Enterprise Social Link Recommendation

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

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

Philip S. Yu
University of Illinois at Chicago
Chicago, IL, USA
Institute for Data Science
Tsinghua University, China
psyu@cs.uic.edu

ABSTRACT

Many companies have started to use Enterprise Social Networks (ESNs), such as Yammer, to facilitate collaboration and communication among their employees in the business context. Social link recommendation, which finds and suggests whom one wants to connect with in a company, is crucial for ESNs to promote their usages. Although link recommendation has been studied extensively in external social networks (e.g., Facebook and Twitter), it has not been addressed in ESNs. In this paper, we study this novel problem. Social link recommendation in ESNs is significantly different from that in external social networks, and also has unique challenges: (1) people usually socialize differently in enterprise than in their personal life, but users' social behaviors in enterprise have not been well explored, and (2) there is important business information available in ESNs under the enterprise context, e.g., a company's organizational chart, but how to exploit it for link recommendation is still an open problem. To this end, we mine not only the social graph and user-generated content in ESNs, but also the company's organizational chart, to model enterprise user social behaviors. We develop a supervised link recommendation algorithm using a large scale ESN based on Yammer (with over 100k users), which shows that the proposed techniques perform effectively. Moreover, we find that social graph and organizational chart are complementary to each other for link recommendation in ESNs.

Categories and Subject Descriptors

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

Keywords

Link Recommendation; Enterprise Social Networks; Data Mining

*This work was done when the first author was on a summer internship at Microsoft Research.

CIKM'15, October 19-23, 2015, Melbourne, VIC, Australia
 ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806549.

1. INTRODUCTION

Traditional social networks, e.g., Facebook and Twitter, usually provide services to help people deal with their personal life issues. Recently, to facilitate the collaboration and communication among employees in companies, a new family of online social networks has been adopted inside the firewalls of many corporations. These social networks are called *Enterprise Social Networks* (ESNs) [34]. ESNs can potentially bring significant benefits to both employees and companies. For example, ESNs can help employees meet new colleagues [5], follow internal news and industry trends [32], identify experts [6] and form teams [20], etc.; ESNs can also bridge the generational gaps among employees with more communication [5]. One of the most popular ESNs is Yammer[1]. Over $500,000$ businesses around the world are using Yammer, including 85% of the Fortune 500[2].

Arguably, social links/connections between people inside the company are key to ESNs. First, many services provided by ESNs are based on social connections. Taking Yammer as an example, one needs to follow another user in order to automatically receive relevant contents from the user. Second, as reported in [19], well-established connections among users can attract them to use the network more frequently. Thus, enterprise social link recommendation, which finds *whom one wants to connect with in the company*, is crucial for ESNs to promote their usages. Although link recommendation or link prediction has been studied extensively in external social networks (e.g., Facebook and Twitter) [21, 13, 35], it has not been addressed in ESNs to the best of our knowledge.

Social link recommendation in ESNs is significantly different from that in external social networks, and also has unique challenges:

- People usually socialize differently in workspace than in their personal life because of the professional context [28], but social behaviors in enterprise have not been well explored. As a result, methods proposed for external social networks, e.g., [21, 13, 35, 31], may not work well for ESNs. To support such a claim, we will compare the link prediction method SCAN proposed from external social networks [31] with the model introduced in this paper in the experiment section.

- There is important and unique business information available in ESNs under the enterprise context, e.g.,

[1]https://www.yammer.com/
[2]https://about.yammer.com/why-yammer/

the company's organizational chart [34]. The organizational chart is a diagram (usually a rooted tree) showing organizational relations (e.g., managers to subworkers, directors to managers, etc.) between people within an organization. An example of organizational chart is shown in Figure 1. Intuitively, the organizational relations are related to enterprise social relations. For example, we observe that about 23.4% employees connect with his/her managers in our data, which is an important factor that should be utilized in building the models. However, how to exploit the organizational chart and organizational relations for recommending social links is an open problem.

In this paper, we study this novel problem of Enterprise Social Link Recommendation. We mine not only the social graph and user-generated content in ESNs, but also the company's organizational chart, to model enterprise user social behaviors.

- We first propose different methods to measure social affinity, organizational affinity, social-organizational affinity and geographic affinity for any given user pairs, based on the intuition that closer users tend more likely to connect to each other in ESNs.
- Furthermore, observing that different users may have different social behaviors in ESNs (e.g., one user may like to connect to high-level management people, while another user may mostly like to connect his/her collaborators), we also capture users' personal preferences in multiple dimensions.
- In addition, we further exploit the user-generated content to investigate if it can also help link recommendation in ESNs, since intuitively a user may like to connect to another user who has published high-quality contents.

We develop a supervised link recommendation algorithm based on a large-scale ESN (i.e., Yammer with over 100k users), that can leverage all the proposed measures and heuristics. Our experiments show that the proposed techniques perform effectively, achieving a NDCG score of over 0.61 at the first recommendation position. Moreover, we find that the social graph and the organizational chart are complementary to each other for link recommendation in ESNs.

The rest of this paper is organized as follows. In Section 2, we will briefly describe and analyze the Yammer network as well as giving the terminology definitions and problem formulation. Detailed information about the methods will be introduced in Section 3. In Section 4, we show the experiment results. Finally, in Sections 5-6, we introduce the related works and conclude this paper.

2. ENTERPRISE SOCIAL LINK RECOMMENDATION

We use Yammer [34] in this paper to study the enterprise social link recommendation problem. Yammer provides users with various social services to facilitate their daily workloads, such as following other users, watching their posts and activities, creating/joining groups of their interests, initiating conversations and writing posts, uploading

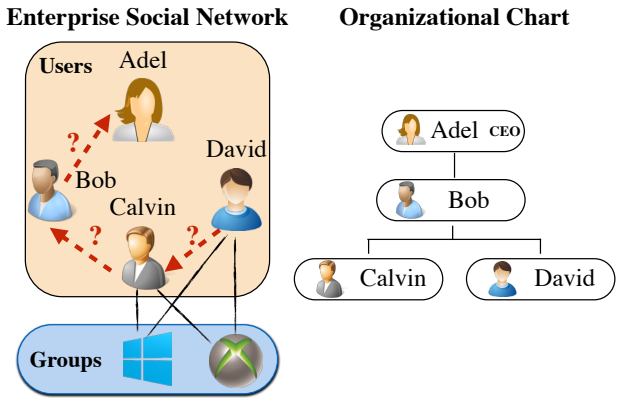


Figure 1: ESN and Organizational Chart (faked examples).

and sharing files, sending online/offline messages to other users, etc. Yammer can be represented as a heterogeneous information network [26].

Definition 1. (Enterprise Social Networks): An enterprise social network can be represented as a heterogeneous information network $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \cup_i \mathcal{V}_i$ and $\mathcal{E} = \cup_j \mathcal{E}_j$ are the sets of different kinds of nodes and links in the enterprise social network. For example, the Yammer can be represented as $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{U} \cup \mathcal{G} \cup \mathcal{C} \cup \mathcal{P}$ contains the nodes of users, groups, conversations and posts, and $\mathcal{E} = \mathcal{E}_{u \to u} \cup \mathcal{E}_{u \to g} \cup \mathcal{E}_{u \to c} \cup \mathcal{E}_{u\ write\ p} \cup \mathcal{E}_{u\ like\ p}$ contains the social links among users, group membership links between users and groups, conversation initiation links between users and conversations, write links and like links between users and posts.

Are the social links between users in Yammer usually professional links that connect users in the same department? To answer this question, we compare the social links from Yammer and the organizational links from the organizational chart. The *organizational chart* [34] denotes a diagram that outlines the internal hierarchic structure of a company and is the common visual depiction of how an organization is structured. In most companies, managers can supervise many subordinates simultaneously, but each subordinate only needs to report to one manager. As a result, the organizational charts are mostly tree-structure diagrams with the CEO at the root. The formal definition of *organizational chart* is given as follows:

Definition 2. (Organizational Chart): The *organizational chart* of a company can be represented as a *rooted tree* (i.e., a tree in which one node has been designated as the root) $T = (\mathcal{N}, \mathcal{L}, root)$, where $\mathcal{N}$ and $\mathcal{L}$ are the sets of nodes and links of T. The *root* of tree T represents the CEO of the company.

By designating the root of the *organizational chart* T, all the links in T will have a natural orientation, i.e., *toward* or *away* from the root. The *tree order* of nodes u, v (i.e., $u \leq v$ iff there is one unique path from the root to v via u) denotes

that u is a manager of v. In addition, nodes in $\mathcal{N}$ can be associated with certain attributes (i.e., the business profile information of employees), which include the *department*, *job title* and *working location*, etc., of the employees.

We examine the social links in Yammer that have an overlap with the organizational links, i.e., one user in a social link is the direct manager, subordinates, or peer of the other user. We observe that, on average, the probabilities for users to follow their direct manager, subordinates and peers are 23.4%, 11.2% and 9.5% respectively. Interestingly, we also observe that subordinates are more likely to follow their managers than managers to follow their subordinates.

On the one hand, this analysis result shows that organizational chart could help enterprise social link recommendation. For example, in Figure 1, we give an example of ESN and organizational chart. In the ESN, we have 4 users, among whom 3 potential links are to be recommended (i.e., the red dashed lines). Considering the scenario of linking Calvin → Bob: merely based on the information in the social network, "Bob" and "Calvin" have nothing in common, neither "common friends" nor "common groups". However, according to their relationship in the organizational chart, we observe that "Bob" is actually the *direct manager* of "Calvin", who should thus be among the recommendation results. On the other hand, organizational chart alone is still not enough and social links among users in ESN are different from organizational links. By analyzing the ESN and organizational chart, we observe the overlap between the organizational links and social links is very small, which accounts for only 13% of the total social links.

Finally, the enterprise social link recommendation problem can be formally defined as follows:

Definition 3. (Enterprise Social Link Recommendation): For the given enterprise social network G and organizational chart T, the sets of users and existing social links among users in the enterprise social network G can be represented as $\mathcal{U}$ and $\mathcal{E}_{u\to u}$ respectively. The set of candidate social links to be recommended can be represented as $\mathcal{E}_{rec} = \mathcal{U}\times\mathcal{U}\setminus\mathcal{E}_{u\to u}$. In the enterprise social link recommendation problem, we use the existing links in $\mathcal{E}_{u\to u}$ and information about these links in both network G and organizational chart T to build a model, which can score and rank the candidate links in $\mathcal{E}_{rec}$. Candidate links that receive higher scores are expected to be more likely actually formed in the ESN.

3. METHODS

In this section, we introduce the Supervised ENterprise SOcial Recommendation (SENSOR) method in details. We will explore both the enterprise social network and the organizational chart to understand and model whom an employee wants to connect with in the company from three dimensions: (1) user-user affinity, (2) user characteristics, and (3) user-generated content.

3.1 User-User Affinity

In social science, "birds of a feather flock together" [23]. "Homophily" [1] is an important principle in social science and that also structures the social ties among users greatly: close users are more likely to follow each other [27] in online social networks. In this paper, the social closeness among users is measured by the user-user *affinity*, which can be computed based on various information sources, e.g., online social behaviors, offline organizational relationships and geographical user distributions.

Heterogeneous social activities from ESN can often capture the social affinity among some users well. However, there are some inactive users and users who just join ESN for a very short period of time. Their sparse information in ESN would be inadequate to calculate their affinity to other users. In this case, fortunately, large amount of information from the offline workplace, e.g., organizational chart, can be exploited to bridge the gap by computing organizational affinity among users. Social affinity and organizational affinity employ information in ESN and organizational chart independently, but do not take advantage of the deep knowledge that can only be obtained by aligning online ESN and offline organizational chart [34]. To remedy such a problem, a set of social-organizational affinity features will be introduced. In addition, the user-user affinity may also be computed based on user geographical distributions, such as country, timezone and even office location, etc., which will be also explored in this section.

3.1.1 Social Affinity

In *enterprise social networks*, users are connected, either directly or indirectly, by various types of connections, such as follow links, shared groups, etc., which could imply social affinity among users. We propose several novel methods to compute social affinity, such as weighted meta path and weighted group membership, inspired by the term weighting techniques in information retrieval [17].

- **Reciprocal Social Link**: Reciprocal links [36] denote the mutual links between two users. The existence of reciprocal links between users is an intuitively important clue indicating the affinity between them. For example, if we want to predict whether link (u_i, u_j) exists or not, then the existence of its *reciprocal link* (u_j, u_i) could boost our confidence of the existence of the link (u_i, u_j). By analyzing the data, we observe that the ratio of reciprocal link number to the total number of links is about 33% in our enterprise social networks, suggesting it is a good source of evidence for social affinity.

- **Weighted Social Meta Path**: Besides the direct social links, user u_i and u_j can often be "connected" by sequences of indirect links, i.e., meta paths as proposed in [26]. In ESN, the meta paths of length 2 only consisting of user nodes include:

$$MP_1 : U \to U \to U, \text{Follower of Follower},$$
$$MP_2 : U \leftarrow U \leftarrow U, \text{Followee of Followee},$$
$$MP_3 : U \leftarrow U \to U, \text{Common Follower},$$
$$MP_4 : U \to U \leftarrow U, \text{Common Followee}.$$

  In this paper, we only consider the above four types of meta paths. Based on these meta paths, a straightforward way of modeling social affinity for potential link (u_i, u_j) is to count the number of meta path instances existing in the network between u_i and u_j [35, 26] formally,

$$|\{p|p \in MP_k\}|, k \in \{1, 2, 3, 4\},$$

  where $p \in MP_k$ denotes that p is an instance of meta path MP_k in the network. For instance, based on

MP_1, such feature extracted for link (u_i, u_j) can also be represented as $|\{u_i \to u \to u_j | u \in \mathcal{U}\}|$.

However, simply counting meta path instances may suffer from some problems: u_i and u_j can be connected by large number of meta path instances merely because of the large out degrees of either u_i, u_j or the intermediate nodes between them, but the affinity between u_i and u_j is not high. For example, given two users u_i and u_j indirectly connected via MP_4 (i.e., $u_i \to u_k \leftarrow u_j$), if the common followee u_k is very popular (such as the CEO in a company), then any two users in the network may be connected by u_k and the existence of path $u_i \to u_k \leftarrow u_j$ may not indicate the affinity between u_i and u_j. To address this problem, we propose to use weighted social meta path to measure social affinity as follows.

For meta paths MP_1 and MP_2: the probabilities of random walking from u_i to u_j (and from u_j to u_i) based on meta paths MP_1 (and MP_2), could intuitively be a better affinity measure than the simple number of path instances. The reason is that the probabilities have been naturally normalized to penalize popular users. The probability of random walking from u_i to u_j, based on meta paths MP_1 can be represented as

$$(1)\ P(MP_1(u_i, u_j)) = \sum_{u_k \in \Gamma_{out}(u_i) \cap \Gamma_{in}(u_j)} P(u_i \to u_k) P(u_k \to u_j) = \sum_{u_k \in \Gamma_{out}(u_i) \cap \Gamma_{in}(u_j)} \frac{1}{|\Gamma_{out}(u_i)|} \frac{1}{|\Gamma_{out}(u_k)|}.$$

Similarly, we can define the probability of reaching u_i from u_j based on meta paths MP_2 as follows

$$(2)\ P(MP_2(u_i, u_j)) = \sum_{u_k \in \Gamma_{in}(u_i) \cap \Gamma_{out}(u_j)} \frac{1}{|\Gamma_{out}(u_k)|} \frac{1}{|\Gamma_{out}(u_j)|},$$

where $\Gamma_{out}(u)$ and $\Gamma_{in}(u)$ denote the set of users that u follows and users who follow u respectively, and $p(u_i \to u_k)$ represents the probability of random walking from u_i to u_k based on social links.

For meta paths MP3 and MP4: the *common followee* and *common follower* of u_i and u_j should also be weighted appropriately to penalize those popular common followees and common followers. To reward/penalize the importance of different common followees (and common followers), we propose to weight them by both an idf-like measure and the pointwise mutual information [24]:

$$(3)\ \text{idf}(MP_3(u_i, u_j)) = \sum_{u_k \in \Gamma_{in}(u_i) \cap \Gamma_{in}(u_j)} \log \frac{|\mathcal{U}|}{|\Gamma_{out}(u_k)|},$$

$$(4)\ \text{idf}(MP_4(u_i, u_j)) = \sum_{u_k \in \Gamma_{out}(u_i) \cap \Gamma_{out}(u_j)} \log \frac{|\mathcal{U}|}{|\Gamma_{in}(u_k)|},$$

$$(5)\ mi(MP_3(u_i, u_j)) = \frac{|\Gamma_{in}(u_i) \cap \Gamma_{in}(u_j)|}{|\mathcal{U}|} \log \frac{\frac{|\Gamma_{in}(u_i) \cap \Gamma_{in}(u_j)|}{|\mathcal{U}|}}{\frac{|\Gamma_{in}(u_i)|}{|\mathcal{U}|} \cdot \frac{|\Gamma_{in}(u_j)|}{|\mathcal{U}|}},$$

$$(6)\ mi(MP_4(u_i, u_j)) = \frac{|\Gamma_{out}(u_i) \cap \Gamma_{out}(u_j)|}{|\mathcal{U}|} \log \frac{\frac{|\Gamma_{out}(u_i) \cap \Gamma_{out}(u_j)|}{|\mathcal{U}|}}{\frac{|\Gamma_{out}(u_i)|}{|\mathcal{U}|} \cdot \frac{|\Gamma_{out}(u_j)|}{|\mathcal{U}|}},$$

where $\mathcal{U}$ is the set of all users in the network.

These 6 proposed measures consider not only the meta path instances between u_i and u_j but also the weight of each meta path and are named weighted meta path based social affinities in this paper.

- **Common Group Membership**: Besides social connections, users can also join groups in enterprise social networks. These groups are created either for professional techniques (e.g., C#, Machine learning, Cloud computing, etc.) or just for users' personal interests (e.g., jogging, swimming, etc.). Intuitively, sharing more groups often suggests that two users have more common interests.

  However, different groups may have different discrimination power. For two users sharing group "Cornell 2013 Alumni" (a small-sized group) are more likely to link to each other than users who share group "Employee News Events" (a large group posting internal news). In order to capture the discrimination of different groups, we propose a concept called "Inverse Membership Frequency" (IMF), inspired by the widely-used IDF [17] in information retrieval. Specifically, if we regard a group as a word and a user as a document. Each user can join many groups, corresponding to that a document is comprised of many words. Specifically, the IMF of group g can be represented as

$$IMF(g) = \log \frac{|\mathcal{U}|}{|\Gamma(g)|},$$

  where $\Gamma(g)$ represents the set of users joining group g. And for any two users u_i and u_j sharing common groups $\Gamma_g(u_i) \cap \Gamma_g(u_j)$, where $\Gamma_g(u)$ denotes the set of groups that u joins, the common-group membership affinity between u_i and u_j can be computed as

$$\sum_{g \in \Gamma_g(u_i) \cap \Gamma_g(u_j)} IMF(g) = \sum_{g \in \Gamma_g(u_i) \cap \Gamma_g(u_j)} \log \frac{|\mathcal{U}|}{|\Gamma(g)|}.$$

3.1.2 Organizational affinity

For inactive and new users, the above introduced social affinity measures would not be very useful due to their sparse activity information in ESNs. However, in the enterprise context, useful alternative information about these users can be obtained, in particular, the organizational chart. We next introduce the organizational affinity, a set of affinity measures among users calculated based on the organizational chart.

Company organizational chart is a tree structured diagram outlining the relationships among users, where users are connected by sequences of "management" links between managers and their subordinates. Given any two users, an intuitive idea to represent their affinity is using the number of required steps to walk to each other along the links in the organizational chart. Colleagues in the same department need less steps to connect from one to another than those in different departments. Based on such an observation, we propose a novel affinity measure "*organizational distance*"

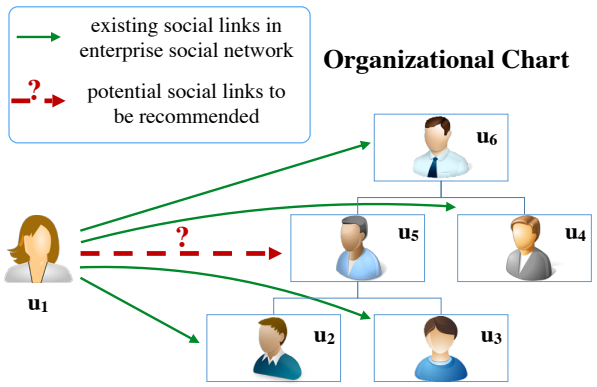


Figure 2: Example of affinity measure extraction across hybrid aligned enterprise graphs

for link (u_i, u_j): the number of steps required to walking from u_i to u_j via the links in the the organizational chart.

However, as mentioned in Section 2, users are more interested in connecting to their managers than their subordinates. As a result, the organizational distance should be asymmetric. Specifically, if we let the *organizational distance* from a user to any of his subordinates be 1, then the *organizational distance* from a user to his direct manager should be $\alpha < 1$. Our preliminary results show that $\alpha = 0.7$ works very well, which is also used throughout our experiments. For instance, based on the organizational chart in Figure 1, the organizational distance from "Calvin" to "Bob" is α, but that from "Adel" to "Bob" is 1.

Besides the links, in the organizational chart, the employees are associated with a set of attributes, which include their departments and job titles. Intuitively, people in the same department have a higher chance to know each other, and people with the same job title are more likely to work together and connect to each other (e.g., software engineer likes to connect to software engineer, while statistician likes to connect to statistician). In addition to the "*organizational distance*", features denoting whether u_i and u_j (1) are in the same department, and (2) share the same title or not are extracted as another two affinity measures from the organizational chart.

3.1.3 Social-Organizational Affinity

The above two categories of affinity measures are extracted from the enterprise social network and the organizational chart *independently*. However, aligning the enterprise social network and organizational chart together [18, 33], many other interesting signals (which were not observable when using either the ESN or the organizational chart alone) arise that could capture the affinity among users from different perspectives.

To illustrate, in Figure 2, we show part of the *organizational chart* involving $\{u_2, u_3, u_4, u_5, u_6\}$ as well as the existing social links between u_1 and u_2, u_3, u_4 and u_6. We observe that the users u_1 follows in ESNs, i.e., u_2, u_3, u_4 and u_6, are all closely working together with u_5. We define the "*groupmates*" concept in this paper to represent such relationship between u_5 and $\{u_2, u_3, u_4, u_6\}$. "Groupmates" is essentially a union set of the user's "*direct manager*", "*peers*" and "*direct subordinates*". For example, based on organizational chart fragment in Figure 2, we can find u_5's "*groupmates*" to be $\{u_2, u_3, u_4, u_6\}$, where u_2 and u_3 are the subordinates of u_5, and u_4 and u_6 are u_5's peer and direct manager respectively.

Then, when predicting the social link (u_i, u_j) in ESNs, if we observe that u_i has followed a large number/proportion of "*groupmates*" of u_j, then u_i is likely to be familiar with u_j's group, and has a high probability to follow u_j as well. For the example in Figure 2, we observe that u_1 has followed 4 of u_5's *groupmates*, then u_1 is likely to follow u_5. Following this intuition, we propose to compute (1) the number of u_j's groupmates followed by u_i, and (2) the percentage of u_j's groupmates followed by u_i as two social-organizational affinity measures across both ESN and organizational chart.

3.1.4 Geographical Affinity

Users who are in the same geo-location tend to have a higher chance to know each other, which thus may also be a possible signal for user-user affinity. We propose to also exploit information about users' working locations, e.g., timezone, country, and longitude/latitude. In this paper, we compute several such geo-affinity features including (1) whether u_i and u_j are in the same timezone, (2) whether u_i and u_j are in the same country, and (3) the geo-distance between u_i and u_j's working locations based on the longitude/latitude coordinate information.

3.2 User Characteristics

To provide personalized recommendations so as to improve user experience, users' personal characteristics play an important role. By analyzing the data, we observe that different users have quite different characteristics. For example, (1) some users like to follow users who have followed them but some other users seldom do so, (2) some users like to follow their managers but some other users like to follow their peers and subordinates, (3) some users like to follow their *groupmates* but some others like to follow employees in other divisions or departments.

To capture users' characteristics and preferences when recommending link (u_i, u_j), a set of user characteristics features are designed for u_i and u_j, which includes: (1) the *reciprocal rate* of users u_i and u_j: $\frac{|\Gamma_{out}(u_i) \cap \Gamma_{in}(u_i)|}{|\Gamma_{in}(u_i)|}$ and $\frac{|\Gamma_{out}(u_j) \cap \Gamma_{in}(u_j)|}{|\Gamma_{in}(u_j)|}$, (2) the probabilities that u_i follows his *managers, peers*, and *subordinates* based on u_i's existing follow links, (3) the sum and average of the *organizational distance* between u_i (u_j), and users that u_i (u_j) follows, (4) the number of users that $u_i(u_j)$ follows and number of followers that $u_i(u_j)$ has, (5) the average number of followers of users that u_i follows, and the average number of followers of users who follow u_j: $\frac{1}{|\Gamma_{out}(u_i)|} \sum_{u_k \in \Gamma_{out}(u_i)} |\Gamma_{in}(u_k)|$ as well as $\frac{1}{|\Gamma_{in}(u_j)|} \sum_{u_k \in \Gamma_{in}(u_j)} |\Gamma_{in}(u_k)|$, and (6) the number of employees that u_i and u_j manage.

3.3 User-Generated Content

User-generated content can reveal important information about users, e.g., the *activeness* and *popularity*. Various types of content can be generated by users via their social activities in enterprise social networks, e.g., writing posts, initiating conversations, posting comments in the conversations, and "*liking*" other users' posts, etc. Based on these

user-generated contents, we compute a set of related features for (u_i, u_j), which include: (1) numbers of posts written by u_i and u_j, (2) numbers of conversations initiated by u_i and u_j, (3) the total and average numbers of "*likes*" that u_i (u_j) receives from other users, and (4) the total and average numbers of comments posted in the conversations initiated by u_i (and u_j).

3.4 Supervised Link Recommendation based on Multiple Additive Regression Tree

Link recommendation (or friend recommendation) services provided by online social networks aim at providing a ranked list of suggested social links (i.e., friends) for users. The setting of the standard link prediction problem [13, 27], essentially does point-wise prediction [3] of the existence of individual links. However, point-wise prediction has been shown ineffective for ranking problem. Besides, there are often much more negative links than positive links in the training data, and pointwise classification-based link prediction suffers from class imbalance problem [14]. Motivated by the effectiveness of the pairwise learning algorithms for learning to rank [22], we solve the link recommendation as a ranking problem rather than a pointwise prediction problem. Our later experiments also justify empirically our choice.

For a certain given user (corresponding query) u_i, our link recommendation model aims at returning a set of friend candidates in the decreasing order of their likelihood that u_i wants to follow a user. The query user u_i together with his potential followees, e.g., u_j, returned by link recommendation model, can be represented as pair (u_i, u_j). In the link recommendation problem, a set of features that depict either the relationships or the characteristics of users u_i and u_j are extracted from the network, which can be represented as vector $\mathbf{x}_{i,j}$. Besides the features, pair (u_i, u_j) is labeled with the relevance scores between the query user u_i and his potential followee u_j, which can be represented as $y_{i,j}$ ($y_{i,j} = 1$ if (u_i, u_j) is connected and 0 otherwise).

Formally, both the positive and negative pairs in the training set $\mathcal{T}$ can be represented as $\mathcal{D} = \{(\mathbf{x}_{i,j}, y_{i,j})\}, (u_i, u_j) \in \mathcal{T}$, where $\mathbf{x}_{i,j} \in \mathbb{R}^k$ of length k is the feature vector extracted for pair (u_i, u_j) and $y_{i,j}$ denotes the correlation between u_i and u_j. Link recommendation can be formalized as building a regression function $h : \mathbb{R}^k \to \mathbb{R}$, such that $h(\mathbf{x}_{i,j}) \approx y_{i,j}$. The regression model will be applied to users in the test set and can return the predicted existence confidence scores $\{h(\mathbf{x}_{i,j})\}$ for pairs consisting of the query user u_i and potential followees u_j in the test set.

We use a state-of-the-art pairwise based regression algorithm, namely MART (Multiple Additive Regression Trees) [29], to develop a regression function. MART is based on the stochastic gradient boosting approach described in [7, 8] which performs gradient descent optimization in the functional space. In our experiments, we used the log-likelihood as the loss function, steepest-descent (gradient descent) as the optimization technique, and binary decision trees as the fitting function.

4. EXPERIMENTS

To evaluate the proposed techniques for enterprise social link recommendation, we conduct extensive experiments on a real-world enterprise social network Yammer used in a large IT company and its organizational chart.

4.1 Dataset

We crawl all the Microsoft employees' information from Yammer and obtain the complete organizational chart involving all these employees in Microsoft during June, 2014 [34]. 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. In summary, it includes more than $100k$ Microsoft employees, and millions of user-generated posts published and the social links.[3]

The social network data contains the complete information of all users till June 20, 2014. We treat the data before May 17, 2014 as the existing graph, which will be purely used to extract features. Users registered before May 17, 2014 are "*old users*" and users registered during May 17, 2014 - June 20, 2014 are "*new users*". Links formed during May 17, 2014 - June 20, 2014 are used as the positive links and randomly split into training, validation and test sets according to a $3:1:1$ ratio. All the non-existing links for each user (both "*old*" and "*new*" users) are used as the negative links (compared with the existing links). The positive training set together with all the negative link set are used to build the model. Parameters of the MART algorithm, such as tree depth, tree number and iteration numbers, etc., are selected with the validation set. The built model will be applied to the test set, the results of which are used in our comparison.

4.2 Experiment Settings

To evaluate the effectiveness of the proposed techniques in recommending social links for users in enterprise social networks, we compare our techniques with several representative baseline methods, including both supervised and unsupervised methods proposed in existing works for link prediction in *external* social networks [31]. Now, we summarize all the comparison methods as follows:

- *Enterprise Social Network + Organizational Chart*: We can use information of users in both enterprise social networks and the company internal organizational chart all about the same company to recommend social links. Our method is a general method and can be potentially applied to the enterprise social networks of other companies.
- *Enterprise Social Network Only*: To examine whether employees' information in the company internal organizational chart is helpful for improving link recommendation results. We introduce a baseline method called SENSOR-N (SENSOR Network), which uses information purely from the enterprise social network.
- *Organizational Chart Only*: One may wonder whether organizational chart alone is enough to predict employees' social links in an enterprise social network? To answer this question, we compare SENSOR with another baseline SENSOR-C (SENSOR Chart), which only utilizes information from the organizational chart.
- *Classification Based Link Prediction*: Existing supervised link prediction methods all use point-wise regression/classification algorithms, which do prediction for

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

each candidate link one by one. As discussed before, we hypothesize that pair-wise learning to rank algorithms would work better, observing their effectiveness in Web search [29]. To verify this hypothesis, we also compare our method with SCAN [31]. SCAN is a representative point-wise supervised learning method which use SVM for training the model. We use SCAN to develop a baseline link recommendation based on the same set of features (including those from both enterprise social networks and organizational chart).

- *Unsupervised Methods*: All the above methods are supervised link recommendation methods. For completeness, in the experiments, some traditional unsupervised link prediction methods are also used as the unsupervised baseline methods, which include "Common Neighbor" (CN), "Jaccard's Coefficient" (JC), "Adamic Adar" (AA). More detailed descriptions about these unsupervised predictors are available in [14].

We use the widely-accepted Normalized Discounted Cumulative Gain (NDCG) [16] to compare different methods. NDCG is a common evaluation metric to measure the performance of ranking methods. NDCG@1, NDCG@2, NDCG@3, NDCG@4 are used in our experiments, since the accuracy of top results is more important in a recommendation scenario like this.

4.3 Experiment Results

The comparison results are given in Figure 3, where Subfigures 3(a)-3(d) show the performance of all the comparison methods measured respectively by NDCG@1, NDCG@2, NDCG@3, and NDCG@4.

According to the results shown in Figure 3, SENSOR which utilizes information from both enterprise social network and the organizational chart performs the best among all the comparison methods across all metrics. First, SENSOR achieves NDCG scores higher than SENSOR-N by 16.8%, and SENSOR-C by 13.7% at different ranking positions 1, 2, 3, and 4 respectively. Recall that SENSOR-N only uses information from ESN. This suggests that incorporating information from organizational chart can improve the link recommendation results significantly. Meanwhile, SENSOR-C recommends social links merely based on information in organizational chart. This shows that social links are still different from those professional and business links implied by the organizational chart, and organizational chart alone is not enough. Furthermore, SENSOR performs much better than SCAN, which uses the same feature set but is using a point-wise learning algorithm instead. Possible reasons are (1) SCAN is a point-wise prediction algorithm, which has been shown less effective than pairwise prediction algorithm like MART [7] for ranking; and (2) SCAN suffers from class imbalance problem, while in real world applications like in our case, there tend to always be much more negative labels than positive labels. In addition, SENSOR can also achieve dramatically better performance than traditional unsupervised link prediction methods CN, JC and AA.

4.4 Feature Ablation Analysis

Different dimensions of features have been extracted from the network, e.g., user affinity, user personal preference, and the quality of user-generated content. To investigate the effectiveness of different kinds of features, we also do a feature ablation analysis. We remove one category of features while keeping all other categories each time, and the results achieved are given in Table 1.

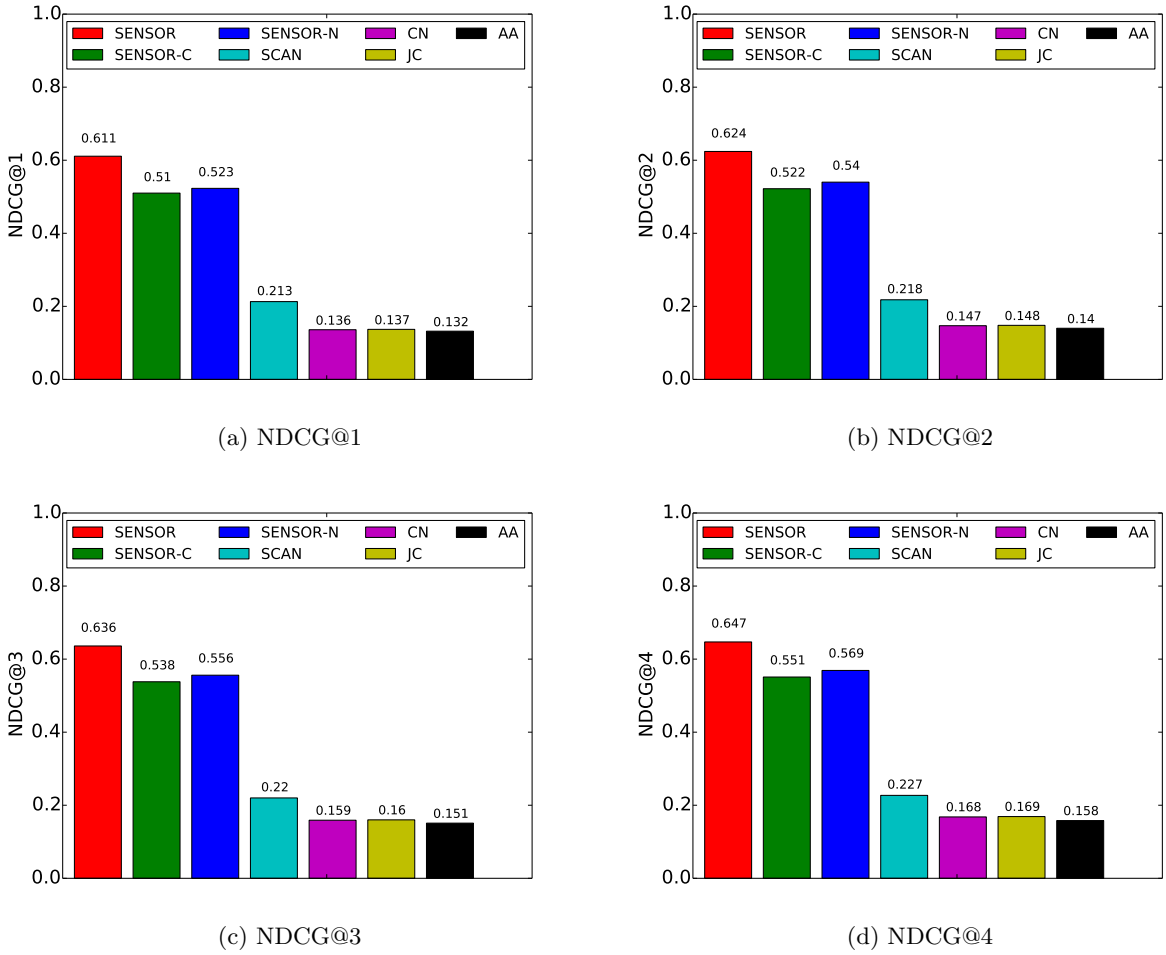
First, we check the user affinity features. By comparing the results achieved by SENSOR with features "*All*", and "*All - affinity features*" (including "*All - social affinity*", "*All - org affinity*", "*All - hybrid affinity*" and "*All - geo affinity*"), we observe that, as expected, by using "*All*" the features, SENSOR performs the best. When excluding the "*social affinity*", "*org affinity*" and "*hybrid affinity*" features, the performance of SENSOR will degrade a lot. For example, with "*All*" features, SENSOR can achieve 0.611 NDCG@1 but when excluding "*social affinity*", "*org affinity*" and "*hybrid affinity*" features, the performance will drop 38.9%, 10.3% and 4.1% respectively. This suggests that the user affinity implied by social networks and organizational chart are both useful, and, what is more, they complement to each other. However, when excluding the "*geo-affinity*" features, the performance of SENSOR does not decrease much. This shows that enterprise social connections are not closely related to users' geographical locations.

Second, we look into the "*user characteristics*" features. As shown in Table 1, when excluding the *user characteristics* features, SENSOR's performance will degrade a lot. For example, Its NDCG@1 is 24.5% lower than that achieved by SENSOR with "*All*" features. This confirms that users in enterprise social networks do have different personal preferences, which significantly influences their decisions of who to connect with in company.

Third, we examine the features based on user-generated content. When excluding this set of features, SENSOR performs almost the same as when using "All" features, which shows that *user generated content* does not help recommend social links in enterprise social networks. One possible reason is that the links in ESN are mostly social links rather than content-based links: a user may not always follow another user who has published high-quality content.

4.5 Link Recommendation for New Users

It has been shown that new users often suffer from the cold start problem, since we do not have data for new users [25]. Does the cold start problem also exist in enterprise social link recommendation? In Figure 4, we give the recommendation results of SENSOR and SENSOR-N (which does not use the organizational chart) in recommending social links for "*new*" users and "*old*" users. The x axis of the figure denotes the number of existing links users have and users with zero existing links are "*new*" users while users with existing links represent "*old*" users. First of all, we see that SENSOR performs better than SENSOR-N consistently for users with different numbers of existing following links. This confirms again that the organizational chart helps link recommendation. Specifically, for users with fewer existing links, SENSOR performs even better than SENSOR-N, which suggests that the organizational chart information is more important in recommending links for "*new*" users. As users become "old" and have more existing links, the advantages of SENSOR over SENSOR-N will decrease steadily. In other words, for "*old*" users who have already had much information in the enterprise social network, organizational chart may not help much.



(a) NDCG@1

(b) NDCG@2

(c) NDCG@3

(d) NDCG@4

Figure 3: Link recommendation results. The improvements of SENSOR over SENSOR-n, SENSOR-c, SCAN, CN, JC, AA are all statistically significant at the 0.01 level using Wilcoxon non-directional test.

One possible reason is that *"new" users* tend to first connect with employees in the same group or department, and then connect with friends beyond the organizational boundaries. This would also lead the link recommendation problem for "new" users to be relatively easier than for "old" users. This has also been confirmed by the results in Figure 4, which shows that we can even achieve higher NDCG for new users. In summary, there is no "cold start" (but a "hot start" instead) for new users in enterprise social link recommendation, thanks to the knowledge from the organizational chart.

4.6 Interesting Findings

We summarize some interesting findings we have observed:

- Organizational chart is useful.
- Users tend to first connect with colleagues within the same group and then to employees beyond the organizational boundaries.
- Social affinity and organizational affinity are both useful, but geo-affinity is useless.
- User-generated content does not help, suggesting social links rather than content links.

5. RELATED WORK

Enterprise social networks [34] can help employees in companies get reliable information [30, 6, 12]. Yarosh et al. [30] explore the importance of different types of information in expert searching based on a small-sized questionnaire dataset and useful information is used to improve the usefulness of supporting systems. Ehrlich et al. [6] propose to search for experts in enterprise with both text and social network analysis techniques. They propose to examine the users' dynamic profile information and get the social distance to the expert before deciding how to initiate the contact. Meanwhile, Enterprise social networks can lead to other benefits to companies and DiMicco et al. [5] study the motivations for social networking at work. Users in enterprise social networks will connect and learn from each other through personal and professional sharing. DiMicco et al. [4] propose to study people sensemaking and relation building on an enterprise social network site. Based on the heterogeneous information in enterprise social networks, Zhang et

Table 1: Feature ablation analysis

	All	All - social affinity	All - org affinity	All - hybrid affinity	All - geo affinity	All - user characteristics	All - user generated content
NDCG@1	0.611	0.373	0.548	0.586	0.605	0.461	0.614
NDCG@2	0.624	0.382	0.561	0.600	0.619	0.486	0.621
NDCG@3	0.636	0.395	0.580	0.616	0.631	0.504	0.636
NDCG@4	0.647	0.408	0.594	0.628	0.641	0.517	0.645

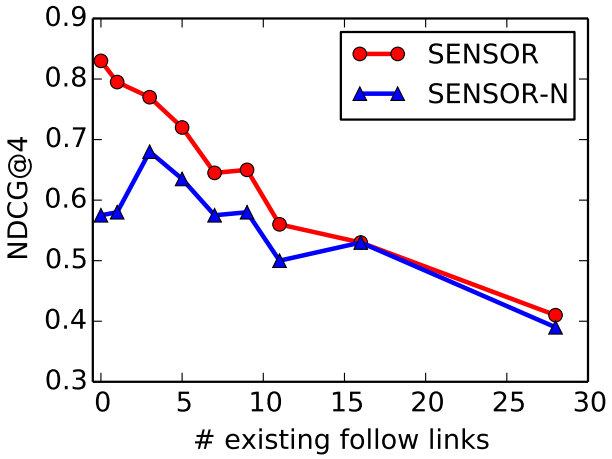


Figure 4: Link recommendation for new users and old users

al. propose to infer the complete organizational chart based on an unsupervised learning framework CREATE in [34].

Social connections among users in enterprise social networks usually have multiple facets. Wu et al. [28] study the multiplexity of social connections among users in enterprise social networks, which include both professional and personal closeness. Zhang et al. [32] give a case study about the early adoption and use of micro-blogging in 500 companies, and attempt to understand how knowledge workers are likely to use micro-bolgging in the enterprise. Hsieh et al. [15] study the problem of computing edge affinity between two users on a social network. Some friend recommendation works have been done either on small-sized enterprise social networks, like Beehive [2], or survey based dataset [11, 2, 10, 9]. Chen et al. [2] study people recommendations designed to help users find known, offline contacts and discover new friends on social networking sites. They evaluated four recommender algorithms in an enterprise social networking site using a personalized survey of 500 users and a field study of 3,000 users. Guy et al. explore the personalized recommendation of social software items [11], close friends [9] and even strangers [10] based on user studies involving less than 500 users in enterprises. Different from these works, we are doing link recommendation using a real-world large-scale ESN.

Link prediction in online social networks first proposed by Liben-Nowell et al. [21] has become an important research topic in recent years. In [21], Liben-Nowell et al. propose various unsupervised link predictors to calculate the existence scores of potential links. Meanwhile, Hasan et al. [13] propose a supervised learning framework to predict potential links in online social networks. These works are all based on one single homogeneous networks. Sun et al. [26] propose a meta path based prediction model to predict co-author relationships in the heterogeneous bibliographic network.

Nowadays, link prediction in multiple heterogeneous social networks simultaneously has attracted much attention. Kong et al. [18] notice that users nowadays are usually involved in multiple social networks at the same time to enjoy specific social services provided by different networks. Zhang et al. [35] study the social link recommendations in multiple partially aligned social networks simultaneously.

6. CONCLUSION

In this paper, we study a new problem of enterprise social link recommendation. To understand and model whom an employee wants to connect with in the company, new techniques are introduced to explore both enterprise social network and organizational chart from three dimensions: (1) user-user affinity, (2) user characteristics, and (3) user generated content. A supervised machine-learned recommendation algorithm, SENSOR, is developed. Evaluation using a real-world large-scale enterprise social network shows that the proposed techniques perform effectively in recommending social links for enterprise social networks.

We plan to further investigate the relationship between ESN and organizational chart. To name a few, e.g., organizational chart inference based on ESN as well as hybrid information diffusion across ESN and organizational chart, will be studied in the future.

7. ACKNOWLEDGEMENT

We thank Ariel Fuxman and Panayiotis Tsaparas for their useful discussions. We also thank Dhyanesh Narayanan, John Pigeon and Chris Slemp for helping crawl the data.

8. REFERENCES

[1] L. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM TWEB*, 2012.

[2] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: Recommending people on social networking sites. In *CHI*, 2009.

[3] W. Chen, T. Liu, Y. Lan, Z. Ma, and H. Li. Ranking Measures and Loss Functions in Learning to Rank. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*. 2009.

[4] J. DiMicco, W. Geyer, D. Millen, C. Dugan, and B. Brownholtz. People sensemaking and relationship building on an enterprise social network site. In *HICSS*, 2009.

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

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

[7] J. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2000.

[8] J. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 2002.

[9] I. Guy, I. Ronen, and E. Wilcox. Do you know?: Recommending people to invite into your social network. In *IUI*, 2009.

[10] I. Guy, S. Ur, I. Ronen, A. Perer, and M. Jacovi. Do you want to know?: Recommending strangers in the enterprise. In *CSCW*, 2011.

[11] I. Guy, N. Zwerdling, D. Carmel, I. Ronen, E. Uziel, S. Yogev, and S. Ofek-Koifman. Personalized recommendation of social software items based on social relations. In *RecSys*, 2009.

[12] S. Han, D. He, J. Jiang, and Z. Yue. Supporting exploratory people search: A study of factor transparency and user control. In *CIKM*, 2013.

[13] M. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM*, 2006.

[14] M. Hasan and M. Zaki. A survey of link prediction in social networks. In C. Aggarwal, editor, *Social Network Data Analytics*. 2011.

[15] C. Hsieh, M. Tiwari, D. Agarwal, X. Huang, and S. Shah. Organizational overlap on social networks and its applications. In *WWW*, 2013.

[16] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 2002.

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

[18] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, 2013.

[19] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

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

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

[22] T. Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[23] M. McPherson, L. Lovin, and J. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 2001.

[24] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6), 2003.

[25] A. Schein, A. Popescul, L. Ungar, and D. Pennock. Methods and metrics for cold-start recommendations. In *SIGIR*, 2002.

[26] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *ASONAM*, 2011.

[27] J. Tang, H. Gao, X. Hu, and H. Liu. Exploiting homophily effect for trust prediction. In *WSDM*, 2013.

[28] A. Wu, J. DiMicco, and D. Millen. Detecting professional versus personal closeness using an enterprise social network site. In *CHI*, 2010.

[29] Q. Wu, C. Burges, K. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 2010.

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

[31] J. Zhang, X. Kong, and P. Yu. Transfer heterogeneous links across location-based social networks. In *WSDM*, 2014.

[32] J. Zhang, Y. Qu, J. Cody, and Y. Wu. A case study of micro-blogging in the enterprise: User, value, and related issues. In *CHI*, 2010.

[33] J. Zhang and P. Yu. Integrated anchor and social link predictions across partially aligned social networks. In *IJCAI*, 2015.

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

[35] J. Zhang, P. Yu, and Z. Zhou. Meta-path based multi-network collective link prediction. In *KDD*, 2014.

[36] Y. Zhu, X. Zhang, G. Sun, M. Tang, T. Zhou, and Z. Zhang. Influence of reciprocal links in social networks. *PLoS ONE*, 9(7), 2014.