

Predicting Social Links for New Users across Aligned Heterogeneous Social Networks

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

Xiangnan Kong
 University of Illinois at Chicago
 Chicago, IL, USA
 xkong4@uic.edu

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

Abstract—Online social networks have gained great success in recent years and many of them involve multiple kinds of nodes and complex relationships. Among these relationships, social links among users are of great importance. Many existing link prediction methods focus on predicting social links that will appear in the future among all users based upon a snapshot of the social network. In real-world social networks, many new users are joining in the service every day. Predicting links for new users are more important. Different from conventional link prediction problems, link prediction for new users are more challenging due to the following reasons: (1) differences in information distributions between new users and the existing active users (i.e., old users); (2) lack of information from the new users in the network. In order to solve the above problems, we need to accommodate the differences in information distributions between old users and new users within the network and transfer additional information from other sources for the new users. We notice that users nowadays are normally involved in multiple social networks to enjoy more online services at the same time, such as Facebook, Twitter and Foursquare. New users in one social network (e.g., Foursquares) might have already joined and been active in another network (e.g., Twitter) for a long time. We propose a link prediction method called SCAN-PS (Supervised Cross Aligned Networks link prediction with Personalized Sampling), to solve the link prediction problem for new users with information transferred from both the existing active users in the target network and other source networks through aligned accounts. We proposed a within-target-network personalized sampling method to process the existing active users' information in order to accommodate the differences in information distributions before the intra-network knowledge transfer. SCAN-PS can also exploit information in other source networks, where the user accounts are aligned with the target network. In this way, SCAN-PS could solve the cold start problem when information of these new users is total absent in the target network. Extensive experiments conducted on Twitter and Foursquare, two real-world aligned heterogeneous social networks, demonstrate that SCAN-PS outperforms other link prediction methods for new users under different degrees of newness consistently and works well with the cold start problems.

I. INTRODUCTION

Online social networks are becoming more and more popular in recent years. Many of these networks involve multiple kinds of nodes, such as users, posts, locations, et al., and complex relationships among the nodes, such as social links and location check-ins. Among these relationships, social link prediction is crucial for many social networks because it

will lead to more connections among users. Meanwhile, well-established online social relationships will attract users to use the network more frequently [10].

Many of previous works on link prediction focus on predicting potential links that will appear among all the users, based upon a snapshot of the social network. These works treat all users equally and try to predict social links for all users in the network. However, in real-world social networks, many new users are joining in the service every day. Predicting social links for new users are more important than for those existing active users in the network as it will leave the first impression on the new users. First impression often has lasting impact on a new user and may decide whether he will become an active user. A bad first impression can turn a new user away. So it is important to make meaningful recommendation to a new user to create a good first impression and attract him to participate more. For simplicity, we refer users that have been actively using the the network for a long time as “old users”. It has been shown in previous works that there is a negative correlation between the age of nodes in the network and their link attachment rates. The distribution of linkage formation probability follows a power-law decay with the age of nodes [8]. So, new users are more likely to accept the recommended links compared with existing old users and predicting links for new users could lead to more social connections.

In this paper, we study the problem of predicting social links for new users, who have created their accounts for just a short period of time. The link prediction problem for new users is different from traditional link prediction problems. Conventional supervised link prediction methods implicitly or explicitly assume that the information are identically distributed over all the nodes in the network without considering the joining time of the users. The models trained over one part of the network can be directly used to predict links in other parts of the network. However, in real-world social networks, the information distributions of the new users could be very different from old users. New users may have only a few activities or even no activities (i.e., no social links or other auxiliary information) in the network. While, old users usually have abundant activities and auxiliary information in the network. In Figures 1 and 2, we show the degree distributions of users who registered their accounts within three months and old users who registered more than three

months before in Twitter and Foursquare respectively. We find that the social link distributions of new users and old users are totally different from each other in both Foursquare and Twitter. As a result, conventional supervised link prediction models trained over old users based upon structural features, such as common neighbors, may not work well on the new users.

Another challenging problem in link prediction for new users is that information owned by new users can be very rare or even totally missing. Conventional methods based upon one single network will not work well due to the lack of historical data about the new users. In order to solve this problem, we need to transfer additional information about the new users from other sources. Nowadays, people are usually involved in multiple social networks to enjoy more services. For example, people will join Foursquare to search for nearby restaurants to have dinner with their family. Meanwhile, they tend to use Facebook to socialize with their friends and involve in Twitter to post comments about recent news. The accounts of the same user in different networks can be linked through account alignments. For example, when users register their Foursquare accounts, they can use their Facebook or Twitter accounts to sign in the Foursquare network. In this paper, we name such links among accounts of the same user as “anchor links” [9], which could help align user’ accounts across multiple social networks. For example, in Figure 3, there are many users in two networks respectively. We find that the accounts in these two networks are actually owned by 6 different users in reality and we add an *anchor link* between each pair of user accounts corresponding to the same user. Via the *anchor links*, we could locate users’ corresponding accounts in the other networks.

New users in one social network (i.e., *target* network) might have been using other social networks (i.e., *source* networks) for a long time. These user accounts in the source networks can provide additional information about the new users in the target network. This additional information is crucial for link prediction for new users, especially when the new users have little activities or no activities in the target network (i.e., cold start problem). For example, in Figure 3, we have two social networks, i.e., the target network and the source network, with aligned user accounts. In the target network, there are many old users with abundant social links and auxiliary information, such as posts, spatial and temporal activities. In addition, there are also some new users, i.e., user u_1^t and u_2^t , in the target network. These two new users have just created their accounts in the target network and have not yet created many social links or auxiliary information. However, we can see that there is abundant information about these two new users in the source network, because of their “anchor linked” user accounts u_1^s and u_2^s in the source network. In this paper, we propose to exploit the new users’ information in source networks to help improve the link prediction performances in the target network.

The problem of social link prediction for new users by using aligned social networks has not been studied yet. It is novel and totally different from existing link prediction problems,

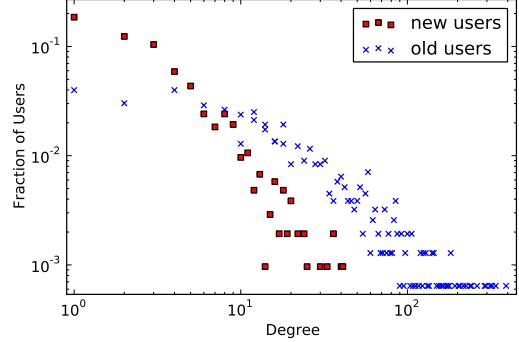


Fig. 1. Degree distributions of users in Foursquare network.

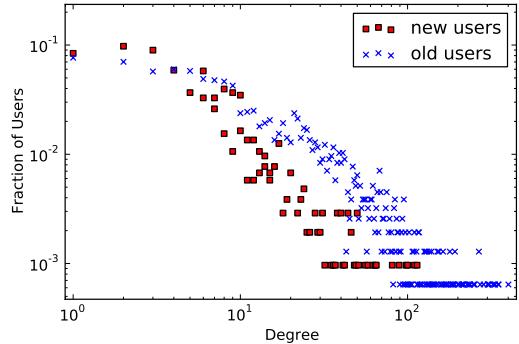


Fig. 2. Degree distributions of users in Twitter network.

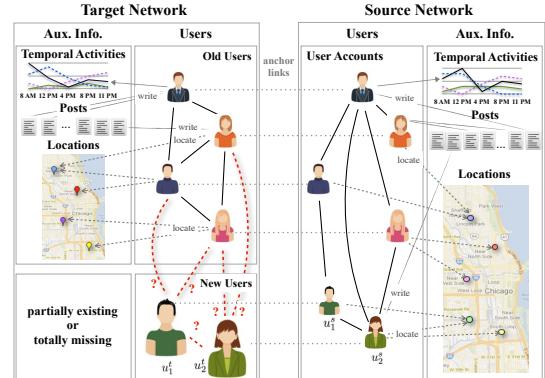


Fig. 3. Example of predicting social links across two aligned heterogeneous online social networks.

e.g., link prediction via personalized sampling [5], pseudo cold start link prediction problem [6], [11], [22], link prediction via network transfer [14], [19] and the traditional transfer learning problems in feature space [3], [4], [12], [15], [21]. A more detailed comparison of all these problems are shown in Table I and Figure 4.

In spite of its significance, social link prediction for new users across aligned social networks is very challenging to solve due to the following reasons:

- 1) *Differences in information distributions.* In order to use the old users’ information in the target network, we need to overcome the problem of the differences in

TABLE I
SUMMARY OF RELATED PROBLEMS.

Property	Predicting Links across Aligned Networks with Personalized Sampling	Link Prediction via Biased Sampling [5]	Pseudo Cold Start Link Prediction [6], [11], [22]	Link Prediction via Network Transfer [14], [19]	Transfer Learning in Feature Space [3], [4], [12], [15], [21]
# networks	multiple	multiple	single	multiple	multiple domains
network type	heterogeneous	homogeneous	heterogeneous	heterogeneous	heterogeneous
network alignment	yes	no	no	no	n/a
target	new users	incomplete network	incomplete network	incomplete network	n/a
sampling	within network	across network	n/a	n/a	n/a
can handle cold start problem?	cold start	n/a	pseudo cold start	n/a	n/a
knowledge to transfer	network structure through anchor links	network structure via attribute info.	n/a	triad patterns in network structure	knowledge in feature space
transfer route	intra-network and inter-network	inter-network	n/a	inter-network	intra-domain or inter-domain

information distributions between old users and new users.

- 2) *No auxiliary information.* Another key part of the problem we want to study is the cold start link prediction problem caused by the lack of information about these new users. We need to find other information sources for such problem and provide the new users with high-quality social link prediction results.
- 3) *Aligned social networks.* Previous works on transfer learning focus on transferring knowledge between two domains via shared feature space or between two networks through shared triad linkage structures. No works have been done on aligned social networks yet.

In order to solve these problems, we propose a novel supervised cross aligned networks link recommendation method, SCAN-PS. Different from previous works, SCAN-PS extracts heterogeneous features from other aligned networks to improve link prediction results for new users in the target network. We analyze the problem about the differences in information distributions between new users and old users in details and propose a within-network personalized sampling method to accommodate that difference. What's more, SCAN-PS could also solve the cold start social link prediction problem assisted by other aligned source networks. Intra and inter network information transfers are conducted simultaneously to make full use of the information contained in these aligned networks to improve the prediction result.

II. PROBLEM FORMULATION

The problem studied in this paper is social link prediction for new users. We propose a supervised method based on aligned heterogeneous networks. In this section, we first define the concept of heterogeneous social network and aligned heterogeneous networks and then present the formulation of the social link prediction problem.

Definition 1 (Heterogeneous Networks): Let $G = (V, E)$ be a network containing different kinds of information, where the set $V = \bigcup_i V_i$ contains multiple kinds of nodes, where $V_i, i \in \{1, 2, \dots, |V|\}$ is the set of nodes of the same kind, $E = \bigcup_i E_i$ contains multiple types of links among the nodes,

where $E_i, i \in \{1, 2, \dots, |E|\}$ is the set of links of the same type.

In this paper, a heterogeneous social network is defined as $G = (V, E)$, where $V = U \cup L \cup T \cup W$ is the set of nodes in the network and $E = E_{social} \cup E_{locate} \cup E_{time} \cup E_{word}$ is the set of heterogeneous links in the network. U, L, T and W are the set of users, locations, time and word respectively, while $E_{social}, E_{locate}, E_{time}$ and E_{word} are the sets of links between friends, locations, timestamps, words and users in U .

Definition 2 (Aligned Heterogeneous Networks): Let $\mathcal{G} = (G_{set}, A_{set})$ be aligned heterogeneous social networks, where $G_{set} = \{G^1, G^2, \dots, G^n\}$ is the set of heterogeneous social networks, whose size is $n = |G_{set}|$, and $A_{set} = \{A^{1,2}, A^{1,3}, \dots, A^{1,n}, A^{2,1}, \dots, A^{n,n-1}\}$ is the set of directed anchor links between pairwise networks in G_{set} and $A^{i,j} \subseteq U^i \times U^j$ is the set of anchor links between G^i and G^j , where U^i and U^j are the user sets in graph G^i and G^j .

Definition 3(Anchor Link): Link (u_m^i, u_n^j) is an *anchor link* between G^i and G^j iff. $(u_m^i \in U^i) \wedge (u_n^j \in U^j) \wedge (u_m^i \text{ and } u_n^j \text{ are accounts owned by the same user})$.

Social Link Recommendation is a traditional problem first proposed by [13] and has been studied for many years. Different from prior works, in this paper, we want to study this problem for new users in the target network by using aligned heterogeneous social networks. Let $\mathcal{G} = (\{G^t, G^s\}, \{A^{t,s}, A^{s,t}\})$ be two aligned heterogeneous social networks, where G^t is the target network and G^s is an aligned source network. $A^{t,s}$ and $A^{s,t}$ are the sets of directed anchor links between G^t and G^s . We want to predict social links for the new users in the target network. Let $U^t = U_{new}^t \cup U_{old}^t$ be the user set in G^t , where U_{new}^t, U_{old}^t are the sets of new users and old users and $U_{new}^t \cap U_{old}^t = \emptyset$. What we want to predict is a subset of potential social links between the new users and all other users: $L \subseteq U_{new}^t \times U^t$. In other words, we want to build a function $f : L \rightarrow \{0, 1\}$, which could decide whether certain links related to new users exist in the target network or not.

III. PROPOSED METHODS

In this section, we will describe the sets of features extracted from the aligned heterogeneous networks first. Next, we will analyze the problem of the differences in information

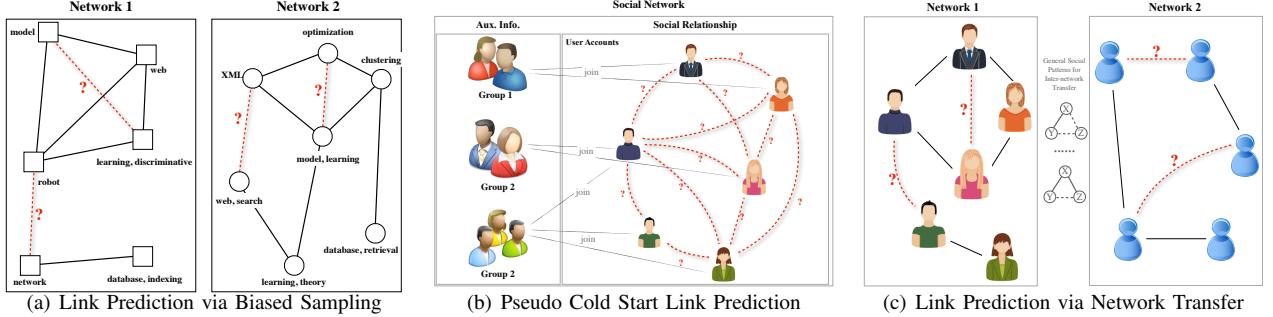


Fig. 4. Comparison of three other different link prediction problems

distributions between old users and new users in the target network and propose a personalized sampling method to solve it. Finally, we will propose our supervised cross network link prediction method SCAN-PS to predict potential social links for new users in the target network with information in aligned heterogeneous networks.

A. Heterogeneous Features

The networks used by us are heterogeneous and involve different nodes and complex dependency relationships. To make full use of information inside the networks, we extract four different categories of features, which are (1) social features, (2) spatial distribution features, (3) temporal distribution features and (4) text content features. More detailed information about these features is listed below.

- Social Features:** Social relationships are what we intend to predict for new users and the existing social information in the network should be utilized definitely. We extract three different social features from the social information in the networks, which are “*common neighbour*” (CN), “*Jaccard’s Coefficient*” (JC) and “*Adamic/Adar Measure*” (AA) [1].

Common neighbour: $CN(u_i, u_j)$ shows the number of shared neighbours of user u_i and u_j in the network. Let $\Gamma(u)$ denote the set of neighbours of user u in the network, then:

$$CN(u_i, u_j) = |\Gamma(u_i) \cap \Gamma(u_j)|$$

Jaccard’s coefficient: $JC(u_i, u_j)$ takes all the neighbours of u_i and u_j into account, considering that $CN(u_i, u_j)$ could be very large because each one has a lot of neighbours rather than they are strongly related to each other.

$$JC(u_i, u_j) = \frac{|\Gamma(u_i) \cap \Gamma(u_j)|}{|\Gamma(u_i) \cup \Gamma(u_j)|}$$

Adamic/Adar Measure: $AA(u_i, u_j)$ further gives each common neighbour u_k a weight $\frac{1}{\log |\Gamma(u_k)|}$ to measure u_k ’s significance, which is defined as follows:

$$AA(u_i, u_j) = \sum_{u_k \in (\Gamma(u_i) \cap \Gamma(u_j))} \frac{1}{\log |\Gamma(u_k)|}$$

- Spatial Distribution Features:** Lots of social networks have provided users the service to allow users to display their locations along with the online posts. We organize the locations that two users u_i and u_j have visited in two location vectors: $l(u_i)$ and $l(u_j)$. These two vectors are of the same length and each column corresponds to a specific location. The values stored in these two vectors are the times that the users has visited specific places. We extract several features from the location information: (1) inner product of these two location vectors, which is $l(u_i) \cdot l(u_j)$, (2) cosine similarity of these two location vectors: $\frac{l(u_i) \cdot l(u_j)}{\|l(u_i)\| \cdot \|l(u_j)\|}$, which takes the total number of locations that these two users have been to into account, (3) the Euclidean distance of two users’ location vectors: $(\sum_k (l(u_i)_k - l(u_j)_k)^2)^{1/2}$. In addition, we also extend CN and JC to get the number and ratio of shared locations that two users have both visited, which are (4) $CNLOCATION(u_i, u_j) = |\Phi(u_i) \cap \Phi(u_j)|$ and (5) $JCLOCATION(u_i, u_j) = \frac{|\Phi(u_i) \cap \Phi(u_j)|}{|\Phi(u_i) \cup \Phi(u_j)|}$, where $\Phi(u)$ denotes the set of locations that user u has visited. (6) The average geographic distance of all the locations that users u_i and u_j have been to is also used as a feature.
- Temporal Distribution Features:** Besides the location information, we could also get the timestamps of online posts, which could reveal users’ activity patterns. We divide each day into 24 slots and save the number of posts that a user posted at each hour in vector $t(u)$, whose length is 24. Similar to the location features, we extract: (1) the number of shared time slots when publishing online posts, (2) the inner product of $t(u_i)$ and $t(u_j)$, which is $t(u_i) \cdot t(u_j)$, (3) cosine similarity of these two temporal distribution vectors: $\frac{t(u_i) \cdot t(u_j)}{\|t(u_i)\| \cdot \|t(u_j)\|}$, (4) the Euclidean distance of these two vectors: $(\sum_k (t(u_i)_k - t(u_j)_k)^2)^{1/2}$ and (5) extend JC in time distribution: $\frac{t(u_i) \cdot t(u_j)}{\|t(u_i)\|^2 + \|t(u_j)\|^2 - t(u_i) \cdot t(u_j)}$.
- Text Usage Features:** Text content of posts could be used as a hint to show the two users’ similarity in their word usage habits. In addition, text content could also reveal some personal information of a user. Consider, for example, the fact that a user usually post tweets about food in Twitter strongly implies that the user like food a

lot. For this reason, two users with similar word usage habits are also likely to share similar hobbies and are likely to make friends with each other. We transform the words used by two users u_i and u_j into two bag-of-words vectors: $w(u_i)$ and $w(u_j)$ weighted by TF-IDF. Similar features are extracted from these two vectors: (1) the number of shared words used online, (2) inner product of two word vectors: $w(u_i) \cdot w(u_j)$, (3) cosine similarities of these two word vectors: $\frac{w(u_i) \cdot w(u_j)}{\|w(u_i)\| \cdot \|w(u_j)\|}$, (4) the Euclidean distance of these two word vectors: $(\sum_k (w(u_i)_k - w(u_j)_k)^2)^{1/2}$ and (5) extended JC in text usage: $\frac{w(u_i) \cdot w(u_j)}{\|w(u_i)\|^2 + \|w(u_j)\|^2 - w(u_i) \cdot w(u_j)}$.

B. Link Prediction within Target Network

After extracting these four categories of features, next we will propose to use them to build supervised methods to predict links for new users in the target network. Before doing that, we notice that the new users' information distribution can be totally different from that of the old users in the target network. However, information of both new users and old users is so important that should be utilized. In this section, we will analyze the differences in information distributions of new users and old users in the target network and propose a personalized within-network sampling method to process old users' information to accommodate the differences. Then, we will revise the traditional supervised link prediction method by using the old users' sampled information in the target network to improve the prediction results.

1) *Sampling Old Users' Information*: A natural challenge inherent in the usage of the target network to predict social links for new users is the differences in information distributions of new users and old users as mentioned before. To address this problem, we propose a method to accommodate old users' and new users' sub-network by using a within-network personalized sampling method to process old users' information. Totally different from the link prediction with sampling problem studied in [5], we are conducting personalized sampling within the target network, which contains heterogeneous information, rather across multiple non-aligned homogeneous networks. And the link prediction target are the new users in the target network in our problem.

By sampling the old users' sub-network, we want to meet the following objectives:

- *Maximizing Relevance*: We aim at maximizing the relevance of the old users' sub-network and the new users' sub-network to accommodate differences in information distributions of new users and old users in the heterogeneous target network.
- *Information Diversity*: Diversity of old users' information after sampling is still of great significance and should be preserved.
- *Structure Maintenance*: Some old users possessing sparse social links should have higher probability to survive after sampling to maintain their links so as to maintain the network structure.

Let the heterogeneous target network be $G^t = \{V^t, E^t\}$, and $U^t = U_{old}^t \cup U_{new}^t \subset V^t$ is the set of user nodes (i.e., set of old users and new users) in the target network. Personalized sampling is conducted on the old users' part: $G_{old}^t = \{V_{old}^t, E_{old}^t\}$, in which each node is sampled independently with the sampling rate distribution vector $\delta = (\delta_1, \delta_2, \dots, \delta_n)$, where $n = |U_{old}^t|$, $\sum_{i=1}^n \delta_i = 1$ and $\delta_i \geq 0$. Old users' heterogeneous sub-network after sampling is denoted as $\bar{G}_{old}^t = \{\bar{V}_{old}^t, \bar{E}_{old}^t\}$.

We aim at making the old users' sub-network as relevant to new users' as possible. To measure the similarity score of a user u_i and a heterogeneous network G , we define a relevance function as follows:

$$R(u_i, G) = \frac{1}{|U|} \sum_{u_j \in U} S(u_i, u_j)$$

where set U is the user set of network G and $S(u_i, u_j)$ measures the similarity between user u_i and u_j in the network. Each user has social relationships as well as other auxiliary information and $S(u_i, u_j)$ is defined as the average of similarity scores of these two parts:

$$S(u_i, u_j) = \frac{1}{2}(S_{aux}(u_i, u_j) + S_{social}(u_i, u_j))$$

In our problem settings, the auxiliary information of each users could also be divided into 3 categories: *location*, *temporal*, and *text*. So, $S_{aux}(u_i, u_j)$ is defined as the mean of these three aspects.

$$S_{aux}(u_i, u_j) = \frac{1}{3}(S_{text}(u_i, u_j) + S_{loc}(u_i, u_j) + S_{temp}(u_i, u_j))$$

There are many different methods measuring the similarities of these auxiliary information in different aspects, e.g. cosine similarity. As to the social similarity, Jaccard's Coefficient can be used to depict how similar two users are in their social relationships.

The relevance between the sampled old users' network and the new users' network could be defined as the expectation value of function $R(\bar{u}_{old}^t, G_{new}^t)$:

$$\begin{aligned} R(\bar{G}_{old}^t, G_{new}^t) &= \mathbb{E}(R(\bar{u}_{old}^t, G_{new}^t)) \\ &= \frac{1}{|U_{new}^t|} \sum_{j=1}^{|U_{new}^t|} \mathbb{E}(S(\bar{u}_{old}^t, u_{new,j}^t)) \\ &= \frac{1}{|U_{new}^t|} \sum_{j=1}^{|U_{new}^t|} \sum_{i=1}^{|U_{old}^t|} \delta_i \cdot S(\bar{u}_{old,i}^t, u_{new,j}^t) \\ &= \delta' s \end{aligned}$$

where vector s equals:

$$\frac{1}{|U_{new}^t|} \left[\sum_{j=1}^{|U_{new}^t|} S(\bar{u}_{old,1}^t, u_{new,j}^t), \dots, \sum_{j=1}^{|U_{new}^t|} S(\bar{u}_{old,n}^t, u_{new,j}^t) \right]^T$$

and $|U_{old}^t| = n$. Besides the relevance, we also need to ensure that the diversity of information in the sampled old users' sub-network could be preserved. Similarly, it also includes

diversities of the auxiliary information and social relationships. The diversity of auxiliary information is determined by the sampling rate δ_i , which could be defined with the averaged *Simpson Index* [17] over the old users' sub-network.

$$D_{aux}(\bar{G}_{old}^t) = \frac{1}{|U_{old}^t|} \cdot \sum_{i=1}^{|U_{old}^t|} \delta_i^2$$

As to the diversity in the social relationship, we could get the existence probability of a certain social link (u_i, u_j) after sampling to be proportional to $\delta_i \cdot \delta_j$. So, the diversity of social links in the sampled network could be defined as average existence probabilities of all the links in the old users' sub-network.

$$D_{social}(\bar{G}_{old}^t) = \frac{1}{|S_{old}^t|} \cdot \sum_{i=1}^{|U_{old}^t|} \sum_{j=1}^{|U_{old}^t|} \delta_i \cdot \delta_j \times I(u_i, u_j)$$

where $|S_{old}^t|$ is the size of social link set of old users' sub-network and $I(u_i, u_j)$ is an indicator function $I : (u_i, u_j) \rightarrow \{0, 1\}$ to show whether a certain social link exists or not originally before sampling. For example, if link (u_i, u_j) is a social link in the target network originally before sampling, then $I(u_i, u_j) = 1$, otherwise it equals to 0.

Considering these two terms simultaneously, we could have the diversity of information in the sampled old users' sub-network to be the average diversities of these two parts:

$$\begin{aligned} D(\bar{G}_{old}^t) &= \frac{1}{2}(D_{social}(\bar{G}_{old}^t) + D_{aux}(\bar{G}_{old}^t)) \\ &= \frac{1}{2}\left(\sum_{i=1}^{|U_{old}^t|} \sum_{j=1}^{|U_{old}^t|} \frac{1}{|S_{old}^t|} \cdot \delta_i \cdot \delta_j \times I(u_i, u_j)\right. \\ &\quad \left.+ \sum_{i=1}^{|U_{old}^t|} \frac{1}{|U_{old}^t|} \cdot \delta_i^2\right) \\ &= \boldsymbol{\delta}' \cdot \left(\frac{1}{2|S_{old}^t|} \cdot \mathbf{A}_{old}^t + \frac{1}{2|U_{old}^t|} \cdot \mathbf{I}_{|U_{old}^t|}\right) \cdot \boldsymbol{\delta} \end{aligned}$$

where matrix $\mathbf{I}_{|U_{old}^t|}$ is the diagonal identity matrix of size $|U_{old}^t| \times |U_{old}^t|$ and \mathbf{A}_{old}^t is the adjacency matrix of old users' sub-network.

To ensure that the structure of the original old users' subnetwork is not destroyed, we need to ensure that users with few links could also preserve their links. So, we could add a regularization term to increase the sampling rate for these users as well as their neighbours by maximizing the following terms:

$$Reg(\bar{G}_{old}^t) = \min\{\mathcal{N}_i, \min_{u_j \in \mathcal{N}_i} \{\mathcal{N}_j\}\} \times \delta_i^2 = \boldsymbol{\delta}' \cdot \mathbf{M} \cdot \boldsymbol{\delta}$$

where matrix \mathbf{M} is a diagonal matrix with element $\mathbf{M}_{i,i} = \min\{\mathcal{N}_i, \min_{u_j \in \mathcal{N}_i} \{\mathcal{N}_j\}\} = \min\{\mathcal{N}_i, \{\mathcal{N}_i | u_j \in \mathcal{N}_i\}\}$ and $\mathcal{N}_j = |\Gamma(u_j)|$ is the size of user u_j 's neighbour set. So, if a user or his/her neighbours have few links, then this user as well as his/her neighbours should have higher sampling rate so as to preserve the links between them.

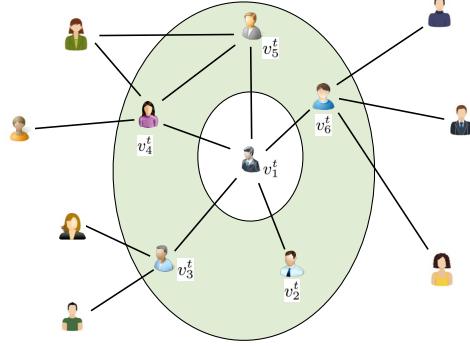


Fig. 5. Personalized sampling preserving network structures.

For example, in Figure 5, we have 6 users. To decide the sampling rate of user v_1^t , we need to consider his/her social structure. We find that since v_1^t 's neighbour v_2^t has no other neighbour except v_1^t . To preserve the social link between v_1^t and v_2^t we need to increase the sampling rate of v_2^t . However, the existence probability of link (v_1^t, v_2^t) is also decided by the sampling rate of user v_1^t , which also needs to be increased too.

Combining the diversity term and the structure preservation term, we could define the regularized diversity of information after sampling to be

$$D_{Reg}(\bar{G}_{old}^t) = D(\bar{G}_{old}^t) + Reg(\bar{G}_{old}^t) = \boldsymbol{\delta}' \cdot \mathbf{N} \cdot \boldsymbol{\delta}$$

$$\text{where } \mathbf{N} = \frac{1}{2|U_{old}^t|} \cdot \mathbf{I}_{|U_{old}^t|} + \frac{1}{2|S_{old}^t|} \cdot \mathbf{A}_{old}^t + \mathbf{M}.$$

The optimal value of $\boldsymbol{\delta}$ should be able to maximize the relevance of new users' sub-network and old users' as well as the regularized diversity of old users' information in the target network

$$\begin{aligned} \boldsymbol{\delta} &= \arg \max_{\boldsymbol{\delta}} R(\bar{G}_{old}^t, G_{new}^t) + \theta \cdot D_{Reg}(\bar{G}_{old}^t) \\ &= \arg \max_{\boldsymbol{\delta}} \boldsymbol{\delta}' \mathbf{s} + \theta \cdot \boldsymbol{\delta}' \cdot \mathbf{N} \cdot \boldsymbol{\delta} \\ &\text{s.t., } \sum_{i=1}^{|U_{old}^t|} \delta_i = 1 \text{ and } \delta_i \geq 0. \end{aligned}$$

where, parameter θ is used to weight the importance of term regularized information diversity.

2) TRAD: A traditional supervised link prediction method TRAD (Traditional Link Prediction) for our task is to use the existing links in the target network to train a classifier and apply it to classify the potential social links for new users. In method TRAD, only the target network is used, which consists of new users and unsampled old users. To overcome the difference in information distribution between new users and old users in the target network, we revise it a little and get method: TRAD-PS (Traditional Link Prediction with Personalized Sampling). TRAD-PS consists of two steps: (1) personalized sampling of the old users' sub-network with the previous method; (2) usage of similar techniques as TRAD to predict links based on the sampled network.

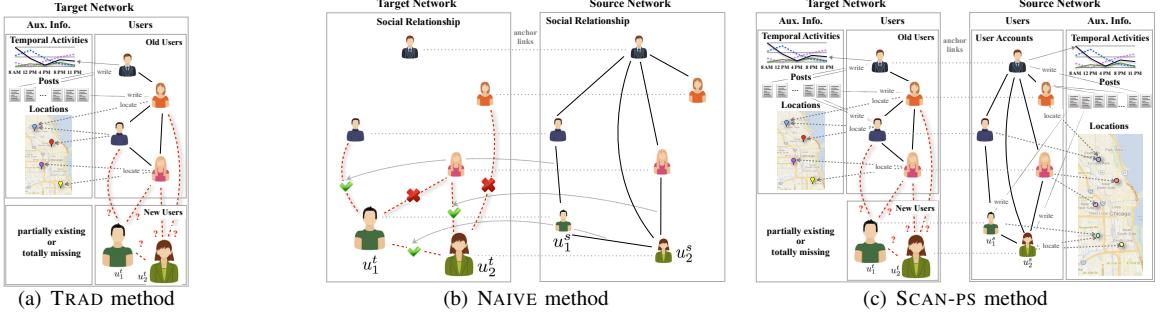


Fig. 6. Different methods to predict social link for new users.

Theoretically, TRAD and TRAD-PS could work well by using information in the target network. However, considering the fact it is impossible for new users to possess large amount of information actually, TRAD and TRAD-PS would suffer from the long-standing cold start problem caused by the lack of historical information indicating these new users' preferences. This problem will be even worse when dealing with brand-new users, who have no information at all in the target network.

For example, in Figure 6(a), user u_1^t and u_2^t are two new users in the target network, who possess very few social links with other users and little auxiliary information. We can not get any information about these two new users and the information we could use is that possessed by other old users. As a result, the links that TRAD and TRAD-PS predicted could hardly be of high quality.

In order to deal with such problem, we will use aligned networks simultaneously in the next section.

C. Cold-Start Link Prediction

In our problem settings, we have two aligned social networks and the methods proposed in previous section using the target network may suffer from the cold start problems when processing brand-new users. In this section, we will propose two methods to utilize the aligned source network to help solve the problem and improve the prediction results.

1) NAIVE: Suppose we have a new user u_i^t in the target network, a naive way to use the aligned source network to recommend social links for user u_i^t is to recommend all the corresponding social links related to this user's aligned account u_i^s in the aligned source network to him/her. Based on this intuition, we propose a cold start link prediction method NAIVE (Naive Link Prediction). To clarify how NAIVE works in reality, we will give an example next. And before that, we formally define a new term *pseudo label* to denote the existence of corresponding links in the aligned source network.

Definition 4 (Pseudo Label): The pseudo label of a link (u_i^t, u_j^t) in the target denotes the existence of its corresponding link (u_i^s, u_j^s) in the aligned source network and it is 1 if (u_i^s, u_j^s) exists and 0 otherwise.

For example, in Figure 6(b), to decide whether to recommend u_1^t to u_2^t in the target network or not, we could find their aligned accounts: u_1^s and u_2^s , and their social link: (u_1^s, u_2^s) in

the aligned source network with the help of *anchor links*. We find that u_1^s and u_2^s are friends in the aligned source network and link (u_1^s, u_2^s) exists in the aligned source network. As a result, the pseudo label of link (u_1^t, u_2^t) is 1 and in the target network, we could recommend u_2^t to u_1^t . And that is the reason why the social link between u_1^t and u_2^t is predicted to be existent by method NAIVE. Other links in Figure 6(b) are predicted in a similar way.

Method NAIVE is very simple and could work well in our task even when these new users are brand new, which means that we could overcome the cold start problem by using this method. However, it may still suffer from some disadvantages: (1) the social structures of different networks are not always identical which will degrade the performance of NAIVE a lot; (2) NAIVE only utilizes these new users' social linkage information in the source network and ignores all other information.

2) SCAN-PS: To overcome all these disadvantages mentioned above, a new method SCAN-PS (Supervised Cross Aligned Networks Link Prediction with Personalized Sampling) is proposed. As shown in Figure 6(c), it could use heterogeneous information existing in both the target network and the aligned source and it is built across two aligned social networks. Taking advantage of the anchor links, we could locate the users' aligned accounts and their information in the aligned source network exactly. If two aligned networks are used simultaneously, different categories of features are extracted from aligned networks. To use multiple networks, these feature vectors extracted for the corresponding links in aligned networks are merged into an expanded feature vector. The expanded feature vector together with the labels from the target network are used to build a cross-network classifier to decide the existence of social links related to these new users in the target network. This is how method SCAN-PS works. SCAN-PS is quite stable and could overcome the cold start problem for the reason that the information about all these users in the aligned source network doesn't change much with the variation of the target network and we get the information showing of these new users' preferences from the information he/she leaves in the aligned source network. As the old users' information inside the target network is also used in SCAN-PS, personalized sampling is also conducted to preprocess the old

TABLE II
PROPERTIES OF THE HETEROGENEOUS SOCIAL NETWORKS

property	network	
	Twitter	Foursquare
# node	user	5,223
	tweet/tip	9,490,707
	location	297,182
# link	friend/follow	164,920
	write	9,490,707
	locate	615,515
		48,756

users' information in the target network.

In addition to features mentioned before, method SCAN-PS also utilizes the information used by NAIVE, which is the *pseudo label* defined before, by regarding it as an extra feature.

- **An Extra Feature:** We use the pseudo label as an extract feature to denote the existence of the corresponding links in the aligned source network.

Compared with SCAN-PS with NAIVE, SCAN-PS has many advantages: (1) SCAN-PS utilizes multiple categories of information; (2) SCAN-PS can make use of the information hidden in the old users' network by incorporating them into the training set; (3) SCAN-PS doesn't rely on the assumption that the social relationships in different networks are identical, which is very risky actually.

Compared with TRAD and TRAD-PS, SCAN-PS can solve the cold start problem as it could have access to information owned by these new users in other aligned source networks. Similar to TRAD and TRAD-PS, these new users' information is used if they are not very new and other old users' information in the target is also preprocessed by using the within-network personalized sampling method before the intra-network knowledge transfer.

IV. EXPERIMENTS

A. Data Preparation

In previous sections, we propose method SCAN-PS to solve the challenges mentioned in the introduction section. To test whether our method is applicable in reality, we conduct extensive experiments on two real-world aligned heterogeneous social networks: Twitter and Foursquare. A more detailed description about these two social networks datasets is summarized in Table II. We crawl these two datasets by using the APIs provided by Twitter and Foursquare.

- **Twitter:** The first social network used by us is Twitter, a famous online microblogging network containing lots of heterogeneous information. 5,223 users together with 9,490,707 tweets posted by them are crawled. Each tweet could contain the text content, timestamps and locations. About 615,515 tweets are found to possess location information(latitude and longitude), which accounts for about 6.5% of all the tweets.
- **Foursquare:** Another social network used by us is Foursquare, which is a well-known location-based social

network(LBSN). We crawled 5,392 users and their tips, the number of which is 48,756. Similar to tweets, content information, timestamps as well as the specific locations(latitude and longitude) are also available. Different from Twitter, in Foursquare, each tip is related to a location and the number of locate links is equal to the number of tips.

Both of these two social networks contain social links, which are used as the ground truth in the experiments. The anchor links between these two networks is acquired by crawling the hyperlink of the users' Twitter account in their Foursquare homepages.

B. Experiment Settings

Comparison Methods: To evaluate the effectiveness of SCAN-PS in predicting social links for new users, we compare SCAN-PS with many baseline methods, including both supervised and unsupervised methods. To ensure the fairness of the comparisons, LibSVM [2] of linear kernel with default parameter is used as the base classifier for all supervised methods. The evaluation methods used by us are: AUC and Accuracy. Next, we will summary all the comparative methods first and then give the description of the experiment settings and the evaluation method.

- *Source Network + Target Network:* SCAN-PS could use the information in the aligned source network and the target network at the same time. Old users' information in the target network is used by SCAN-PS and it is processed with personalized sampling method before being transferred. To show personalized sampling of old users' information is helpful for our task, we compare it with another weaker baseline method, which use old users' information without sampling. The method is named as SCAN (Supervised Cross Aligned Networks Link Prediction).
- *Target Network Only:* Some other supervised baseline methods are built only with the information in target network. Method NEW is built only with new users' information, while method OLD only uses the information old users' information in the target network. Method TRAD could use all the information in the target network. To show that personalized sampling of old users' information is helpful, methods TRAD-PS and OLD-PS are used to compare with TRAD and OLD respectively.
- *Source Network Only:* To show that using two networks simultaneously is better than using one network. Besides those methods using target network only, we also compare SCAN-PS with another baseline method SOURCE, built with all the information in the target network.
- *Unsupervised Methods:* NAIVE and some traditional unsupervised social link prediction methods are also used as the unsupervised baseline methods to be compared with SCAN-PS. The other unsupervised baseline methods include *Common Neighbour* (CN), *Jaccard Coefficient* (JC) and *Adamic Adar* (AA). NAIVE uses social information in the aligned source network only, while all other

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT SOCIAL LINK PREDICTION METHODS FOR USERS OF DIFFERENT DEGREES OF NEWNESS. TARGET NETWORK: FOURSQUARE. SOURCE NETWORK: TWITTER. (DEGREE OF NEWNESS DENOTES THE RATIO OF INFORMATION OWNED BY USERS)

measure	method	REMAINING INFORMATION RATIO							
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
AUC	SCAN-PS	0.783±0.009	0.839±0.008	0.864±0.013	0.883±0.008	0.902±0.011	0.910±0.009	0.912±0.003	0.913±0.012
	SCAN	0.768±0.013	0.808±0.007	0.833±0.009	0.846±0.006	0.854±0.005	0.860±0.008	0.869±0.009	0.882±0.006
	SOURCE	0.761±0.008	0.768±0.015	0.800±0.014	0.802±0.011	0.806±0.003	0.815±0.011	0.820±0.006	0.820±0.007
	TRAD-PS	0.553±0.007	0.626±0.003	0.69±0.012	0.681±0.012	0.701±0.008	0.701±0.007	0.735±0.014	0.736±0.013
	OLD-PS	0.554±0.016	0.567±0.01	0.564±0.022	0.571±0.012	0.558±0.005	0.578±0.009	0.570±0.015	0.575±0.010
	TRAD	0.555±0.006	0.593±0.007	0.622±0.009	0.646±0.012	0.658±0.006	0.671±0.016	0.681±0.010	0.708±0.011
	OLD	0.550±0.008	0.510±0.010	0.527±0.008	0.541±0.015	0.551±0.006	0.571±0.012	0.574±0.010	0.568±0.009
	NEW	0.495±0.018	0.616±0.011	0.631±0.005	0.646±0.006	0.653±0.009	0.656±0.004	0.670±0.010	0.675±0.009
Acc.	CN	0.500±0.000	0.523±0.005	0.536±0.004	0.552±0.006	0.562±0.004	0.573±0.005	0.576±0.007	0.587±0.003
	JC	0.500±0.000	0.523±0.005	0.534±0.006	0.554±0.007	0.562±0.010	0.572±0.005	0.575±0.009	0.587±0.004
	AA	0.500±0.000	0.521±0.004	0.531±0.003	0.548±0.006	0.556±0.004	0.566±0.004	0.569±0.006	0.583±0.002
	SCAN-PS	0.747±0.005	0.772±0.010	0.802±0.007	0.811±0.009	0.813±0.012	0.821±0.008	0.826±0.005	0.834±0.008
	SCAN	0.732±0.014	0.746±0.008	0.763±0.010	0.778±0.007	0.791±0.008	0.790±0.009	0.794±0.009	0.803±0.009
	SOURCE	0.695±0.011	0.712±0.011	0.716±0.015	0.733±0.009	0.738±0.003	0.735±0.012	0.745±0.009	0.740±0.006
	TRAD-PS	0.506±0.004	0.600±0.006	0.610±0.009	0.625±0.005	0.628±0.005	0.632±0.009	0.645±0.006	0.653±0.007
	OLD-PS	0.506±0.002	0.504±0.002	0.505±0.004	0.512±0.026	0.518±0.006	0.535±0.010	0.520±0.015	0.524±0.026
Acc.	TRAD	0.506±0.002	0.524±0.006	0.540±0.004	0.559±0.006	0.586±0.009	0.599±0.007	0.624±0.012	0.635±0.009
	OLD	0.503±0.002	0.503±0.002	0.503±0.004	0.505±0.003	0.505±0.003	0.515±0.004	0.509±0.005	0.516±0.003
	NEW	0.478±0.010	0.563±0.009	0.581±0.004	0.591±0.007	0.602±0.009	0.604±0.006	0.615±0.010	0.628±0.005
	NAIVE	0.616±0.009	0.608±0.004	0.622±0.003	0.616±0.008	0.619±0.009	0.613±0.003	0.615±0.009	0.614±0.008

TABLE IV

PERFORMANCE COMPARISON OF DIFFERENT SOCIAL LINK PREDICTION METHODS FOR USERS OF DIFFERENT DEGREES OF NEWNESS. TARGET NETWORK: TWITTER. SOURCE NETWORK: FOURSQUARE. (DEGREE OF NEWNESS DENOTES THE RATIO OF INFORMATION OWNED BY USERS)

measure	method	REMAINING INFORMATION RATIO							
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
AUC	SCAN-PS	0.608±0.006	0.832±0.005	0.859±0.004	0.886±0.003	0.890±0.003	0.899±0.004	0.911±0.005	0.910±0.005
	SCAN	0.602±0.005	0.788±0.005	0.827±0.003	0.851±0.005	0.850±0.007	0.854±0.003	0.870±0.004	0.884±0.002
	SOURCE	0.621±0.007	0.736±0.005	0.734±0.005	0.743±0.006	0.745±0.004	0.743±0.001	0.749±0.003	0.749±0.008
	TRAD-PS	0.526±0.004	0.772±0.006	0.785±0.002	0.807±0.006	0.822±0.005	0.837±0.002	0.841±0.003	0.857±0.004
	OLD-PS	0.530±0.003	0.680±0.007	0.653±0.006	0.644±0.007	0.635±0.007	0.640±0.002	0.627±0.004	0.542±0.010
	TRAD	0.456±0.003	0.697±0.007	0.772±0.004	0.801±0.006	0.820±0.004	0.833±0.005	0.846±0.005	0.858±0.004
	OLD	0.423±0.002	0.519±0.004	0.528±0.005	0.568±0.006	0.600±0.006	0.629±0.003	0.653±0.003	0.674±0.004
	NEW	0.492±0.013	0.766±0.008	0.788±0.003	0.806±0.005	0.822±0.004	0.834±0.004	0.842±0.005	0.851±0.004
Acc.	CN	0.500±0.000	0.731±0.006	0.786±0.001	0.814±0.006	0.821±0.005	0.830±0.005	0.837±0.003	0.839±0.003
	JC	0.500±0.000	0.716±0.007	0.760±0.002	0.789±0.006	0.794±0.006	0.804±0.007	0.810±0.003	0.813±0.003
	AA	0.500±0.000	0.728±0.005	0.782±0.002	0.811±0.004	0.818±0.005	0.828±0.007	0.835±0.003	0.837±0.003
	SCAN-PS	0.588±0.001	0.769±0.004	0.793±0.005	0.815±0.004	0.822±0.002	0.848±0.004	0.860±0.005	0.868±0.004
	SCAN	0.582±0.004	0.685±0.007	0.715±0.004	0.731±0.004	0.753±0.008	0.776±0.004	0.791±0.004	0.817±0.003
	SOURCE	0.573±0.006	0.669±0.005	0.676±0.003	0.680±0.005	0.684±0.002	0.683±0.004	0.686±0.003	0.686±0.008
	TRAD-PS	0.505±0.002	0.710±0.001	0.705±0.005	0.741±0.006	0.753±0.005	0.765±0.003	0.769±0.003	0.778±0.005
	OLD-PS	0.515±0.003	0.501±0.013	0.503±0.002	0.502±0.010	0.512±0.002	0.502±0.002	0.503±0.052	0.501±0.003
Acc.	TRAD	0.503±0.002	0.545±0.005	0.625±0.002	0.680±0.009	0.723±0.002	0.745±0.003	0.763±0.004	0.767±0.005
	OLD	0.516±0.006	0.500±0.002	0.513±0.001	0.504±0.002	0.503±0.002	0.510±0.002	0.500±0.001	0.503±0.002
	NEW	0.488±0.008	0.661±0.006	0.707±0.003	0.731±0.004	0.743±0.004	0.758±0.005	0.765±0.003	0.775±0.004
	NAIVE	0.552±0.003	0.552±0.002	0.553±0.002	0.552±0.004	0.554±0.003	0.553±0.004	0.553±0.002	0.552±0.003

three methods are based on the target network without sampling.

Experiment Setting: To get two fully aligned networks, 1000 users in each of these two networks with full anchor links are randomly sampled with breadth-first-search and these users' complete social links and other auxiliary information is preserved. Then, we randomly sample 20% of these 1000 users in the target network as new users and the remaining are regarded as old users. All the social links related to new users are grouped into a positive link set and equivalent number of non-existent social links related these new users are organized into a negative link set. We partition these

two link sets into two groups by 5-fold cross validation: four folds are used as the training set and the remaining one fold is used as the testing set. To get different degree of newness, all the information, i.e., social links and other auxiliary information, owned by these new users inside the network are randomly sample with a certain rate denoting the novelty. If the old users' information is used, we use the within-network personalized sampling method to preprocess the old users' information inside the target network before the intra-network transfer. The personalized sampling vector δ is learnt from the target network. All the existent social links related to the old users after sampling and equivalent

number of nonexistent links in the target network are added to the training set. Heterogeneous features of each positive and negative link are extracted from the aligned networks. If two networks are used simultaneously, the feature vectors extracted from each social network are merged into expanded ones. There are two networks in our dataset and we choose Foursquare as the target network and Twitter as the source network first. And, then use them in a reverse way.

Evaluation Methods: Evaluation methods utilized by us are AUC and Accuracy. Since the three unsupervised methods *CN*, *JC*, *AA* could only predict a real-number score to measure the confidence about the existence of a certain social link, we only use AUC to evaluate these methods' performance. And *NAIVE* could only predict the labels without confidence, so it is evaluated only by Accuracy. All other methods are evaluated by both AUC and Accuracy.

C. Experiment Result

In Table III, we show the performance of all the methods under the evaluation of AUC and Accuracy when Foursquare is used as the target network and Twitter is used as the source network. By comparing method *OLD-PS* with method *OLD* and comparing method *TRAD-PS* with method *TRAD*, we could find that sampling the old users' information could improve our prediction performance. By comparing method *TRAD* with *OLD* and *NEW*, we find that the performance of *TRAD* is even worse than *OLD*, which means that old users' information without sampling could degrade the prediction performance. Comparison of *TRAD-PS* with *OLD-PS* and *NEW* reveals that using the sampled old users' and the new users' information simultaneously could lead to a better prediction results. As the information owned by these new users increases, i.e., the remaining information ratio increases, the effectiveness of sampling decreases continuously as the new users and old users without sampling are becoming more and more similar. By comparing the results of methods *SCAN*, *SOURCE* and *TRAD*, we find that using two networks at the same time could achieve better performance than using a single one. *SCAN-PS* performs better than *SCAN* indicates that personalized sampling could still work when two aligned networks are used simultaneously.

From the result, we also could find that most of these methods will fail to work because of the cold start problem when remaining information ratio is 0.0, which means that the users are brand new. However, method *SCAN-PS*, *SCAN*, *SOURCE* and *NAIVE* could still work well because these method could get information about new users from another aligned source network. It could support the intuition of this paper that using another aligned network could help cure the cold start problem. In Table IV, similarly results could be gotten when Twitter is used as the target network and Foursquare is used as the aligned source network.

V. RELATED WORK

Link prediction and recommendation first proposed in [13] is such a significant task in graph mining that it provides

researchers with the opportunities to study the network. Hasan et al. [7] is the first to study the link prediction problem as a supervised problem. Today, many social networks are heterogeneous and to conduct the link prediction in these networks, Sun et al. [18] propose a meta path-based prediction model to predict co-author relationship in the heterogeneous bibliographic network.

Mosting existing researches approach link prediction problem with a single source of information. However, Tang et al. [19] focus on inferring the particular type of links over multiple heterogeneous networks and develop a framework for classifying the type of social ties. To deal with the differences in information distributions of multiple networks, Qi et al. [5] propose to use biased cross-network sampling to do link prediction across networks. Kong et al. [9] propose to infer the anchor links between two heterogeneous networks with a two-phase methods.

Location-based social networks(LBSNs) are becoming quite hot in recent years and many works have been done on predicting links on these networks. Scellato et al. [16] predict social links by using heterogeneous information in the network. Wang et al. [20] try to predict social links by considering the moving pattern of users.

Cold start link prediction problem start to appear in recently years. Leroy et al. [11] propose a two-phase method based on the bootstrap probability graph to deal with the problem with auxiliary information. Ge et al. [6] solve similar problem with similar methods. But they assume that most of the links are missing while multiple heterogeneous information sources are available, which doesn't conform with real cold start problem in reality and they are called pseudo cold start problem as a result.

VI. CONCLUSION

In this paper, we study the link prediction problem for new users and propose a supervised method *SCAN-PS* to solve this problem by using information in multiple aligned heterogeneous social networks. A within-network personalized sampling method is proposed to address the differences in information distributions of new users and old users. Information from the aligned source network and that owned by the old users in the target network is transferred to help improve the prediction result. Extensive experiments results show that *SCAN-PS* works well for users of different degrees of novelty and can also solve the cold start problem.

REFERENCES

- [1] L. Adamic and E. Adar. Friends and neighbors on the web. *SOCIAL NETWORKS*, 25:211–230, 2001.
- [2] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [3] P. Cremonesi, A. Tripodi, and R. Turrin. Cross-domain recommender systems. In *ICDMW*, pages 496–503, 2011.
- [4] W. Dai, Y. Chen, G. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2009.
- [5] C. Aggarwal G. Qi and T. Huang. Link prediction across networks by biased cross-network sampling. In *ICDE*, pages 793–804, 2013.

- [6] L. Ge and A. Zhang. Pseudo cold start link prediction with multiple sources in social networks. In *SDM*, pages 768–779, 2012.
- [7] M. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM*, 2006.
- [8] K. Klemm and V. M. Eguíluz. Highly clustered scale-free networks. *Physical Review E*, 2002.
- [9] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, 2013.
- [10] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.
- [11] V. Leroy, B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *KDD*, pages 393–402, 2010.
- [12] B. Li, Q. Yang, and X. Xue. Can movies and books collaborate?: Cross-domain collaborative filtering for sparsity reduction. In *IJCAI*, pages 2052–2057, 2009.
- [13] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, pages 556–559, 2003.
- [14] Z. Lu, B. Savas, W. Tang, and I. Dhillon. Supervised link prediction using multiple sources. In *ICDM*, pages 923–928, 2010.
- [15] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, and Q. He. Transfer learning from multiple source domains via consensus regularization. In *CIKM*, pages 103–112, 2008.
- [16] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD*, pages 1046–1054, 2011.
- [17] E. H. Simpson. Measurement of diversity. In *Nature*, pages 688–688, 1949.
- [18] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *ASONAM*, pages 121–128, 2011.
- [19] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogeneous networks. In *WSDM*, pages 743–752, 2012.
- [20] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. Barabasi. Human mobility, social ties, and link prediction. In *KDD*, pages 1100–1108, 2011.
- [21] D. Xing, W. Dai, G. Xue, and Y. Yu. Bridged refinement for transfer learning. In *ECML-PKDD*, pages 324–335, 2007.
- [22] Z. Zhang, C. Liu, Y. Zhang, and T. Zhou. Solving the cold-start problem in recommender systems with social tags. *EPL (Europhysics Letters)*, 92(2), 2010.