Enterprise Community Detection via Hierarchical Information Fusion

Jiawei Zhang[†], Philip S. Yu[†], Yuanhua Lv[‡]
[†]University of Illinois at Chicago, Chicago, IL, USA
[‡]Microsoft Research, Redmond, WA, USA
jzhan9@uic.edu, psyu@cs.uic.edu, yuanhual@microsoft.com

Abstract—Employees in companies can be divided into different social communities, and those who frequently socialize with each other are treated as close friends and will be grouped in the same community. In the enterprise context, a large amount of information about the employees is available in both (1) offline company internal sources and (2) online enterprise social networks (ESNs). What's more, each of the information sources can also contain multiple categories of employees' socialization activity information at the same time. In this paper, we propose to detect the social communities of the employees in companies based on these different information sources simultaneously, and the problem is formally called the "Enterprise Community Detection" (ECD) problem. To address the problem, a novel community detection framework named "HeterogeneoUs Multi-sOurce ClusteRing" (HUMOR) is introduced in this paper. Based on the various *enterprise social intimacy* measures introduced in this paper, HUMOR detects a set of *micro community structures* of the employees based on these different categories of information available in the online and offline sources respectively. To obtain the (globally) consistent community structure of employees in the company, HUMOR further fuses these *micro community structures* via two phases: (1) *intra-fusion* of *micro community structures* to obtain the online and offline (locally) consistent communities respectively, and (2) *inter-fusion* of the online and offline consistent communities to achieve the (globally) consistent community structure of employees. Extensive experiments conducted on real-world enterprise datasets demonstrate our method can perform very well in addressing the ECD problem evaluated by various metrics (like rand index, mutual information, etc.).

Keywords-Community Detection; Enterprise Social Network; Data Mining

I. INTRODUCTION

People in social organizations (e.g., schools, companies, and even countries) can generally be divided into different social communities, where individuals who frequently socialize with each other will be in the same community, while those who rarely interact with each other will be in different communities instead. Meanwhile, detecting the social communities of people within social organizations is formally called the *community detection* problem. Depending on the specific application settings, the community detection problem has already been studied in social organizations of various granularities in different areas. In this paper, we will take the "company" as an example, and propose to analyze the community structures formed by employees within companies.

Multiple source information mining and knowledge discovery is an important research topic in big data studies. Based on the professional context in enterprises, various company internal information about the employees is available, which can provide crucial signals about the community structures formed by the employees. For instance, employees in companies are usually divided into different hierarchies, where people in management are at higher levels, while the regular employees are at lower levels. The whole management structure of the employees in a company can be characterized with the *enterprise organizational chart* visually [32]. Generally, employees who are close in the company *organizational chart* (e.g., colleagues in the same group) tend to be more familiar with each other and are highly likely to be in the same community. What's more, employees in companies all have their job titles, which describe their duties, responsibilities, and relative ranks in the company. Employees who have common job titles may need to cooperate to do projects together, and are more likely to be in the same community. In addition, for large international companies with branches all around the world, from the company we can know the employees' workplace information and their office locations as well. Employees who work at the same place will have more chance to meet in the real world and are more likely to be in the same community, especially compared with those working in different countries. Therefore, the company internal information provides very important clues for us to understand employees' social community structures within companies.

Meanwhile, nowadays, to facilitate the communication and cooperation among employees within companies, a new family of online social networks has been adopted in many companies, which are called the "*enterprise social networks*" (ESNs) [32]. A representative example of online ESNs is Yammer. Over 500,000 businesses around the world are now using Yammer, including 85% of the Fortune 500. ESNs can provide employees with various types of professional services to help them deal with daily work issues, e.g., identification of cooperators with required expertise; document sharing among the team members; and instant notifications of some important internal events (like management board changes, and newly released products). What's more, in online ESNs, employees can also perform various personal and casual social activities, like (1) making new friends, (2) chatting with other employees on various personal and work-related topics, (3) socializing with other employees and writing comments on their posts, as well as (4) creating personal interest groups and inviting their
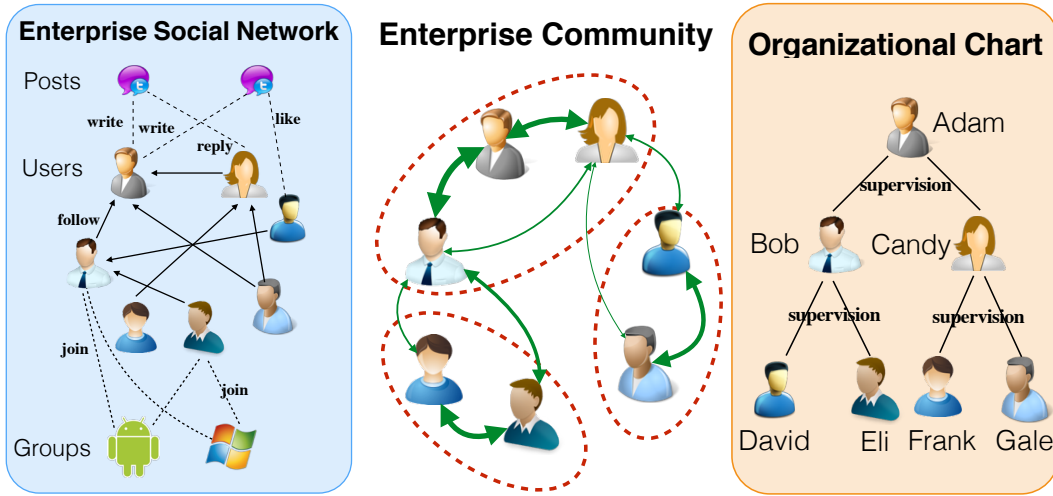
Fig. 1. An example of enterprise community detection (red circles: detected community structures; green links: enterprise intimacy scores calculated among employees, and wider links denote larger intimacy scores).

friends to join the groups. Generally, employees who socialize with each other more frequently in online ESNs are more close and will be partitioned into the same communities. Therefore, the personal information about employees' in online ESNs is an important complement of their company internal professional information, which together will provide us with a more comprehensive understanding about employees' community structures in companies.

Problem Studied: In this paper, we propose to study the social community structures of employees in companies based on both the company internal professional information and the personal information available in the online ESNs. Formally, the problem is named as the ECD (Enterprise Community Detection) problem. The ECD aims at fusing multiple information sources for the community detection problem, which is an important problem for big data researches. To illustrate the ECD problem more clearly, we also provide an example in Figure 1. Based on the information available in both online ESNs and offline organizational chart, we can compute the enterprise social intimacy scores among the employees (denoted as the width of the green links among the employees). These 7 employees are divided into 3 communities, and those with large intimacy scores are grouped into the same communities respectively.

The ECD problem is an important research problem, which can be the prerequisite for many concrete applications in the real-world companies, like *team formation for company internal projects* [15], and *friend recommendations in ESNs* [30], etc. Generally, employees who are in the same communities will know each other much better than those in other communities and the *communication costs* among them will be relatively lower, which is a very important concern in *team formation* problems [15]. Meanwhile, for employees who are in the same communities, we can recommend them as potential friends to each other in the online ESNs, and they are more

likely to accept the recommendations [30].

Besides its importance, the ECD problem is a new problem, which has never been studied before. To the best of our knowledge, we are the first to propose the *enterprise community detection* (ECD) problem in the enterprise context. The ECD problem is different from existing works on community detection problems, as the existing methods fail to consider the enterprise context nor the information about employees in internal company sources and online ESNs. Therefore, the existing clustering and community detection methods cannot be applied to address the ECD problem directly.

In addition, the ECD problem is very challenging to address due to the following reasons:

- *Online ESNs based Community Extraction*: Different types of social interactions about the employees in online ESNs, e.g., friendship, interest groups memberships, posts and employee generated contents, provide important information about the employees closeness in the online world. A clear definition about the *enterprise social intimacy* and the *social community structure* of the employees based on the information in ESNs is desired for addressing the ECD problem.
- *Internal Information based Community Extraction*: Various categories of internal company information, e.g., the *company organizational chart*, *employees' job titles*, and *employees' working locations*, all can indicate the closeness among the employees in the company from different aspects. A formal definition of the *enterprise social intimacy* and the *social community structure* of the employees can be the prerequisite for utilizing the company internal information to address the ECD problem.
- *Community Structure Fusion*: Distinct employees' community structures can be detected with these various types of information about employees in both online ENS and company internal sources respectively. An effective

fusion of these information sources can provide us with a more comprehensive understanding about employees' social communities in companies, which is still an open problem so far.

To address the above challenges, a new enterprise social community detection framework HUMOR (HeterogeneoUs Multi-sOurce ClusteRing) is introduced in this paper. A new concept named *enterprise social intimacy* is formally defined to measure the closeness among employees in this paper. HUMOR calculates the intimacy scores among employees based on the heterogeneous information available in online ESNs and offline company internal sources respectively. Based on each category of the information, HUMOR can detect unique community structures involving the employees, which are called the *micro enterprise communities* in this paper. Framework HUMOR obtains the globally consistent enterprise community structure by fusing these detected *micro enterprise communities* with two hierarchical phases: (1) *intra-fusion* of the *micro enterprise communities* detected in online ESNs and company internal sources respectively, and (2) *inter-fusion* of the enterprise community structures between online ESNs and company internal sources.

The remaining part of this paper is organized as follows. At first, we introduce several important concepts and define the ECD problem in Section II. The framework HUMOR is introduced in detail in Section III, and evaluated in Section IV. Finally, we talk about the related works in Section V and conclude this paper in Section VI.

II. PROBLEM FORMULATION

In this section, we will first define several important concepts used in this paper, based on which, we will introduce the formulation of the ECD problem next.

A. Terminology Definition

Enterprise social networks studied in this paper are a new type of social networks launched in the firewalls of companies especially, which can be formally defined as follows.

Definition 1 (Enterprise Social Network (ESN)): An *enterprise social network* can be represented as a heterogeneous information network $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes, and $\mathcal{E}$ represents the complex links in the network.

In this paper, we will take Yammer as an example of online ESNs. Users in Yammer can perform various kinds of social activities, e.g., (1) follow other users, (2) create and join social groups of their interests, (3) write posts, and (4) comment on/reply/like posts written by others. Therefore, according to the definition, the Yammer network studied in this paper can be represented as $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{U} \cup \mathcal{G} \cup \mathcal{P}$ contains the node sets of the users, social groups and posts, and $\mathcal{E} = \mathcal{E}_{u,u} \cup \mathcal{E}_{u,g} \cup \mathcal{E}_{u,p}^w \cup \mathcal{E}_{u,p}^c \cup \mathcal{E}_{u,p}^l$ involves the links among users, those between users and groups, and the write/comment/like links between users and posts respectively. Besides the online ESNs, a large amount of information (e.g., the organizational chart, and various other attribute information) about the employees is available inside the company,

which can be formally represented as an attribute augmented organizational chart.

Definition 2 (Attribute Augmented Organizational Chart): An *attribute augmented organizational chart* can be represented as a rooted tree $T = (\mathcal{N}, \mathcal{L}, root, \mathcal{A})$, where $\mathcal{N}$ denotes the set of employees in the company, $\mathcal{L}$ represents the management links from managers to subordinates, and $root \in \mathcal{N}$ indicates the CEO of the company. Set $\mathcal{A}$ is the set of attributes attached to employees in $\mathcal{N}$. For instance, in this paper, each employee in the company is associated with both the *job title* and *workplace* attributes. Therefore, set $\mathcal{A}$ can be represented as $\mathcal{A} = \mathcal{A}_t \cup \mathcal{A}_l$, where job title attribute set $\mathcal{A}_t = \{A_t(u_1), A_t(u_2), \cdots, A_t(u_{|\mathcal{N}|})\}$ involves the job title attribute of all the employees and employee workplace attribute set $\mathcal{A}_l = \{A_l(u_1), A_l(u_2), \cdots, A_l(u_{|\mathcal{N}|})\}$ contains the workplace attribute of the employees.

Generally, the organizational chart and company internal attribute information sources involve all the employees in the company and we treat the employee node set $\mathcal{N}$ in the above definition as the complete employee set. Among all the employees, some of them may be using the online ESNs. In other words, the users joining in the online ESNs is actually a subset of the whole employees, i.e., $\mathcal{U} \subseteq \mathcal{N}$, which will be considered when conducting the *inter-fusion* of community structures between online ESNs and company internal sources.

B. Problem Definition

Based on the concepts defined above, we can define the ECD problem formally in this section.

Problem Definition (The ECD Problem): Given the online ESN G and the offline attribute augmented organizational chat T, the ECD problem aims at inferring the community structure of employees in the company. More specifically, we aim at partition the employee set $\mathcal{N}$ in the company into K disjoint social communities $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$. In this paper, we don't consider that case that each employee is involved in multiple communities. Therefore, for these detected community structures we have $C_i \cap C_j = \emptyset, \forall i, j \in \{1, 2, \cdots, K\}, i \neq j$ and $\bigcup_i C_i = \mathcal{N}$. Generally, the employees grouped in each community (e.g., C_i) tend to interact with each other more frequently and have larger closeness scores with each other compared with those in other communities (i.e., $\mathcal{C} \setminus C_i$).

III. PROPOSED METHODS

In this section, we will talk about the enterprise community detection framework HUMOR in great detail. At first, we will introduce the definition of *ESNs based enterprise social intimacy* measure and calculate the intimacy scores among employees based on the different types of social interaction information in online ESNs in Section III-A1. With these calculated intimacy scores, a set of *micro enterprise communities* can be extracted, which will be fused with the *intra-fusion* step by HUMOR in Section III-A2. Similarly, in Section III-B1, we will calculate the *company internal information based intimacy* scores among employees based on the company internal information, and fuse the detected *micro enterprise*

communities with the *intra-fusion* step in Section III-B2. Finally, to obtain the globally consistent enterprise community structure of the employees, we will formulate the *inter-fusion* step of the HUMOR framework as a joint optimization problem, where will be introduced in Section III-C.

A. ESNs based Community Detection

In this part, we will introduce the *enterprise intimacy* concept based on ESNs. With the various categories of social information in online ESNs, we can calculate different types of enterprise intimacy scores among employees and detect the corresponding micro enterprise communities respectively.

1) Enterprise Social Intimacy in Online ESN: Generally, social intimacy is a measure of how people closely interact with each other. At an individual level, specifically, the enterprise social intimacy involves the quality and number of connections one has with other employees in the company.

Definition 3 (ESNs based Enterprise Social Intimacy): Based on the social interactions information available in the online *enterprise social network* $G = (\mathcal{V}, \mathcal{E})$, the *ESNs based enterprise social intimacy* between employees u and v denotes how close u and v are in the online ESNs G, which can be represented as $EI^g(u, v)$ in this paper.

In this section, we will mis-use "employees" and "users" to represent the individuals involved in online ESNs. As introduced in the previous section, employees in online ESNs can perform various types of social activities, based on which, different enterprise social intimacy measures can be calculated.

Social Connection based Enterprise Social Intimacy

According to the definition introduced in Section II, the set of social connections among the employees in online ESN G can be represented as set $\mathcal{E}_{u,u}$. The connections among employees $(u, v) \in \mathcal{E}_{u,u}$ is a directed link, which denotes that employee u follows v in network G. Since the *enterprise intimacy* concept is symmetrical, when calculating the intimacy scores among employees based on the social connection information, we will neglect the directions of the social links in this paper. Formally, for a certain employee user $u \in \mathcal{U}$, we can represent the set of neighbor employees that are connected with u as set $\Gamma_s^g(u) = \{v | v \in \mathcal{U}, (u, v) \in \mathcal{E}_{u,u} \vee (v, u) \in \mathcal{E}_{u,u}\}$, where the subscript s denotes that it is based on the social connection information and the superscript g indicates it is extracted from the online ESN G. Similarly, we can also represent the set of neighbors of employee v as set $\Gamma_s^g(v)$. Furthermore, the common neighbors shared by u and v can be denotes as set $\Gamma_s^g(u) \cap \Gamma_s^g(v)$, which is a strong indicator about the enterprise social intimacy between employees u and v, and close employees would share more common neighbors.

However, simply counting the common neighbors may suffer from some problems: employees u and v can be connected by a large number of common neighbors merely because of the large social degrees of either u, v or the intermediate nodes between them, but the affinity between u and v is not high. To address this problem, we propose to weight the degrees of employees u and v to penalize those popular ones in the closeness calculation. To achieve such a goal, we propose the following pointwise mutual information based social intimacy between employees u and v:

$$EI_s^g(u, v) = \frac{|\Gamma_s^g(u) \cap \Gamma_s^g(v)|}{|\mathcal{U}|} \log \frac{\frac{|\Gamma_s^g(u) \cap \Gamma_s^g(v)|}{|\mathcal{U}|}}{\frac{|\Gamma_s^g(u)|}{|\mathcal{U}|} \cdot \frac{|\Gamma_s^g(v)|}{|\mathcal{U}|}}.$$

Group Membership based Enterprise Social Intimacy

Besides social connections, employee users can also create/join their personal interest groups in ESNs. Intuitively, sharing more common interest groups often suggests that two employees have more common interests and are closer to each other. Based on the group membership information in online ESNs, we propose to calculate the group participation activity based enterprise social intimacy among employees. As introduced before, in the online ESN G, the employees' group membership links can be represented as set $\mathcal{E}_{u,g}$, and the set of groups that employee u is involved in can be represented as $\Gamma_g^g(u) \subseteq \mathcal{G}$. Similarly, we can represent group set for v as $\Gamma_g^g(v) \subseteq \mathcal{G}$, and the groups shared by u and v can be represented as $\Gamma_g^g(u) \cap \Gamma_g^g(v)$.

However, in online ESNs, different groups may have different discrimination power, where larger groups involving a large number of members can hardly show the intimacy among the participants. More specifically, we can represent the set of employees involved in a group g in the online ESN as set $\Gamma_g^g(g) \subset \mathcal{U}$. For any two different social groups g_1 and g_2 both containing employees u and v, group g_1 will have a greater discrimination power than g_2 if $\Gamma_g^g(g_1) < \Gamma_g^g(g_2)$. In order to capture the discrimination of different groups, we propose to apply the "Inverse Membership Frequency" (IMF) introduced in [30], inspired by the IDF measure [13] widely used in information retrieval, and the IMF of group g can be represented as $\text{IMF}(g) = \log \frac{|\mathcal{U}|}{|\Gamma_g^g(g)|}$.

Furthermore, the ESN group membership based enterprise social intimacy between any two employees u and v can be represented as the sum of the shared groups' IMF measures

$$EI_g^g(u, v) = \sum_{g \in \Gamma_g^g(u) \cap \Gamma_g^g(v)} \text{IMF}(g) = \sum_{g \in \Gamma_g^g(u) \cap \Gamma_g^g(v)} \log \frac{|\mathcal{U}|}{|\Gamma_g^g(g)|}.$$

Text Content based Enterprise Social Intimacy

In addition to the social connection and group membership information, employees in online ESNs can also write/reply/comment on/like posts, which will generate lots of text content related information. In this paper, we will not go deep into dealing with the text words, and simply calculate the enterprise social intimacy based on the correlated posts shared by different employees. For instance, we can represent the sets of posts that employee u and v have written/replied/commented on/liked as $\Gamma_p^g(u) \subseteq \mathcal{P}$ and $\Gamma_p^g(v) \subseteq \mathcal{P}$ respectively. The set of shared posts that u and v both correlate to can be represented as $\Gamma_p^g(u) \cap \Gamma_p^g(v)$. Generally, the more interactions employees have via the posts online, the closer they would be, but the total number of posts employees are correlated to (i.e., the degrees of u and v) also need to be considered carefully. Therefore, based on the user generated content (i.e., post) information in online ESNs, can represent

the enterprise social intimacy between u and v as the Jaccard's Coefficient (JC) of the post sets about employees u and v:

$$EI_p^g(u,v) = \text{JC}(\Gamma_p^g(u), \Gamma_p^g(v)) = \frac{|\Gamma_p^g(u) \cap \Gamma_p^g(v)|}{|\Gamma_p^g(u) \cup \Gamma_p^g(v)|}.$$

2) ESNs based Enterprise Community Detection and Intra-Fusion: **Micro Community Structure Detection in ESNs**

Based on each enterprise social intimacy measure introduced in the previous section, a concrete enterprise social intimacy matrix can be constructed. For instance, according to the social connection based enterprise social intimacy, we can define the enterprise social intimacy matrix $\mathbf{A}_s^g \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{U}|}$, where entry $A_s^g(i,j) = EI_s^g(u_i, u_j)$, $u_i, u_j \in \mathcal{U}$ denotes the social connection information based enterprise social intimacy between employees u_i and u_j. In a similar way, we can represent the enterprise social intimacy scores among employees based on the group participation and user generated content information as matrices $\mathbf{A}_g^g, \mathbf{A}_p^g \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{U}|}$ respectively.

Based on each of the enterprise social intimacy matrix (e.g., $\mathbf{A}_s^g$), various methods can be applied to detect the community structure of the employees. In this paper, we propose to use the non-negative matrix factorization (NMF) method to partition the intimacy matrix $\mathbf{A}_s^g$ due to its outstanding performance and wide applications in clustering problems [28]. The NMF method aims at minimizing the following objective equation to infer the community structure hidden factor matrix $\mathbf{U}$:

$$J_C(\mathbf{A}_s^g) = \left\| \mathbf{A}_s^g - \mathbf{U}_s \mathbf{U}_s^\top \right\|_F^2.$$

Here, the entry $U_s(i,j)$ in the hidden factor matrix represents the confidence of employee $u_i \in \mathcal{U}$ belonging to the j_{th} community $C_j \in \mathcal{C}$. Formally, the enterprise social community structure described by matrix $\mathbf{U}_s$ is called the *micro enterprise community structure* in this paper.

Similarly, we can define the objective equations $J_C(\mathbf{A}_g^g)$ and $J_C(\mathbf{A}_p^g)$ of the enterprise social intimacy matrix based on the group membership and user generated content information. The corresponding hidden factor matrices $\mathbf{U}_g$ and $\mathbf{U}_p$ can be obtained by minimizing the following two objective equations

$$J_C(\mathbf{A}_g^g) = \left\| \mathbf{A}_g^g - \mathbf{U}_g \mathbf{U}_g^\top \right\|_F^2, \;\; J_C(\mathbf{A}_p^g) = \left\| \mathbf{A}_p^g - \mathbf{U}_p \mathbf{U}_p^\top \right\|_F^2.$$

Intra-Fusion of Micro Community Structures in ESNs

The detected hidden factor matrices $\mathbf{U}_s, \mathbf{U}_g, \mathbf{U}_p$ characterize the community structures of the employee users in ESNs from different aspects, which can be different from each other. Meanwhile, these hidden factor matrices are all about community structures of the same set of employees in the online ESNs information source, which should be consistent with the real-world community structure of the employees. To achieve such a goal, in this paper, an *consistent hidden factor matrix* $\mathbf{U} \in \mathbb{R}^{|\mathcal{U}| \times K}$ is introduced to represent the consistent community structure of the employees in online ESNs. To infer $\mathbf{U}$, we add a set of extra regularization terms to minimize the differences of community structure described by $\mathbf{U}$ from those described by $\mathbf{U}_s, \mathbf{U}_g, \mathbf{U}_p$ respectively:

$$J_R(\mathbf{A}_s^g, \mathbf{A}_g^g, \mathbf{A}_p^g) = \|\mathbf{U}_s - \mathbf{U}\|_F^2 + \|\mathbf{U}_g - \mathbf{U}\|_F^2 + \|\mathbf{U}_p - \mathbf{U}\|_F^2.$$

Formally, the process of inferring the consistent community structure matrix $\mathbf{U}$ is called the *intra-fusion* phrase in the HUMOR framework. Based on the above remarks, the optimal consistent community structure matrix $\mathbf{U}^*$ which can minimize the matrix decomposition costs and regularization terms simultaneously can be obtained be resolving the following objective equation:

$$\mathbf{U}^* = \arg_{\mathbf{U}} \min_{\mathbf{U}, \mathbf{U}_s, \mathbf{U}_g, \mathbf{U}_p} \sum_{i \in \{s,g,p\}} J_C(\mathbf{A}_i^g) + \alpha \cdot J_R(\mathbf{A}_s^g, \mathbf{A}_g^g, \mathbf{A}_p^g),$$

where α represents the weight of the consistency regularizer and it is assigned with value 1 in the experiments.

B. Company Internal Information based Enterprise Community Detection

Besides the information available in online ESNs, employees also have a large amount of information available in the offline company information, which can also help identify the social community structures of the employees.

1) Internal Information based Enterprise Intimacy: In this section, we will measure the enterprise social intimacy among employees based on the company internal information, which include the organizational chart, job titles and workplaces.

Organizational Chart based Enterprise Intimacy

Company organizational chart is a tree structured diagram outlining the management relationships among employees, where employees are connected by the "management" links between managers and their subordinates. For any two given employees u and v, an intuitive idea to represent their intimacy is using the number of required steps to walk between them along the "management" links in the organizational chart. Generally, employees connected by less "management links" (e.g., colleagues in the same group) are closer to each other in the organizational chart. Based on such an observation, we propose to follow the closeness measure "organizational intimacy" introduced in [30] for employees u and v as the inverse of required number of steps to walk from u to v via the links in the organizational chart.

Formally, let $\text{step}(u,v)$ denote the number of steps needed to walk from u to v along the links in the adjusted organizational chart, based on which the enterprise social intimacy $EI_c^t(u,v)$ between u and v can be represented as

$$EI_c^t(u,v) = \frac{1}{\text{step}(u,v)},$$

where the subscripts c denotes that $EI_c^t(u,v)$ is based on the <u>c</u>hart information, and the superscript t denotes that the information is from the company internal information in T.

Job Title based Enterprise Intimacy

Employees in the enterprise have different job titles to indicate their duties and relative ranks in the companies. From the point of view of job content, employees sharing common functionality on job titles will have very similar workload and work categories, and they will have more chance to cooperate with each other. To use the job title information to calculate the enterprise social intimacy among employees in enterprises, we

propose to extract the root job terms from all job titles, which can categorize what the job is actually about. For instance, the root job term of both "Senior SDE" and "SDE" are both "SDE", but that of "Senior Researcher" will be "Researcher" instead. Formally, for any two employees, if their job titles are of the same job categories (i.e., sharing the common root job term), then we will apply the Jaccard's Coefficient based method to calculate their enterprise intimacy; otherwise, their enterprise social intimacy will be 0 instead. Let $\text{job}(u)$ and $\text{root}(u)$ represent the bag of word representation and the root job term of employee u's job title respectively. We can represent the job title based enterprise social intimacy between employees u and v as:

$$EI_t^t(u, v) = I\left(\text{root}(u), \text{root}(v)\right) \frac{|\text{job}(u) \cap \text{job}(v)|}{|\text{job}(u) \cup \text{job}(v)|},$$

where $I\left(\text{root}(u), \text{root}(v)\right) = 1$ iff $\text{root}(u) = \text{root}(v)$.

Here, we need to clarify that we don't mean employees with similar job titles will be close to each other for certain. Actually, the job title provides an important signal for us to infer their potential closeness. In many companies, employees taking similar category of workload will work in the same workplace (e.g., HR people on one floor, SDEs work together at another floor), and they have more chance to meet each other. The job title together with other categories of information will provide more comprehensive knowledge about the closeness among employees at workplace.

Workplace based Enterprise Intimacy

Besides the organizational chart and job title information, we can also know the workplaces of employees from the company, which can help denote the geographical intimacy among employees. Specifically, employees' workplace information includes both the *country* and *time zone* of employees' offices. Generally, employees working in closer locations (e.g., in the same country, or the same time zone) have more chance to interact with each other. Therefore, in this paper, we propose to measure the enterprise intimacy among employees by checking whether they are in the same country, and the same time zone or not. Let $\text{country}(u)$ and $\text{time zone}(u)$ represent the country and time zone of employee u's offices. The workplace based enterprise intimacy between employees u and v can be represented as

$$EI_l^t(u, v) = \frac{1}{2}\Big(I\big(\text{country}(u), \text{country}(v)\big) + I\big(\text{zone}(u), \text{zone}(v)\big)\Big),$$

where $I\big(\text{country}(u), \text{country}(v)\big) = 1$ iff $\text{country}(u) = \text{country}(v)$, and similarly for function $I\big(\text{zone}(u), \text{zone}(v)\big)$.

2) Internal Information based Community Detection and Intra-Fusion: Based on the above enterprise social intimacy measures, we can represent the intimacy scores among employees calculated based on organizational chart, job title and workplace information in the offline enterprise as adjacency matrices $\mathbf{A}_c^t, \mathbf{A}_t^t, \mathbf{A}_l^t \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$ respectively. Similarly, the non-negative matrix factorization (NMF) method can be applied to partition the intimacy matrices to obtain their corresponding community hidden factor matrices, and

the decomposition cost functions are listed as follows:

$$J_C(\mathbf{A}_c^t) = \left\|\mathbf{A}_c^t - \mathbf{V}_c\mathbf{V}_c^\top\right\|_F^2, J_C(\mathbf{A}_t^t) = \left\|\mathbf{A}_t^t - \mathbf{V}_t\mathbf{V}_t^\top\right\|_F^2,$$
$$J_C(\mathbf{A}_l^t) = \left\|\mathbf{A}_l^t - \mathbf{V}_l\mathbf{V}_l^\top\right\|_F^2.$$

Meanwhile, these community hidden factor matrices $\mathbf{V}_c, \mathbf{V}_t$ and $\mathbf{V}_l$ all describe the *micro enterprise community structures* of the employees in the offline enterprise from different perspectives, which should be consistent with each other as well. To define the consistency regularization term, a new consistent community hidden factor matrix $\mathbf{V} \in \mathbb{R}^{|\mathcal{N}| \times K}$ is introduced, and the consistency regularization term between $\mathbf{V}$ and $\mathbf{V}_c, \mathbf{V}_t, \mathbf{V}_l$ can be represented as follows:

$$J_R(\mathbf{A}_c^t, \mathbf{A}_t^t, \mathbf{A}_l^t) = \|\mathbf{V}_c - \mathbf{V}\|_F^2 + \|\mathbf{V}_t - \mathbf{V}\|_F^2 + \|\mathbf{V}_l - \mathbf{V}\|_F^2.$$

Based on the above remarks and the *intra-fusion* strategy, the optimal consistent community structure matrix $\mathbf{V}^*$ of employees in the offline enterprise can be obtained by resolving the following objective equation:

$$\mathbf{V}^* = \arg_{\mathbf{V}} \min_{\mathbf{V}, \mathbf{V}_c, \mathbf{V}_t, \mathbf{V}_l} \sum_{i \in \{c, t, l\}} J_C(\mathbf{A}_i^t) + \alpha \cdot J_R(\mathbf{A}_c^t, \mathbf{A}_t^t, \mathbf{A}_l^t),$$

where the same parameter α (with value 1 in the experiments) is also used to represents the weight of the consistency regularization term here.

C. Inter-Fusion of Enterprise Communities

Generally, the company contains the complete information about all the employees, and some of whom can get involved in the online ESNs as well. In this section, we will introduce the joint optimization objective function for detecting the communities via the *inter-fusion* of information available in the online ESNs and offline company.

For the common employees shared by the online ESNs and offline company internal information sources, the detected community structures about them should be consistent with each other. To achieve such a goal, we propose to regularize the detected community structures from online ESNs and offline company internal information sources. Before introducing the regularization term, we first define the binary employee transition matrix $\mathbf{T} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{N}|}$, where entry $T(i, j) = 1$ iff employee u_i and u_j are actually the same employee in the online ESNs and offline enterprise respectively. Based on the transitional matrix, we can map the employees as well as their hidden community matrix from the online ESN to the offline enterprise by simply multiplying it with the transition matrix. For instance, given the hidden community structure matrix $\mathbf{U}$ of the employees online, we can represent the mapped community structure matrix to the offline company as $\mathbf{T}^\top\mathbf{U}$. The difference of the employees' community structure mapped from the online ESNs (i.e., $\mathbf{T}^\top\mathbf{U}$) and that obtained from the offline enterprise (i.e., $\mathbf{V}$) can be represented as inter-source community regularization term

$$J_D(\mathbf{U}, \mathbf{V}) = \left\|(\mathbf{T}^\top\mathbf{U})(\mathbf{T}^\top\mathbf{U})^\top - \mathbf{V}\mathbf{V}^\top\right\|_F^2,$$

where the non-zero entries in matrices $(\mathbf{T}^\top\mathbf{U})(\mathbf{T}^\top\mathbf{U})^\top$ and $\mathbf{V}\mathbf{V}^\top$ denote the confidence scores for the corresponding employees to be in the same community based on the information available in online ESNs and offline company internal data respectively.

By adding the inter-source community regularization term, we can redefine the enterprise community detection objective equation. Formally, the process of adding the inter-source community regularization term $J_D(\mathbf{U},\mathbf{V})$ to the objective function is named as the *inter-fusion* procedure in this paper. The new objective equation can be represented as:

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i\in\{s,g,p\}} J_C(\mathbf{A}_i^g) + \alpha \cdot J_R(\mathbf{A}_s^g, \mathbf{A}_g^g, \mathbf{A}_p^g)$$
$$+ \sum_{i\in\{c,t,l\}} J_C(\mathbf{A}_i^t) + \alpha \cdot J_R(\mathbf{A}_c^t, \mathbf{A}_t^t, \mathbf{A}_l^t) + \beta \cdot J_D(\mathbf{U},\mathbf{V}),$$

where parameter β denotes the weight of the inter-source community regularization term, whose sensitivity analysis is also available in Section IV.

By replacing the cost and regularization functions with the matrix representations, we can obtain the final joint objective function involving 8 hidden factor matrix variables, each of which can represent the community structure of the employees from different aspects. However, simultaneous inference of the optimal values for the matrix variables can be very computational hard and time consuming.

To simplify the problem, in this paper, we propose to constrain the *intra-fusion* regularization terms $J_R(\mathbf{A}_1^g, \mathbf{A}_2^g, \mathbf{A}_3^g)$ as well as $J_R(\mathbf{A}_c^t, \mathbf{A}_t^t, \mathbf{A}_l^t)$ to be both 0, i.e.,

$$\mathbf{U}_s = \mathbf{U}_g = \mathbf{U}_p = \mathbf{U},$$
$$\mathbf{V}_c = \mathbf{V}_t = \mathbf{V}_l = \mathbf{V}.$$

And the simplified objective function will contain two variables $\mathbf{U}$ and $\mathbf{V}$ only, which will be much more efficient to address. The simplified objective function can be represented as

$$\min_{\mathbf{U},\mathbf{V}} \left\|\mathbf{A}_s^g - \mathbf{U}\mathbf{U}^\top\right\|_F^2 + \left\|\mathbf{A}_g^g - \mathbf{U}\mathbf{U}^\top\right\|_F^2 + \left\|\mathbf{A}_p^g - \mathbf{U}\mathbf{U}^\top\right\|_F^2$$
$$+ \left\|\mathbf{A}_c^t - \mathbf{V}\mathbf{V}^\top\right\|_F^2 + \left\|\mathbf{A}_t^t - \mathbf{V}\mathbf{V}^\top\right\|_F^2 + \left\|\mathbf{A}_l^t - \mathbf{V}\mathbf{V}^\top\right\|_F^2$$
$$+ \beta \cdot \left\|(\mathbf{T}^\top\mathbf{U})(\mathbf{T}^\top\mathbf{U})^\top - \mathbf{V}\mathbf{V}^\top\right\|_F^2.$$

The objective equation is not actually jointly convex and no closed-form solution exists. In this paper, we propose to solve with an alternative updating approach. We will fix one variable (e.g., $\mathbf{V}$) and update the other variable (e.g., $\mathbf{U}$) iteratively and alternatively, and such a process continues until all the variables converge.

Let $\mathcal{L}(\mathbf{U},\mathbf{V})$ denote the objective function. In iteration τ, the updating equations can be represented as

$$\mathbf{U}^{(\tau)} = \mathbf{U}^{(\tau-1)} - \eta_1 \cdot \frac{\partial\mathcal{L}(\mathbf{U}^{(\tau-1)}, \mathbf{V}^{(\tau-1)})}{\partial(U)}$$
$$= \mathbf{U}^{(\tau-1)} - 2\eta_1\Big(6\mathbf{U}^{(\tau-1)}(\mathbf{U}^{(\tau-1)})^\top\mathbf{U}^{(\tau-1)} - (\mathbf{A}_s^g + \mathbf{A}_g^g + \mathbf{A}_p^g)\mathbf{U}^{(\tau-1)}$$
$$- ((\mathbf{A}_s^g)^\top + (\mathbf{A}_g^g)^\top + (\mathbf{A}_p^g)^\top)\mathbf{U}^{(\tau-1)} - 2\mathbf{T}\mathbf{V}^{(\tau-1)}(\mathbf{V}^{(\tau-1)})^\top\mathbf{T}^\top\mathbf{U}^{(\tau-1)}$$
$$+ 2\mathbf{T}\mathbf{T}^\top\mathbf{U}^{(\tau-1)}(\mathbf{U}^{(\tau-1)})^\top\mathbf{T}\mathbf{T}^\top\mathbf{U}^{(\tau-1)}\Big).$$

	property	**Yammer**
# node	user	119,989
	group	10,783
	conversation	688,230
	post	12,768,795
# link	follow	496,068
	group membership	518,529
	initiate	688,230
	write	12,768,795
	like	250,953

$$\mathbf{V}^{(\tau)} = \mathbf{V}^{(\tau-1)} - \eta_2 \cdot \frac{\partial\mathcal{L}(\mathbf{U}^{(\tau)}, \mathbf{V}^{(\tau-1)})}{\partial(V)}$$
$$= \mathbf{V}^{(\tau-1)} - 2\eta_2\Big(8\mathbf{V}^{(\tau-1)}(\mathbf{V}^{(\tau-1)})^\top\mathbf{V}^{(\tau-1)} - (\mathbf{A}_c^t + \mathbf{A}_t^t + \mathbf{A}_l^t)\mathbf{V}^{(\tau-1)}$$
$$- ((\mathbf{A}_c^t)^\top + (\mathbf{A}_t^t)^\top + (\mathbf{A}_l^t)^\top)\mathbf{V}^{(\tau-1)} - 2\mathbf{T}^\top\mathbf{U}^{(\tau)}(\mathbf{U}^{(\tau)})^\top\mathbf{T}\mathbf{V}^{(\tau-1)}\Big).$$

where parameters η_1 and η_2 denote the gradient descent steps in updating matrices $\mathbf{U}$ and $\mathbf{V}$ respectively. The optimal learning rates η_1 and η_2 in each iteration steps obtaining the minimum $\mathcal{L}(\mathbf{U},\mathbf{V})$ can be represented as

$$\eta_1^{(\tau)} = \arg_{\eta_1} \min \mathcal{L}(\mathbf{U}^{(\tau)}, \mathbf{V}^{(\tau)}),$$
$$\eta_2^{(\tau)} = \arg_{\eta_2} \min \mathcal{L}(\mathbf{U}^{(\tau)}, \mathbf{V}^{(\tau)}).$$

The functions can be addressed by taking derivative of $\mathcal{L}(\cdot)$ with regards to η_i (or η_2) and make it equal to 0, we can obtain a cubic equation involving η_i (or η_2). Multiple roots may exist when addressing the equation and the representation of the roots is very complicated. In this paper, for simplicity, we propose to assign η_i and η_2 with a small constant value (i.e., 0.05 in the experiments). The above alternative updating scheme involves the multiplication of matrices. Let n be the number of employees in the company and τ be the required rounds to achieve convergence, the time complexity of the alternative updating method can be represented $O(\tau n^{2.373})$, where the Optimized Coppersmith-Winograd algorithm [5] is applied in the matrix product calculation. In addition, some pre-computation can be applied to effective reduce the real running time of the framework by storing the results of matrices $(\mathbf{A}_s^g)^\top + (\mathbf{A}_g^g)^\top + (\mathbf{A}_p^g)^\top$, $\mathbf{T}\mathbf{T}^\top$, etc., in advance.

IV. EXPERIMENTS

To test the effectiveness of the proposed framework HUMOR in detecting the communities in companies, extensive experiments have been done on real-world enterprise datasets. In this section, we will first describe the datasets used in this paper, and then introduce the experiment settings in detail. Finally, we will show the experiment results together with detailed explanation and give the parameter sensitivity analysis.

A. Dataset Description

The enterprise datasets used in this paper is about the Microsoft company. We crawled all the data about the employees from Yammer as well as the complete organizational chart

	property	**MSR**
# node	employee	688
	team	142
# link	team member	1,089

during June, 2014, whose statistical information is given in Table I. Brief descriptions about both the Yammer enterprise social network data and the Microsoft organizational chart are available as follows.

- *Enterprise Social Network*: We crawl all the Microsoft employees' information from Yammer and obtain the complete organizational chart involving all these employees in Microsoft during June, 2014 [32], [30]. The social network data covers all the user-generated content (such as posts, replies, topics, etc.) and social graphs (such as user-user following links, user-group memberships, user-topic following links, etc.) by then that are set to be public.

- *Organizational Chart*: Besides the online *enterprise social network* and *enterprise organizational chart* datasets, we can also obtain the job titles and the specific workplaces of all the employees from Microsoft, which together with the *company organizational chart* is treated as the company internal information in this paper.

These above datasets are used as the online and offline information sources for building the models only (e.g., to calculate the intimacy scores). Meanwhile, to evaluate the proposed framework, we further crawled a dataset containing employees together their corresponding research teams from the official website of Microsoft Research (MSR)[1]. The statistical information about the MSR dataset is available in Table II. Here, we need to clarify that the research teams crawled from MSR are not the same as the organizational chart. Each research team denotes an on-going project carried in MSR, in which the team members may not necessary from the same departments. In the research teams, researchers with different skills and from different departments (in the organizational chart) are invited to cooperate together to carry out certain research/production projects that the team focuses on. For many research teams crawled from MSR, actually the team members can actually come from the campuses in different countries (e.g., MSR@US Redmond, MSR@US Silicon Valley, MSR@CN Beijing, MSR@UK Cambridge).

After necessary data cleaning and employee ID alignment with the *enterprise social network* and *organization chart* datasets, finally we obtain 688 employees together with 142 research teams in the MSR department. Each team in the dataset is treated as a community, and the employee-team membership is used as the ground truth of the enterprise community structure.

[1]http://research.microsoft.com/en-us/

B. Experiment Settings

In this part, we will introduce the experiment settings, which covers the detailed experiment setup, comparison methods and the evaluation metrics.

1) Experiment Setup: Based on the social connection, group participation and employee generated post information available in the online ESNs, we can construct the adjacency matrices $\mathbf{A}_s^g$, $\mathbf{A}_g^g$ and $\mathbf{A}_p^g$ involving the all the target employees $\mathcal{U}$ to be partitioned. Meanwhile, based on the organizational chart, job titles, and specific work location information available in the company internal sources, we can construct the adjacency matrices $\mathbf{A}_c^t$, $\mathbf{A}_t^t$ and $\mathbf{A}_l^t$ respectively. By incorporating these constructed matrices into the objective function and obtaining the final community hidden factor matrix $\mathbf{V}$, we can infer the community structures of the target employee set by applying some traditional clustering models (e.g., kmeans) to matrix $\mathbf{V}$.

In the experiments, we will first apply the HUMOR model to the small-sized MSR dataset, whose performance can be measured with the ground truth community structure, i.e., the research teams (each teams is treated as a community). In addition, we also apply the HUMOR model to all whole employees in Microsoft, and the performance of HUMOR is measured with some traditional evaluation metrics, which will be introduced at the end of this part.

2) Comparison Methods: To show the advantages of the HUMOR framework, we compare HUMOR with both traditional and state-of-art community detection algorithms, which can be divided into three categories:

Multi-Source Community Detection

- HUMOR: HUMOR is the method proposed in this paper, which utilizes information in both online ESNs and offline company internal information sources to infer the community structure of all the employees in the company.

Single-Source (ESNs Only) Community Detection

To show that the internal information sources are helpful for providing extra social information about the employees, we apply 4 other different baseline methods utilizing the information available in the online ESNs only but no information from the company internal information sources.

- HUMOR-ESN: We also propose a simplified version of HUMOR which just infer the community structure of the employees based on information in online ESNs only.

- INF-ESN: Method INF-ESN is the influence based clustering method proposed in [33], which proposes to propagate information among the employees (only the social link information) within one single information source (e.g., online ESNs), where employees with strong mutual influences will be grouped in the same cluster.

- CUT-ESN: Community detection method "normalized cut" [23] aims at minimizing the normalized cut of links between different clusters can be used to detect the communities. Method CUT-ESN that minimizes the cut of social links among employees is used as a comparison method in the experiments.
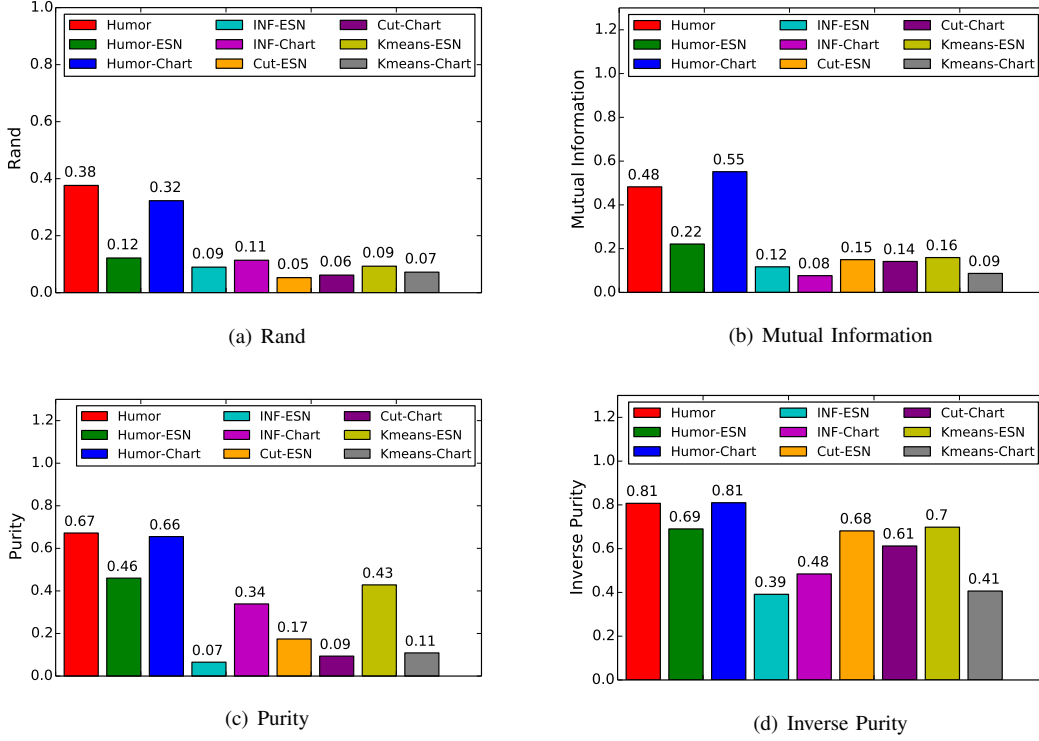
(a) Rand

(b) Mutual Information

(c) Purity

(d) Inverse Purity

Fig. 2. Experiment result of comparison methods evaluated by ground truth with Rand, Mutual Information, Purity and Inverse Purity metrics.

- KMEANS-ESN: In addition, for completeness, the traditional clustering method KMeans (KMEANS-ESN) is also used as a baseline method, where the social adjacency matrix involving the employees in online ESNs is used as the input of KMEANS-ESN.

Single-Source (Company Internal Information Only) Community Detection

To show that the online ESN is useful in providing extra social information about the employees, we also apply 4 different baseline methods utilizing the information available in the company internal information sources only.

- HUMOR-CHART: HUMOR-CHART is the simplified version of HUMOR corresponding to HUMOR-ESN, which infers the employees' community structure based on the information available in the company internal information sources.
- INF-CHART: Similar to INF-ESN, method INF-CHART builds the influence propagation model based on the organizational chart information among all the employees in the company internal information sources.
- CUT-CHART: Method CUT-CHART aims at minimizing the cut of management links between employees in the organizational chart in the company internal information sources.
- KMEANS-CHART: Method KMEANS-CHART uses the adjacency matrix constructed based on the organizational chart as the input and infers the employees community structure.

3) Evaluation Metrics: To measure the performance of the comparison methods, different metrics are applied in this paper.

For the MSR dataset, we have both the predicted community structure $\mathcal{C}_{pred}$ (outputted by the methods) and the real community structure $\mathcal{C}_{true}$ (i.e., the ground truth) about these employees. To evaluate the performance of the methods (i.e., compare $\mathcal{C}_{pred}$ and $\mathcal{C}_{true}$), the metrics used in this paper include *Rand* [18], *Mutual Information* (MI) [18], *Purity* [1] and *Inverse Purity* [1].

In addition, for the community structure $\mathcal{C}_{pred}$ outputted by different comparison methods, we will use 4 other widely applied metrics *normalized-dbi* [6], *silhouette index* [21], *density* [22], and *entropy* [18] in this paper. Metrics *ndbi*, *silhouette* will use the similarity score among the employees, *density* counts the number of edges in each of the community, and *entropy* measure the distribution of the community sizes. For *ndbi* and *silhouette*, we calculate the average of the 6 *enterprise intimacy* scores defined in this paper as the real-world similarities among the employees, while for *density*, we count the ratio sum of social links in ESNs and management links in organizational chart as the real-world links among employees.

C. Experiment Results

According to the experiments, the alternative updating scheme works very well and both the matrix **U** and **V** can converge within 30 iterations. The experiment results obtained based on the converged matrices are available in Figures 2-3,

(a) Density

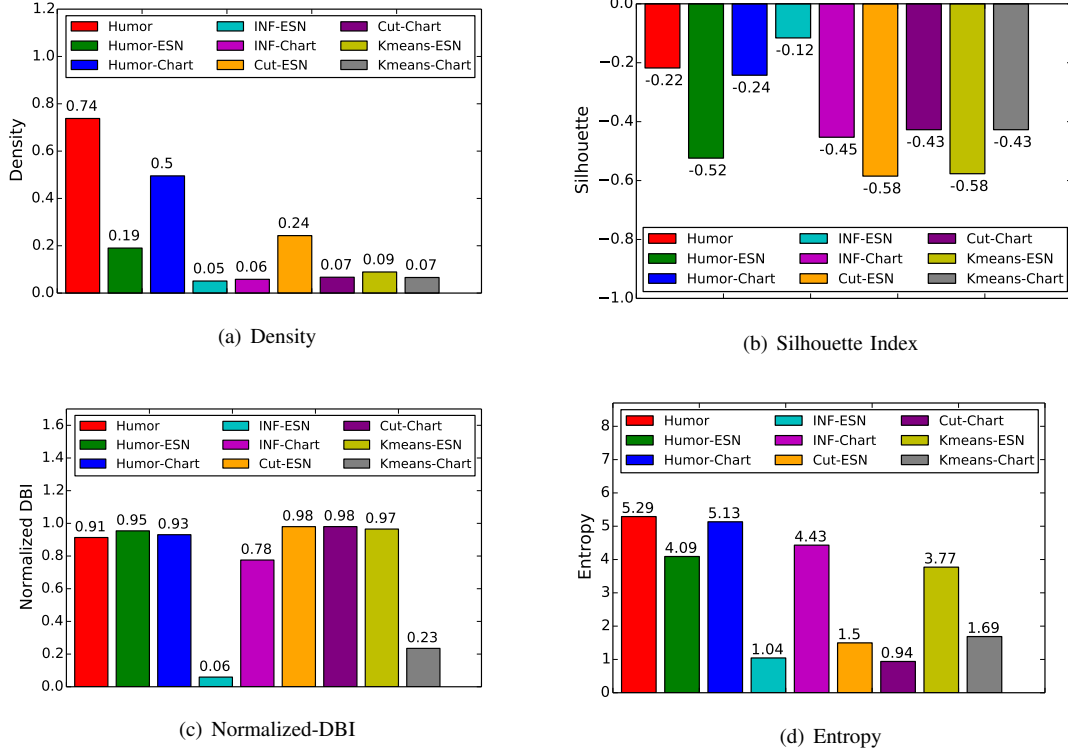(b) Silhouette Index

(c) Normalized-DBI

(d) Entropy

Fig. 3. Experiment result of comparison methods evaluated by the Density, Silhouette Index, Normalized-DBI and Entropy metrics.

where plots in Figure 2 show the results achieved by different methods compared with the ground truth and plots in Figure 3 show the performance of different methods evaluated by some traditional frequently-used clustering metrics.

According to the plots in Figure 2, the enterprise community detection framework HUMOR introduced in this paper can perform very well in detecting the real-world community structures. For instance, as shown in Figure 2(a), the *Rand* score achieved by HUMOR is 0.38, which denotes among all the potential employee pairs, 38% of them are either correctly grouped in the same community or correctly divided into different communities. The *Rand* score obtained by HUMOR is more than 3 times larger than the scores achieved by methods INF-ESN, INF-CHART, CUT-ESN, CUT-CHART, KMEANS-ESN and KMEANS-CHART respectively. Similar results can be observed when the evaluation metrics are *Mutual Information*, *Purity* and *Inverse Purity* in Figures 2(b)-2(d) respectively.

In addition, by comparing HUMOR with HUMOR-ESN and HUMOR-CHART, based on information from both sources, HUMOR can outperform the single-source community detection methods HUMOR-ESN and HUMOR-CHART with great advantages. For instance, the *Rand* score achieved by HUMOR is 19% higher than the score achieved by HUMOR-CHART, and over 3 times greater than the Rand score of HUMOR-ESN. For other metrics, HUMOR and HUMOR-CHART have comparable performance, and HUMOR can outperform HUMOR-ESN consistently. Meanwhile, according to the results, method HUMOR-CHART based on the chart performs much better than

the method HUMOR-ESN based on the ESNs. One potential explanation can be that ESNs provide the information for a subset of the total employees in the company only, but the company internal information can cover all the employees in the company. Lots of the employees actually either have no accounts or have no social activities in the ESNs, and the community structure of these users is hard to infer merely based on the information in ESNs.

In Figure 3, we show the results obtained by the comparison methods evaluated by traditional metrics, like *Density*, *Silhouette*, *Normalized-DBI* and *Entropy*. From the results, we can observe that HUMOR can perform much better than the other baseline methods when evaluated by *Density*, *Silhouette Index* and *Entropy*, and have comparable performance with some baseline methods when evaluated by *Normalized-DBI*. For instance, the *Density* score obtained by HUMOR is 0.74, which denotes 74% of the following links and management links are preserved within the communities. The density score obtained by HUMOR is 3 times larger than that obtained by CUT-ESN and over 10 times greater than those obtained by INF-ESN, INF-CHART, CUT-CHART, KMEANS-ESN and KMEANS-CHART. Similar results can be observed for the *Silhouette Index* and *Entropy* metrics in Figure 3(b) and Figure 3(d).

D. Parameter Analysis

In this part, we will analyze the sensitivity of the community number parameter K on the performance of methods HUMOR, HUMOR-ESN and HUMOR-CHART. The results evaluated by
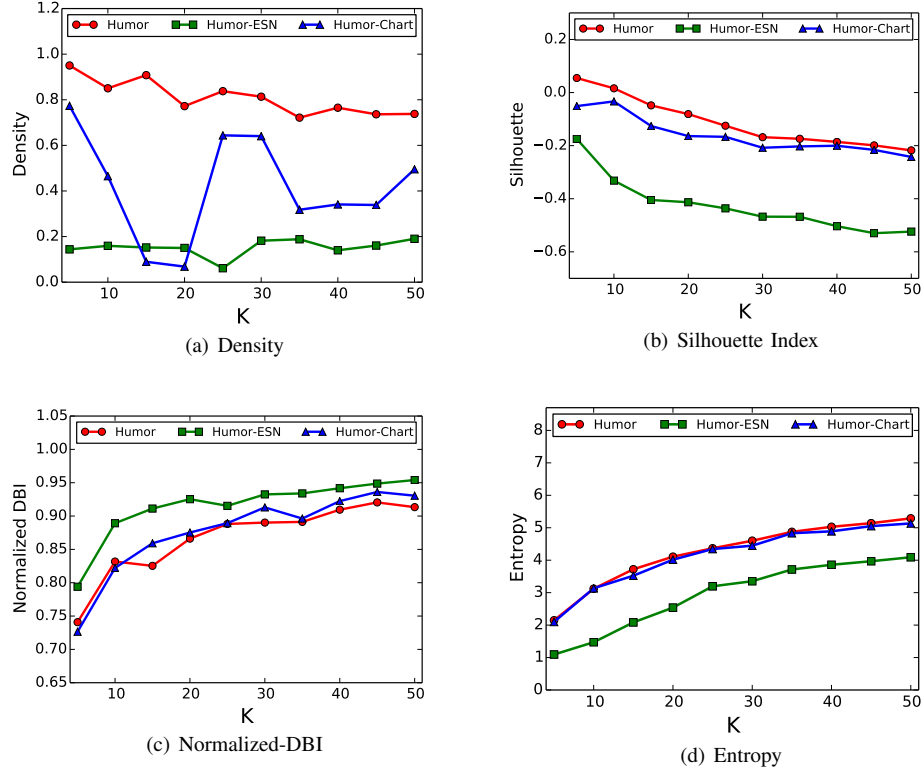
Fig. 4. Parameter analysis of HUMOR, HUMOR-ESN and HUMOR-CHART evaluated by the Density, Silhouette Index, Normalized-DBI and Entropy metrics.

Density, *Silhouette*, *Normalized-DBI* and *Entropy* are available in Figure 4. From the results, we can observe that the performance of these methods depend on parameter K a lot and will change significantly as K varies. For instance, as shown in Figure 4(a) and Figure 4(b), the *Density* and *Silhouette Index* obtained by these three methods decrease steadily as K increases, but the slopes of the curves tend to be flatter. Similar observations are shown in Figures 4(c)-4(d). As K increases, the performance of these methods gets better when evaluated by *Normalized-DBI* and *Entropy*, and the curve tend to be more smooth.

V. RELATED WORK

Clustering aims at grouping similar objects in the same cluster and many different clustering methods have also been proposed. One type is the hierarchical clustering methods [9], which include agglomerative hierarchical clustering methods [4] and divisive hierarchical clustering methods [4]. Meanwhile, according to the manner that the similarity measure is calculated, the hierarchical clustering methods can be further divided into single-link clustering [24], complete-link clustering [14] and average-link clustering [27]. Another type of clustering methods is partition-based methods, which include K-Means for instances with numerical attributes [11], K-Medoids for instances with categorical attributes [19], probabilistic clustering [3], as well as density-based clustering methods [2]. Other clustering methods include grid-based clustering methods [20], constraint-based clustering [26] and fuzzy clustering [10].

Clustering method has also been widely used to detect communities in networks. Newman et al. introduce a modularity function measuring the quality of a division of networks [17]. Shi et al. introduce the concept of normalized cut and discover that the eigenvectors of the Laplace matrix provide a solution to the normalized cut objective function [23]. In addition, many community detection works have been done on heterogeneous online social networks. Sun et al. [25] propose to study the clustering problem with complete link information but incomplete attribute information. Lin et al. [16] try to detect the communities in networks with incomplete relational information but complete attribute information. In recent years, many works have propose to detect communities across aligned networks. Zhang et al. [31] propose to detect the communities in emerging networks with extremely sparse relational information and little attribute information by propagating information from other aligned networks (with abundant of information). Jin et al. [12] propose to study the synergistic partition problem of multiple large scale partially aligned social networks based on the Map-Reduce.

Enterprise social networks [32] can help employees in companies get reliable information [29], [8]. Yarosh et al. [29] explore the importance of different information types in expert searching based on a small-sized questionnaire dataset. Meanwhile, Enterprise social networks can lead to

other benefits to companies and DiMicco et al. [7] study the motivations for social networking at work. 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 [32]. By analyzing the employees' various social behaviors at workplace, Zhang et al. propose to recommend friends for the employees in the online enterprise social networks in [30].

VI. CONCLUSION

In this paper, we have studied the enterprise community detection problem based on the enterprise information about the employees in both company internal information sources and online ESNs involving the employees. An integrated community detection framework HUMOR has been proposed in this paper. HUMOR can detect the *micro community structures* about the employees based on each category of the *enterprise intimacy* scores calculated based on one type of enterprise information. In addition, the *micro community structures* about the employees are further fused in HUMOR with the *intra-fusion* and *inter-fusion* steps. Extensive experiments have been done on the real-world enterprise datasets (including both the company internal data and the online ESNs data), and the results have demonstrated the outstanding performance of framework HUMOR.

REFERENCES

[1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 2009.

[2] J. Banfield and A. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 1993.

[3] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD*, 2004.

[4] P. Cimiano, A. Hotho, and S. Staab. Comparing conceptual, divise and agglomerative clustering for learning taxonomies from text. In *ECAI*, 2004.

[5] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 1990.

[6] D. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.

[7] J. DiMicco, D. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Muller. Motivations for social networking at work. In *CSCW*, 2008.

[8] K. Ehrlich, C. Lin, and V. Griffiths-Fisher. Searching for experts in the enterprise: Combining text and social network analysis. In *GROUP*, 2007.

[9] T. Hastie, R. Tibshirani, and J. Friedman. Hierarchical clustering. In *The Elements of Statistical Learning*. 2009.

[10] F. Hopner, F. Hoppner, and F. Klawonn. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. 1999.

[11] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowledge Discovery*, 1998.

[12] S. Jin, J. Zhang, P. Yu, S. Yang, and A. Li. Synergistic partitioning in multiple large scale social networks. In *IEEE BigData*, 2014.

[13] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML*, 1997.

[14] B. King. Step-Wise Clustering Procedures. *Journal of The American Statistical Association*, 1967.

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

[16] W. Lin, X. Kong, P. Yu, Q. Wu, Y. Jia, and C. Li. Community detection in incomplete information networks. In *WWW*, 2012.

[17] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, 2004.

[18] N. Nguyen and R. Caruana. Consensus clusterings. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, 2007.

[19] H. Park and C. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, 2009.

[20] N. Park and W. Lee. Statistical grid-based clustering over data streams. *SIGMOD Record*, 2004.

[21] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987.

[22] S. Schaeffer. Survey: Graph clustering. *Computer Science Review*, 2007.

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.

[24] P. Sneath and R. Sokal. *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. 1973.

[25] Y. Sun, C. Aggarwal, and J. Han. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *VLDB*, 2012.

[26] A. Tung, R. Ng, L. Lakshmanan, and J. Han. Constraint-based clustering in large databases. In *ICDT*, 2001.

[27] J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 1963.

[28] J. Yang and J. Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *WSDM*, 2013.

[29] S. Yarosh, T. Matthews, and M. Zhou. Asking the right person: Supporting expertise selection in the enterprise. In *CHI*, 2012.

[30] J. Zhang, Y. Lv, and P. Yu. Friend recommendation in enterprise social networks. In *CIKM*, 2015.

[31] J. Zhang and P. Yu. Community detection for emerging networks. In *SDM*, 2015.

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

[33] Y. Zhou and L. Liu. Social influence based clustering of heterogeneous information networks. In *KDD*, 2013.