2013 IEEE 13th International Conference on Data Mining

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*—Nowadsys, many new users are keeping joining in the online social networks every day and these new users usually have very few social connections and very sparse auxiliary information in the network. Prediction social links for new users is very important. Different from conventional link prediction problems, link prediction for new users is more challenging due to the lack of information from the new users in the network. Meanwhile, in recent years, users are usually involved in multiple social networks simultaneously to enjoy the specific services offered by different social networks. The shared users of multiple networks can act as the "anchors" aligned the networks they participate in. In this paper, we propose a link prediction method called SCAN-PS (Supervised Cross Aligned Networks link prediction with Personalized Sampling), to solve the social link prediction problem for new users. SCAN-PS can use information transferred from both the existing active users in the target network and other source networks through aligned accounts. In addition, 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 two real-world aligned heterogeneous social networks demonstrate that SCAN-PS can perform well in predicting social links for new users.**

***Index Terms*—link prediction; data mining;**

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 [5].

Many of previous works on link prediction treat all users in the network equally and focus on predicting potential links that will appear among all the users, based upon a snapshot of the social 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. 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 also been shown in previous works that the distribution of linkage formation probability follows a power-law decay with the age of nodes [3]. So, new users are more likely to accept the recommended links compared with old users and predicting links for new users could lead to more social connections.

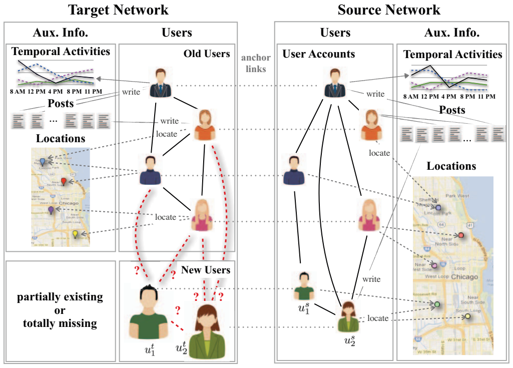

Fig. 1. Example of predicting social links across two aligned heterogeneous online social networks.

In this paper, we study the problem of predicting social links for new users, who have created their accounts for just a short time. The link prediction problem for new users is different from traditional link prediction problems which 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. 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

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.134

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. 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" [4], which could help align user' accounts across multiple social networks.

For example, in Figure 1, 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. 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 source 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). In this paper, we propose to exploit the new users' information in source networks to help improve the link prediction results in the target network.

The problem of social link prediction for new users by using aligned social networks has not been studied yet. We show a detailed comparison of this problem with many correlated link prediction problems in the full version of this paper [11]. However, in spite of its significance and novelty, 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 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.
3) *Aligned social networks*. Previous works on transfer learning focus on transferring knowledge between two domains via shared feature space [6], [8] or between two networks through shared triad linkage structures [7], [10]. 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. 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. 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. A full version of the paper is available in [11].

II. PROBLEM FORMULATION

The problem studied in this paper is social link prediction for new users based on aligned heterogeneous networks. In this section, we first define the concept of aligned heterogeneous networks and then present the formulation of the social link prediction for new users 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 $V_i, i \in \{1, 2, \cdots, |V|\}$ is the set of nodes of the i_{th} kind, $E = \bigcup_j E_j$ contains multiple types of links among the nodes, where $E_j, j \in \{1, 2, \cdots, |E|\}$ is the set of links of the j_{th} type.

Definition 2 (Aligned Heterogeneous Networks): Let $\mathcal{G} = (G_{set}, A_{set})$ be aligned heterogeneous social networks, where $G_{set} = \{G^1, G^2, \cdots, G^n\}$ is the set of heterogeneous social networks, whose size is $n = |G_{set}|$, and $A_{set} = \{A^{1,2}, A^{1,3}, \cdots, A^{1,n}, A^{2,1}, \cdots, A^{n,n-1}\}$ 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^i_m, u^j_n) is an *anchor link* between G^i and G^j iff. $(u^i_m \in U^i) \wedge (u^j_n \in U^j) \wedge (u^i_m$ and u^j_n are accounts owned by the same user).

Social Link Recommendation: 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 $U^t = U^t_{new} \cup U^t_{old}$ be the user set in G^t, where U^t_{new}, U^t_{old} are the sets of new users and old users and $U^t_{new} \cap U^t_{old} = \emptyset$. What we want to predict is a subset of potential social links between the new users and all other users: $L \subseteq U^t_{new} \times U^t$. In other words, we want to build a function $f : L \to \{0, 1\}$, which could decide whether certain links related to new users exist in the target network or not.

III. PROPOSED METHODS

New users possess little information in the target network. To solve this problem, we propose to transfer information from the old users in the target network. However, the information distribution of new users and old users can be totally different. In this section, we will analyze the problem of the differences in information distribution and propose a personalized within-network sampling method to process the old users' information to accommodate the difference.

A. Sampling Old Users' Information

Different from the link prediction with sampling problem studied in [2], 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 G^t.
- *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 $G^t = \{V^t, E^t\}$, where $U^t = U^t_{old} \cup U^t_{new} \subset V^t$ 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: $G^t_{old} = \{V^t_{old}, E^t_{old}\}$, in which each node is sampled independently with the sampling rate distribution vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_n)$, where $n = |U^t_{old}|$, $\sum_{i=1}^{n} \delta_i = 1$ and $\delta_i \geq 0$. Old users' heterogeneous sub-network after sampling is denoted as $\bar{G}^t_{old} = \{\bar{V}^t_{old}, \bar{E}^t_{old}\}$.

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 based on 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 $R(\bar{u}^t_{old}, G^t_{new})$ (derivation is available in [11]):

$$R(\bar{G}^t_{old}, G^t_{new}) = \mathbb{E}(R(\bar{u}^t_{old}, G^t_{new})) = \boldsymbol{\delta' s}$$

where $|U^t_{old}| = n$ and vector $\boldsymbol{s}$ equals:

$$\frac{1}{|U^t_{new}|}[\sum_{j=1}^{|U^t_{new}|} S(\bar{u}^t_{old,1}, u^t_{new,j}), \cdots, \sum_{j=1}^{|U^t_{new}|} S(\bar{u}^t_{old,n}, u^t_{new,j})]^T.$$

Besides the relevance, we also need to ensure that the diversity of information in the sampled old users' sub-network could be preserved. The diversity of auxiliary information is determined by the sampling rate δ_i, which could be define with the averaged *Simpson Index* [9] over the old users' sub-network, $D_{aux}(\bar{G}^t_{old}) = \frac{1}{|U^t_{old}|} \cdot \sum_{i=1}^{|U^t_{old}|} \delta_i^2$. 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 $D_{social}(\bar{G}^t_{old}) = \frac{1}{|S^t_{old}|} \cdot \sum_{i=1}^{|U^t_{old}|} \sum_{j=1}^{|U^t_{old}|} \delta_i \cdot \delta_j \times I(u_i, u_j)$, where $|S^t_{old}|$ is the size of social link set of old users' sub-network and $I(u_i, u_j)$ is an indicator function $I : (u_i, u_j) \to \{0, 1\}$ to show whether a certain social link exists or not originally before sampling.

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:

$$\begin{aligned} D(\bar{G}^t_{old}) &= \frac{1}{2}(D_{social}(\bar{G}^t_{old}) + D_{aux}(\bar{G}^t_{old})) \\ &= \boldsymbol{\delta'} \cdot (\frac{1}{2|S^t_{old}|} \cdot \mathbf{A^t_{old}} + \frac{1}{2\,|U^t_{old}|} \cdot \mathbf{I}_{|\mathbf{U^t_{old}}|}) \cdot \boldsymbol{\delta} \end{aligned}$$

where $\mathbf{I}_{|\mathbf{U^t_{old}}|} \in \mathcal{R}^{|U^t_{old}| \times |U^t_{old}|}$ is the diagonal identity matrix and $\mathbf{A^t_{old}}$ is the adjacency matrix of old users' sub-network.

To ensure that the structure of the old users' sub-network is not severely destroyed, we could add a regularization term to increase the sampling rate for these users as well as their neighbours by maximizing the following terms:

$$Reg(\bar{G}^t_{old}) = \min\{\mathcal{N}_i, \min_{u_j \in \mathcal{N}_i}\{\mathcal{N}_j\}\} \times \delta_i^2 = \boldsymbol{\delta'} \cdot \mathbf{M} \cdot \boldsymbol{\delta}$$

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.

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

$$D_{Reg}(\bar{G}^t_{old}) = D(\bar{G}^t_{old}) + Reg(\bar{G}^t_{old}) = \boldsymbol{\delta'} \cdot \mathbf{N} \cdot \boldsymbol{\delta}$$

where $\mathbf{N} = \frac{\mathbf{1}}{\mathbf{2|U^t_{old}|}} \cdot \mathbf{I}_{|\mathbf{U^t_{old}}|} + \frac{\mathbf{1}}{\mathbf{2|S^t_{old}|}} \cdot \mathbf{A^t_{old}} + \mathbf{M}$.

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

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

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

B. Cold-Start Link Prediction across Aligned Networks

In our problem settings, we have two aligned social networks. Link prediction methods porposed based on one single network can suffer from the cold start problem a lot. In this section, we will propose two methods to utilize the aligned source network to help solve the problem.

1) NAIVE*:* Suppose we have a new user u^t_i in the target network, a naive way to use the aligned source network to recommend social links for user u^t_i is to recommend all the corresponding social links related to this user's aligned account u^s_i in the aligned source network to him/her.

Definition 4 (Pseudo Label): The pseudo label of a link (u^t_i, u^t_j) in the target denotes the existence of its corresponding link (u^s_i, u^s_j) in the aligned source network and it is 1 if (u^s_i, u^s_j) exists and 0 otherwise.

Based on this intuition, we propose a cold start link prediction method NAIVE (Naive Link Prediction). NAIVE just use the pseudo labels as the final prediction results of links in the target network. 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 but ignores all other information.

2) SCAN-PS*:* To overcome all these disadvantages, a new method SCAN-PS (Supervised Cross Aligned Networks Link Prediction with Personalized Sampling) is proposed. SCAN-PS could use heterogeneous information existing in both the target network and the aligned source and it is built across two aligned social networks. By 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. A more detailed description about the extracted features is available in [11]. In addition to features described in [11], method SCAN-PS also utilizes the information used by NAIVE, which is the *pseudo label* defined before, by regarding it as another feature.

To use the information in multiple networks, 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 users' information in the target network.

TABLE I
PROPERTIES OF THE HETEROGENEOUS SOCIAL NETWORKS

		network	
	property	**Twitter**	**Foursquare**
# node	user	5,223	5,392
	tweet/tip	9,490,707	48,756
	location	297,182	38,921
# link	friend/follow	164,920	31,312
	write	9,490,707	48,756
	locate	615,515	48,756

IV. EXPERIMENTS

A. Data Preparation

The datasets used in this paper are crawled with the methods proposed in [4] and include two different heterogeneous online social networks: Foursquare and Twitter. As mentioned in the problem formulation section, users in both Foursquare and Twitter can follow other users, publish online posts, which can contain text information, timestamps and attached location check-ins.

A more detailed statistical information about these two social networks datasets is summarized in Table I. 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 [1] 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

TABLE II
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)

		Remaining Information Ratio							
measure	method	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
	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
Acc.	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
	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 III
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)

		Remaining Information Ratio							
measure	method	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
	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
Acc.	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
	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

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. Method, Scan (Supervised Cross Aligned Networks Link Prediction), is used as a baseline in the experiment, which is the same as Scan-ps except that Scan doesn't have the sampling part.

- *Target Network Only*: Some other supervised baseline methods are built only with the information in target network, which include New, Trad, Trad-ps, Old-ps and Old. A more detailed information about these baseline methods is available in [11].
- *Source Network Only*: To show that using two networks simultaneously is better than using one network. We also compare Scan-ps with another baseline method Source, built with the information in the source network only.
- *Unsupervised Methods*: Naive and some traditional unsupervised social link prediction methods are also used as the unsupervised baseline methods, which include *Common Neighbour* (CN), *Jaccard Coefficient* (JC) and *Adamic Adar* (AA). Naive uses social information in the aligned source network, while all other 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 $\boldsymbol{\delta}$ 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 II, 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. Compared with OLD and NEW, TRAD can perform even worse, 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, 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 that using a single one. SCAN-PS performs better that SCAN indicates that personalized sampling could still work when two aligned networks are used simultaneously.

The results also show 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 III, similarly results could be gotten when Twitter is used as the target network and Foursquare is used as the aligned source network.

V. 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.

VI. ACKNOWLEDGMENTS

This work is supported in part by NSF through grants CNS-1115234, DBI-0960443, and OISE-1129076, US Department of Army through grant W911NF-12-1-0066, and Huawei Grant.

REFERENCES

[1] 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.

[2] C. Aggarwal G. Qi and T. Huang. Link prediction across networks by biased cross-network sampling. In *ICDE*, pages 793–804, 2013.

[3] K. Klemm and V. M. Eguíluz. Highly clustered scale-free networks. Physical Review E, 2002.

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

[5] 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.

[6] 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.

[7] Z. Lu, B. Savas, W. Tang, and I. Dhillon. Supervised link prediction using multiple sources. In *ICDM*, pages 923–928, 2010.

[8] 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.

[9] E. H. Simpson. Measurement of diversity. Nature, pages 688–688, 1949.

[10] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM*, pages 743–752, 2012.

[11] J. Zhang, X. Kong, and P. Yu. Predicting social links for new users across aligned heterogeneous social networks. CORR arXiv:1310.3492, 2013.