Towards Realistic Team Formation in Social Networks based on Densest Subgraphs

Syama Rangapuram
Max Planck Institute for
Computer Science
Saarbrücken, Germany
srangapu@mpi-inf.mpg.de

Thomas Bühler
Faculty of Mathematics
and Computer Science
Saarland University
Saarbrücken, Germany
tb@cs.uni-saarland.de

Matthias Hein
Faculty of Mathematics
and Computer Science
Saarland University
Saarbrücken, Germany
hein@cs.uni-saarland.de

ABSTRACT

Given a task $\mathcal{T}$, a set of experts V with multiple skills and a social network $G(V, W)$ reflecting the compatibility among the experts, *team formation* is the problem of identifying a team $C \subseteq V$ that is both competent in performing the task $\mathcal{T}$ and compatible in working together. Existing methods for this problem make too restrictive assumptions and thus cannot model practical scenarios. The goal of this paper is to consider the team formation problem in a realistic setting and present a novel formulation based on densest subgraphs. Our formulation allows modeling of many natural requirements such as (i) inclusion of a designated team leader and/or a group of given experts, (ii) restriction of the size or more generally cost of the team (iii) enforcing *locality* of the team, e.g., in a geographical sense or social sense, etc. The proposed formulation leads to a generalized version of the classical densest subgraph problem with cardinality constraints (DSP), which is an NP hard problem and has many applications in social network analysis. In this paper, we present a new method for (approximately) solving the generalized DSP (GDSP). Our method, **FORTE**, is based on solving an *equivalent* continuous relaxation of GDSP. The solution found by our method has a quality guarantee and always satisfies the constraints of GDSP. Experiments show that the proposed formulation (GDSP) is useful in modeling a broader range of team formation problems and that our method produces more coherent and compact teams of high quality. We also show, with the help of an LP relaxation of GDSP, that our method gives close to optimal solutions to GDSP.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Graph algorithms*

General Terms

Algorithms, Experimentation, Theory

Keywords

Team formation, Social networks, Densest subgraphs

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

1. INTRODUCTION

Given a set of skill requirements (called task $\mathcal{T}$), a set of experts who have expertise in one or more skill, along with a social or professional network of the experts, the team formation problem is to identify a competent and highly collaborative team. This problem in the context of a social network was first introduced by [18] and has attracted recent interest in the data mining community [15, 2, 12]. A closely related and well-studied problem in operations research is the assignment problem. Here, given a set of agents and a set of tasks, the goal is to find an agent-task assignment minimizing the cost of the assignment such that exactly one agent is assigned to a task and every task is assigned to some agent. This problem can be modeled as a maximum weight matching problem in a weighted bipartite graph. In contrast to the assignment problem, the team formation problem considers the underlying social network, which for example models the previous collaborations among the experts, while forming teams. The advantage of using such a social network is that the teams that have worked together previously are expected to have less communication overhead and work more effectively as a team.

The criteria explored in the literature so far for measuring the effectiveness of teams are based on the shortest path distances, density, and the cost of the minimum spanning tree of the subgraph induced by the team. Here the density of a subgraph is defined as the ratio of the total weight of the edges within the subgraph over the size of the subgraph. Teams that are well connected have high density values. Methods based on minimizing diameter (largest shortest path between any two vertices) or cost of the spanning tree have the main advantage that the teams they yield are always connected (provided the underlying social network is connected). However, diameter or spanning tree based objectives are not robust to the changes (addition/deletion of edges) in the social network. As demonstrated in [12] using various performance measures, the density based objective performs better in identifying well connected teams. On the other hand, maximizing density may give a team whose subgraph is disconnected. This happens especially when there are small groups of people who are highly connected with each other but are sparsely connected to the rest of the graph.

Existing methods make either strong assumptions on the problem that do not hold in practice or are not capable of incorporating more intuitive constraints such as bounding the total size of the team. The goal of this paper is to consider the team formation problem in a more realistic setting and present a novel formulation based on a generalization of the densest subgraph problem. Our formulation allows modeling of many realistic requirements such as (i) inclusion of a designated team leader and/or a group of given experts, (ii) restriction on the size or more generally cost of the team (iii) enforcing *locality* of the team, e.g., in a geographical sense or social sense, etc. In fact most of the future directions pointed out by [12] are covered in our formulation.

2. RELATED WORK

The first work [18] in the team formation problem in the presence of a social network presents greedy algorithms for minimizing the diameter and the cost of the minimum spanning tree (MST) induced by the team. While the greedy algorithm for minimizing the diameter has an approximation guarantee of two, no guarantee is proven for the MST algorithm. However, [18] impose the strong assumption that a skill requirement of a task can be fulfilled by a single person; thus a more natural requirement such as "at least k experts of skill s are needed for the task" cannot be handled by their method. This shortcoming has been addressed in [12], which presents a 2-approximation algorithm for a slightly more general problem that can accommodate the above requirement. However, both algorithms cannot handle an upper bound constraint on the team size. On the other hand, the solutions obtained by all these algorithms (including the MST algorithm) can be shown to be connected subgraphs if the underlying social graph is connected.

Two new formulations are proposed in [15] based on the shortest path distances between the nodes of the graph. The first formulation assumes that experts from each skill have to communicate with every expert from the other skill and thus minimizes the sum of the pairwise shortest path distances between experts belonging to different skills. They prove that this problem is NP-hard and provide a greedy algorithm with an approximation guarantee of two. The second formulation, solvable optimally in polynomial time, assumes that there is a designated team leader who has to communicate with every expert in the team and minimizes the sum of the distances only to the leader. The main shortcoming of this work is its restrictive assumption that *exactly* one expert is sufficient for each skill, which implies that the size of the found teams is always upper bounded by the number of skills in the given task, noting that an expert is allowed to have multiple skills. They exploit this assumption and (are the first to) produce top-k teams that can perform the given task. However, although based on the shortest path distances, neither of the two formulations does guarantee that the solution obtained is connected.

In contrast to the distance or diameter based cost functions, [12] explore the usefulness of the density based objective in finding strongly connected teams. Using various performance measures, the superiority of the density based objective function over the diameter objective is demonstrated. The setting considered in [12] is the most general one until now but the resulting problem is shown to be NP hard. The greedy algorithms that they propose have approximation guarantees (of factor 3) for two special cases. The teams found by their algorithms are often quite large and it is not straightforward to modify their algorithms to integrate an additional upper bound constraint on the team size. Another disadvantage is that subgraphs that maximize the density under the given constraints need not necessarily be connected.

Recently [2] considered an *online* team formation problem where tasks arrive in a sequential manner and teams have to be formed minimizing the (maximum) load on any expert across the tasks while bounding the coordination cost (a free parameter) within a team for any given task. Approximation algorithms are provided for two variants of coordinate costs: diameter cost and Steiner cost (cost of the minimum Steiner tree where the team members are the terminal nodes). While this work focusses more on the load balancing aspect, it also makes the strong assumption that a skill is covered by the team if there exists at least one expert having that skill.

All of the above methods allow only binary skill level, i.e., an expert has a skill level of either one or zero.

We point out that many methods have been developed in the operations research community for the team formation problem, [5, 9, 21, 20], but none of them explicitly considers the underlying social or professional connections among the experts. There is also literature discussing the social aspects of the team formation [10] and their influence on the evolution of communities, e.g., [4].

3. REALISTIC TEAM FORMATION IN SOCIAL NETWORKS

Now we formally define the *Team Formation* problem that we address in this paper. Let V be the set of n experts and $G(V, W)$ be the weighted, undirected graph reflecting the relationship or previous collaboration of the experts V. Then non-negative, symmetric weight $w_{ij} \in W$ connecting two experts i and j reflects the level of compatibility between them. The set of skills is given by $\mathcal{A} = \{a_1, \ldots, a_p\}$. Each expert is assumed to possess one or more skills. The non-negative matrix $M \in \mathbb{R}^{n \times p}$ specifies the skill levels of all experts in each skill. Note that we define the skill level on a continuous scale. If an expert i does not have skill j, then $M_{ij} = 0$. Moreover, we use the notation $M_j \in \mathbb{R}^{n \times 1}$ for the $j-$th column of M, i.e. the vector of skill levels corresponding to skill j. A task $\mathcal{T}$ is given by the set of triples $\{(a_j, \kappa_j, \iota_j)\}_{j=1}^{p}$, where $a_j \in \mathcal{A}$, specifying that at least κ_j and at most ι_j of skill a_j is required to finish the given task.

Generalized team formation problem. Given a task $\mathcal{T}$, the generalized team formation problem is defined as finding a team $C \subseteq V$ of experts maximizing the *collaborative compatibility* and satisfying the following constraints:

- **Inclusion of a specified group:** a predetermined group of experts $S \subset V$ should be in C.
- **Skill requirement:** at least κ_j and at most ι_j of skill a_j is required to finish the task $\mathcal{T}$.
- **Bound on the team size:** the size of the team should be smaller than or equal to b, i.e., $|C| \leq b$.
- **Budget constraint:** total budget for finishing the task is bounded by B, i.e., $\sum_{i \in C} c_i \leq B$, where $c_i \in \mathbb{R}_+$ is the cost incurred on expert i.
- **Distance based constraint:** the distance (measured according to some non-negative, symmetric function, dist) between any pair of experts in C should not be larger than d_0, i.e., $\text{dist}(u, v) \leq d_0, \forall u, v \in C$.

Discussion of our generalized constraints. In contrast to existing methods, we also allow an upper bound on each skill and on the total team size. If the skill matrix is only allowed to be binary as in previous work, this translates into upper and lower bounds on the number of experts required for each skill. Using vertex weights, we can in fact encode more generic constraints, e.g., having a limit on the total budget of the team. It is not straightforward to extend existing methods to include any upper bound constraints. Up to our knowledge we are the first to integrate upper bound constraints, in particular on the size of the team, into the team formation problem. We think that the latter constraint is essential for realistic team formation.

Our general setting also allows a group of experts around whom the team has to be formed. This constraint often applies as the team leader is usually fixed before forming the team. Another important generalization is the inclusion of *distance* constraints for any general distance function[1]. Such a constraint can be used to enforce locality of the team e.g. in a geographical sense (the distance could be travel time) or social sense (distance in the network). Another potential application are mutual incompatibilities of team members e.g. on a personal level, which can be addressed by assigning a high distance to experts who are mutually incompatible and thus should not be put together in the same team.

We emphasize that all constraints considered in the literature are special instances of the above constraint set.

Measure of collaborative compatiblity. In this paper we use as a measure of collaborative compatibility a generalized form of the density of subgraphs, defined as

$$\text{density}(C) := \frac{\text{assoc}(C)}{\text{vol}_g(C)} = \frac{\sum_{i,j \in C} w_{ij}}{\sum_{i \in C} g_i}, \qquad (1)$$

where w_{ij} is the non-negative weight of the edge between i and j and $\text{vol}_g(C)$ is defined as $\sum_{i \in C} g_i$, with g_i being the positive weight of the vertex i. We recover the original density formulation, via $g_i = 1, \forall i \in V$. We use the relation, $\text{assoc}(C) = \text{vol}_d(C) - \text{cut}(C, V \backslash C)$, where $d_i = \sum_{j=1}^{n} w_{ij}$ is the degree of vertex i and $\text{cut}(A, B) := \sum_{i \in A, j \in B} w_{ij}$.

Discussion of density based objective. As pointed out in [12], the density based objective possesses useful properties like strict monotonicity and robustness. In case of the density based objective, if an edge gets added (because of a new collaboration) or deleted (because of newly found incompatibility) the density of the subgraphs involving this edge necessarily increases resp. decreases, which is not true for the diameter based objective. In contrast to density based objective, the impact of small changes in graph structure is more severe in the case of diameter objective [12].

The generalized density that we use here leads to further modeling freedom as it enables to give weights to the experts according to their expertise. By giving smaller weight to those with high expertise, one can obtain solutions that not only satisfy the given skill requirements but also give preference to the more competent team members (i.e. the ones having smaller weights).

Problem Formulation. Using the notation introduced above, an instance of the team formation problem based on the generalized density can be formulated as

$$\begin{aligned} &\max_{C \subseteq V} \frac{\text{assoc}(C)}{\text{vol}_g(C)} && (2) \\ \text{subject to}: \ & S \subseteq C \\ & \kappa_j \leq \text{vol}_{M_j}(C) \leq \iota_j, \quad \forall j \in \{1, \ldots, p\} \\ & |C| \leq b \\ & \text{vol}_c(C) \leq B \\ & \text{dist}(u, v) \leq d_0, \quad \forall u, v \in C, \end{aligned}$$

Note that the upper bound constraints on the team size and the budget can be rewritten as skill constraints and can be incorporated into the skill matrix M accordingly. Thus, without loss of generality, we omit the budget and size constraints from now on, for the sake of brevity. Moreover, since S is required to be part of the solution, we can assume that $dist(u, v) \leq d_0, \forall u, v \in S$, otherwise the above problem is infeasible. The distance constraint also implies that any $u \in V$ for which $\text{dist}(u, s) > d_0$, for some $s \in S$, cannot be a part of the solution. Thus, we again assume wlog that there is no such $u \in V$; otherwise such vertices can be eliminated without changing the solution of problem (2).

Our formulation (2) is a generalized version of the classical densest subgraph problem (DSP), which has many applications in graph analysis, e.g., see [19]. The simplest version of DSP is the problem of finding a densest subgraph (without any constraints on the solution), which can be solved optimally in polynomial time [13]. The densest-k-subgraph problem, which requires the solution to contain exactly k vertices, is a notoriously hard problem in this class and has been shown not to admit a polynomial time approximation scheme [16]. Recently, it has been shown that the densest subgraph problem with an upper bound on the size is as hard as the densest-k-subgraph problem [17]. However, the densest subgraph problem with a lower bound constraint has a 2-approximation algorithm [17]. It is based on solving a sequence of unconstrained densest subgraph problems. They also show that there exists a linear programming relaxation for this problem achieving the same approximation guarantee.

Recently [12] considered the following generalized version of the densest subgraph problem with lower bound constraints in the context of team formation problem:

$$\begin{aligned} &\max_{C \subseteq V} \frac{\text{assoc}(C)}{\text{vol}_g(C)} && (3) \\ \text{subject to}: \ & \text{vol}_{M_j}(C) \geq \kappa_j, \quad \forall j \in \{1, \ldots, p\} \end{aligned}$$

where M is the *binary* skill matrix. They extend the greedy method of [17] and show that it achieves a 3-approximation guarantee for some special cases of this problem. [8] recently improved the approximation guarantee of the greedy algorithm of [12] for problem (3) to a factor 2. The time complexity of this greedy algorithm is $O(kn^3)$, where n is the number of experts and $k := \sum_{j=1}^{m} k_j$ is the minimum number of experts required.

Direct integration of subset constraint. The subset constraint can be integrated into the objective by directly working on the subgraph G' induced by the vertex set $V' =$

[1] The distance function need not satisfy the triangle inequality.

$V \backslash S$. Note that any $C \subset V$ that contains S can be written as $C = A \cup S$, for $A \subset V'$. We now reformulate the team formation problem on the subgraph G'. We introduce the notation $m = |V'|$, and we assume wlog that the first m entries of V are the ones in V'.

The terms in problem (2) can be rewritten as

$$\begin{aligned}\mathrm{assoc}(C) &= \mathrm{assoc}(A) + \mathrm{assoc}(S) + 2\ \mathrm{cut}(A, S),\\ &= \mathrm{vol}_d(A) - \mathrm{cut}(A, V \backslash A) + \mathrm{assoc}(S) + 2\ \mathrm{cut}(A, S)\\ &= \mathrm{vol}_d(A) - \mathrm{cut}(A, V' \backslash A) + \mathrm{assoc}(S) + \mathrm{cut}(A, S)\\ \mathrm{vol}_g(C) &= \mathrm{vol}_g(A) + \mathrm{vol}_g(S)\end{aligned}$$

Moreover, note that we can write: $\mathrm{cut}(A, S) = \mathrm{vol}_{d^S}(A)$, where $d_i^S = \sum_{j \in S} w_{ij}$ denotes the degree of vertex i restricted to the subset S in the original graph. Using the abbreviations, $\mu_S = \mathrm{assoc}(S)$, $\nu_S = \mathrm{vol}_g(S)$, $\mathrm{assoc}_S(A) = \mathrm{vol}_d(A) - \mathrm{cut}(A, V' \backslash A) + \mu_S + \mathrm{vol}_{d^S}(A)$, we rewrite the team formation problem (2) as

$$\begin{aligned}&\max_{A \subseteq V',\ A \neq \emptyset} \frac{\mathrm{assoc}_S(A)}{\mathrm{vol}_g(A) + \nu_S} \qquad \text{(GDSP)}\\ &\text{subject to}: k_j \leq \mathrm{vol}_{M_j}(A) \leq l_j, \quad \forall j \in \{1, \ldots, p\}\\ &\qquad \mathrm{dist}(u, v) \leq d_0, \quad \forall u, v \in A,\end{aligned}$$

where for all $j = 1, \ldots, p$, the bounds were updated as $k_j = \kappa_j - \mathrm{vol}_{M_j}(S)$, $l_j = \iota_j - \mathrm{vol}_{M_j}(S)$. Note that here we already used the assumption: $\mathrm{dist}(u, s) \leq d_0, \forall u \in V, \forall s \in S$. The constraint, $A \neq \emptyset$, has been introduced for technical reasons required for the formulation of the continuous problem in Section 4.2. The equivalence of problem (GDSP) to (2) follows by considering either S (if feasible) or the set $A^* \cup S$, where A^* is an optimal solution of (GDSP), depending on whichever has higher density.

To the best of our knowledge there is no greedy algorithm with an approximation guarantee to solve problem (GDSP). Instead of designing a greedy approximation algorithm for this discrete optimization problem, we derive an *equivalent* continuous optimization problem in Section 4. That is, we reformulate the discrete problem in continuous space while preserving the optimality of the solutions of the discrete problem. The rationale behind this approach is that the continuous formulation is more flexible and allows us to choose from a larger set of methods for its solution than for the discrete one. Although the resulting continuous problem is as hard as the original discrete problem, recent progress in continuous optimization [14] allow us to find a locally optimal solution very efficiently.

4. DERIVATION OF FORTE

In this section we present our method, *Formation Of Realistic Teams* (**FORTE**, for short) to solve the team formation problem, which is rewritten as (GDSP), using the continuous relaxation. We derive **FORTE** in three steps:

i. Derive an equivalent unconstrained discrete problem (4) of the team formation problem (GDSP) via an *exact penalty* approach.

ii. Derive an equivalent continuous relaxation (6) of the unconstrained problem (4) by using the concept of *Lovasz extensions.*

iii. Compute the solution of the continuous problem (6) using the recent method RatioDCA from *fractional programming.*

4.1 Equivalent Unconstrained Problem

A general technique in constrained optimization is to transform the constrained problem into an equivalent unconstrained problem by adding to the objective a penalty term, which is controlled by a parameter $\gamma \geq 0$. The penalty term is zero if the constraints are satisfied at the given input and strictly positive otherwise. The choice of the regularization parameter γ influences the tradeoff between satisfying the constraints and having a low objective value. Large values of γ tend to enforce the satisfaction of constraints. In the following we show that for the team formation problem (GDSP) there exists a value of γ that guarantees the satisfaction of all constraints.

Let us define the penalty term for constraints of the team formation problem (GDSP) as

$$\mathrm{pen}(A) := \begin{cases} \sum_{j=1}^{p} \max\{0, \mathrm{vol}_{M_j}(A) - l_j\} & \\ + \sum_{j=1}^{p} \max\{0, k_j - \mathrm{vol}_{M_j}(A)\} & \\ + \sum_{u,v \in A} \max\{0,\ \mathrm{dist}(u, v) - d_0\} & A \neq \emptyset \\ 0 & A = \emptyset. \end{cases}$$

Note that the above penalty function is zero only when A satisfies the constraints; otherwise it is strictly positive and increases with increasing infeasibility. The special treatment of the empty set is again a technicality required later for the Lovasz extensions, see Section 4.2. For the same reason, we also replace the constant terms μ_S and ν_S in (GDSP) by $\mu_S\,\mathrm{unit}(A)$ and $\nu_S\,\mathrm{unit}(A)$ respectively, where $\mathrm{unit}(A) := 1, A \neq \emptyset$ and $\mathrm{unit}(\emptyset) = 0$.

The following theorem shows that there exists an unconstrained problem equivalent to the constrained optimization problem (GDSP).

THEOREM 1. *The constrained problem (GDSP) is equivalent to the unconstrained problem*

$$\min_{\emptyset \neq A \subseteq V} \frac{\mathrm{vol}_g(A) + \nu_S\,\mathrm{unit}(A) + \gamma\,\mathrm{pen}(A)}{\mathrm{assoc}_S(A)} \qquad (4)$$

for $\gamma > \frac{\mathrm{vol}_d(V)}{\theta}\ \frac{\mathrm{vol}_g(A_0) + \nu_S}{\mathrm{assoc}_S(A_0)}$, *where* A_0 *is any feasible set of problem (GDSP) such that* $\mathrm{assoc}_S(A_0) > 0$ *and* θ *is the minimum value of infeasibility, i.e.,* $\mathrm{pen}(A) \geq \theta$, *if* A *is infeasible.*

PROOF. *We define* $\mathrm{spvol}(A) := \frac{\mathrm{vol}_g(A) + \nu_S\,\mathrm{unit}(A)}{\mathrm{assoc}_S(A)}$. *Note that maximizing (GDSP) is the same as minimizing* $\mathrm{spvol}(A)$ *subject to the constraints of (GDSP). For any feasible subset A, the objective of (4) is equal to* $\mathrm{spvol}(A)$, *since the penalty term is zero. Thus, if we show that all minimizers of (4) satisfy the constraints then the equivalence follows. Suppose, for the sake of contradiction, that* $A^* (\neq \emptyset$, *if* $S = \emptyset)$ *is a minimizer of (4) and that* A^* *is infeasible for problem (GDSP). Since* $\nu_S \geq 0$ *and* $g_i > 0, \forall i$, *we have under the given condition on* γ,

$$\begin{aligned}&\frac{\mathrm{vol}_g(A^*) + \nu_S + \gamma\ \mathrm{pen}(A^*)}{\mathrm{assoc}_S(A^*)} > \frac{\gamma\ \mathrm{pen}(A^*)}{\mathrm{assoc}_S(A^*)}\\ &\geq \frac{\gamma\,\theta}{\max_{A \subseteq V} \mathrm{assoc}_S(A)} \geq \frac{\gamma\,\theta}{\mathrm{vol}_d(V)} > \frac{\mathrm{vol}_g(A_0) + \nu_S}{\mathrm{assoc}_S(A_0)},\end{aligned}$$

which leads to a contradiction because the last term is the objective value of (4) at A_0. □

4.2 Equivalent Continuous Problem

We will now derive a tight continuous relaxation of problem (4). This will lead us to a minimization problem over $\mathbb{R}^m$, which then can be handled more easily than the original discrete problem. The connection between the discrete and the continuous space is achieved via thresholding. Given a vector $f \in \mathbb{R}^m$, one can define the sets

$$A_i := \{j \in V | f_j \geq f_i\}\,, \tag{5}$$

by thresholding f at the value f_i. In order to go from functions on sets to functions on continuous space, we make use of the concept of Lovasz extensions.

Definition 1. *(Lovasz extension) Let $R : 2^V \to \mathbb{R}$ be a set function with $R(\emptyset) = 0$, and let $f \in \mathbb{R}^m$ be ordered in ascending order $f_1 \leq f_2 \leq \cdots \leq f_m$. The Lovasz extension $R^L : \mathbb{R}^m \to \mathbb{R}$ of R is defined by*

$$R^L(f) = \sum_{i=1}^{m-1} R(A_{i+1})\,(f_{i+1} - f_i) + R(V)f_1.$$

Note that $R^L(\mathbf{1}_A) = R(A)$ for all $A \subset V$, i.e. R^L is indeed an extension of R from 2^V to $\mathbb{R}^V$ ($|V| = m$). In the following, given a set function R, we will denote its Lovasz extension by R^L. The explicit forms of the Lovasz extensions used in the derivation will be dealt with in Section 4.3.

In the following theorem we show the equivalence for GDSP. A more general result showing equivalence for fractional set programs can be found in [7].

Theorem 2. *The unconstrained discrete problem (4) is equivalent to the continuous problem*

$$\min_{f\in\mathbb{R}_+^{V'}} \frac{\mathrm{vol}_g^L(f) + \nu_S\,\mathrm{unit}^L(A) + \gamma\,\mathrm{pen}^L(f)}{\mathrm{assoc}_S^L(f)} \tag{6}$$

for any $\gamma \geq 0$. Moreover, optimal thresholding of a minimizer $f^* \in \mathbb{R}_+^m$,

$$A^* := \min_{A_i=\left\{j\in V'|f_j^*\geq f_i^*\right\},\, i=1,\ldots,m} \frac{\mathrm{vol}_g(A_i) + \nu_S + \gamma\,\mathrm{pen}(A_i)}{\mathrm{assoc}_S(A_i)},$$

yields a set A^* that is optimal for problem (4).

Proof. *Let $R(A) = \mathrm{vol}_g(A) + \nu_S\,\mathrm{unit}(A) + \gamma\,\mathrm{pen}(A)$. Then we have*

$$\min_{A\subset V'} \frac{R(A)}{\mathrm{assoc}_S(A)} = \min_{A\subset V'} \frac{R^L(\mathbf{1}_A)}{\mathrm{assoc}_S^L(\mathbf{1}_A)} \geq \min_{f\in\mathbb{R}_+^{V'}} \frac{R^L(f)}{\mathrm{assoc}_S^L(f)},$$

where in the first step we used the fact that $R^L(f)$ and $\mathrm{assoc}^L(f)$ are extensions of $R(A)$ and $\mathrm{assoc}(A)$, respectively. Below we first show that the above inequality also holds in the other direction, which then establishes that the optimum values of both problems are the same. The proof of the reverse direction will also imply that a set minimizer of the problem (4) can be obtained from any minimizer f^* of (6) via optimal thresholding.

We first show that the optimal thresholding of any $f \in \mathbb{R}_+^m$ yields a set A such that $\mathbf{1}_A$ has an objective value at least as good as the one of f. This holds because

$$\begin{aligned}
R^L(f) &= \sum_{i=1}^{m-1} R(A_{i+1})(f_{i+1} - f_i) + f_1 R(V') \\
&= \sum_{i=1}^{m-1} \frac{R(A_{i+1})}{\mathrm{assoc}_S(A_{i+1})}\mathrm{assoc}_S(A_{i+1})(f_{i+1} - f_i) \\
&\qquad + \frac{R(V')}{\mathrm{assoc}_S(V')}\mathrm{assoc}_S(V')f_1 \\
&\geq \min_{j=1,\ldots m} \frac{R(A_j)}{\mathrm{assoc}_S(A_j)} \\
&\qquad \Big(\sum_{i=1}^{m-1} \mathrm{assoc}_S(A_{i+1})(f_{i+1} - f_i) + \mathrm{assoc}_S(V')f_1\Big) \\
&= \min_{j=1,\ldots m} \frac{R(A_j)}{\mathrm{assoc}_S(A_j)}\mathrm{assoc}_S^L(f)
\end{aligned}$$

The third step follows from the fact that f is non-negative ($f_1 \geq 0$) and ordered in ascending order, i.e., $f_{i+1} - f_i \geq 0, \forall i = 1, \ldots, m-1$. Since $\mathrm{assoc}_S^L(f)$ is non-negative, the final step implies that

$$\frac{R^L(f)}{\mathrm{assoc}_S^L(f)} \geq \min_{j=1,\ldots m} \frac{R(A_j)}{\mathrm{assoc}_S(A_j)}. \tag{7}$$

Thus we have

$$\min_{f\in\mathbb{R}_+^{V'}} \frac{R^L(f)}{\mathrm{assoc}_S^L(f)} \geq \min_{A\subset V'} \frac{R(A)}{\mathrm{assoc}_S(A)}.$$

From inequality (7), it follows that optimal thresholding of f^* yields a set that is a minimizer of problem (4). □

Corollary 1. *The team formation problem (GDSP) is equivalent to the problem (6) if γ is chosen according to the condition given in Theorem 1.*

Proof. *This directly follows from Theorems 1 and 2.* □

While the continuous problem is as hard as the original discrete problem, recent ideas from continuous optimization [14] allow us to derive in the next section an algorithm for obtaining locally optimal solutions very efficiently.

4.3 Algorithm for the Continuous Problem

We now describe an algorithm for (approximately) solving the continuous optimization problem (6). The idea is to make use of the fact that the fractional optimization problem (6) has a special structure: as we will show in this section, it can be written as a special ratio of difference of convex (d.c.) functions, i.e. it has the form

$$\min_{f\in\mathbb{R}_+^{V}} \frac{R_1(f) - R_2(f)}{S_1(f) - S_2(f)} := Q(f), \tag{8}$$

where the functions R_1, R_2, S_1 and S_2 are positively one-homogeneous convex functions[2] and numerator and denominator are nonnegative. This reformulation then allows us to use a recent first order method called RatioDCA [14, 7].

In order to find the explicit form of the convex functions, we first need to rewrite the penalty term as $\mathrm{pen}(A) =$

[2] A function f is said to be positively one-homogeneous if $f(\alpha x) = \alpha f(x), \alpha \geq 0$.

$\mathrm{pen}_1(A) - \mathrm{pen}_2(A)$, where

$$\begin{aligned} \mathrm{pen}_1(A) &= \textstyle\sum_{j=1}^{p} \mathrm{vol}_{M_j}(A) + \sum_{j=1}^{p} k_j \,\mathrm{unit}(A), \\ \mathrm{pen}_2(A) &= \textstyle\sum_{j=1}^{p} \min\{l_j, \mathrm{vol}_{M_j}(A)\} + \sum_{j=1}^{p} \min\{k_j, \mathrm{vol}_{M_j}(A)\} \\ &\quad - \textstyle\sum_{u,v\in A} \max\{0,\ \mathrm{dist}(u,v) - d_0\}. \end{aligned}$$

Using this decomposition of $\mathrm{pen}(A)$, we can now write down the functions R_1, R_2, S_1 and S_2 as

$$\begin{aligned} R_1(f) &= \mathrm{vol}_\rho^L(f) + \sigma \max_i\{f_i\} \\ R_2(f) &= \gamma\, \mathrm{pen}_2^L(f) \\ S_1(f) &= \mathrm{vol}_d^L(f) + \mathrm{vol}_{d^S}^L(f) + \mu_S \max_i\{f_i\} \\ S_2(f) &= \mathrm{cut}^L(f). \end{aligned}$$

where $\rho := g + \gamma \sum_{j=1}^{p} M_j$, $\sigma := \nu_S + \gamma \sum_{j=1}^{p} k_j$, $\mathrm{pen}_2^L(f)$ denotes the Lovasz extension of $\mathrm{pen}_2(A)$, and

$$\begin{aligned} \mathrm{vol}_h^L(f) &= \langle (h_i)_{i=1}^m, f\rangle\,, \text{where } h \in \mathbb{R}^n, \\ \mathrm{cut}^L(f) &= \tfrac{1}{2} \textstyle\sum_{i,j=1}^{m} w_{ij}\, |f_i - f_j|. \end{aligned}$$

Lemma 1. *Using the functions R_1, R_2, S_1 and S_2 defined above, the problem (6) can be rewritten in the form* (8). *The functions R_1, R_2, S_1 and S_2 are convex and positively one-homogeneous, and $R_1 - R_2$ and $S_1 - S_2$ are nonnegative.*

Proof. The denominator of (6) is given as $\mathrm{assoc}_S^L(f) = \mathrm{vol}_d^L(f) - \mathrm{cut}^L(f) + \mathrm{vol}_{d^S}^L(f) + \mu_S\, \mathrm{unit}^L(f)$, and the numerator is given as $\mathrm{vol}_g^L(f) + \nu_S\, \mathrm{unit}^L(A) + \gamma\, \mathrm{pen}^L(f)$. Using Prop.2.1 in [3] and the decomposition of $\mathrm{pen}(A)$ introduced earlier in this section, we can decompose $\mathrm{pen}^L(f) = \mathrm{pen}_1^L(f) - \mathrm{pen}_2^L(f)$. The Lovasz extension of $\mathrm{pen}_1(A)$ is given as $\mathrm{pen}_1^L(f) = \sum_{j=1}^{p} \mathrm{vol}_{M_j}^L(f) + \sum_{j=1}^{p} k_j \max_i\{f_i\}$, and let $\mathrm{pen}_2^L(f)$ denote the Lovasz extension of $\mathrm{pen}_2(A)$ (an explicit form is not necessary, as shown later in this section). The equality between (6) and (8) then follows by simple rearranging of the terms.

The nonnegativity of the functions $R_1 - R_2$ and $S_1 - S_2$ follows from the nonnegativity of denominator and numerator of (6) and the definition of the Lovasz extension. Moreover, the Lovasz extensions of any set function is positively one-homogeneous [3].

Finally, the convexity of R_1 and S_1 follows as they are a non-negative combination of the convex functions $\max_i\{f_i\}$ and $\langle (h_i)_{i=1}^m, f\rangle$ for some $h \in \mathbb{R}^n$. The function $S_2(f) = \mathrm{cut}^L(f)$ is well-known to be convex [3]. To show the convexity of R_2, we will show that the function $\mathrm{pen}_2(A)$ is submodular[3]. The convexity then follows from the fact that a set function is submodular if and only if its Lovasz extension is convex [3]. For the proof of the submodularity of the first two sums one uses the fact that the pointwise minimum of a constant and a increasing submodular function is again submodular. Writing $D_{uv} := \max\{0,\ \mathrm{dist}(u,v) - d_0\}$, the last sum can be written as $-\sum_{u,v\in A} D_{uv} = -\sum_{u\in A, v\in V'} D_{uv} + \sum_{u\in A, v\in V'\setminus A} D_{uv}$. Using $(d_D)_i = \sum_j D_{ij}$, we can write its Lovasz extension as $-\mathrm{vol}_{d_D}(f) + \frac{1}{2}\sum_{i,j\in V'} D_{ij}\, |f_i - f_j|$, which is a sum of a linear term and a convex term. □

The reformulation of the problem in the form (8) enables us to apply a modification of the recently proposed RatioDCA [14, 7], a method for the *local* minimization of objectives of the form (8) on the whole $\mathbb{R}^m$. Given an initialization f_0, the above algorithm solves a sequence of convex optimization problems (line 3). Note that we do not need an explicit description of the terms $S_1(f)$ and $R_2(f)$, but only elements of their sudifferential $s_1(f) \in \partial S_1(f)$ resp. $r_2(f) \in \partial R_2(f)$. The explicit forms of the subgradients are given in the appendix. The convex problem (line 3) then has the form

RatioDCA [14] Minimization of a non-negative ratio of one-homogeneous d.c functions over $\mathbb{R}_+^m$

```
1: Initialization: $f^0 \in \mathbb{R}_+^m$, $\lambda^0 = Q(f^0)$
2: repeat
3:     $f^{l+1} = \arg\min_{u\in\mathbb{R}_+^m,\ \|u\|_2\le 1} R_1(u) + \lambda^l S_2(u) - \langle u, r_2(f^l) + \lambda^l s_1(f^l)\rangle$
       where  $r_2(f^l) \in \partial R_2(f^l)$, $s_1(f^l) \in \partial S_1(f^l)$
4:     $\lambda^{l+1} = Q(f^{l+1})$
5: until $\frac{|\lambda^{l+1} - \lambda^l|}{\lambda^l} < \epsilon$
```

$$\min_{f\in\mathbb{R}_+^m} \frac{\lambda^l}{2} \sum_{i,j=1}^{m} w_{ij}\, |f_i - f_j| + \langle f, c\rangle + \sigma\ \max_i\{f_i\}, \qquad (9)$$

where $c = \rho - r_2(f^l) - \lambda^l s_1(f^l)$. Note that (9) is a *non-smooth* problem. However, there exists an equivalent smooth dual problem, which we give below.

Lemma 2. *The problem* (9) *is equivalent to*

$$\min_{\substack{\|\alpha\|_\infty \le 1 \\ \alpha_{ij} = -\alpha_{ji}}} \ \min_{v\in S_m} \frac{1}{2} \left\| P_{\mathbb{R}_+^m}\left(-c - \frac{\lambda^l}{2} A\alpha - \sigma v\right)\right\|_2^2,$$

where $A : \mathbb{R}^E \mapsto \mathbb{R}^V$ with $(A\alpha)_i := \sum_j w_{ij}(\alpha_{ij} - \alpha_{ji})$, $P_{\mathbb{R}_+^m}$ denotes the projection on the positive orthant and S_m is the simplex $S_m = \{v \in \mathbb{R}^m \,|\, v_i \ge 0, \sum_{i=1}^m v_i = 1\}$.

Proof. First we use the homogenity of the objective in the inner problem to eliminate the norm constraint. This yields the equivalent problem

$$\min_{u\in\mathbb{R}_+^n} \sigma \max_i u_i + \frac{1}{2}\|u\|_2^2 + \langle u, c\rangle + \frac{\lambda^l}{2}\sum_{i,j=1}^{n} w_{ij} |u_i - u_j|.$$

We derive the dual problem as follows:

$$\begin{aligned} &\min_{u\in\mathbb{R}_+^n} \frac{\lambda^l}{2}\sum_{i,j=1}^{n} w_{ij}\, |u_i - u_j| + \sigma \max u_i + \frac{1}{2}\|u\|_2^2 + \langle u, c\rangle \\ &= \min_{u\in\mathbb{R}_+^n} \Big\{ \max_{\substack{\|\alpha\|_\infty \le 1 \\ \alpha_{ij} = -\alpha_{ji}}} \frac{\lambda^l}{2} \sum_{i,j=1}^{n} w_{ij}\,(u_i - u_j)\,\alpha_{ij} \\ &\qquad + \max_{v\in S_n} \sigma\,\langle u, v\rangle + \frac{1}{2}\|u\|_2^2 + \langle u, c\rangle \Big\} \\ &= \max_{\substack{\|\alpha\|_\infty \le 1 \\ \alpha_{ij} = -\alpha_{ji} \\ v\in S_n}} \ \min_{u\in\mathbb{R}_+^n} \frac{1}{2}\|u\|_2^2 + \left\langle u, c + \frac{\lambda^l}{2} A\alpha + \sigma v \right\rangle, \end{aligned}$$

where $(A\alpha)_i := \sum_j w_{ij}(\alpha_{ij} - \alpha_{ji})$. The optimization over u has the solution $u = P_{\mathbb{R}_+^n}(-c - \frac{\lambda^l}{2} A\alpha - \sigma v)$. Plugging u into the objective and using that $\langle P_{\mathbb{R}_+^n}(x), x\rangle = \|P_{\mathbb{R}_+^n}(x)\|_2^2$, we obtain the result. □

[3] A set function $R : 2^V \to \mathbb{R}$ is submodular if for all $A, B \subset V$, $R(A \cup B) + R(A \cap B) \le R(A) + R(B)$.

The smooth dual problem can be solved very efficiently using recent scalable first order methods like FISTA [6], which has a guaranteed convergence rate of $O(\frac{1}{k^2})$, where k is the number of steps done in FISTA. The main part in the calculation of FISTA consists of a matrix-vector multiplication. As the social network is typically sparse, this operation costs $O(m)$, where m is the number of non-zeros of W.

RatioDCA [14], produces a strictly decreasing sequence f^l, i.e., $Q(f^{l+1}) < Q(f^l)$, or terminates. This is a typical property of fast local methods in non-convex optimization. Moreover, the convex problem need not be solved to full accuracy; we can terminate the convex problem early, if the current f^l produces already sufficent descent in Q. As the number of required steps in the RatioDCA typically ranges between 5-20, the full method scales to large networks. Note that convergence to the global optimum of (8) cannot be guaranteed due to the non-convex nature of the problem. However, we have the following quality guarantee for the team formation problem.

THEOREM 3. *Let A_0 be a feasible set for the problem* (GDSP) *and γ is chosen as in Theorem 1. Let f^* denote the result of RatioDCA after initializing with the vector $\mathbf{1}_{A_0}$, and let A_{f^*} denote the set found by optimal thresholding of f^*. Either RatioDCA terminates after one iteration, or produces A_{f^*} which satisfies all the constraints of the team formation problem (GDSP) and*

$$\frac{\mathrm{assoc}_S(A_{f^*})}{\mathrm{vol}_g(A_{f^*}) + \nu_S} > \frac{\mathrm{assoc}_S(A_0)}{\mathrm{vol}_g(A_0) + \nu_S}.$$

PROOF. *RatioDCA generates a decreasing sequence $\{f^l\}$ such that $Q(f^{l+1}) < Q(f^l)$ until it terminates [14]. We have $Q(f^1) < Q(\mathbf{1}_{A_0})$, if the algorithm does not stop in one step. As shown in Theorem (2) optimal thresholding of f^1 yields a set A_f that achieves smaller objective on the corresponding set function. Since the chosen value of γ guarantees the satisfaction of the constraints, A_f has to be feasible.* □

5. LP RELAXATION OF GDSP

Recall that our team formation problem based on the density objective is rewritten as the following GDSP after integrating the subset constraint:

$$\begin{aligned} \max_{A \subseteq V'} \ & \frac{\mathrm{assoc}_S(A)}{\mathrm{vol}_g(A) + \nu_S} \qquad (10) \\ \text{subject to}: \ & k_j \le \mathrm{vol}_{M_j}(A) \le l_j, \quad \forall j \in \{1, \dots, p\} \\ & \mathrm{dist}(u, v) \le d_0, \quad \forall u, v \in A \end{aligned}$$

Note that here we do not require the additional constraint, $A \neq \emptyset$, that we added to (GDSP). In this section we show that there exists a Linear programming (LP) relaxation for this problem. The LP relaxation can be solved optimally in polynomial time and provides an upper bound on the optimum value of GDSP. In practice such an upper bound is useful to check the quality of the solutions found by approximation algorithms.

THEOREM 4. *The following LP is a relaxation of the Generalized Densest Subgraph Problem (10).*

$$\begin{aligned} \max_{t \in \mathbb{R},\ f \in \mathbb{R}^{V'},\ \alpha \in \mathbb{R}^{E'}} \ & \sum_{i,j=1}^{m} w_{ij}\alpha_{ij} + 2\ \left\langle d^S, f \right\rangle + t\mu_S \qquad (11) \\ \text{subject to}: \ & tk_j \le \langle M_j, f \rangle \le tl_j, \quad \forall j \in \{1, \dots, p\} \\ & f_u + f_v \le t, \quad \forall u, v : \mathrm{dist}(u, v) > d_0 \\ & t \ge 0, \quad \alpha_{ij} \le f_i,\ \alpha_{ij} \le f_j,\ \forall (i, j) \in E' \\ & 0 \le f_i \le t,\ \forall i \in V',\ \alpha_{ij} \ge 0,\ \forall (i, j) \in E' \\ & \langle g, f \rangle + t\nu_S = 1. \end{aligned}$$

where $V' = V \backslash S$, E' is the set of edges induced by V'.

PROOF. *The following problem is equivalent to (10), because (i) for every feasible set A of* (10), *there exist corresponding feasible y, X given by $y = \mathbf{1}_A$, $X_{ij} = \min\{y_i, y_j\}$, with the same objective value and (ii) an optimal solution of the following problem always satisfies $X^*_{ij} = \min\{y^*_i, y^*_j\}$.*

$$\begin{aligned} \max_{y \in \{0,\ 1\}^{V'},\ X \in \{0,\ 1\}^{E'}} \ & \frac{2\sum_{i<j} w_{ij}X_{ij} + 2\ \left\langle d^S, y \right\rangle + \mu_S}{\langle g, y \rangle + \nu_S} \\ \text{subject to}: \ & k_j \le \langle M_j, y \rangle \le l_j, \quad \forall j \in \{1, \dots, p\} \\ & y_u + y_v \le 1, \quad \forall u, v : \mathrm{dist}(u, v) > d_0 \\ & X_{ij} \le y_i, \quad X_{ij} \le y_j, \quad \forall (i, j) \in E' \end{aligned}$$

Relaxing the integrality constraints and using the substitution, $X_{ij} = \frac{\alpha_{ij}}{t}$ and $y_i = \frac{f_i}{t}$, we obtain the relaxation:

$$\begin{aligned} \max_{t \in \mathbb{R},\ f \in \mathbb{R}^{V'},\ \alpha \in \mathbb{R}^{E'}} \ & \frac{2\sum_{i<j} w_{ij}\alpha_{ij} + 2\ \left\langle d^S, f \right\rangle + t\mu_S}{\langle g, f \rangle + t\nu_S} \\ \text{subject to}: \ & tk_j \le \langle M_j, f \rangle \le tl_j, \quad \forall j \in \{1, \dots, p\} \\ & f_u + f_v \le t, \quad \forall u, v : \mathrm{dist}(u, v) > d_0 \\ & t \ge 0, \quad \alpha_{ij} \le f_i,\ \alpha_{ij} \le f_j,\ \forall (i, j) \in E' \\ & 0 \le f_i \le t,\ \forall i \in V',\ \alpha_{ij} \ge 0,\ \forall (i, j) \in E' \end{aligned}$$

Since this problem is invariant under scaling, we can fix the scale by setting the denominator to 1, which yields the equivalent LP stated in the theorem. □

Note that the solution f^* of the LP (11) is, in general, not integral, i.e., $f^* \notin \{0, 1\}^{V'}$. One can use standard techniques of randomized rounding or optimal thresholding to derive an integral solution from f^*. However, the resulting integral solution may not necessarily give a subset that satisfies the constraints of (10). In the special case when there are only lower bound constraints, i.e., problem (3), one can obtain a feasible set A for problem (3) by thresholding f^* (see (5)) according to the objective of (10). This is possible in this special case because there is always a threshold f^*_i which yields a non-empty subset A_i (in the worst case the full set V') satisfying all the lower bound constraints. In our experiments on problem (3), we derived a feasible set from the solution of LP in this fashion by choosing the threshold that yields a subset that satisfies the constraints and has the highest objective value.

Note that the LP relaxation (11) is vacuous with respect to upper bound constraints in the sense that given $f \in \mathbb{R}^m$ that does not satisfy the upper bound constraints of the LP (11) one can construct $\tilde{f}$, feasible for the LP by rescaling f without changing the objective of the LP. This implies that one can always transform the solution of the unconstrained problem into a feasible solution when there are *only* upper bound constraints. However, in the presence of lower bound or subset constraints, such a rescaling does not yield a feasible solution and hence the LP relaxation is useful on the instances of (10) with at least one lower bound or a subset constraint (i.e., $\nu_S > 0$).

6. EXPERIMENTS

We now empirically show that **FORTE** consistently produces high quality compact teams. We also show that the quality guarantee given by Theorem 3 is useful in practice as our method often improves a given sub-optimal solution.

6.1 Experimental Setup

Since we are not aware of any publicly available real world datasets for the team formation problem, we use, as in [12], a scientific collaboration network extracted from the DBLP database. Similar to [12], we restrict ourselves to four fields of computer science: Databases (DB), Theory (T), Data Mining (DM), Artificial Intelligence (AI). Conferences that we consider for each field are given as follows: DB = {SIGMOD, VLDB, ICDE, ICDT, PODS}, T = {SODA, FOCS, STOC, STACS, ICALP, ESA}, DM = {WWW, KDD, SDM, PKDD, ICDM, WSDM}, AI = {IJCAI, NIPS, ICML, COLT, UAI, CVPR}.

For our team formation problem, the skill set is given by $\mathcal{A}$ ={DB, T, DM, AI}. Any author who has at least three publications in any of the above 23 conferences is considered to be an expert. In our DBLP co-author graph, a vertex corresponds to an expert and an edge between two experts indicates prior collaboration between them. The weight of the edge is the number of shared publications. Since the resulting co-author graph is disconnected, we take its largest connected component (of size 9264) for our experiments.

Directly solving the non-convex problem (6) for the value of γ given in Theorem 1 often yields poor results. Hence in our implementation of **FORTE** we adopt the following strategy. We first solve the unconstrained version of problem (6) (i.e., $\gamma = 0$) and then iteratively solve (6) for increasing values of γ until all constraints are satisfied. In each iteration, we increase γ only for those constraints which were infeasible in the previous iteration; in this way, each penalty term is regulated by different value of γ. Moreover, the solution obtained in the previous iteration of γ is used as the starting point for the current iteration.

6.2 Quantitative Evaluation

In this section we perform a quantitative evaluation of our method in the special case of the team formation problem with lower bound constraints and $g_i = 1 \, \forall i$ (problem (3)). We evaluate the performance of our method against the greedy method proposed in [12], refered to as **mdAlk**. Similar to the experiments of [12], an expert is defined to have a skill level of 1 in skill j, if he/she has a publication in any of the conferences corresponding to the skill j. As done in [12], we create random tasks for different values of skill size, $k = \{3, 8, 13, 18, 23, 28\}$. For each value of k we sample k skills with replacement from the skill set $\mathcal{A}$ = {DB, T, DM, AI}. For example if $k = 3$, a sample might contain {DB, DB, T}, which means that the random task requires at least two experts from the skill DB and one expert from the skill T.

In Figure 1, we show for each method the densities, sizes and runtimes for the different skill sizes k, averaged over 10 random runs. In the first plot, we also show the optimal values of the LP relaxation in (11). Note that this provides an upper bound on the optimal value of (GDSP). We can obtain feasible solutions from the LP relaxation of (GDSP) via thresholding (see Section 5), which are shown in the plot as **LPfeas**. Furthermore, the plots contain the results obtained when the solutions of **LPfeas** and **mdAlk** are used as the initializations for **FORTE** (in each of the γ iteration).

The plots show that **FORTE** always produces teams of higher densities and smaller sizes compared to **mdAlk** and **LPfeas**. Furthermore, **LPfeas** produces better results than the greedy method in several cases in terms of densities and sizes of the obtained teams. The results of **mdAlk+FORTE** and **LPfeas+FORTE** further show that our method is able improve the sub-optimal solutions of **mdAlk** and **LPfeas** significantly and achieves almost similar results as that of **FORTE** which was started with the unconstrained solution of (6). Under the worst-case assumption that the upper bound on (GDSP) computed using the LP is the optimal value, the solution of **FORTE** is 94% − 99% optimal (depending on k).

6.3 Qualitative Evaluation

In this experiment, we assess the quality of the teams obtained for several tasks with different skill requirements. Here we consider the team formation problem (GDSP) in its more general setting. We use the generalized density objective of (1) where each vertex is given a rank r_i, which we define based on the number of publications of the corresponding expert. For each skill, we rank the experts according to the number of his/her publications in the conferences corresponding to the skill. In this way each expert gets four different rankings; the total rank of an expert is then the minimum of these four ranks. The main advantage of such a ranking is that the experts that have higher skill are given preference, thus producing more competent teams. Note that we choose a relative measure like rank as the vertex weights instead of an absolute quantity like number of publications, since the distribution of the number of publications varies between different fields. In practice such a ranking is always available and hence, in our opinion, should be incorporated.

Furthermore, in order to identify the main area of expertise of each expert, we consider his/her relative number of publications. Each expert is defined to have a skill level of 1 in skill j if he has more than 25% of his/her publications in the conferences corresponding to skill j. As a distance function between authors, we use the shortest path on the *unweighted version* of the DBLP graph, i.e. two experts are at a distance of two, if the shortest path between the corresponding vertices in the unweighted DBLP graph contains two edges. Note that in general the distance function can come from other general sources beyond the input graph, but here we had to rely on the graph distance because of lack of other information.

In order to assess the *competence* of the found teams, we use the list of the 10000 most cited authors of Citeseer [1]. Note that in contrast to the skill-based ranking discussed above, this list is only used in the evaluation and *not* in the construction of the graph. We compute the average inverse rank as in [12] as $AIR := 1000 \cdot \sum_{i=1}^{k} \frac{1}{R_i}$, where k is the size of the team and R_i is the rank of expert i on the Citeseer list of 10000 most cited authors. For authors not contained

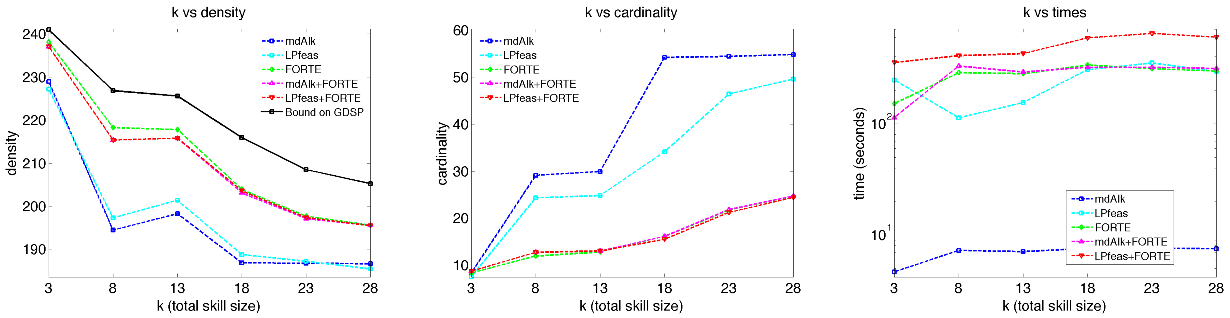

Figure 1: Densities, team sizes and runtimes of mdAlk, our method (FORTE), a feasible point constructed from the LP (LPfeas), and FORTE initialized with LPfeas and mdAlk, averaged over 10 trials. All versions of (FORTE) significantly outperform mdAlk, and LPfeas both in terms of densities and sizes of the teams found. The densities of FORTE are close to the upper bound on the optimum of the GDSP given by the LP.

on the list we set $R_i = 10001$. We also report the densities of the teams found in order to assess their *compatibility*.

We create several tasks with various constraints and compare the teams produced by **FORTE**, **mdAlk** and **LPfeas** (feasible solution derived from the LP relaxation). Note that in our implementation we extended the **mdAlk** algorithm of [12] to incorporate general vertex weights, using Dinkelbach's method from fractional programming [11]. The results for these tasks are shown in Table 1. We report the upper bound given by the LP relaxation, density value, AIR as well as number and sizes of the connected components. Furthermore, we give the names and the Citeseer ranks of the team members who have rank at most 1000. Note that **mdAlk** could only be applied to some of the tasks and **LPfeas** failed to find a feasible team in several cases.

As a first task we show the unconstrained solution where we maximize density without any constraints. Note that this problem is optimally solvable in polynomial time and all methods find the optimal solution. The second task asks for at least three experts with skill DB. Here again all methods return the same team, which is indeed optimal since the LP bound agrees with the density of the obtained team.

Next we illustrate the usefulness of the additional modeling freedom of our formulation by giving an example task where obtaining meaningful, connected teams is not possible with the lower bound constraints alone. Consider a task where we need at least four experts having skill AI (Task 3). For this, all methods return the same disconnected team of size seven where only four members have the skill AI. The other three experts possess skills DB and DM and are densely connected among themselves. One can see from the LP bound that this team is again optimal. This example illustrates the major drawback of the density based objective which while preferring higher density subgraphs compromises on the connectivity of the solution. Our further experiments revealed that the subgraph corresponding to the skill AI is less densely connected (relative to the other skills) and forming coherent teams in this case is difficult without specifying additional requirements. With the help of subset and distance based constraints supported by **FORTE**, we can now impose the team requirements more precisely and obtain meaningful teams. In Task 4, we require that Andrew Y. Ng is the team leader and that all experts of the team should be within a distance of two from each other in terms of the underlying co-author graph. The result of our method is a densely connected and highly ranked team of size four with a density of 3.89. Note that this is very close to the LP bound of 3.91. The feasible solution obtained by **LPfeas** is worse than our result both in terms of density and AIR. The greedy method **mdAlk** cannot be applied to this task because of the distance constraint. In Task 5 we choose Bernhard Schoelkopf as the team leader while keeping the constraints from the previous task. Out of the three methods, only **FORTE** can solve this problem. It produces a large disconnected team, many members of which are highly skilled experts from the skill DM and have strong connections among themselves. To filter these densely connected members of high expertise, we introduce a budget constraint in Task 6, where we define the cost of the team as the total number of publications of its members. Again this task can be solved only by **FORTE** which produces a compact team of four well-known AI experts. A slightly better solution is obtained when **FORTE** is initialized with the infeasible solution of the LP relaxation as shown (only in this task). This is an indication that on more difficult instances of (GDSP), it pays off to run **FORTE** with more than one starting point to get the best results. The solution of the LP, possibly infeasible, is a good starting point apart from the unconstrained solution of (6).

Tasks 7, 8 and 9 provide some additional teams found by **FORTE** for other tasks involving upper and lower bound constraints on different skills. As noted in Section 5 the LP bound is loose in the presence of upper bound constraints and this is also the reason why it was not possible to derive a feasible solution from the LP relaxation in these cases. In fact the LP bounds for these tasks remain the same even if the upper bound constraints are dropped from these tasks.

7. CONCLUSIONS

By incorporating various realistic constraints we have made a step forward towards a realistic formulation of the team formation problem. Our method finds qualitatively better teams that are more compact and have higher densities than those found by the greedy method [12]. Our linear programming relaxation not only allows us to check the solution quality but also provides a good starting point for our non-convex method. However, arguably, a potential downside of a density-based approach is that it does not guarantee connected components. A further extension of our approach could aim at incorporating "connectedness" or a relaxed version of it as an additional constraint.

Task	**FORTE**	**mdAlk**	**LPfeas**
Task 1: Unconstrained (LP bound: 32.7)	#Comps: 1 (2) Density: 32.7 AIR: 11.1 Jiawei Han (54), Philip S. Yu (279)	#Comps: 1 (2) Density: 32.7 AIR: 11.1 Jiawei Han (54), Philip S. Yu (279)	#Comps: 1 (2) Density: 32.7 AIR: 11.1 Jiawei Han (54), Philip S. Yu (279)
Task 2: DB≥3 (LP bound: 29.8)	#Comps: 1 (3) Density: 29.8 AIR: 7.56 Jiawei Han (54), Philip S. Yu (279) **(+1)**	#Comps: 1 (3) Density: 29.8 AIR: 7.56 Jiawei Han (54), Philip S. Yu (279) **(+1)**	#Comps: 1 (3) Density: 29.8 AIR: 7.56 Jiawei Han (54), Philip S. Yu (279) **(+1)**
Task 3: AI≥4 (LP bound: 16.6)	#Comps: 3 (1,3,3) Density: 16.6 AIR: 10.3 Michael I. Jordan (28), *Jiawei Han (54)*, Daphne Koller (127), *Philip S. Yu (279)*, Andrew Y. Ng (345), Bernhard Schoelkopf (364) **(+1)**	#Comps: 3 (1,3,3) Density: 16.6 AIR: 10.3 Michael I. Jordan (28), *Jiawei Han (54)*, Daphne Koller (127), *Philip S. Yu (279)*, Andrew Y. Ng (345), Bernhard Schoelkopf (364) **(+1)**	#Comps: 3 (1,3,3) Density: 16.6 AIR: 10.3 Michael I. Jordan (28), *Jiawei Han (54)*, Daphne Koller (127), *Philip S. Yu (279)*, Andrew Y. Ng (345), Bernhard Schoelkopf (364) **(+1)**
Task 4: AI≥4, $\mathrm{dist}_G(u,v) \le 2$, S={Andrew Ng} (LP bound: 3.91)	#Comps: 1 (4) Density: 3.89 AIR: 14.2 Michael I. Jordan (28), Sebastian Thrun (97), Daphne Koller (127), Andrew Y. Ng (345)		#Comps: 1 (6) Density: 3.5 AIR: 12.5 Michael I. Jordan (28), Geoffrey E. Hinton (61), Sebastian Thrun (97), Daphne Koller (127), Andrew Y. Ng (345), Zoubin Ghahramani (577)
Task 5: AI≥4, $\mathrm{dist}_G(u,v) \le 2$, S={B.Schoelkopf} (LP bound: 6.11)	#Comps: 2 (11,1) Density: 3.54 AIR: 3.94 *Jiawei Han (54)*, *Christos Faloutsos (140)*, Thomas S. Huang (146), *Philip S. Yu (279)*, *Zheng Chen (308)*, Bernhard Schoelkopf (364), *Wei-Ying Ma (523)*, *Ke Wang (580)* **(+4)**		
Task 6: AI≥4, $\mathrm{dist}_G(u,v) \le 2$, S={B.Schoelkopf}, $\sum_i c_i \le 255$ (LP bound: 2.06)	#Comps: 1 (4) Density: 1.24 AIR: 1.82 Alex J. Smola (335), Bernhard Schoelkopf (364) **(+2)** LP+FORTE: #Comps: 2 (2,2) Density: 1.77 AIR: 2.73 Robert E. Schapire (293), Alex J. Smola (335), Bernhard Schoelkopf (364), Yoram Singer (568)		
Task 7: 3≤DB≤6, DM≥10, (LP bound: 11.3)	#Comps: 1 (10) Density: 9.52 AIR: 4.96 Haixun Wang (50), Jiawei Han (54), Philip S. Yu (279), Zheng Chen (308), Ke Wang (580) **(+5)**		
Task 8: 2≤DB≤5, 10≤DM≤15, 5≤AI≤10 (LP bound: 10.7)	#Comps: 3 (1,12,3) Density: 7.4 AIR: 5.06 Michael I. Jordan (28), Jiawei Han (54), Daphne Koller (127), Philip S. Yu (279), Zheng Chen (308), Andrew Y. Ng (345), Bernhard Schoelkopf (364), Wei-Ying Ma (523), Divyakant Agrawal (591) **(+7)**		
Task 9: AI≤2, T≥2, |C|≤6 (LP bound: 19)	#Comps: 3 (2,2,2) Density: 6.17 AIR: 1.53 Didier Dubois (426), Micha Sharir (447), *Divyakant Agrawal (591)*, Henri Prade (713), Pankaj K. Agarwal (770) **(+1)**		

Table 1: Teams formed by FORTE, mdAlk and LPfeas for various tasks. We list the number and sizes of the found components, the (generalized) maximum density as well as the average inverse rank (AIR) based on the Citeseer list. Finally, we give name and rank of each team member with rank at most 1000. Experts who do not have the skill required by the task but are still included in the team are shown in *italic font.*

Acknowledgements

We gratefully acknowledge support from the Excellence Cluster MMCI at Saarland University funded by the German Research Foundation (DFG) and the project NOLEPRO funded by the European Research Council (ERC).

APPENDIX

The subgradient of $S_1(f)$ is given by $s_1(f) = d + d^S + \mu_S I_{max}(f)$, where $I_{max}(f)$ is the indicator function of the largest entry of f. For the subgradient of R_2, using Prop. 2.2. in [3], we obtain for the subgradient $t_{(l_j,M_j)}$ of the terms of the form $\min\{l_j, \mathrm{vol}_{M_j}(A)\}$,

$$\left(t_{(l_j,M_j)}(f)\right)_i = \begin{cases} 0 & \mathrm{vol}_{M_j}(A_{i+1}) > l_j \\ l_j - \mathrm{vol}_{M_j}(A_{i+1}) & \mathrm{vol}_{M_j}(A_i) \ge l_j, \\ & \mathrm{vol}_{M_j}(A_{i+1}) \le l_j \\ M_{ij} & \mathrm{vol}_{M_j}(A_i) < l_j \end{cases}.$$

Defining $D_{uv} := \max\{0,\ \mathrm{dist}(u,v) - d_0\}$, an element of the subgradient of the second term of R_2 is given as $d_D - p(f)$, where $(d_D)_i = \sum_j D_{ij}$ and $p(f)_i \in \left\{ \sum_{j=1}^m D_{ij} u_{ij} \,|\, u_{ij} = -u_{ji}, u_{ij} \in \mathrm{sign}(f_i - f_j) \right\}$, where $\mathrm{sgn}(x) := +1$, if $x > 0$; -1 if $x < 0$; $[-1,1]$, if $x = 0$. In total, we obtain for the subgradient $r_2(f)$ of $R_2(f)$,

$$r_2(f) = \gamma \sum_{j=1}^p t_{(l_j,M_j)}(f) + \gamma \sum_{j=1}^p t_{(k_j,M_j)}(f) + \gamma(p(f) - d_D).$$

1. REFERENCES

[1] Citeseer statistics – Most cited authors in computer science. *http://citeseerx.ist.psu.edu/stats/authors?all=true*.

[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *WWW*, pages 839–848, 2012.

[3] F. Bach. Learning with submodular functions: A convex optimization perspective. *CoRR*, abs/1111.6453, 2011.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, pages 44–54, 2006.

[5] A. Baykasoglu, T. Dereli, and S. Das. Project team selection using fuzzy optimization approach. *Cybern. Syst.*, 38(2):155–185, 2007.

[6] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Processing*, 18(11):2419–2434, 2009.

[7] T. Bühler, S. Rangapuram, M. Hein, and S. Setzer. Constrained fractional set programs and their application in local clustering and community detection. In *ICML*, pages 624–632, 2013.

[8] V. T. Chakaravarthy, N. Modani, S. R. Natarajan, S. Roy, and Y. Sabharwal. Density functions subject to a co-matroid constraint. In *FSTTCS*, pages 236–248, 2012.

[9] S. J. Chen and L. Lin. Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. *IEEE Trans. Engineering Management*, 51(2):111–124, 2004.

[10] N. Contractor. Some assembly required: leveraging web science to understand and enable team assembly. *Physical and Engineering Sciences*, 371(1987), 2013.

[11] W. Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967.

[12] A. Gajewar and A. D. Sarma. Multi-skill collaborative teams based on densest subgraphs. In *SDM*, pages 165–176, 2012.

[13] A. V. Goldberg. Finding a maximum density subgraph. Technical Report UCB/CSD-84-171, EECS Department, University of California, Berkeley, 1984.

[14] M. Hein and S. Setzer. Beyond spectral clustering - tight relaxations of balanced graph cuts. In *NIPS*, pages 2366–2374, 2011.

[15] M. Kargar and A. An. Discovering top-k teams of experts with/without a leader in social networks. In *CIKM*, pages 985–994, 2011.

[16] S. Khot. Ruling out ptas for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM J. Comput.*, 36(4), 2006.

[17] S. Khuller and B. Saha. On finding dense subgraphs. In *ICALP*, pages 597–608, 2009.

[18] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *KDD*, pages 467–476, 2009.

[19] B. Saha, A. Hoch, S. Khuller, L. Raschid, and X.-N. Zhang. Dense subgraphs with restrictions and applications to gene annotation graphs. In *RECOMB*, pages 456–472, 2010.

[20] H. Wi, S. Oh, J. Mun, and M. Jung. A team formation model based on knowledge and collaboration. *Expert Syst. Appl.*, 36(5):9121–9134, 2009.

[21] A. Zzkarian and A. Kusiak. Forming teams: an analytic approach. *IIE Trans.*, 31(1):85–97, 2004.