_k\}$, in which $g_i \subseteq V$, $i \in 1, \ldots, k$. We are asked to find a subtree $G' = (V', E')$ of $G = (V, E)$, $V' \subseteq V$, $E' \subseteq E$, such that $V' \cap g_i \neq \phi$ and the cost $\sum_{e \in E'} c(e) < \delta$.

Now we are transforming an instance of the *Group Steiner Tree* problem to an instance of our generalized team formation problem as follows. We associate each subset g_i in the GST with a skill s_j. The expertise query of the task R aims to satisfy not only the skills abut also the number of experts of each required skill, specifically, $R = \{(s_1, k_1), \ldots, (s_q, k_q)\}$. We know that if our problem is solved, it is natural that the corresponding basic team formation problem, which targets at satisfying $R = \{(s_1, 1), \ldots, (s_q, 1)\}$, is also solved. However, it is not true the other way around. That is, if the basic team formation is NP-hard, the generalized version is NP-hard as well.

For each node $v \in V$ in the GST problem, we create an expert i_v with a set of skills $X_v = \{s_i | v \in g_i\}$. The graph in the generalized team formation problem, with the number of required skills equal to 1, can be mapped into a GST problem, in which the cost function c is defined as the sum of edge weights in the generalized team formation instance of the GST problem. Then it is easy to conclude the GST has a solution if and only if a solution also exists for the generalized team formation problem. The problem is proved to be NP. □

4 Generalized Enhanced-Steiner algorithm

The Enhanced-Steiner algorithm [14] was proposed to solve the original team formation problem (i.e., the basic task). In this section, we modify this algorithm to deal with the generalized tasks. The *generalized Enhanced-Steiner algorithm* is regarded as a novel and strong baseline in the evaluation. To describe this algorithm, some definitions are given in advance.

Definition 4.1 Given two nodes $i, j \in V$, the *distance* dist(i, j) between two nodes i and j is the sum of edge weights along the shortest path. Meanwhile, path(i, j) is the sequence of nodes along the shortest path.

Definition 4.2 The *distance* between a node i and a set of nodes V' is defined as dist $(i, V') = \min_{j \in V'}$dist$(i, j)$. Likewise, path$(i, V')$ is the set of nodes along the shortest path from i to j.

Based on these two definitions, Algorithm 1 incrementally finds and adds selected team members into the solution set. In the algorithm, two sets of nodes, U and V', are maintained. U contains the skill nodes that have yet to be satisfied in terms of the number of experts needed. V' is the current solution set which contains the selected expert nodes. The algorithm repeats

several rounds until the number of experts for each required skill is sufficient (line 4). At each round, a skill node v^* from U that has the minimum distance to V' is selected (line 7). Then the number of required experts for the skill represented by node v^* is decreased by one (line 9). Moreover, for each node j along the shortest path $path(v^*, V')$, if j possesses a required skill w and is not yet added into V', the required number of experts for w is decreased by one (line 10 & 11), and all the nodes along the shortest path from v^* to V' are added to the solution set V' (line 12).

Algorithm 1. Generalized Enhanced-Steiner for generalized tasks.

Input: $G=(V,E)$, the skill sets $\{X_1, ..., X_n\}$ of individuals;
a task $R=(S, K)=\{(s_i, k_i) \mid 1 \le i \le q\}$.
Output: Team $V' \subseteq V$ and its induced subgraph $G[V']$.

1: $H=(V_H, E_H) \leftarrow EnhancedGraph(G, S)$.
2: $V' \leftarrow v$, where v is a *random* node from R.
3: $k_v = k_v - 1$, where v is a skill node.
4: **while** $U \neq \phi$ **do**
5: **for each** $s_j \in V'$ **do**
6: **if** $k_j \le 0$ **then** $U = S \setminus s_j$.
7: $v^* \leftarrow \text{argmin}_{u \in U}\, dist(u, V')$ in H
8: **if** $path\ (v^*, U) \neq \phi$ **then**
9: $k_{v^*} = k_{v^*} - 1$
10: **for each** w, $w \in X_j$, $j \in path\ (v^*, V')$ & $j \notin V'$ **do**
11: $k_w = k_w - 1$
12: $V' \leftarrow V' \cup path\ (v^*, V')$
13: $E_H = E_H \setminus \{EdgesInPath(v^*, V') \setminus E\}$
14: $V' \leftarrow V' \setminus \{s_1, ..., s_q\}$.

The expected running time of this generalized Enhanced-Steiner tree algorithm is similar to the original Enhanced-Steiner Algorithm. The time complexity of the original algorithm is $O(n_s \times |E|)$, where n_s is the number of required skill nodes. The difference between the original and generalized versions lies in the number of times the main while loop is executed (Line 4–13). In the worst case, the while loop has to be executed up to n_p times, where n_p is the total number of desired experts of all required skills, $n_p = \sum_{i=1...q} k_i$, where k_i is the required number of experts for skill s_i. Therefore, the worst-case time complexity of the generalized algorithm is $O(n_p \times |E|)$.

5 Better initial node selection strategy

Both the original Enhanced-Steiner algorithm and the proposed generalized Enhanced-Steiner algorithm start from selecting a skill node randomly from the enhanced graph. Instead of selecting a seed node randomly, in this section, we propose a more effective strategy to select the seed node based on the neighborhood structure of skill nodes. This comes from the observation that the higher the *neighborhood density* of a skill node, the better chance it is involved in a lower-cost communication subgraph, since such skill node could have shorter distances to other nodes when traversing the enhanced graph. To achieve this goal, we propose the *ε-neighborhood density* by the following definitions.

Definition 5.1 Given two nodes $v,\ w \in V$, the *length* $l(v,\ w)$ between nodes is defined as the number of edges in the shortest path between v and w.

Definition 5.2 The *ε-neighborhood* of a node $v \in V$, is defined as the set of nodes $N_\varepsilon(v) = \{w_i | 1 \leq l(v,\ w_i) \leq \varepsilon,\ w_i \in V\}$.

Definition 5.3 Given a set of nodes $U \subseteq V$, the *density* of U is defined as

$$\text{Density}(U) = \frac{2|\{e_{ij} | e_{ij} \in E, v_i, v_j \in U\}|}{|U| \times (|U - 1|)}$$

In other words, the density of U is the ratio of the number of edges between each pair of nodes in U to the maximal possible number of edges.

Definition 5.4 The *ε-neighborhood density* of a node $v \in V$ is defined as $\text{Density}(N_\varepsilon(v))$, where $N_\varepsilon(v)$ is v's ε-neighborhood.

Consequently, line 2 of Algorithm 1 can be improved by selecting a seed skill node with the highest density. Let us take Fig. 1 for example again: the 2-neighborhood densities of $s_1,\ s_2,\ s_3$, and s_4 are 0.33, 1.0, 0.67, and 0.33, respectively. Starting from s_2 ($s_3,\ s_4$), the solution is {4, 5, 6} with communication cost 0.3 (0.3, 0.4). When starting from s_1, the resulting team is {3, 4, 5} with cost 0.6.

Using the *ε-neighborhood density* to improve the effectiveness of the generalized Enhanced-Steiner algorithm, the time complexity only increases slightly to $O(|R|+n_p \times |E|)$ since we need to compute the density of each skill node before entering the while loop. Since $|R|$ is much small than $|E|$, the worst-case time complexity is approximated to be $O(n_p \times |E|)$.

6 Grouping-based approach to team formation

The generalized Enhanced-Steiner algorithm suffers from poor efficiency when the user-specified task consists of too many skills or when the collaborative social network contains too many individuals and collaboration relationships. To address such issue, we propose a grouping-based method for generalized team-formation tasks.

The central idea of the grouping-based approach is to aggregate the raw collaborative network into a compact structure, called the *group graph*, which keeps only relevant individuals and potential interactions between the groups for the required skills. Group graph is able to boost the time efficiency because it reduces the search space. Moreover, as shown in the evaluation section, the group graph is proved to be capable of guiding the graph traversals to avoid redundant communication cost and decrease the cardinality of the compiled team.

Our grouping-based method consists of four stages. The first is skill-based individual grouping, which collects individuals possessing the same required skills into groups. The second is constructing the group graph, in which linkages capture the individuals' interactions between groups. In the third stage, we apply the modified Enhanced-Steiner algorithm on the group graph, to discover the subgraph of groups which strongly connects the required skills. Finally, a *role composition* algorithm is developed to finalize the team satisfying both required skills and the corresponding specified number of experts.

6.1 Skill-based individual grouping

The first step is to group individuals in the collaborative network according to the required skills. A *group*, with respect to one of the required skills, say s_i, is a connected subgraph in

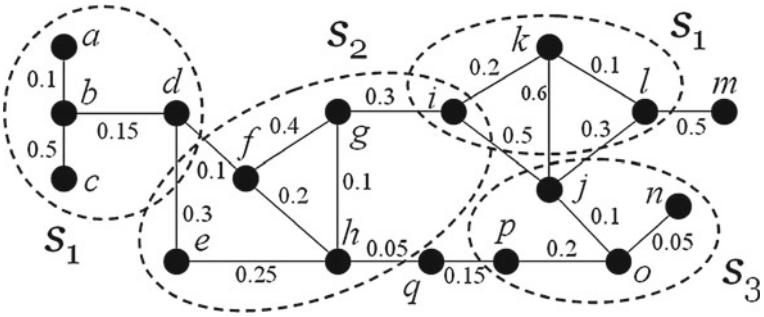


Fig. 2 An example of skill-based individual grouping

which each individual node possesses the skill s_i. Figure 2 shows an example of the skill-based grouping for the required skills s_1, s_2, and s_3. Each group is enclosed by a dotted circle. It can be observed that there are two groups correspond to skill s_1. The two groups are separated components since individuals in them have no past collaborations. Besides, node i belongs to two groups because i is skilled in both s_1 and s_2. Node m does not belong to any groups, because m contains no required skills and thus will not be considered for further processing.

In general, most of individuals in the network tend to be good at more than one skill. Therefore, generated groups overlap with each other. Such overlapping provides a potential for reducing the cardinality of the built team because we can try to find experts who satisfy multiple required skills from the overlapping areas. Moreover, grouping can improve the efficiency of forming teams. We will elaborate the details in the following.

6.2 Group graph construction

We have aggregated individuals into groups based on the required skills. The next step is to exploit the underlying interactions between groups. The group-based interactions are essential for finding effective connections between team members. As observed in Fig. 2, there are three kinds of relationships between groups. First, two groups are overlapped with node i since individual i is good at both skill s_1 and s_2. Second, the s_1 group at left connects directly with the s_2 group due to the collaboration relationships between individuals d and f as well as d and e. The similar relationship exists (a) between the s_1 group (right) and the s_3 group, and (b) between the s_2 group and the s_3 group. Third, the s_2 group and the s_3 group have indirect communication by an inter-mediator q.

To precisely reflect the communication costs between individuals in the original network, we aim at modeling the overlapped, direct-connected, and indirect-connected relationships between groups into the group graph. We associate each group interaction with a weight value to capture the communication cost between groups. Such cost not only reflects the correlation between different required skills but also guide to later group search process to form effective teams. By integrating skill-based groups, interactions between groups, and weights on group relationships, we construct a group graph to condense the information about the required skills in the collaborative social network. We formally define the group graph, group nodes, and group links below.

Definition 6.1 A *group graph* $H = (V_H, E_H)$ is a weighted graph and is constructed according to the required skills from the collaborative social network $G = (V, E)$, where

V_H is a finite set of *group nodes*, $E_H \subseteq V_H \times V_H$ is a finite set of *group links*, and each edge $(\text{grp}_{s_i}, \text{grp}_{s_j}) \in E_H$ is associated with a weight w_{ij}^H.

Definition 6.2 *Group nodes* are defined according to the required skills. For a certain required skill $s \in S$, a group node $\text{grp}_s \in V_H$ contains a set of nodes $V'(\text{grp}_s) \subseteq V$ in G and satisfies the following conditions: (1) $\forall u \in V'(\text{grp}_s)$, $s \subseteq X_u$ and (2) nodes in $V'(\text{grp}_s)$ need to form an induced connected subgraph $G[V'(\text{grp}_s)]$.

Definition 6.3 A *group link* $e_H \in E_H$ is defined as the connection between two group nodes grp_{s_i} and grp_{s_j} in G_H. The corresponding induced subgraphs of the two group nodes, $G[V'(\text{grp}_{s_i})]$ and $G[V'(\text{grp}_{s_j})]$, need to be reachable to each other. Note that two induced subgraph in G, connected by a group link, can be overlapped, direct-connected, or indirect-connected.

A group graph H can be regarded as a super-level graph of the raw collaborative network G. Each group node in H is a super-node containing a set of individual nodes in G. To encode the interactions between groups into the group graph, we associate each group link with a weight. Such weight is derived from aggregating the communication costs between individuals from two end groups of G. Given two groups, each of which contains a set of individuals in G, we employ the distance measure in single-link hierarchical clustering to compute edge weights in H. Specifically, for a group link $e_H = (\text{grp}_{s_i}, \text{grp}_{s_j})$, the minimum shortest length between individuals in grp_{s_i} and grp_{s_j} from the expertise graph $G = (V, E)$ is defined as the weight of edge e_H. The calculation can be formulated as

$$\text{weight}(e_H) = |\min\{\text{dist}_G(u, v)\}|,$$

where $u \in V'(\text{grp}_{s_i}) \subseteq V$, $v \in V'(\text{grp}_{s_j}) \subseteq V$, and $e_H \in E_H$. The value $\text{dist}_G(u, v)$ represents the shortest distance between node u and v in G. Moreover, we need to keep track of the mapping between each group link and its corresponding *minimum shortest path*. We denote this mapping as $\text{MSP}(e_H) = \text{path}(u, v)$. Figure 3a shows the group graph for the collaborative network in Fig. 2. Zero weight indicates that two groups are overlapped. The *MSP* of the group link $e_H = (s_2, s_3)$ is the path containing edges (h, q) and (q, p) in G of Fig. 2.

6.3 Applying Ehanced-Steiner algorithm on the group graph

The group graph provides two merits allowing us to find effective connections between groups for the required skills efficiently. First, the group graph reduces the search space of the collaborative network. Second, since the costs between groups are minimized, the group graph can guide the graph search by traversing the lower-cost links and the more effective nodes (e.g. overlapped nodes) to connect groups. Therefore, the group graph can avoid redundant costs and decrease the cardinality of team members and unnecessary inter-mediators.

We apply the Enhanced-Steiner algorithm to the group graph. Similar to the original and generalized Enhanced-Steiner algorithms, an enhanced graph is constructed by adding and

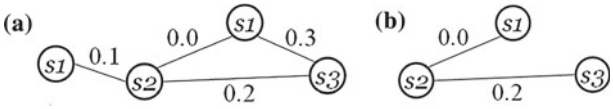


Fig. 3 **a** Group graph construction from Fig. 2. **b** The effective subgraph of (**a**)

connecting the required skills to the group nodes possessing such skills. Then by adopting Algorithm 0 with density-based seed selection, an effective group-level subgraph that connects groups with minimum communication cost can be derived. We denote such *effective subgraph of groups* as $H[V'_H]$, where $V'_H \subseteq V_H$. The corresponding effective subgraph of groups shown in Fig. 3a is shown in Fig. 3b.

6.4 The role composition method

Now we have obtained the effective connection subgraph of groups satisfying all required skills with minimum communication cost between groups. However, we have not yet considered the requirement of the minimum number of experts for each required skill. In this final stage, we present a role composition method to decode the connection subgraph of groups $H[V'_H]$ and compose a team with specific number of skilled members having the required skills who can communicate effectively.

The rationale of our role composition method lies in the members of the final team V' who acts with different functional roles in their communication network $G[V']$. An *inter-mediator* who does not possess any of the required skills acts as the mediator between two skilled groups. For example, in Fig. 4, the individual q acts as an inter-mediator between skilled groups s_2 and s_3. *Connectors* deal with coordinating people between different skilled groups. The connectors can be individuals who are good at multiple skills, or individuals who communicate directly with the inter-mediator. For example in Fig. 4, the individual i is a connector who is good at both s_1 and s_2. Both h and p are also connectors who communicate directly with *inter-mediator* q. *Intra-mediators* are individuals who communicate with more than one connectors directly. For the example in Fig. 4, the individual g is an intra-mediator connecting two connectors i and h within the skilled group s_2. The other individuals who belong to skilled groups are regarded as *collaborators*. Hence, the other individuals in Fig. 4, namely *f, l, k, n,* and *o*, are collaborators. Recall that we have recorded the mapping from group links to the corresponding minimum shortest path between two groups, it is easy to find the roles from the communication network quickly.

Based on the observed roles within/among groups, including connectors, inter-mediators, intra-mediators, and collaborators, we present a role composition algorithm in the following to recommend a team for the given generalized tasks. We first find the connectors and inter-mediators (line 1–3), and then we find the intra-mediators by connecting the connectors within a group (line 4–8). Finally we check the specified number of experts for each required skill (line 9). If the number of requirement of any one of the required skills is not met, we will add

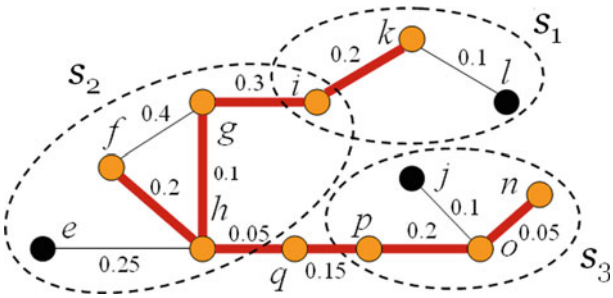


Fig. 4 The final team of Fig. 2 by role composition method

more collaborators in the corresponding skilled group by calculating the shortest paths in the corresponding groups until the given generalized task is satisfied (line 10–12). Figure 4 shows the result subgraph of team for the generalized task $\{< s_1, 2 >, < s_2, 4 >, < s_3, 3 >\}$, in which those highlighted nodes and edges belong to the final subgraph.

Algorithm 2. Role Composition.

Input: An effective subgraph of groups $H[V'_H] = (V'_H, E'_H)$;
the mapping from groups link to minimum shortest path *MSP*;
Output: Team $V' \subseteq V$ and its induced subgraph $G[V']$.

1: **for each** $e'_H \in E'_H$ **do**
2: $V' \leftarrow V' \cup MSP(e'_H)$. //add connectors and inter-mediators
3: Update k_i and k_j of end vertices grp_{si} and grp_{sj} of e'_H.
4: **for each** $pair = <x, y>$ of nodes in V' **do**
5: **if** $x \in grp_{si}$ and $y \in grp_{si}$, where $grp_{si} \in V'_H$ **then**
6: // $G[grp_{si}]$ is the induced graph of individuals in grp_{si}.
7: $V' \leftarrow V' \cup \{Path(x, y)$ in $G[grp_{si}]\}$.//add intra-mediators
8: Update k_i of s_i.
9: **while** a skill s_i whose specified number k_i is not met **do**
10: $v^* = \arg\min_{u \in grp_{si} \setminus V' \text{ and } grp_{si} \in V'_H} \{dist(u, V') \text{ in } G\}$.
11: $V' \leftarrow V' \cup v^*$. // add collaborators
12: Update k_i of s_i.

The time complexity of the grouping-based approach consists of the following parts. (1) The running time of the skill-based individual grouping is $O(|V|)$. (2) For constructing the group graph, the bottleneck lies in the computation of the shortest paths between the inter-mediator nodes when finding the group links. It is natural to see that if the weighted shortest paths have higher values, they are less likely to be selected into the resulting Steiner tree. Therefore, we simplify this part by considering the r-step paths between groups. That is, we record only those paths whose lengths equal to or less than r in *MSP* ($r = 2$ in the later experiments we conducted). As a result, the time complexity of this part turns out to be $O(|V|)$ since we only need to scan the nodes to test their neighbors for finding *MSP*s of at most 2 steps between groups. (3) Applying the Steiner tree algorithm on the group graph has running time of $O(|R| \times |E_H|)$. In the worst case, the time complexity is $O(|V|^3)$ because $|R| = O(|V|)$ and $|E_H| = O(|V|^2)$. However, in practice, the required skill set R and the edge set E_H are much smaller than the worst-case scenarios. (4) The time complexity of the role composition algorithm is $O(|E_H| + |V'|^2 + n_p \times |V'| \times |grp_{s_i}|) = O(|E_H| + |V'|^2)$. Likewise, the worst case would be $O(|V|^2 + |V|^2) = O(|V|^2)$ (due to $|E_H| = O(|V|^2)$ and $|V'| = O|V|$). However, the real-world query E_H and V' are very small. In summary, the overall time complexity of the proposed grouping method is $O(|V| + |R| \times |E_H| + |V'|^2)$.

7 Evaluation

In this section, we report the performance of our proposed methods to find teams for the generalized task. We show that our methods can organize teams with low communication cost, low cardinality, and less inter-mediators. In addition, our solution to compose teams of experts is efficient in terms of running time.

7.1 The DBLP dataset

We use the DBLP bibliography database to extract the connective and expertise data. The snapshot on December 30, 2008 of data mining- related conferences (including KDD, ICDM, SDM, PAKDD, PKDD, ICML, CIKM, WWW, and SIGIR) is used. We construct the collaborative social network using coauthorships. The set of connected persons consists of authors that cowork at least three papers. The skill set X_i of each author i consists of the set of terms occurring in at least four paper titles that he has published. Totally there are 5,482 authors, 11,905 skills, and 10,339 edges. The weights on edges are generated according to the probability of collaboration as $w(i,\ j) = 1 - |P_i \cap P_j|/|P_i \cup P_j|$, where P_i is the set of papers of i.

7.2 Experiment design

We conduct a series of experiments to demonstrate the effectiveness and efficiency of the proposed algorithms, comparing to some other methods. The evaluation consists of four parts. The first and the second are designed to compose the teams for **Generalized** and **Basic Tasks**, respectively. The third is to study the performance when the required skills are compiled from diverse or irrelevant areas (denoted by **Cross-domain Tasks**). The fourth is to show the time efficiency of different methods. When tackling the generalized tasks, we compute the effectiveness of two families of methods: the generalized Enhanced-Steiner tree algorithm and the proposed grouping-based method. When dealing with the basic tasks, we further consider another family of method: the Lappas' Steiner tree algorithm [14]. The performance measures include (a) the **communication cost** (lower communication cost indicates better team collaboration), (b) the **cardinality** of a team (smaller cardinality indicates lower cost of such team composition), (c) the **number of inter-mediators** (more inter-mediators mean such team contains more irrelevant people), (d) the **number of common publications** shared by at least two experts in the team (more common publications among the experts indicate that they tend to communicate well with each other), and (e) the **skill count**, defined as the average frequency that a skill appears in papers of the experts in the team, averaged over all the required skills and over all the experts in the team (higher skill count value means the team members tend to be good at the required skills).

To generate the expertise queries for the experiments, instead of randomly selecting the required skills from the entire skill set [14], which is a bit unreasonable in the real-world scenario, we consider each paper as one task and regard the keywords in each paper as the corresponding required skills. In other words, we simulate the process of recruiting experts who possess different skills to work on the research topic described in a paper. Papers published in years 2005–2008 are used to construct query skill sets, and the numbers of required skill sets are 1,467, 1,712, 1,755, and 1,824, respectively, in these four years. To solve the generalized tasks, the next step is to associate a specific number of experts with each required skill. Since there is no information about the number of experts required by each skill, we generate such numbers according to a simple rule: If a certain skill is more popular (i.e., more people are good at such skill), the project leader will want to select more individuals for such skill because this skill is commonly required in the real world. Specifically, every query task is controlled by two parameters: (a) the number of required skills t and (b) a fixed ratio $r \in [0, 1]$ which determines the specified number associated with each required skill. Specifically, we randomly pick t required skills from the keywords appearing in all paper titles. If a keyword (i.e., a required skill) appears F times in all paper titles, we round off F^*r to be the required number of experts for such skill. In addition, the experiments below have

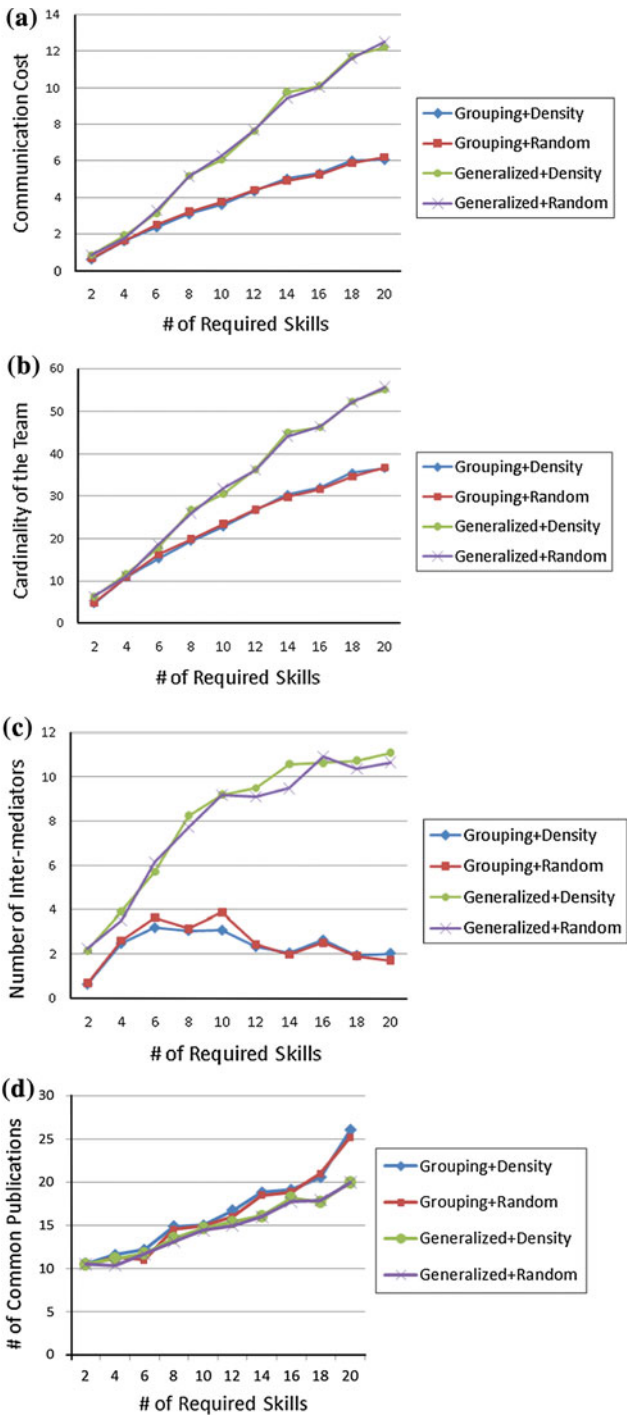


Fig. 5 Experimental results for generalized tasks under the performance measures of **a** Average communication cost, **b** average cardinality of teams, **c** average number of inter-mediator, **d** average number of common publications, and **e** average skill count

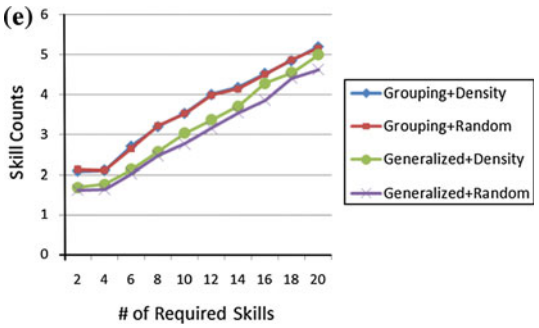


Fig. 5 continued

the following settings: we vary $t = 2, 4, \ldots, 20$ and set $r = 0.02$. For every (t, r) pair, we generate 100 random generalized tasks and report the average values of the results obtained.

In the following experiments, we compute and compare the performance over the methods of three families. The first is the proposed grouping-based approach with random seed selection (**Grouping + Random**) and with density-based seed selection (**Grouping + Density**). The second is the devised generalized Enhanced-Steiner algorithm with random seed selection (**Generalized + Random**) and with density-based seed selection (**Generalized + Density**). The third is the Lappas's Enhanced-Steiner algorithm [14] with random seed selection (**Lappas**) and with density-based seed selection (**Lappas + Density**). For the experiments of generalized tasks, we compare only the first two families (because the original Lappas' method cannot handle the generalized tasks). For the basic tasks, we compare all the three families.

7.3 Experimental results

7.3.1 Generalized task

Figure 5 exhibits the results of the proposed grouping-based methods and the devised generalized Enhanced-Steiner algorithms under the five performance measures. Figure 5a, b shows that our proposed grouping-based method outperforms the generalized algorithm on both communication cost and the cardinality of the team. In particular, as the number of required skills increase, the advantage of the grouping-based method is more significant. It is because of that the grouping can not only reduce the costs for individuals possessing the same skill but also minimize the number of inter-mediators (i.e., irrelevant persons) between groups, as evidenced by Fig. 5c. Besides, we can observe that in our grouping-based method, the number of inter-mediators does not grow as the number of required skills increases. For the last two performance measures, as shown in Fig. 5d, e, our grouping-based approach outperforms the generalized Enhanced-Steiner method. That says, our method produces smaller teams with higher number of common publications (which indicates they can enhance the experience of cooperation) and more expertise at the required skills. In short, for generalized tasks, using our proposed grouping-based method can find more effective teams.

7.3.2 Basic task

The basic tasks can be regarded as the special cases of the generalized tasks. We exploit the proposed grouping-based approach, the devised generalized Enhance-Steiner algorithm,

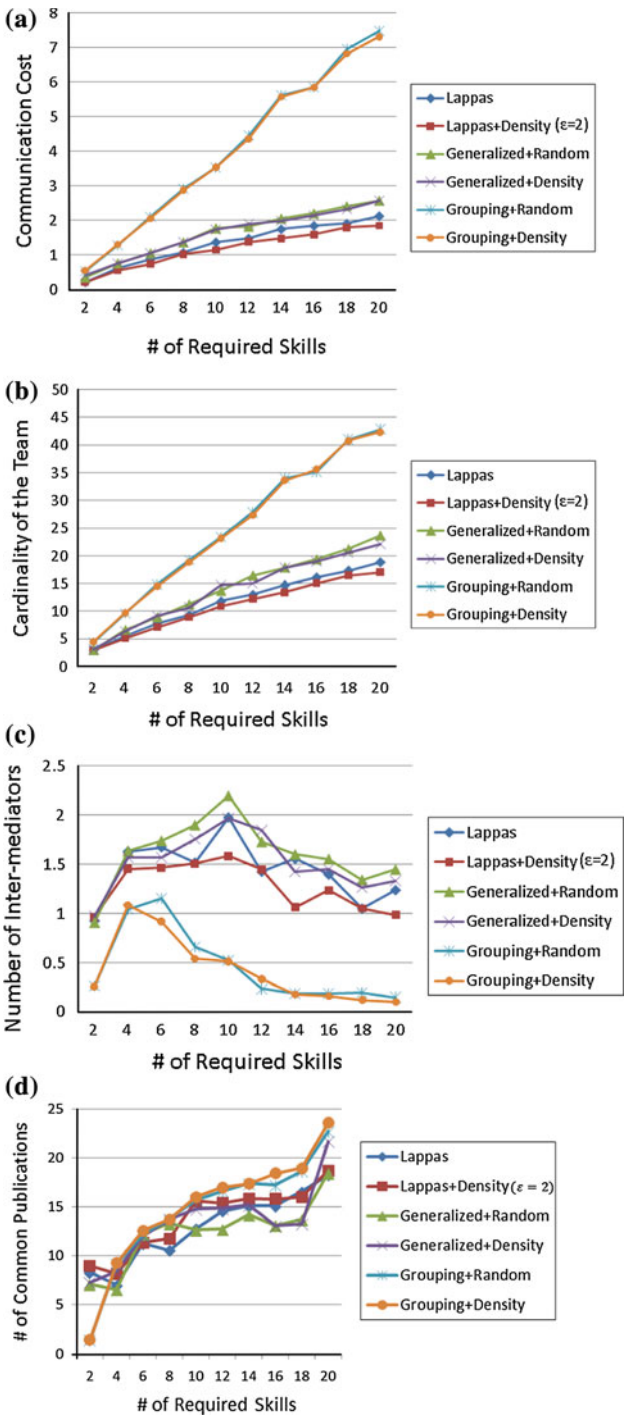


Fig. 6 Experimental results for basic tasks under the performance measures of **a** Average communication cost, **b** average cardinality of teams, **c** average number of inter-mediator, **d** average number of common publications, and **e** average skill count

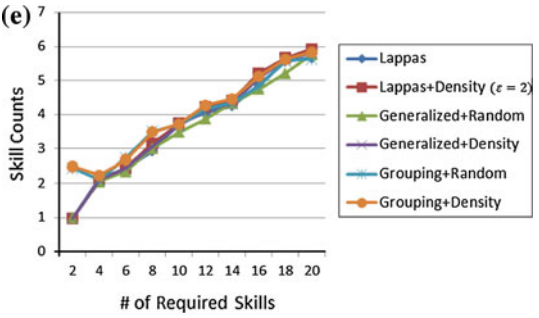


Fig. 6 continued

and the original Lappas' method to find teams of experts for basic tasks. Figure 6 shows the effectiveness under the five performance measures. The experimental results, as shown in Fig. 6a, b, demonstrate that the Lappas' and the generalized approaches perform better than the grouping-based method on the communication cost and the cardinality of the constructed team. It is because that the grouping action aims at gathering those possessed the same skills together to minimize the cost among them while sacrificing a bit on cost between groups. However, for basic tasks, only one expert is needed for each required skill. Consequently, more cost is introduced due to the grouping. Nevertheless, as evidenced in Fig. 6c, the proposed grouping-based approach faithfully minimizes the number of inter-mediators. The generalized Enhanced-Steiner algorithm performs a bit worse because it would include more inter-mediators which cause the raise of the communication cost and the cardinality. Though the grouping-based method performs worse in terms of the communication cost and cardinality, it can still ensure the team members having better collaboration experience and having higher expertise on the required skills, as presented in Fig. 6d, e. Furthermore, from Fig. 6a–e, the proposed density-based seed selection can have better performance than the random-based seed selection under different measures and different methods.

7.3.3 Cross-domain task

We also demonstrate the ability of the proposed methods for handling skills compiled from diverse research fields or irrelevant topics. We manually construct four generalized cross-domain tasks whose number of required skill is 6, as listed in Table 2. The experimental results on the communication cost, the cardinality, and the number of inter-mediators are shown in Fig. 7, in which the **Avg.** represents the average results of #RequiredSkills=6 in

Table 2 The list of manually compiled generalized cross-domain tasks, in which skills are either from either multiple disciplinary research fields or irrelevant topics

ID	Generalized cross-domain Tasks (# of required skills = 6)
A	{visualization = 3, gene = 2, graph = 5, video = 2, convex = 2, crawler = 1}
B	{music = 1, biological = 2, translation = 6, image = 4, kernel = 3, coding = 2}
C	{biological = 3, editing = 1, entropy = 2, security = 3, language = 8, video = 3}
D	{speech = 3, interactive = 7, query = 3, social = 1, optimization = 2, security = 2}

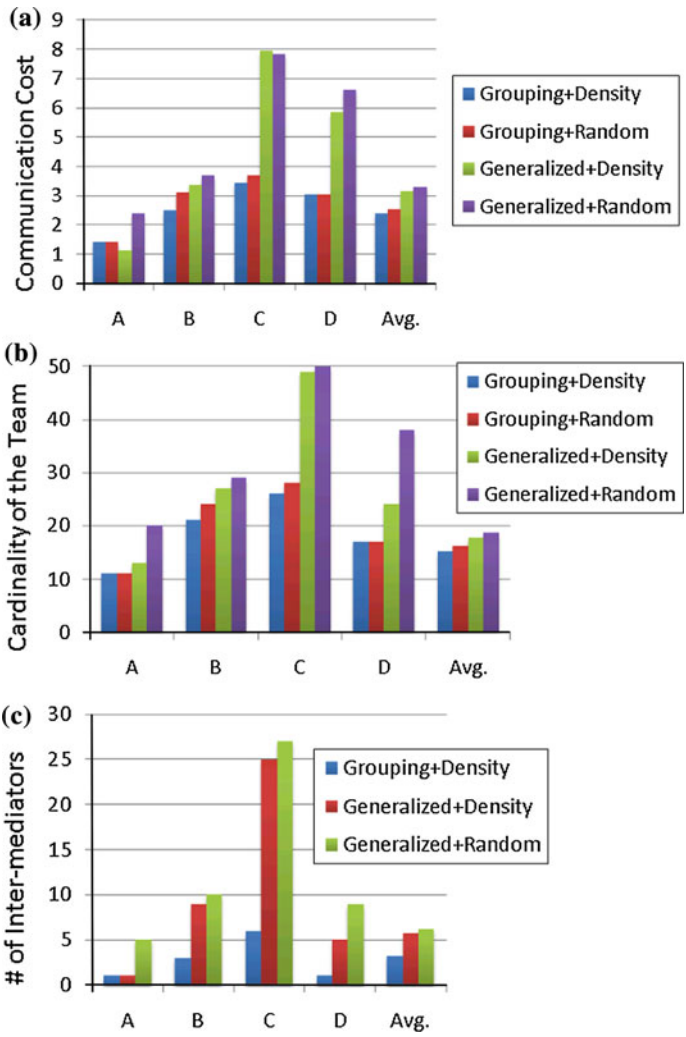


Fig. 7 Experimental results for cross-domain tasks in Table 2 under the performance measures of **a** communication cost, **b** cardinality of the team, and **c** the number of inter-mediators

Fig. 5. We can see that the proposed grouping-based method outperforms the generalized Enhanced-Steiner algorithm. Such result indicates the grouping is capable of finding teams of experts effectively for skills distributed in highly diverse fields.

7.3.4 Time efficiency

Figure 8 shows the time efficiency of the proposed grouping-based methods and the generalized Enhanced-Steiner algorithm when tackling the generalized tasks. We can find that the grouping-based method is more efficient than the generalized Enhanced-Steiner algorithm in general. As the number of required skills increases, the average execution time (in second) of the group-based method grows linearly and slowly while the execution time of the

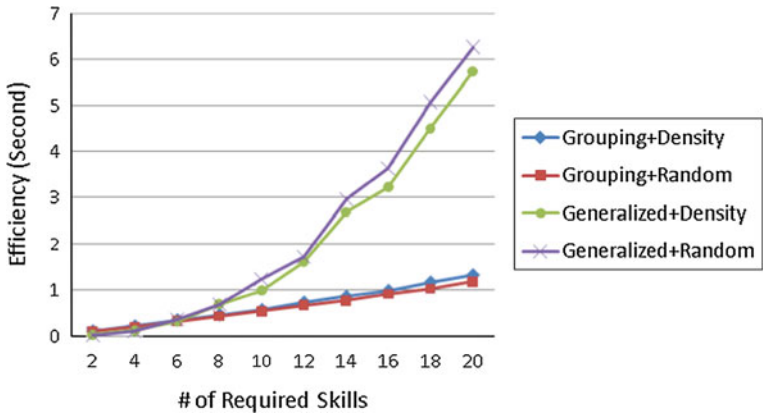


Fig. 8 Average execution time run in second when using the proposed grouping-based methods and the generalized Enhanced-Steiner algorithms

Fig. 9 The positive correlation between the number of groups and the execution time as the number of required skills increases

generalized Enhanced-Steiner algorithm rises very quickly. We believe it is attributed to the grouping action, which produces the group graph as a compact representation, drastically reduces the search space. We confirm such conjecture by plotting the correlation between the skill-based node grouping and the time efficiency in Fig. 9. Such result shows that the number of groups is highly related to the execution time. In short, our grouping-based method can find effective teams for generalized tasks in an efficient manner.

8 Conclusion

We tackle the problem of finding teams of experts for generalized tasks consisting of a set of required skills, where each of which is associated with a specified number of experts. To compose an effective team of experts, we modify the Enhanced-Steiner algorithm to deal

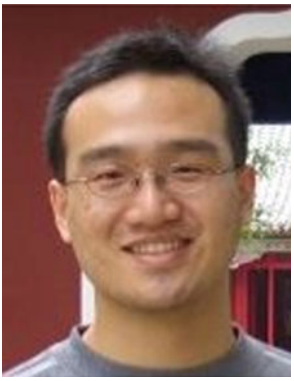
with generalized tasks. To improve the algorithm, we propose a density-based method to select the seed node more wisely. More importantly, we propose a grouping-based approach with the role composition algorithm to further boost the effectiveness of the constructed teams and boost the efficiency in the team formation process. Experimental results show our proposed algorithms can effectively find teams of experts for generalized, basic, and diverse tasks under five performance measures. Our grouping-based method has been proved to be very efficient as well.

References

1. Agustín-Blas LE, Salcedo-Sanz S, Ortiz-García EG, Portilla-Figueras A, Pérez-Bellido AM, Jiménez-Fernández S (2010) Team formation based on group technology: a hybrid grouping genetic algorithm approach. Comput Oper Res 38:484–495
2. Anagnostopoulos A, Becchetti L, Castillo C, Gionis A, Leonardi S (2010) Power in unity: forming teams in large-scale community systems. In: Proceedings of ACM international conference on information and knowledge management (CIKM'10), pp 599–608
3. Anagnostopoulos A, Becchetti L, Castillo C, Gionis A, Leonardi S (2012) Online team formation in social networks. In: Proceedings of ACM international conference on World Wide Web (WWW'12), pp 839–848
4. Cheatham M, Cleereman K (2006) Application of social network analysis to collaborative team formation. In: Proceedings of international symposium on collaborative technologies and systems, pp 306–311
5. Chen SJ, Lin L (2004) Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. IEEE Trans Eng Manag 51(2):111–124
6. Cheng J, Ke Y, Ng W (2009a) Efficient processing of group-oriented connection queries in a large graph. In: Proceedings of ACM international conference on information and knowledge management (CIKM'09), pp 1481–1484
7. Cheng J, Ke Y, Ng W, Yu JX (2009b) Context-aware object connection discovery in large graphs. In: Proceedings of IEEE international conference on data engineering (ICDE'09), pp 856–867
8. Faloutsos C, McCurley KS, Tomkins A (2004) Fast discovery of connection subgraph. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD'04), pp 118–127
9. Fitzpatrick EL, Askin RG (2005) Forming effective worker teams with multi-functional skill requirements. Comput Ind Eng 48(3):593–608
10. Gaston M, Simmons J, DesJardins M (2004) Adapting network structures for efficient team formation. In: Proceedings of the AAAI fall symposium on artificial multi-agent learning
11. Kargar M, An A (2011) 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
12. Kargar M, An A, Zihayat M (2012) Efficient bi-objective team formation in social networks. Mach Learn Knowl Discov Databases 7524:483–498, LNCS
13. Kasneci G, Ramanath M, Sozio M, Suchanek FM, Weikum G (2009) STAR: Steiner-tree approximation in relationship graphs. In: Proceedings of IEEE international conference on data engineering (ICDE'09), pp 868–879
14. Lappas T, Liu K, Terzi E (2009) Finding a team of experts in social networks. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD'09), pp 467–475
15. Li C-T, Shan M-K, Lin S-D (2011) Context-based people search in labeled social networks. In: Proceedings of ACM international conference on information and knowledge management (CIKM'11), pp 1607–1612
16. Li C-T, Shan M-K (2012) Composing activity groups in social networks. In: Proceedings of ACM international conference on information and knowledge management (CIKM'12), pp 2375–2378
17. Liu W, Sun W, Chen C, Huang Y, Jing Y, Chen K (2012) Circle of friend query in geo-social networks. In: Proceedings of international conference on database systems for advanced applications (DASFAA'12), pp 126–137
18. Majumder A, Datta S, Naidu KVM (2012) Capacitated team formation problem on social networks. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD'12), pp 1005–1013
19. Reich G, Widmayer P (1990) Beyond Steiner's problem: a VLSI oriented generalization. In: Proceedings of international workshop on graph-theoretic concepts in computer science, pp 196–210

20. Sorkhi M, Hashemi S, Hamzeh A (2011) An effective expert team formation in social networks based on skill grading. In: Proceedings of IEEE international conference on data mining workshops (ICDMW'11), pp 366–372
21. Sozio M, Gionis A (2010) 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
22. Tong H, Faloutsos C, Pan J-Y (2006) Fast random walk with restart and its application. In: Proceedings of IEEE international conference on data mining (ICDM'06), pp 613–622
23. Tong H, Faloutsos C (2006) Center-piece subgraph: problem definition and fast solution. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD'06), pp 404–413
24. Tong H, Qu H, Jamjoom H, Faloutsos C (2009) iPoG: fast interactive proximity querying on graphs. In: Proceedings of ACM international conference on information and knowledge management (CIKM'09), pp 1673–1676
25. Tong H, Faloutsos C, Gallagher B, Eliassi-Rad T (2007) Fast best-effort pattern matching in large attributed graphs. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD'07), pp 737–746
26. Wi H, Oh S, Mun J (2009) Jung M (2009) A team formation model based on knowledge and collaboration. Expert Syst Appl 36(5):9121–9134
27. Yang D-N, Chen Y-L, Lee W-C, Chen M-S (2011) On social-temporal group query with acquaintance constraint. Proc VLDB Endow 4(6):397–408
28. Yang D-N, Shen C-Y, Lee W-C, Chen M-S (2012) 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

Author Biographies



Cheng-Te Li received his B.S. and Ph.D. degrees from Graduate Institute of Networking and Multimedia, National Taiwan University, in 2009 and 2013, respectively. His research interests include social and information networks, data mining, and social media analytics. His international recognition includes Facebook Fellowship 2012 Finalist Award, ACM KDD Cup 2012 First Prize (member of NTU team), IEEE/ACM ASONAM 2011 Best Paper Award, and Microsoft Research Asia Fellowship 2010.

Man-Kwan Shan received the B.S. degree in computer engineering and the M.S. degree in computer and information science both from National Chiao Tung University, Taiwan, in 1986 and 1988, respectively. From 1988 to 1990, he served as a lecture in the Army Communications and Electronics School. Then, he worked as a lecture at the Computer Center of National Chiao Tung University, where he supervised the Research and Development Division. He received the Ph.D. degree in Computer Science and Information Engineering from National Chiao Tung University in 1998. Then he joined the Department of Computer Science at National Chengchi University as an assistant professor. He became an associated professor in 2003 and a professor in 2011. His current research interests include data mining, multimedia systems, and bioinformatics.

Shou-De Lin is currently an associate professor in the CSIE department of National Taiwan University. He holds a BS in EE department from National Taiwan University, an MS-EE from the University of Michigan, and an MS in Computational Linguistics and Ph.D. in Computer Science both from the University of Southern California. He joined National Taiwan University in 2007. Before joining NTU, he was a postdoctoral research fellow at the Los Alamos National Lab. Prof. Lin's research includes knowledge discovery and data mining, social network analysis, natural language processing, and machine learning. His international recognition includes the best paper award in IEEE WI 2003, Google Research Award in 2007, Microsoft research award in 2008, merit paper award in TAAI 2010, and best paper award in ASONAM 2011. He is the all-time winners in ACM KDD Cup, leading or coleading the NTU team to win championships in 2008, 2010, 2011, 2012, 2013, ranked 2nd in 2003 and 3rd in 2009.