Enterprise Employee Training via Project Team Formation

Jiawei Zhang
University of Illinois at Chicago
Chicago, IL, USA
jzhan9@uic.edu

Philip S. Yu
University of Illinois at Chicago,
Chicago, IL, USA
psyu@cs.uic.edu

Yuanhua Lv
Microsoft Research
Redmond, WA, USA
yuanhual@microsoft.com

ABSTRACT

Professional career training for novice employees at elementary levels to help them master necessary working skills is critical for both achieving employees' professional success and enhancing the enterprise growth. Besides adopting professional services from external career training agencies, companies can actually train the employees more effectively by involving them in various internal projects carried out in the company. In this paper, we will study the "Employee Training" (ET) problem by assigning the employees to various concrete company internal projects. From the company perspective, besides training the employees, another important objective of carrying out these projects is to finish them successfully. The successful accomplishment of the projects depends on various issues, like the skill qualification of the built teams and the effective collaboration among the team members. To achieve these two objectives simultaneously, a novel framework named "Team foRmAtion based employee traINing" (TRAIN) is proposed in this paper. TRAIN formulates the ET problem as a joint optimization problem, where the objective function considers the employees' overall skill gain and the team internal communication costs at the same time. To ensure the success of the projects, a new team skill qualification constraint is proposed and added to the optimization problem. Extensive experiments conducted on the real-world enterprise employee project team dataset demonstrate the effectiveness of TRAIN in addressing the problem.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

Keywords

Employee Training; Team Formation; Enterprise Social Network; Data Mining

1. INTRODUCTION

Formally, the concept *training* denotes a function of human resource management concerned with organizational activities, which aims at improving the performance of individuals in the organizational settings [17, 16, 18]. Providing necessary professional career training for employees is very common in companies, which can help novice employees at elementary levels to master necessary working skills under the instructions of people from a high professional practitioner position. As the labour market becomes more saturated, employees are required to be proficient at skills of their expertise. Meanwhile, companies are also increasingly investing in the future of employee career success through training organizations and subsidized apprenticeship or traineeship initiatives. In companies, necessary professional career training of the employees is critical for both polishing individuals' professional skills and enhancing the enterprise growth.

Besides adopting professional services from external career training agencies, companies themselves can also train employees by involving them in projects carried out in the companies [17, 16, 18]. By collaborating with other employees to finish projects, employees can learn necessary professional skills from each other, which will effectively reduce the unnecessary expenditure and is also more practical for companies. Generally, each project carried out in companies needs a group of employees with certain expertise. Consider, for example, given a new data mining project requiring skills {*software engineering, data mining, human-computer interaction, algorithm, data visualization*}, the project manager needs to build a team of scientists and engineers with all these skills before carrying out the project. What's more, for the success of the projects, besides the team skill qualification issues, effective co-operation and communications among the team members is another important factor that the team manager needs to consider carefully in the team building.

Problem Studied: In this paper, we will study the employee training problem by partitioning employees into various concrete company internal projects. Formally, the problem is called the "Employee Training" (ET) problem. The objectives of the ET problem include both training the employees and finishing the projects simultaneously.

The ET problem is an important yet interesting problem. Besides training employees in companies [17, 16, 18], it can also be applied in many other concrete real-world problems. For instance, we can educate students in schools by involving them into study groups to do homework and projects together. Nowadays, education in many countries worldwide suffers from some serious problems: teachers just report a speech verbatim, while students will only memorize what the teachers say from the books. Students educated in this way lack both teamwork skills and the abilities to think independently. New ways to educate students will be helpful for both students' individual prosperity as well as the societal growth [2]. In addition, ET can also be applied in the training of (1) soldiers about new weapons and methods of warfare, (2) farmers about the latest cultivation methods, (3) novice and veteran players (in computer games) about the advanced game skills, and even (4) animals (e.g., police and rescue dogs) about various hunting and rescue skills.

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom
©2017 ACM. ISBN 978-1-4503-4675-7/17/02 $15.00
DOI: http://dx.doi.org/10.1145/3018661.3018682 .

Albeit its importance, ET is a novel problem and we are the first to study it in the enterprise context. ET is totally different from existing works, like (1) "*Grouping students in educational settings*" [2], which studies how to put students of different abilities levels together to let them learn from each other. The objective of [2] is to maximize the improvement of the students only, but it doesn't consider the enterprise context nor the success of the projects; (2) "*Expert team formation in social networks*" [7], which aims at selecting a subset of experts to form a qualified team for certain projects. There exists two significant difference between ET and [7]: Firstly, ET aims at partitioning the employees into different teams for several projects carried out in the company and all the employees will be assigned to at least one specific project, while [7] merely selects one team from the whole company for an input project; Secondly, besides the success of projects, another objective covered in ET is the training of employees, which is not touched in [7]; and (3) "*Social community detection*" [10], which merely focuses on partitioning close individuals into the same groups, and is totally different from ET in both the objective and problem setting. Besides these 3 works, more information about other related works is available in Section 5 and a recent survey paper [12].

In addition to its importance and novelty, the ET problem is very challenging to address due to the following reasons:

- *Skills Gain Quantification*: Grouping novice and senior employees together so that they can learn from each other to get trained is one of the main objective of the ET problem. A clear quantification of the skills gain of the employees can be the prerequisite for defining the ET problem in this paper.
- *Team Skills Qualification*: To finish the projects successfully, the expertise of the team members need to meet the skill requirements of the projects in both the skill categories and the skill proficiency levels. Skill qualification check of the formed teams is an important guarantee for the success of the projects from the technical perspective.
- *Enterprise Closeness Measure*: Another important factor determining the success of the projects is the effective collaboration and communication among the team members. A comprehensive measure of the enterprise closeness among the employees can help define communication costs better.
- *NP-Hardness*: The ET problem itself is a difficult problem, which is shown to be NP-hard. Therefore, no efficient algorithms exist that can address the ET problem in polynomial time, if P≠NP.

To resolve all the above challenges covered in the ET problem, a novel framework named "Team foRmAtion based employee traINing" (TRAIN) in proposed in this paper. TRAIN introduces many new concepts, e.g., *individual/team skill mastery level*, *individual skill improvement space*, and *individual skill gain* to help quantify the employees' skill improvement by joining in the project teams. A new team skill qualification constraint (of both skill categories and skill proficiency levels) is proposed when building the teams for each project in TRAIN. The closeness among the employees can be calculated with both their personal contacts in company-internal online social networks and their relative management relationships in the organizational chart [17]. The ET problem is mapped into a constrained integer programming problem by TRAIN, and TRAIN addresses the objective function approximately in two phases: constraint relaxation and post-processing of the results.

The remaining part of this paper is organized as follows. We describe the definitions of several concepts and formulate the ET problem in Section 2. The method is introduced in Section 3, which will be evaluated in Section 4. Finally, we talk about the related works in Section 5 and conclude this paper in Section 6.

2. PROBLEM FORMULATION

In this section, we will first define several important concepts used in this paper, and then provide the formulation of the ET problem.

2.1 Terminology Definition

In companies [17, 16, 18], employees usually have professional skills and expertise at different proficiency levels. Generally, senior employees who have gotten trained for a long time will know many skills very thoroughly. Meanwhile, the novice employees who have never participated in any real system/product development projects before may be not experienced enough, and the skills they have can be limited and not deep. Let $\mathcal{U} = \{u_1, u_2, \cdots, u_n\}$ be the set of employees in the company, and the whole skills involved in the ET problem can be represented as set $\mathcal{S} = \{s_1, s_2, \cdots, s_m\}$. To denote the proficiency levels of skills that the employees can master, we introduce the *employee skill mastery* concept in this paper.

Definition 1 (Employee Skill Mastery): Formally, the set of skills that employee $u_i \in \mathcal{U}$ knows can be represented as set $S(u_i) \subset \mathcal{S}$. For each skill s^j that u_i knows (i.e., $s^j \in S(u_i)$), we introduce the concept of *employee skill mastery level*, $level(u_i, s^j) = \xi_i^j \in [0, 1]$, to represent how proficiently u_i can master s^j.

Based on the above definition, given a skill s^j with certain required mastery level $\pi^j \in [0, 1]$, we can represent the set of qualified employees (i.e., employees who know skill s^j with at least level π^j) as s^j's support set $Sup(s^j, \pi^j) = \{u_i | u_i \in \mathcal{U} \wedge s^j \in S(u_i) \wedge \xi_i^j \geq \pi^j\}$. Meanwhile, regardless of the skill mastery level, the set of employees who have skill s^j can be represented as the support set of skill s^j: $Sup(s^j) = \{u_i | u_i \in \mathcal{U} \wedge s^j \in S(u_i)\}$. In companies, there usually exist many projects to be carried out at the same time, each of which will have a set of needed skills with required proficiency levels.

Definition 2 (Skill Specified Project): Formally, the set of projects carried out in the company can be represented as $\mathcal{P} = \{p_1, p_2, \cdots, p_k\}$. For each project $p_i \in \mathcal{P}$, the company will pre-specify the required skills to finish the project, which can be represented as set $S(p_i)$. For skill s^j needed in project p_i (i.e., $s^j \in S(p_i)$), its required proficiency level can be represented as $\pi_i^j \in [0, 1]$, which indicates the intrinsic challenges of the project in terms of skill s^j.

To ensure the success of the projects to be carried out, effective communications and co-operation among the team members are another important factor. Generally, employees who are familiar with each other may communicate and co-operate with each other much better. In the context of enterprises, the communication costs among the employees can be measured by using the enterprise information from various sources, e.g., company internal organizational chart which outlines the management structure of the whole company [17, 16, 18].

Definition 3 (Chart Based Communication Cost): Let the rooted tree $T = (\mathcal{U}, \mathcal{L}, root)$ denote the enterprise internal organizational chart involving employees $\mathcal{U}$, where $\mathcal{L}$ represents the directed management links from managers to subordinates and $root \in \mathcal{U}$ indicates the CEO of the company. The communication cost between employees $u, v \in \mathcal{U}$ calculated based on the organizational chart T, i.e., $cost_c(u, v)$, is called the *chart based communication cost* in this paper.

In addition, employees at workplaces nowadays are usually involved in a new type of online social networks, called *enterprise online social network* (ESNs) [17, 16, 18], to enjoy various kinds of provided professional services to deal with their daily working issues. Based on the social interaction among employees in the ESNs, we can represent the communication costs among employees to be:

Definition 4 (ESNs based Communication Cost): Let graph $G =$

$(\mathcal{U}, \mathcal{E})$ represent the enterprise social network launched in a company, where $\mathcal{E}$ denotes the set of interaction links (e.g., friendship links) among the employees in $\mathcal{U}$. The communication cost calculate between the employees $u, v \in \mathcal{U}$ based on the ESN G, i.e., $cost_e(u, v)$, is called the *ESN based communication cost* in this paper.

Generally, *chart based communication cost* denotes the professional closeness among employees at the workplace, while *ESN based communication cost* represents the casual interactions of employees in personal life. Depending on the specific information used in the cost calculation, the *chart based communication cost* and the *ESN based communication cost* measures can have different concrete representations, which will be introduced in Section 3.2.2 in detail.

Definition 5 (Enterprise Communication Cost): Let $u, v \in \mathcal{U}$ be two employees in the company. We can represent the *enterprise communication cost* between u and v as a linear combination of the costs calculated based on the offline organizational chart and online ESNs:

$$cost(u, v) = \omega \cdot cost_c(u, v) + (1 - \omega) \cdot cost_e(u, v),$$

where $\omega \in [0, 1]$ represents the weight of term $cost_c(u, v)$.

2.2 Problem Formulation

Based on the definitions introduced above, we will define the ET problem formally as follows:

Definition 6 (The ET Problem): For the given employee set $\mathcal{U}$ and project set $\mathcal{P}$, the ET problem studies how to partition employees $\mathcal{U}$ into teams $\mathcal{T} = \{T_1, T_2, \cdots, T_k\}$ involving both novice employees and senior employees for all projects in $\mathcal{P}$, where $T_i \subset \mathcal{U}$ denotes the team built for project p_i. Two objectives are covered in the ET simultaneously, which are (1) training the employees to help them improve skills, and (2) finishing the projects successfully with the least communication costs among the team members.

By discarding the training objective, the ET problem can be reduced to the traditional team formation problem [7], which is proved to be NP-hard when the communication costs are defined to be either the *diameter* or the *MST (minimum spanning tree) based cost*. Meanwhile, by neglecting the success of the projects objective, the ET will be reduced to the *l-Group* problem studied in [2], which has also been proved to be NP-hard when the skill improvement measure is to count the people having the improvement space in the group. Meanwhile, the ET is a more generalized problem about the training and collaboration problems in enterprises, and it is also much more difficult and challenging than the other two problems respectively.

Let $train(T_i, p_i)$ denote the skill gain of team members in T_i by participating in project p_i, function $qualify(T_i, p_i) = 1$ represent that team T_i can meet the skill requirements of project p_i, and $cost(T_i)$ be the communication cost among members in T_i. Formally, the ET problem aims at inferring the optimal teams $\mathcal{T}^*$ which can maximize the following objective function

$$\mathcal{T}^* = \arg \max_{\mathcal{T}=\{T_1, T_2, \cdots, T_k\}} \sum_{p_i \in \mathcal{P}} train(T_i, p_i) - \beta \cdot \sum_{p_i \in \mathcal{P}} cost(T_i),$$

$$s.t.\ qualify(T_i, p_i) = 1, \forall p_i \in \mathcal{P}, \text{ and } \bigcup_{p_i \in \mathcal{P}} T_i = \mathcal{U},$$

where parameter β denotes the weight of the cost term. The concrete representations of $train(T_i, p_i)$, function $qualify(T_i, p_i)$, as well as $cost(T_i)$ will be talked about in the following section when introducing the TRAIN framework.

3. PROPOSED METHODS

In this section, we will introduce the framework TRAIN in detail. We will first talk about the skill mastery levels of groups, based on which we will introduce the concrete measure about employees' skill gain in Section 3.1. The concrete representation of the team qualification function and the communication cost measure among the team members will be introduced in Section 3.2. Finally, we will provide the joint objective function of the ET problem, and a linear programming algorithm is proposed to address the function in Section 3.3.

3.1 Objective 1: Employee Skill Improvement

One main objective of the ET problem is to train the employees to help improve their skill levels. In this part, we will focus on quantifying the employees skill improvement by co-operating with others. We will first introduce the skill mastery levels of a team involving multiple employees, based on which we will define the employees' skill improvement concept. Finally, we will give the concrete representation of $train(T_i, p_i)$ used in defining the ET problem in Section 2.

3.1.1 Skill Mastery Level of Teams

In real world, knowledge structures of skills are quite diverse, which can be either *horizontal* or *hierarchical* [9]. Generally, *horizontal*-knowledge structure is characterized as having weak "*verticality*" internal relations among ideas. Meanwhile, the *hierarchical* knowledge structure aims to bring a broadening base of empirical phenomena and develop through the integration with previous knowledge.

For the *horizontal*-structured skills, different employees can know a different subset of the knowledge. Generally, the knowledge that senior people know is broader than that of junior people, but cannot totally cover what junior people know. For skills belonging to such a category, adding more employees to a team can always introduce new ideas and increase the skill mastery level of the team.

However, for the *hierarchical*-structured skills, different employees can know a subset of the knowledge, where the elementary knowledge that the junior people know is a subset of the advanced knowledge mastered by the senior people. For skills belonging to such a category, the skill mastery level of the whole team is determined by the employee with the highest skill mastery. Therefore, the *level* of skill s^k that team T can master could be simply represented as $level(T, s^k) = \max\{level(u_i, s^k)\}_{u_i \in T}$.

The framework TRAIN proposed in this paper works for both of these two structured skills, and in this paper, we will take the *horizontal-structured skill* as an example to illustrate the TRAIN framework.

Definition 7 (Team Skill Mastery Level): Given a team of employees $T = \{u_1, u_2, \cdots, u_j\}$ with mastery levels $\{\xi_i^k\}_{u_i \in T}$ respectively of skill s^k. Formally, by taking the *horizontal-structured* skill mastered by the employees to be i.i.d. (i.e., independent and identically distributed), the mastery levels of team T about skill s^k can be defined as the *team skill mastery level*:

$$level(T, s^k) = 1 - \prod_{u_i \in T} (1 - \xi_i^k).$$

According to the definition, the skill mastery level of a small team formed by u_i and u_j, i.e., $\{u_i, u_j\}$, about the same skill s^k can be represented as

$$level(\{u_i, u_j\}, s^k) = 1 - (1 - \xi_i^k)(1 - \xi_j^k).$$

With some simple derivations, the equations $level(\{u_i, u_j\}, s^k) \geq level(u_i, s^k)$ and $level(\{u_i, u_j\}, s^k) \geq level(u_j, s^k)$ can both hold for the above definition. Let's consider, for example, u_i is an expert in skill s^k, by pairing him with a junior employee u_j, u_i can borrow some new ideas from u_j, which can effectively enhance the overall skills of the group $\{u_i, u_j\}$ (i.e., $level(\{u_i, u_j\}, s^k) > level(u_i, s^k)$). In addition, by involving employees knowing very little about skill s^k into the group will not change the overall skills of the group. Let's assume there exists another employee u_l who knows nothing about skill s^k (i.e., $\xi_l^k = 0$). By adding u_l into group $\{u_i, u_j\}$, the new overall skill mastery level of the new group $\{u_i, u_j, u_l\}$ is identical to that of team $\{u_i, u_j\}$. It is also reasonable, since employee u_l cannot make any contributions to the team about the skill s^k.

3.1.2 Employee Skill Improvement in Teams

The overall team skill mastery level is an important characteristic of the teams, which denotes both the competitiveness of the group (it will be used in defining the team skill qualification constraint in Section 3.2.1) and the room for skill improvement of the members in the group. Formally, by grouping the senior employee u_i with the junior employee u_j together, the skill improvement space for the junior employee u_j can be represented as

$$\begin{aligned} space(u_j, \{u_i, u_j\}, s^k) &= level(\{u_i, u_j\}, s^k) - level(u_j, s^k) \\ &= 1 - (1 - \xi_i^k)(1 - \xi_j^k) - \xi_j^k \\ &= \xi_i^k - \xi_i^k \cdot \xi_j^k. \end{aligned}$$

Similarly, the skill improvement space for senior employee u_i will be

$$space(u_i, \{u_i, u_j\}, s^k) = \xi_j^k - \xi_i^k \cdot \xi_j^k.$$

According to the above definition of *skill improvement space*, we observe that the junior employee u_j will have larger improvement space than the senior employee u_i by pairing them into a team.

Meanwhile, in the real scenarios, employees' actual skill mastery level improvement is positively correlated with both the learning space as well as their learning abilities, which can be represented with the following *skill improvement* definition.

Definition 8 (Employee Skill Improvement): Let $\alpha_i \in [0, 1]$ be the learning ability of employee u_i, and $space(u_i, \{u_i, u_j\}, s^k)$ be the skill improvement space for u_i by grouping him and employee u_j into a team $\{u_i, u_j\}$. The real *skill improvement* for u_i regarding skill s^k can be defined as

$$improvement(u_i, \{u_i, u_j\}, s^k) = \alpha_i \cdot space(u_i, \{u_i, u_j\}, s^k).$$

Generally, different employees have various learning abilities in the real world. Meanwhile, to simplify the problem setting, we regard the learning abilities of all the employees to be the same in this paper, which can be denoted as parameter $\alpha \in [0, 1]$, and employees' skill improvement will be determined by the improvement space only.

3.1.3 The Skill Gain Measure

Furthermore, based on the *employee skill improvement* definition, the representation of the skill gain measure $train(T_i, p_i)$ for team T_i and project p_i (used define the ET problem in Section 2) can be represented as

$$\begin{aligned} & train(T_i, p_i) \\ &= \sum_{s^j \in S(p_i)} \sum_{u_k \in T_i \cap Sup(s^j)} improvement(u_k, T_i, s^j) \\ &= \sum_{s^j \in S(p_i)} \sum_{u_k \in T_i \cap Sup(s^j)} \alpha \cdot space(u_k, T_i, s^j) \\ &= \sum_{s^j \in S(p_i)} \sum_{u_k \in T_i \cap Sup(s^j)} \alpha \cdot \Big(\big(1 - \prod_{u_l \in Sup(s^j) \cap T_i} (1 - \xi_l^j)\big) - \xi_k^j \Big), \end{aligned}$$

where $Sup(s^j)$ denotes the support set of skill s^j regardless of the mastery levels as introduced in Section 2.

3.2 Objective 2: Success of Projects

Besides training the employees, another main objective of ET is to finish the projects successfully. In this part, we will first give the concrete representation of the team skill qualification function $qualify(T_i, p_i)$ used in defining the ET problem. Next, we will talk about the communication costs among the employees and give the concrete representation of the $cost(T_i)$ measure.

3.2.1 Team Skill Qualification

As defined in Section 2, the set of projects carried out in the company can be represented as set $\mathcal{P} = \{p_1, p_2, \cdots, p_k\}$. For each project $p_i \in \mathcal{P}$, the needed skills together with the required minimum skill proficiency levels can be represented as set $\{s^j : \pi_i^j\}_{s^j \in S(p_i)}$. Meanwhile, for each project $p_i \in \mathcal{P}$, one unique team T_i will be formed. The skill mastery level of team T_i in each skill s^j required by p_i (i.e., $s^j \in S(p_i)$) can be represented as $level(T_i, s^j)$ as introduced in the previous subsection. Based on these notations, team T_i built for project p_i is qualified in both the skill categories and proficiency iff $\forall s^j \in S(p_i), s^j \in \bigcup_{u_i \in T_i} S(u_i) \wedge level(T_i, s^j) \geq \pi_i^j$. Therefore, we can define the qualification measure of team $\mathcal{T}_i$ in carrying out project p_i with the following *team project qualification* function:

Definition 9 (Team Project Qualification): The qualification of team T_i in carrying out project p_i can be represented with the function $Qualify(T_i, p_i) \in \{0, 1\}$, where $Qualify(T_i, p_i) = 1$ iff $s^j \in \bigcup_{u_i \in T_i} S(u_i) \wedge level(T_i, s^j) \geq \pi_i^j, \forall s^j \in S(p_i)$.

3.2.2 Communication Costs among Employees

Generally, in the company, employees who have frequent interactions with each other in workplace and real-life will know each other much better, and the communication costs among them will be relatively lower. As introduced in Section 2, the communication costs among employees can be calculated based on the information available in both online ESNs and offline organizational chart. In this part, we will first give the concrete representations of the *chart based communication cost* $cost_c(u, v)$ and *ESNs based communication cost* $cost_e(u, v)$ among the employees first. Based on these two communication cost measures, we will introduce the integrated communication cost measure $cost(T_i)$ used in the problem definition.

Chart based Communication Cost

The company internal organizational structure can provide valuable information in indicating the closeness among employees. Generally, employees within the same group are closer to each other compared with those in other groups. As proposed in [16], the closeness between employees u and v can be measured by the reciprocal of steps required to walk between them along the management links in the chart, i.e.,

$$Closeness_c(u, v) = \frac{1}{step(u, v)}$$

Generally, the closeness among the groupmates is relatively greater than that between employees who are in different departments. Furthermore, the communication cost between close employees will be smaller than that between employees who are not close to each other. In this paper, we propose to define the *chart based communication cost* $cost_c(u, v)$ as follows:

$$cost_c(u, v) = 1 - Closeness_c(u, v) = 1 - \frac{1}{step(u, v)}.$$

ESNs based Communication Cost

Besides the organizational chart, employees are involved in online ESNs nowadays, in which they can have extensive social interactions. In this paper, we will take the friendship connections as an example of the social interactions in ESNs. Based on the social connections among employees, we can obtain the neighbors that employee u is connected with as set $\Gamma(u) = \{w | w \in \mathcal{U}, (u, w) \in \mathcal{E} \vee (w, u) \in \mathcal{E}\}$. In a similar way, we can denote the neighbor set of employee v as $\Gamma(v)$. Generally, the more common friend two employees shares (i.e., $\Gamma(u) \cap \Gamma(v)$), the more likely that they may know each other [8].

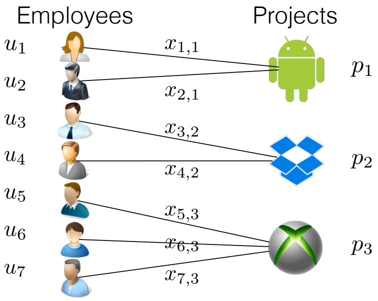


Figure 1: An example of employee project matching.

Meanwhile, different from the traditional online social networks, in the enterprise context of ESNs, almost all the employees like to follow the people in high management levels a lot. For instance, any two employees in the online ESN may share a large number of common friends, like the CEO and EVPs of the company, even through they are not close to each other. To address that problem, in this paper, we propose to follow the IDF-based social closeness measure introduced in [16], which assigns each common neighbor a weight inversely correlated with their social degrees:

$$Closeness_e(u,v) = \frac{1}{|\Gamma(u) \cap \Gamma(v)|} \cdot \sum_{w \in \Gamma(u) \cap \Gamma(v)} \log \frac{|\mathcal{U}|}{|\Gamma(w)|}.$$

Furthermore, the *ESNs based communication cost* measure can be represented as

$$\begin{aligned} cost_e(u,v) &= 1 - Closeness_e(u,v) \\ &= 1 - \frac{1}{|\Gamma(u) \cap \Gamma(v)|} \cdot \sum_{w \in \Gamma(u) \cap \Gamma(v)} \log \frac{|\mathcal{U}|}{|\Gamma(w)|}. \end{aligned}$$

Communication Cost Measure

Based on the representation of $cost_c(u,v)$ and $cost_e(u,v)$, we can define the communication cost between u and v as a linear combination of them as introduced in Section 2:

$$cost(u,v) = \omega \cdot cost_c(u,v) + (1-\omega) \cdot cost_e(u,v).$$

In addition, the communication cost $cost(T_i)$ introduced by the team T_i (used in the problem definition in Section 2) can be represented as the sum of communication cost between any pairs of employees in team T_i, i.e.,

$$cost(T_i) = \frac{1}{2} \sum_{u,v \in T_i, u \neq v} \text{cost}(u,v).$$

Generally, larger teams will lead to higher communications, as the communications between any pairs in the team members will be counted. Therefore, in the ET problem, minimization of the cost measure will favor smaller teams involving close team members, and can effectively help avoid forming groups containing a unrealistically large number employees.

3.3 Joint Optimization Function

Based on the above descriptions, in this part, we can obtain the concrete representation of the objective function of the ET problem. Actually, as shown in Figure 1, the ET problem studied in this paper is to resolve the matching problem between employees and the projects. If employee $u_i \in \mathcal{U}$ is assigned to the team of project $p_j \in \mathcal{P}$, we can add one link between u_i and p_j to denote that $u_i \in T_j$. Therefore, all the potential employee-project assignment links can be represented as matching set $\mathcal{M} = \mathcal{U} \times \mathcal{P}$. For each potential employee-project matching pair (u_i, p_j) in $\mathcal{M}$, we introduce a variable $x_{i,j} \in \{0,1\}$ to denote whether u_i is assigned to project p_j or not and further rewrite the objective function based on these variables in this section.

3.3.1 *Objective 1: Employee Skill Improvement*

For each project p_j, all the employees assigned to p_j in the result can be represented as $\{u_i | u_i \in \mathcal{U} \wedge x_{i,j} = 1\}$, and the group skill mastery level of skill s^k required by p_j can be denoted as

$$1 - \prod_{u_i \in Sup(s^k)} (1 - x_{i,j} \cdot \xi_i^k).$$

Therefore, the skill improvement of all team members involved in project p_i who knows skill s^k will be

$$\sum_{u_l \in Sup(s^k)} x_{l,j} \cdot \alpha \cdot \Big(\big(1 - \prod_{u_i \in Sup(s^k)} (1 - x_{i,j} \cdot \xi_i^k)\big) - \xi_l^k \Big).$$

3.3.2 *Objective 2: Success of Projects*

Team Project Skill Qualification

For each skill s^k required by project p_j (i.e., $s^k \in S(p_j)$), the team project qualification requirements can be represented as the following constraint

$$1 - \prod_{u_i \in Sup(s^k)} (1 - x_{i,j} \cdot \xi_i^k) \geq \pi_j^k.$$

Communication Cost among Employees

What's more, for any two employees u_i and u_l, if u_i and u_l are both involved in project p_j (i.e., $x_{i,j} = x_{l,j} = 1$), then the communication cost between u_i and u_l will influence the effective cooperations between them. The communication cost among all the employees involved in the team of project p_j can be represented as

$$\sum_{u_i, u_l \in \mathcal{U}, u_i \neq u_l} \frac{1}{2} \cdot x_{i,j} \cdot x_{l,j} \cdot \text{cost}(u_i, u_l).$$

3.3.3 *Joint Optimization Function*

Based on the remarks, the joint optimization function of the ET problem can be represented as

$$\begin{aligned} \max_x \sum_{p_j \in \mathcal{P}} \sum_{s^k \in S(p_j)} \sum_{u_i \in Sup(s^k)} & x_{i,j} \alpha \Big(\big(1 - \prod_{u_l \in Sup(s^k)} (1 - x_{l,j} \xi_l^k)\big) \\ & - \xi_i^k \Big) - \beta \cdot \sum_{p_j \in \mathcal{P}} \sum_{u_i, u_l \in \mathcal{U}, u_i \neq u_l} \frac{1}{2} \cdot x_{i,j} \cdot x_{l,j} \cdot \text{cost}(u_i, u_l) \\ s.t.\ \ 1 - & \prod_{u_l \in Sup(s^k)} (1 - x_{l,j} \cdot \xi_l^k) \geq \pi_j^k, \forall p_j \in \mathcal{P}, \\ & \sum_{p_j \in \mathcal{P}} x_{i,j} \geq 1, \forall u_i \in \mathcal{U}, \\ & x_{i,j} \in \{0,1\}, \forall u_i \in \mathcal{U}, p_j \in \mathcal{P}, \end{aligned}$$

where constraint $\sum_{p_j \in \mathcal{P}} x_{i,j} \geq 1$ denotes all the employees are involved in the team of at least one project.

The objective function is a non-linear integer programming (IP) problem, which is shown to be NP-hard [11] and no polynomial-time solutions exist that can solve the problem efficiently. In this paper, we propose to resolve the problem with two steps: (1) interger constraint relaxation, and (2) result post-processing. As proposed in [6], such a linear programming (LP) relaxation of the IP problem can help address the problem in polynomial time with regarding to its input size. In addition, the obtained solution provides a optimistic approximation of the optimal result, the introduced result difference is tightly bounded according to the error bound analysis provided in [13].

We propose to relax the integer constraint on the variables, and allow them to take real values in range $[0, 1]$. The optimization objective function after the relaxation can be solved with some open-source non-linear programming toolkits like, Scipy[1], effectively. Meanwhile, based on the obtained solution $\{x_{i,j}\}_{(u_i,p_j)\in\mathcal{M}}$ denoting the confidence scores of the employee-project links, we need to determine which employee should be assigned to which project, i.e., the post-processing of the variable results to binary values in $\{0, 1\}$. In this paper, a confidence threshold $\theta \in [0, 1]$ is applied to prune the results. For the variables, e.g., $x_{i,j}$, if $x_{i,j} \geq \theta$, we will map $x_{i,j}$ to value 1; otherwise, we map it to value 0. Based on the obtained solutions $\{x_{i,j}\}_{(u_i,p_j)\in\mathcal{M}}$, for the variables $x_{i,j}$ being assigned with value 1, we will add the corresponding employee u_i into the team of project p_j. Therefore, we can obtain the final optimal team formation result $\mathcal{T} = \{T_1, T_2, \cdots, T_k\}$ of the ET problem.

4. EXPERIMENTS

To test the effectiveness of the proposed framework Train, we will conduct experiments on real-world datasets. In this section, we will first introduce the datasets used in the experiments, and then discuss the experiment setting. The experiment results and parameter analysis will be provided next. Finally, we will give a cast study to help illustrate the experiment results.

4.1 Dataset Description

The data used in the experiments involves 3 different datasets: (1) enterprise project team dataset, (2) enterprise online social networks dataset, and (3) enterprise organization chart dataset.

In this paper, we crawled all the information about the employees and project teams of the Microsoft Research department from its official website[2] on November 15, 2015, which altogether comprise the *enterprise project team dataset*. The Microsoft Research department (i.e., MSR) involves 719 employees and 154 different research teams all around the world. Each employee can participate in at least one research groups, and the number of team membership link in the dataset is $1,089$.

Meanwhile, to get more information about the employees, we also get all the Microsoft employees' social connections from Yammer (an online ESN launched in Microsoft) and obtain the complete organizational chart involving all these employees from Microsoft.[3] For more information about these two datasets, please refer to [17, 16, 18].

In addition to these three datasets, we also obtain the job titles and the expertise about the employees from the Microsoft company internal information sources. Job titles in Microsoft may involve various levels. For instance, just for the "*researcher*" title in MSR, it can have 4 different levels: "*Researcher*", "*Senior Researcher*", "*Principal Researcher*", and "*Distinguished Researcher*". Meanwhile, the obtained employees' expertise information can represent the skills employees have in the company.

4.2 Experiment Settings

4.2.1 Experiment Setups

In the experiments, the job title levels can indicate the employees' experience levels of their skill (i.e., their expertise), and generally senior employees are more experienced than the junior employees. Based on the job title information, we can infer the mastery level of employee u_i in mastering the skill s^k to be a real value $\xi_i^k \in [0, 1]$. More specifically, (1) for regular employees u_i (e.g., with job titles like "Researcher"), value ξ_i^k is randomly sampled from range $[0, 0.25)$; (2) for senior-level employees u_i (e.g., with job titles like "Senior Researcher"), we sample their mastery level randomly from range $[0.25, 0.5)$; (3) for principal-level employees u_i (e.g., with job titles like "Principal Researcher"), we randomly sample value ξ_i^k from range $[0.5, 0.75)$; and (4) for distinguished-level employees u_i (e.g., with job titles like "Distinguished Researcher"), value ξ_i^k is sampled from range $[0.75, 1.0]$.

In addition, with the employees and their skill information, we can obtain required skills and mastery levels for each project team, which should be pre-specified by the company before carrying out the projects. For instance, based on the team T_i of project p_i, we can define the skill needed by p_i to be $S(p_i) \bigcup_{u_j \in T_i} S(u_j)$, whose required skill levels can be obtained based on the group mastery level definition. The employee team member links are not involved in building the models, which is used for evaluation only.

Via some data pre-processing, the majority of the crawled employees of the MSR department can be mapped to the Yammer and organizational chart datasets, and the remaining are pruned as no social interaction information about them is available. Meanwhile, based on the social connection information available in the online ESNs (i.e. Yammer) and the management relationships in the organizational chart, we can calculate the communication costs among employees. In the experiment, the communication cost weight parameter ω is set as 0.5. In addition, in the experiments, the learning abilities of employees (i.e., parameter α) is set as 1.0, which denotes the employees can improve their skills by 100% of the skill learning space of the team. Framework Train is built by fusing these different categories of information to form the potential teams for each project.

4.2.2 Comparison Methods

The ET problem is a new problem, and no existing methods can be applied to address it directly. In this paper, to show the advantages of the framework Train, we extend some methods proposed in other related works and compare them with Train. The comparison methods used in the experiments are listed as follows:

- *The* Train *Framework*: The framework Train introduced in this paper can assign employees to the teams such that employees involved in the teams can get trained and the projects can be finished at the same time. The framework Train with different parameter values of β will be used as different comparison methods in the experiments.

- *Method* IterL&F: IterL&F is the method introduced in [2], which studies the problem of student education by partitioning them into different study groups. Method IterL&F is an iterative heuristic method. In every iteration of IterL&F, one group of size k is formed. The selection of the group is done with IterL&F, where the students who have not yet been assigned to any group are used as the input. No skill qualification nor communication cost issues are considered in IterL&F.

- *Method* RF: RF is an extension to the RarestFirst method proposed in [7], which selects employees with the rarest required skills for one single projects. In the experiments, we extend RF to select the employees for each project, but no skill mastery level information is considered in RF.

- *Method* RF-Level: To show the importance of skill mastery level information in team formation, we also further extend RarestFirst [7] and introduce the method RF-Level in this paper. Method RF-Level keeps selecting the employees with

[1] http://docs.scipy.org/doc/scipy/reference/optimize.html

[2] http://research.microsoft.com/en-us/default.aspx

[3] We are not able to reveal the actual numbers here and throughout the paper for commercial reasons.

Table 1: Performance comparison of different comparison methods evaluated by metrics F1, Accuracy, Avg. Skill Gain (per team member), Avg. Communication Cost (per team member), Qualified Team Ratio and Team Size.

Method	Metrics					
	F1	Accuracy	Avg. Skill Gain	Avg. Communication Cost	Qualified Team Ratio	Team Size
TRAIN($\beta = 0.01$)	**0.49**	0.53	**48.53**	378.04	**1.0**	16.43
TRAIN($\beta = 0.05$)	**0.53**	0.67	**48.48**	242.02	**0.97**	13.03
TRAIN($\beta = 0.1$)	**0.57**	**0.84**	**38.86**	109.74	**0.70**	8.27
TRAIN($\beta = 0.2$)	0.39	**0.89**	24.80	**18.74**	0.63	3.00
ITERL&F	0.49	0.32	36.22	340.34	0.7	21.5
RF	0.31	0.72	24.70	26.52	0.56	1.30
RF-LEVEL	0.31	0.71	24.84	**23.74**	0.48	1.60
KMEANS	0.41	**0.75**	10.46	**10.59**	0.5	4.67

rarest skills for each project until the required skill level can be met or no employees with the skill exist any more.

- *Method* KMEANS: The ET is different from the clustering problems. To support such a claim, we also compare TRAIN with traditional clustering method KMEANS. In KMEANS, employees with lower communication costs are grouped into the same cluster.

4.2.3 Evaluation Metrics

To compare the performance of different methods, different evaluation metrics are used to measure their results. To show that the employees can get trained by involving in the projects, we calculate the average skill improvement per employee in the teams (i.e., Avg. Skill Gain) based on the results outputted by different methods. To ensure the success of the projects, we count the ratio of projects that can meet the skill requirements (i.e., Qualified Team Ratio), as well as the average team-internal communication costs per employee (i.e., Avg. Communication Cost). The sizes of the teams built by different methods is also used as metric, i.e., Team Size. In addition, we also have the real-world team memberships of these projects, which can be used as the ground truth. By comparing the obtained results with the ground truth, some frequently-used metrics, like Accuracy and the multi-class version of F1 score (which is a weighted sum of the F1 score achieved for each class) [1], are applied in the experiments.

4.3 Experiment Result

The experiment results achieved by different comparison methods at $\theta = 0.1$ (i.e., the potential employee-project pairs with at least 0.1 confidence scores are preserved) are given in Table 1, which are evaluated by metrics *F1*, *Accuracy*, *Avg. Skill Gain*, *Communication Cost*, *Qualified Team Ratio* and *Team Size* respectively.

As shown in Table 1, when the evaluation metrics are the traditional *F1* and *Accuracy* measures, generally framework TRAIN with different β values can outperform the other methods. For instance, the F1 and Accuracy scores achieved by TRAIN($\beta = 0.1$) are 0.57 and 0.84 respectively, which is much higher than the scores obtained by ITERL&F (which are 0.49 and 0.32) and RF (which are 0.31 and 0.72). Here, we observe that methods RF, RF-LEVEL and KMEANS can also achieve high Accuracy scores due the class imbalance setting (i.e., the number of non-existent employee-project pairs is larger than that of the existing ones).

Meanwhile, the Accuracy score achieved by ITERL&F is relatively low as the teams formed by ITERL&F are of larger sizes, where a large number of redundant employees are involved for each project team. Involving more employees can help increase the total skill gain for all the employees in the team in the ITERL&F method [2]. For the TRAIN framework, its team size is highly dependent on the parameter β. Generally, smaller β (i.e., less weight of the cost term) favors larger-sized teams. The parameter analysis of β on TRAIN will be given in Section 4.4.

As evaluated by the *Avg. Skill Gain* and *Avg. Communication Cost* metrics, the average skill gain of each employees achieved by TRAIN is greater than the other methods. Meanwhile, the efforts that each employee devoted to the communication with other team members in the results of TRAIN is also greater, as the teams built by TRAIN is relatively larger than those formed by RF and KMEANS. For instance, the average skill gain each employee achieved by TRAIN($\beta = 0.1$) is 38.86, which is almost the double of those obtained by RF, RF-LEVEL and KMEANS.

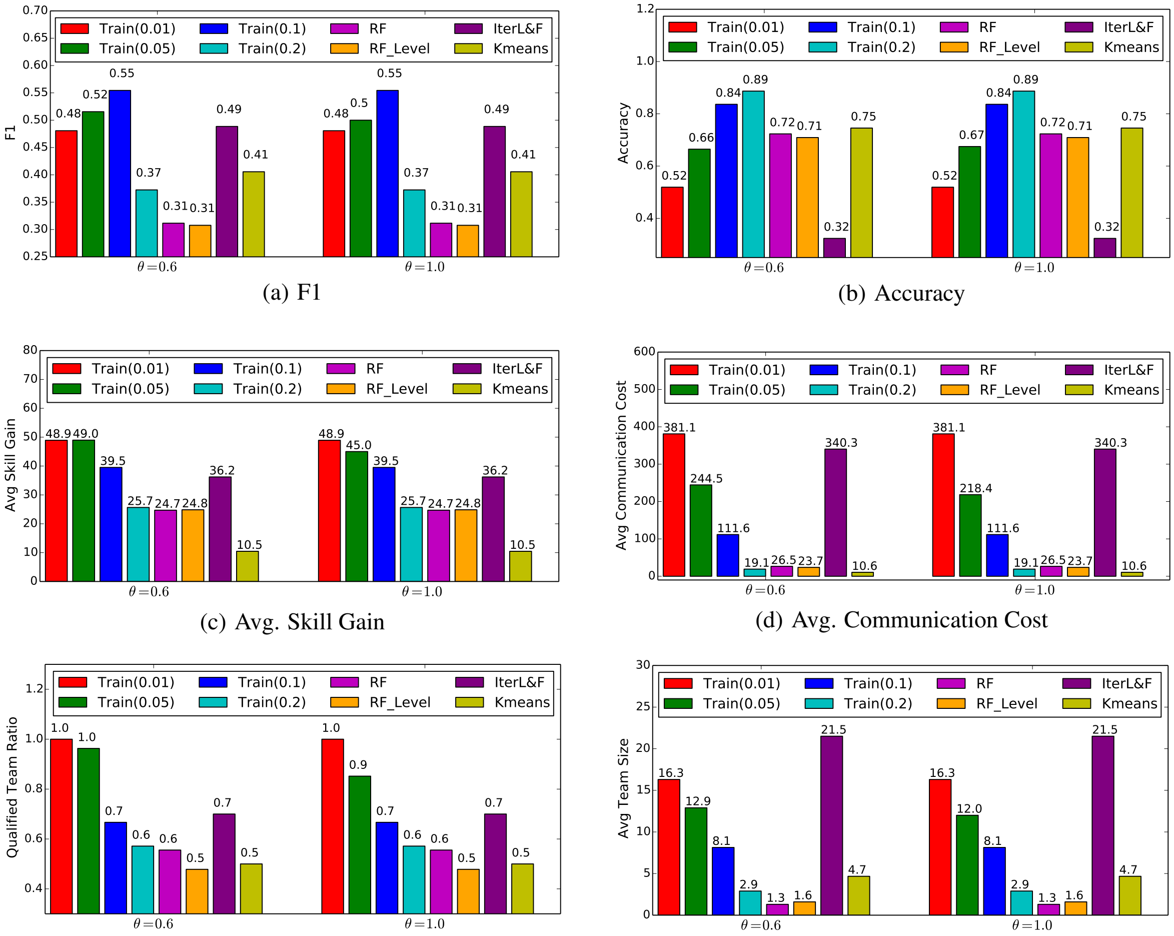
In addition, we also show the qualified ratio of the teams formed by these different comparison methods in terms of the skills and their mastery levels. Generally speaking, involving more people into the team, the more likely that the required skills can be achieved by the whole team. For instance, the *Qualified Team Ratio* achieved by TRAIN($\beta = 0.01$) and ITERL&F are 1.0 and 0.7 respectively, which are larger than that achieved by RF, RF-LEVEL and KMEANS. The reason that TRAIN($\beta = 0.01$), TRAIN($\beta = 0.05$) as well as TRAIN($\beta = 0.1$) can outperform ITERL&F for the metric *Qualified Team Ratio* is that the teams built by TRAIN are of higher quality in terms of the skill satisfaction, as the skill satisfaction is added as a constraint of the objective function of TRAIN, which is not considered in ITERL&F at all.

4.4 Parameter Analysis

Two important parameters are involved in formulating and addressing the ET problem, which are β and θ. In this section, we will first analyze the effects of parameter θ in pruning the redundant employee-project paris, and then study the sensitivity of the weight parameter β on the performance of TRAIN.

To address the objective function of TRAIN, the hard integer constraint on the parameters are relaxed, and the introduced redundant employee-project paris with confidence scores lower that θ are pruned in TRAIN. The sensitivity analysis results of θ at 0.6 and 1.0 are available in Figure 2. By comparing the results in the plots and that in Table 1, we observe that the effects of θ has no effects on methods ITERL&F, RF, RF-LEVEL and KMEANS, as θ is not involved in the model building of these methods. Meanwhile, the influence of θ on TRAIN is also very small. To understand the reasons, we study the output results of framework TRAIN, and we observe that the final values of the variables outputted by the Scipy toolkit are mainly distributed close to the bounds (i.e., 0 and 1). In other words, the pruning effects of parameter $\theta = 0.1$ are actually very similar to the parameter $\theta = 0.9$.

Parameter β denotes the weight of the communication cost term in the objective function of framework TRAIN, where the weight of the skill gain term is constant (i.e., 1). Generally, larger β will give



(a) F1

(b) Accuracy

(c) Avg. Skill Gain

(d) Avg. Communication Cost

(e) Qualified Team Ratio

(f) Team Size

Figure 2: Experiment results of comparison methods evaluated by F1, Accuracy, Avg. Communication Cost, Average Team Size, Qualified Team Ratio and Avg. Skill Gain.

the cost term more weight and tend to favor the small-sized teams, while smaller β will favor the larger-sized teams, which can introduce a large amount of skill gains for employees involved in the teams. The sensitivity analysis result of parameter β is available in Figure 3, where the performance of TRAIN with different β values ($\{0.01, 0.05, 0.1, 0.15, 0.2, 0.3, \cdots, 1.0\}$) are shown and evaluated by metrics F1, Accuracy, Avg. Communication Cost, Avg. Skill Gain, Qualified Team Ratio and Team Size respectively.

As shown in Figure 3(a), when β increases, the F1 score achieved by TRAIN increases first and then decrease consistently. The potential explanation for the observation can be that for too small/large β, the teams formed for each project will be either very large or very small, while the real-world project teams in the dataset are actually of moderate sizes. Meanwhile, the Accuracy scores achieved by TRAIN increases steadily as β increases. The reason can be that, as β increases the team become smaller and the majority of the employee-project pairs will be identified to be non-existent, and the Accuracy score achieved by TRAIN will still be high in such a class-imbalance case. Meanwhile, for the *Avg. Communication Cost* and *Team Size* metrics, as β increases the costs and team sizes will decrease consistently. However, for the *Avg Skill Gain* and *Qualified Team Ratio* metrics, as β increases, they drop first and then keep relatively stable, which achieve the lowest skill gain and qualified team ratio at $\beta = 0.2$.

4.5 Case Study

Besides evaluating the results by these evaluation metrics, we also provide a case study in this part to help illustrate the performance of TRAIN as well as its difference from other comparison methods, like ITERL&F and RF. As shown in Table 2, we aim at building a team for a certain project carried out in the Microsoft Research. 7 employees with specific homepage (containing their IDs) are shown in the project team homepage, and from the Microsoft Research site we can get the required skills of the project, both of which are shown in Table 2. Based on the input project together with the information in the ESNs and the organizational chart of Microsoft, the teams built by the methods TRAIN($\beta = 0.1$), ITERL&F and RF are shown in the table, while the specific skills of all these employees are provided in Table 3. (The project team name is not provided, and the names of all the employees are abbreviated due to the privacy and commercial reasons.)

According to the results, framework TRAIN builds a team of size 8, and among the team members 5 of them (whose names are underlined) are involved in the real-world team and the achieved precision and recall scores are $\frac{5}{8}$ and $\frac{5}{7}$ respectively. In the built team, the skills of team members *A. B.*, *Q. Y.*, *J. T.*, *L. Z.* can already meet the skill requirements of the project, and the remaining young employees (in the italic font and marked with *) are involved to polish their skills from these senior employees.

Table 2: Teams formed by different methods for the input project (skill experience levels: Proficiency ≤ Expert ≤ Mastery).

Required Skill: *Levels*	Real-World Team(size: 7)	Team Built by TRAIN (size: 8)	Team Built by ITERL&F (size: 13)		Team Built by RF(size: 3)
Networking: *Proficiency*	A. B.	A. B.	Z. L.	*S. R.**	J. T.
HCI: *Mastery*	C. W.	*C. W.**	O. R.	*C. N.**	A. B.
SDE: *Mastery*	X. G.	*X. G.**	*C. O.**	*D. C.**	Q. Y.
Security:*Mastery*	Y. X.	L. Z.	*Q. L.**	J. T.	
Multimedia: *Mastery*	Q. Y.	Q. Y.	*Y. X.**		
Collaboration: *Expert*	W. X.	*Q. L.**	*S. A.**		
Management: *Expert*	J. T.	J. T.	*M. C.**		
Hardware: *Expert*		*Y. L.**	K. R.		
Health: *Proficiency*			Y. L.		

Table 3: Skills of employees in the company (skill experience levels: Proficiency ≤ Expert ≤ Mastery).

Employee/Skill	**HCI**	**SDE**	**Security**	**Multimedia**	**Collaboration**	**Management**	**Hardware**	**Health**	**Networking**
A. B.	Mastery		Mastery			Mastery	Mastery		
C. W.		Proficiency							
X. G.		Proficiency							
Y. X.		Expert							
Q. Y.		Mastery		Mastery					
W. X.		Mastery							
J. T.					Expert			Proficiency	Proficiency
K. R.	Mastery	Mastery							
C. O.		Proficiency							
O. R.		Expert							Mastery
Q. L.	Proficiency				Proficiency				Proficiency
S. A.	Proficiency								
M. C.			Proficiency						
S. R.		Proficiency	Proficiency			Proficiency			
L. Z.		Mastery				Mastery	Mastery		Mastery
C. N.		Proficiency		Proficiency					
Z. L.	Mastery	Mastery		Mastery					Mastery
Y. L.				Expert	Expert		Expert		Expert
D. C.							Expert		

Compared with TRAIN, the team build by ITERL&F is of a relatively larger size, which includes 13 employees in all (two are in the real project team). In addition, among the 13 team members, the majority (8 out of 13) of them are actually more junior employees just with the *Proficiency* levels of certain skills. In addition, without considerations about the success of the projects, the team built by ITERL&F cannot necessarily finish the projects. For example, in the example given in the table, the team built by ITERL&F cannot meet the *Mastery* level in Security required by the project.

Different from TRAIN and ITERL&F, the teams built by RF are extremely small, so as to minimize the total communication cost. For instance, for the given project, the team built by RF is of size 3 and all these 3 employees are correctly chosen, i.e., the precision and recall achieved by RF in the example are $\frac{3}{7}$ and 1 respectively. These 3 employees in the team are qualified to carry out the project in terms of the skills. However, the skills of these 3 employees are totally disjoint, while no young employees are involved. In other words, for the team built by RF, none of the team members improve their skills by involving in the project, and the total skill gain of the team members will be 0 in this example.

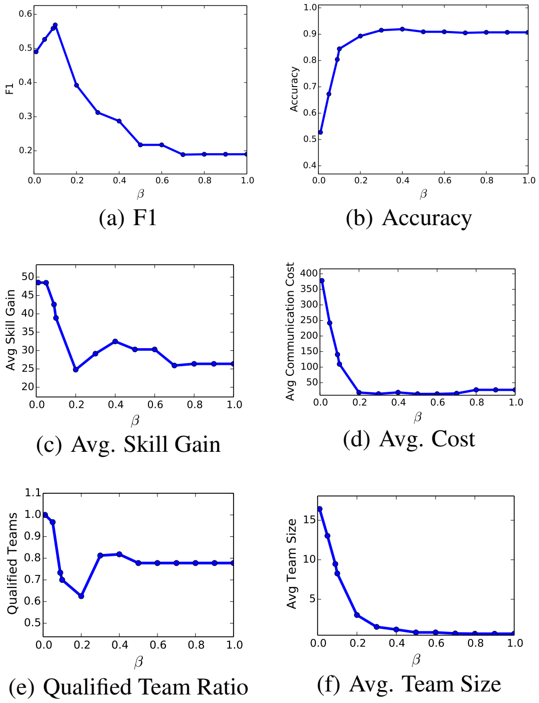
4.6 Experimental Result Insights

In summary, (1) the teams built by ITERL&F are usually of large size so as to maximize the overall skill gain of all the employees, and method ITERL&F can be applied in some education/training agencies that are mainly concerned about individuals' personal development; (2) the teams built by RF are very small so as to minimize the communication costs of the team but of high quality in terms of skill satisfaction, and method RF is helpful for building team for small-sized start-up style companies that focus on developing products in a relatively fast pace; and (3) the teams built by TRAIN is of intermediate size and can balance between the communication cost and employee training objectives, which can be applied in the normal real-world companies that aim at both training employees and finishing the projects successfully.

5. RELATED WORK

Enterprise Social Networks: Abundant information about employees at enterprise context provides researchers with the opportunity to study employees' social behaviors in enterprise social networks [17] with the background knowledge about their professional positions in the company organizational chart [17]. Some prior research works have been done by Zhang et al. to fuse the enterprise context data for synergistic knowledge discovery problems [17, 16, 18]. Based on the heterogeneous information in enterprise social networks, Zhang et al. propose to infer the complete organizational chart based on an unsupervised learning framework CREATE in [17]. By analyzing employees' various online social activities in the context of enterprise, Zhang et al. [16] propose to recommend friends for employees in online ESNs. Generally, employees will spend a large amount of time in companies, and workplace has become an important social occasion for effective communication and information exchange. Zhang et al. propose to study the information diffusion problem at workplace and extract various diffusion channels from both the online and offline information sources [18].

Education, Training and Learning: Although group work is sometimes hailed as an educational panacea, the realities are considerably more complex. Many works have been done on group based education in schools. Ward [14] gives some important observation about group based student instructions in the classroom. A dominant theme in the research findings is that some types of instructional grouping contribute to more positive academic and affective outcomes for students. Other groups, particularly stable, long-term groups based on student ability, have a negative effect upon students. Wieman et al. [15] provides a brief review of different levels of group work and list the potential benefits and negatives, and what requirements research



(a) F1

(b) Accuracy

(c) Avg. Skill Gain

(d) Avg. Cost

(e) Qualified Team Ratio

(f) Avg. Team Size

Figure 3: Parameter analysis of communication cost weight β under the evaluation of F1, Accuracy, Avg. Communication Cost, Avg. Skill Gain, Qualified Team Ratio and Avg. Team Size.

has shown are needed to ensure a high probability of success. Boaler et al. [4] report upon interim data from a four-year longitudinal study that is monitoring the mathematical learning of students in six UK schools. They develops and expands themes arising from a study of two schools that offered "traditional" and "progressive" approaches to the teaching of mathematics [4]. Agrawal et al. [2] study the group based student education problem from the computational perspective, and based on their formulation the problem is shown to be NP-hard.

Team Formation: Initially, the team formation problem is studied in the multi-agent systems [5], which are typically embedded in dynamic environments. Gaston et al. [5] propose to develop a distributed, online network adaptation mechanisms for discovering effective network structures, where several strategies for agent-organized networks are proposed in the context of dynamic team formation. Afterwards, the team formation problem that we study now is first proposed by Lappas et al. in [7], which has become a popular research problem and lots of works have been done already. Lappas et al. [7] formulates the team formation problem as a sub-graph extraction problem, where the communication costs introduced in the extracted sub-graph is minimized. Two different communication cost measures are introduced based on the diameter and the minimum spanning tree of the sub-graph, and the optimal team formation problem is shown to be NP hard [7]. Anagnostopoulos et al. [3] propose the team formation in online social networks, where a sequence of tasks arrives in an online fashion, and each task requires a specific set of skills. The goal is to form a new team upon arrival of each task.

6. CONCLUSION

In this paper, we have studied the ET problem to train the employees in companies by involving them in the company internal projects. Two objectives are covered in the ET problem: (1) maximize the employees' skill gain, and (2) ensure the success of the projects to be carried out. To address the ET problem, a novel employee training framework TRAIN has been proposed and introduced in great detail. TRAIN formulates the ET as an optimization problem, which aims at maximizing the employee skills gain and minimizing the communication costs among employees in each project team. In addition, the team skill qualification is used as a hard constraint to the objective function to guarantee the success of the projects. Extensive experiments have been done on real-world datasets, and the experiment results have demonstrated the effective and advantages of TRAIN in addressing the ET problem.

7. ACKNOWLEDGEMENT

This work is supported in part by NSF through grants IIS-1526499, and CNS-1626432. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

8. REFERENCES

[1] *sklearn.metrics.f1_score*. Software available at `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`.

[2] R. Agrawal, B. Golshan, and E. Terzi. Grouping students in educational settings. In *KDD*, 2014.

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

[4] J. Boaler, D. Wiliam, and M. Brown. Students' experiences of ability grouping - disaffection, polarisation and the construction of failure. *British Educational Research Journal*, 2010.

[5] M. Gaston and M. desJardins. Agent-organized networks for dynamic team formation. In *AAMAS*, 2005.

[6] D. Hochbaum, N. Megiddoy, J. Naorz, and A. Tamir. Tight bounds and 2-approximation algorithms for integer programs with two variables p er inequality. 1992.

[7] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *KDD*, 2009.

[8] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[9] K. Maton. Reclaiming knowers: Advancing bernstein's sociology of knowledge. *Sixth International Basil Bernstein Symposium*, 2010.

[10] M. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 2004.

[11] C. Papadimitriou. On the complexity of integer programming. *Journal of the ACM*, 1981.

[12] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. Yu. A survey of heterogeneous information network analysis. *CoRR*, abs/1511.04854, 2015.

[13] O. Stein. Error bounds for mixed integer linear optimization problems. 2013.

[14] B. Ward. Instructional grouping in the classroom. *SCHOOL IMPROVEMENT RESEARCH SERIES*, 1987.

[15] C. Wieman. Student group work in educational settings. *CWSEI & CU-SEI*, 2008.

[16] J. Zhang, Y. Lv, and P. Yu. Enterprise social link recommendation. In *CIKM*, 2015.

[17] J. Zhang, P. Yu, and Y. Lv. Organizational chart inference. In *KDD*, 2015.

[18] J. Zhang, P. Yu, and Y. Lv. Information diffusion at workplace. In *CIKM*, 2016.