Meta-path based Multi-Network Collective Link Prediction

Jiawei Zhang
Big Data and Social Computing (BDSC) Lab
University of Illinois at Chicago
Chicago, IL, USA
jzhan9@uic.edu

Philip S. Yu
Big Data and Social Computing (BDSC) Lab
University of Illinois at Chicago
Chicago, IL, USA
psyu@cs.uic.edu

Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology,
Nanjing University
Nanjing 210023, China
zhouzh@lamda.nju.edu.cn

ABSTRACT

Online social networks offering various services have become ubiquitous in our daily life. Meanwhile, users nowadays are usually involved in multiple online social networks simultaneously to enjoy specific services provided by different networks. Formally, social networks that share some common users are named as partially aligned networks. In this paper, we want to predict the formation of social links in multiple partially aligned social networks at the same time, which is formally defined as the multi-network link (formation) prediction problem. In multiple partially aligned social networks, users can be extensively correlated with each other by various connections. To categorize these diverse connections among users, 7 "intra-network social meta paths" and 4 categories of "inter-network social meta paths" are proposed in this paper. These "social meta paths" can cover a wide variety of connection information in the network, some of which can be helpful for solving the multi-network link prediction problem but some can be not. To utilize useful connection information, a subset of the most informative "social meta paths" are picked, the process of which is formally defined as "social meta path selection" in this paper. An effective general link formation prediction framework, MLI (Multi-network Link Identifier), is proposed in this paper to solve the multi-network link (formation) prediction problem. Built with heterogenous topological features extracted based on the selected "social meta paths" in the multiple partially aligned social networks, MLI can help refine and disambiguate the prediction results reciprocally in all aligned networks. Extensive experiments conducted on real-world partially aligned heterogeneous networks, Foursquare and Twitter, demonstrate that MLI can solve the multi-network link prediction problem very well.

Categories and Subject Descriptors

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

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623645.

Keywords

Link Prediction; Social Networks; Classification; Transfer Learning; Data Mining

1. INTRODUCTION

Nowadays, online social networks with specific characteristics have become an essential part in our daily life. Users in online social networks can usually enjoy a wide variety of social services, e.g., establish social connections with friends, write online posts, check in at locations. Meanwhile, social networks which provide these various services can usually contain heterogeneous information, which include multiple kinds of information entities, e.g., users, posts and locations, and complex links among these entities, e.g., social links among users [25, 26] and location checkin links between users and locations [26].

Via these diverse links in online social networks, users can be connected with each other closely. Consider, for example, given two users, *Alice* and *Bob*, who both have checked in at the "*Lincoln Memorial*" in Foursquare[1], the location checkin links "*Alice* – *Lincoln Memorial* – *Bob*" can form a path from *Alice* to *Bob* in the network. Formally, the sequences of links starting and ending with users in online social networks are defined as the *social paths*. The length of social paths is defined as the number of links that constitute them. For instance, path "*Alice* – *Lincoln Memorial* – *Bob*" is a social path of length 2 connecting *Alice* and *Bob* in Foursquare.

In all possible connections among users, social links, which are one kind of social paths of length 1 as well, among users have received lots of attention in recent years [21, 22, 23]. The problem of predicting social links to be formed in the near future based on a snapshot of online social networks is formally defined as the *social link (formation) prediction* problem. Many concrete social services in social networks can be cast as *social link prediction* problems, e.g., friend recommendation. Meanwhile, as pointed out in [25], users' "loyalty" to a social network is positively correlated to the number of friends they have in the network. As a result, *social link formation prediction* problem, which can help introduce more social connections for users, can be very important for online social networks.

Traditional link prediction problems which aim at predicting one single kind of links in one network [16, 20, 24, 3] have been studied for many years. Dozens of different link prediction methods have been proposed so far [1, 14, 19, 16, 20, 24, 3]. Conventional link prediction methods usually assume

[1]https://foursquare.com

that there exists sufficient information within the network to compute features (e.g., common neighborhoods [9]) for each pair of nodes. However, as proposed in [25, 26], such assumption can be violated seriously when dealing with social networks containing little information because of the "new network" problems.

The *new network problem* can be encountered when online social networks branch into new geographic areas or social groups [26] and information within the new networks can be too sparse to build effective link prediction models. Meanwhile, Zhang et al. [10, 25, 26] notice that users nowadays can participate in multiple online social networks simultaneously. Users who are involved in a new network can have been using other well-developed networks for a long time, in which they can have plenty of heterogeneous information. To address the new network problem, Zhang et al. [25, 26] propose to transfer information from the well-developed networks to overcome the shortage of information problem in the new network. Formally, networks that share some common users are defined as the "*partially aligned networks*" and the common users shared across these "*partially aligned networks*" are named as the "*anchor users*" [10, 25, 26]. In this paper, we define the unshared users as the "*non-anchor users*" between the *aligned networks*.

Social networks aligned by the "anchor users" can share common information. Meanwhile, as proposed in [15, 23], different online social networks constructed to provide different services usually have distinct characteristics. Moreover, information in various social networks may be of different distributions [15, 23], which is named as the "*network difference problem*" in this paper. The "*network difference problem*" will be a obstacle in link prediction across *multiple partially aligned networks*, as it is likely that information transferred from other aligned networks could deteriorate the prediction performance in a given network.

In this paper, we want to predict the formation of social links in multiple *partially aligned networks* simultaneously, which is formally defined as the *multi-network link prediction* problem in this paper. As introduced at the beginning of this section, the *multi-network link prediction* problem can have very extensive applications in real-world social networks. As a result, the *multi-network link prediction* problem studied in this paper is very important for *multiple partially aligned social networks*.

The *multi-network link prediction* problem studied in this paper is a novel problem and totally different from other existing link prediction problems. Moreover, link prediction methods proposed in [25, 26] cannot be applied to solve the *multi-network link prediction* problem directly because these existing methods: (1) are proposed to transfer useful information for *anchor users* only; (2) fail to consider the *network difference problem*; (3) can only predict links in each network independently. A more detailed comparison of the *multi-network link prediction* problem with these correlated problems, e.g., social link prediction for new users [25], transfer heterogeneous links across networks [26], anchor link prediction [10] and multi-transfer with multiple views and sources [18], is available in Table 1.

Despite of its importance and novelty, the *multi-network link prediction* problem studied in this paper is also very challenging to solve due to the following reasons:

- *lack of features*: Networks studied in this paper can contain different kinds of information. Proper definition of heterogeneous features extracted for social links from the networks is a prerequisite for addressing link formation prediction tasks.
- *partial alignment*: To overcome the "new network problem", we propose to transfer information from other aligned networks. Existing information transferring methods proposed in [25, 26] can only work well for *anchor users*. Method that can transferring information for both *anchor users* and *non-anchor users* is what we desire in this paper.
- *network difference problem*: Different networks usually have different characteristics and information transferred from other *aligned networks* can be different from that of the given network, which could deteriorate the link prediction performance in the given network.
- *simultaneous link prediction in multiple networks*: The *multi-network link prediction* problem covers multiple *link prediction* tasks in multiple *partially aligned networks* simultaneously. Analysis and utilization the correlations among these tasks to enhance the prediction performance in each network mutually is very challenging.

To solve all these above challenges in the *multi-network link prediction* problem, a novel link prediction framework, Mli, is proposed in this paper. Inspired by Sun's [17] work on meta path as a means to capture similarity of nodes, which are not directly connected in heterogeneous information networks, Mli explores the meta path concept to generate useful features. Mli can generate not only intra-network features via "intra-network meta paths", but also inter-network features via "inter-network meta paths" through the *anchor links*. By judiciously selecting the "inter-network meta paths", Mli can take advantage of the commonality among the *multiple partially aligned networks*, while contain the potential negative transfers from network differences. These derived features can greatly improve the effectiveness of Mli in predicting links for each network. Furthermore, Mli is a general link formation prediction framework that solves the *multi-network link prediction* problem and the *link prediction* tasks in different networks can help each other mutually.

The rest of this paper is organized as follows. In Section 2, we formulate the problem. Detailed description of the methods is available in Section 3. We show the experiment results in Section 4. Related works are given in Section 5. Finally, we conclude the paper in Section 6.

2. PROBLEM FORMULATION

In this section, we will give the formal definitions of many important concepts used in this paper and the formulation of the *multi-network link prediction* problem.

2.1 Terminology Definition

Definition 1 (Heterogeneous Social Network): A social network is *heterogeneous* if it contains multiple kinds of nodes and links. *Heterogeneous social networks* can be represented as $G = (V, E)$, where $V = \bigcup_i V_i$ is the union of different node sets and $E = \bigcup_i E_i$ is the union of heterogeneous link sets.

Networks used in this paper are Twitter and Foursquare, which are both heterogeneous social networks. Users in

Table 1: Summary of related problems.

Property	Mutual Social Link Prediction in Multiple Aligned Networks	Transfer Heterogeneous Links across Networks [26]	Multi-Transfer with Multi-View & Multi-Domain [18]	Social Link Prediction for New Users [25]	Inferring Anchor Links across Networks [10]
information sources	multiple **networks**	multiple **networks**	multiple **domains**	multiple **networks**	multiple **networks**
source type	heterogeneous	heterogeneous	multi-view	heterogeneous	heterogeneous
sources aligned?	partially aligned	partially aligned	no	fully aligned	fully aligned
source differences	solved	not solved	solved	not solved	not solved
predicted links	social links in all aligned networks	heterogeneous links in the target network	n/a	social links in the target network	anchor links across networks
settings	semi-supervised learning and transfer learning	supervised learning and transfer learning	supervised learning and transfer learning	supervised learning and transfer learning	supervised learning and transfer learning
knowledge to transfer	network structure via meta paths	network structure through anchor links	domain information via common feature space	network structure through anchor links	network structure through anchor links

both Twitter and Foursquare can make friends with other users, write posts online, which can contain text content, timestamps and attach location check-ins. Both Twitter and Foursquare can be formulated as $G = (V, E)$, where $V = U \quad P \quad L \quad T \quad W$, and U, P, L, T and W are the sets of user, post, location, timestamp and word nodes in the network respectively, while $E = E_{u,u} \quad E_{u,p} \quad E_{p,l} \quad E_{p,t} \quad E_{p,w}$ are the sets of heterogeneous links in G, which include the *social links* among users, *write link* between users and posts, links between posts and locations, timestamps and words.

Definition 2 (Aligned Heterogeneous Social Networks): If two different social networks share some common users, then these two networks are called *aligned networks*. *Multiple aligned heterogeneous social networks* can be formulated as $\mathcal{G} = ((G^1, G^2, \cdots, G^n), (A^{1,2}, A^{1,3}, \cdots, A^{1,n}, A^{2,3}, \cdots, A^{(n-1),n}))$, where $G^i, i \quad \{1, 2, \cdots, n\}$ is a *heterogeneous social network* and $A^{i,j} = \ , i, j \quad \{1, 2, \cdots, n\}$ is the set of undirected *anchor links* between G^i and G^j.

Definition 3 (Anchor Links): Let U^i and U^j be the user sets of G^i and G^j respectively. Link (u^i, v^j) is a undirected *anchor link* between G^i and G^j iff($u^i \quad U^i$) $(v^j \quad U^j)$ (u^i and v^j are the accounts of the same user in G^i and G^j respectively).

Definition 4 (Anchor Users): User $u^i \quad U^i$ is an *anchor user* in G^i between G^i and G^j iff $v^j \quad U^j, (u^i, v^j) \quad A^{i,j}$. The set of *anchor users* in G^i between G^i and G^j can be represented as $U^i_{A^{i,j}} = \{u^i | u^i \quad U^i, \ v^j \quad U^j, (u^i, v^j) \quad A^{i,j}\}$.

Definition 5 (Non-Anchor Users): User $u^i \quad U^i$ is an *non-anchor user* between G^i and G^j iff $u^i \ / \ U^i_{A^{i,j}}$. The set of *non-anchor users* in G^i between G^i and G^j can be represented as $U^i_{\_A^{i,j}} = U^i - U^i_{A^{i,j}}$.

Definition 6 (Full Alignment): Networks G^i and G^j are *fully aligned* if users in both G^i and G^j are all anchor users. In other words, G^i and G^j are *fully aligned* iff($U^i_{A^{i,j}} = U^i$) $(U^j_{A^{i,j}} = U^j)$.

Definition 7 (Partial Alignment): Networks G^i and G^j are *partially aligned* if there exist users in G^i or G^j who are *non-anchor users*. In other words, G^i and G^j are partially aligned iff $((U^i_{\_A^{i,j}} = \ ) \quad (U^j_{\_A^{i,j}} = \ )) \quad (A^{i,j} = \ )$.

Considering that *fully aligned networks* can hardly exist in the real world, different from the strict full alignment assumption of networks proposed in [10, 25], networks used in this paper are *partially aligned* instead. In addition, there exists no restriction about the constraint on *anchor links*, which means that the *anchor links* can be either *one-to-one* [10] or *many-to-many*.

2.2 Multi-PU Link Prediction

Let $G^i, i \quad \{1, 2, \cdots, n\}$ be a *heterogeneous online social network* in the multiple *aligned networks*. The user set and existing social link set of G^i can be represented as U^i and $E^i_{u,u}$ respectively. In network G^i, all the existing links are the formed links and, as a result, the formed links of G^i can be represented as $\mathcal{P}^i$, where $\mathcal{P}^i = E^i_{u,u}$. Furthermore, a large set of unconnected user pairs are referred to as the unconnected links, $\mathcal{U}^i$, and can be extracted from network G^i: $\mathcal{U}^i = U^i \times U^i - \mathcal{P}^i$. However, no information about links that will never be formed can be obtained from the network. In this paper, with $\mathcal{P}^i$ and $\mathcal{U}^i$, we formulate the *link formation prediction* as a *PU link prediction* problem.

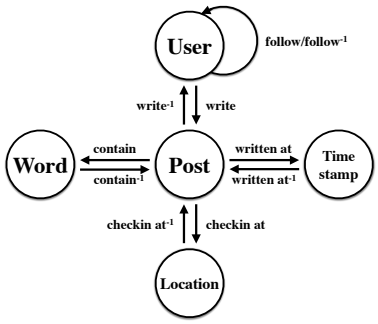


Figure 1: Schema of heterogeneous network.

Similarly, let $\{\mathcal{P}^1, \cdots, \mathcal{P}^n\}$, $\{\mathcal{U}^1, \cdots, \mathcal{U}^n\}$ and $\{\mathcal{L}^1, \cdots, \mathcal{L}^n\}$ be the sets of formed links, unconnected links, and links to be predicted of $G^1, G^2, \cdots, G^n$ respectively. With the formed and unconnected links of $G^1, G^2, \cdots, G^n$, we can solve the *multi-network link prediction* problem as a *multi-PU link prediction* problem.

3. PROPOSED METHODS

In this section, we will introduce Mli to solve the *multi-network link prediction* problem. This section includes 3 parts: (1) social meta path based feature extraction and selection; (2) PU link prediction; (3) multi-network link prediction framework.

3.1 Social Meta Path Definition and Selection

Before talking about the link prediction methods, we will introduce the features extracted from the *partially aligned networks* in this subsection at first.

3.1.1 Intra-Network Social Meta Path

Users in heterogeneous online social network can be extensively connected to each other via different paths. In this part, we will categorize the diverse paths connecting users in one single network with the *intra-network social meta paths* concept.

For a given heterogeneous online social network, e.g., G, to describe its structure more clearly, we define its *schema* to be $S_G = (T, R)$, where T, R are the sets of node types and link types in G. For example, if $G = (V, E)$, where $V = U \quad P \quad L$ contains user, post and location nodes, $E = E_{u,u} \quad E_{u,p} \quad E_{p,l}$ contains the social links, write links and location links, then $S_G = (T, R)$, $T = \{$User, Post, Location$\}$ and $R = \{$Social Link, Write Link, Location Link$\}$. A complete schema of networks studied in this paper is shown in Figure 1. In network G, nodes can be connected with each other via extensive paths consisting of various links. To categorize all possible paths in heterogeneous networks G, we define the concept of *intra-network meta path* based on schema S_G as follows:

Definition 8 (Intra-Network Meta Path): Based on the given the network schema, $S_G = (T, R)$, $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ is defined to be a *meta path* in network G, where $T_i \quad T, i \quad \{1, 2, \cdots, k\}$ and $R_i \quad R, i \quad \{1, 2, \cdots, k-1\}$.

Meanwhile, depending on types of nodes and links that constitute it, Φ can be divided into two different categories.

Definition 9 (Homogeneous and Heterogeneous Intra-Network Meta Path): For a given meta path $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ defined based on S_G, if ($T_1, \cdots, T_k$ are all the same) ($R_1, \cdots, R_{k-1}$ are all the same), then Φ is a *homogeneous meta path*; otherwise P is a *heterogeneous meta path*.

In this paper, we are mainly concerned about meta paths connecting user nodes, which can be defined as the *intra-network social meta path*.

Definition 10 (Intra-Network Social Meta Path): For a given meta path $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ defined based on S_G, if T_1 and T_k are both the "User" node type, then P is defined as a *social meta path*. Depending on whether $T_1, \cdots, T_k$ and $R_1, \cdots, R_{k-1}$ are the same or not, P can be divided into two categories: *homogeneous intra-network social meta path* and *heterogeneous intra-network social meta path*.

Based on the schema of networks studied in this paper, shown in Figure 1, we can define many different kinds of *homogeneous and heterogeneous intra-network social meta paths* for network G, whose physical meanings and notations are listed as follows:

Homogeneous Intra-Network Social Meta Path

- *ID 0. Follow*: User $\xrightarrow{follow}$ User, whose notation is "$U \rightarrow U$" or $\Phi_0(U, U)$.
- *ID 1. Follower of Follower*: User $\xrightarrow{follow}$ User $\xrightarrow{follow}$ User, whose notation is "$U \rightarrow U \rightarrow U$" or $\Phi_1(U, U)$.
- *ID 2. Common Out Neighbor*: User $\xrightarrow{follow}$ User $\xrightarrow{follow^{-1}}$ User, whose notation is "$U \rightarrow U \leftarrow U$" or $\Phi_2(U, U)$.
- *ID 3. Common In Neighbor*: User $\xrightarrow{follow^{-1}}$ User $\xrightarrow{follow}$ User, whose notation is "$U \leftarrow U \rightarrow U$" or $\Phi_3(U, U)$.

Heterogeneous Intra-Network Social Meta Path

- *ID 4. Common Words*: User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Word $\xrightarrow{contain^{-1}}$ Post $\xrightarrow{write^{-1}}$ User, whose notation is "$U \rightarrow P \rightarrow W \leftarrow P \leftarrow U$" or $\Phi_4(U, U)$.
- *ID 5. Common Timestamps*: User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Time $\xrightarrow{contain^{-1}}$ Post $\xrightarrow{write^{-1}}$ User, whose notation is "$U \rightarrow P \rightarrow T \leftarrow P \leftarrow U$" or $\Phi_5(U, U)$.
- *ID 6. Common Location Checkins*: User $\xrightarrow{write}$ Post $\xrightarrow{attach}$ Location $\xrightarrow{attach^{-1}}$ Post $\xrightarrow{write^{-1}}$ User, whose notation is "$U \rightarrow P \rightarrow L \leftarrow P \leftarrow U$" or $\Phi_6(U, U)$.

3.1.2 Social Meta Path based Features

Meta paths introduced in the previous part can actually cover a large number of path instances connecting users in the network. Formally, we denote that node n (or link l) is an instance of node type T (or link type R) in the network as $n \quad T$ (or $l \quad R$). Identity function $I(a, A) = \begin{cases} 1, & \text{if } a \quad A \\ 0, & otherwise, \end{cases}$ can check whether node/link a is an instance of node/link type A in the network. To consider the effect of the unconnected links when extracting features for social links in the network, we formally define the *Intra-Network Social Meta Path based Features* to be:

Definition 11 (Intra-Network Social Meta Path based Features): For a given link (u, v), the feature extracted for it based on meta path $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ from the network is defined to be the expected number of formed path instances between u and v in the network:

$$x(u, v) = I(u, T_1) I(v, T_k) \sum_{n_1 \quad \{u\}, n_2 \quad T_2, \cdots, n_k \quad \{v\}} \prod_{i=1}^{k-1} p(n_i, n_{i+1}) I((n_i, n_{i+1}), R_i),$$

where $p(n_i, n_{i+1}) = 1.0$ if $(n_i, n_{i+1}) \quad E_{u,u}$ and otherwise, $p(n_i, n_{i+1})$ denotes the *formation probability* of link (n_i, n_{i+1}) to be introduced in Subsection 3.2.

Features extracted based on $\Phi = \{\Phi_1, \cdots, \Phi_6\}$ are named as the *intra-network social meta path* based social features. (Φ_0 will be used in Subsection 3.1.4 only.)

3.1.3 Anchor Meta Path

When a network is very new, features extracted based on *intra-network social meta paths* can be very sparse, as there exist few connections in the network. Consider, for example, in Figure 2, we want to predict whether social link (A^1, B^1) in network G^1 will be formed or not. Merely based on the *intra-network social meta paths*, the feature vector of extracted for link (A^1, B^1) will be $\mathbf{0}$. However, we find that A^1 and B^1 can be correlated actually with various inter-network paths, e.g., $B^1 \rightarrow B^2 \rightarrow A^2 \rightarrow A^1$, $B^1 \rightarrow B^2 \rightarrow F^2 \rightarrow A^2 \rightarrow A^1$ and $B^1 \rightarrow B^2 \rightarrow E^2 \rightarrow A^2 \rightarrow A^1$.

By following this idea, we propose to transfer useful information from aligned networks with the following *anchor meta path* and the *inter-network social meta paths* to be introduced in Subsection 3.1.4.

Definition 12 (Anchor Meta Path): Let U^i, U^j be the user nodes of G^i and G^j respectively and $A^{i,j}$ be the anchor links between G^i and G^j. Meta path $\Upsilon = T_1 \xrightarrow{R_1} T_2$ is an *anchor meta path* between network G^i and G^j iff $T_1 = U^i$ and $T_2 = U^j$ and $R_1 = A^{i,j}$. The notation of *anchor meta*

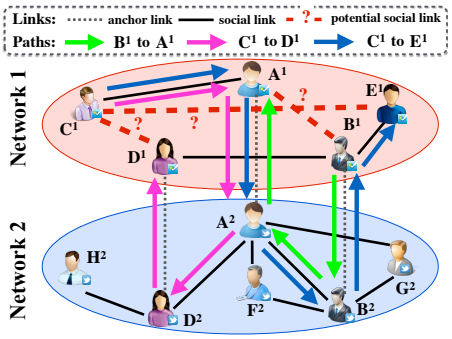


Figure 2: Meta Path across Aligned Networks.

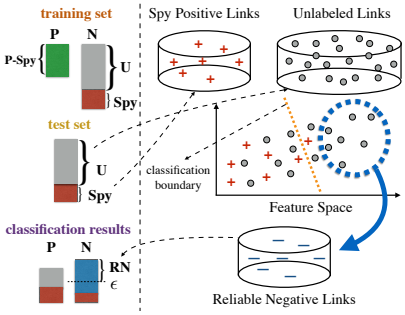


Figure 3: PU Link Prediction.

path from G^i to G^j is $\Upsilon(U^i, U^j)$ and the length of $\Upsilon(U^i, U^j)$ is 1.

3.1.4 Inter-Network Social Meta Paths

Based on the definition of *anchor meta path*, we can describe the inter-network paths from B^1 to A^1 in Figure 2 with the following *inter-network meta path*.

Definition 13 (Inter-Network Meta Path): Meta path $\Psi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ is an *inter-network meta path* across G^i and G^j iff $m \in \{1, 2, \cdots, k-1\}$, $T_m \xleftrightarrow{R_m} T_{m+1} = \Upsilon(U^i, U^j)$.

In this paper, we are concerned about *inter-network meta path* starting and ending with users, which are named as the *inter-network social meta path*. The 4 specific *inter-network social meta paths* used in this paper include:

Category 1: $\Upsilon(U^i, U^j) \circ (\Phi(U^j, U^j) \oplus \Phi_0(U^j, U^j)) \circ \Upsilon(U^j, U^i)$, whose notation is $\Psi_1(U^i, U^i)$;

Category 2.: $(\Phi(U^i, U^i) \oplus \Phi_0(U^i, U^i)) \circ \Upsilon(U^i, U^j) \circ (\Phi(U^j, U^j) \oplus \Phi_0(U^j, U^j)) \circ \Upsilon(U^j, U^i)$, whose notation is $\Psi_2(U^i, U^i)$;

Category 3.: $\Upsilon(U^i, U^j) \circ (\Phi(U^j, U^j) \oplus \Phi_0(U^j, U^j)) \circ \Upsilon(U^j, U^i) \circ (\Phi(U^i, U^i) \oplus \Phi_0(U^i, U^i))$, whose notation is $\Psi_3(U^i, U^i)$;

Category 4.: $(\Phi(U^i, U^i) \oplus \Phi_0(U^i, U^i)) \circ \Upsilon(U^i, U^j) \circ (\Phi(U^j, U^j) \oplus \Phi_0(U^j, U^j)) \circ \Upsilon(U^j, U^i) \circ (\Phi(U^i, U^i) \oplus \Phi_0(U^i, U^i))$, whose notation is $\Psi_4(U^i, U^i)$;

where $\Phi(U^i, U^i) \oplus \Phi_0(U^i, U^i) = \{\Phi_0(U^i, U^i), \cdots, \Phi_6(U^i, U^i)\}$ denote the 7 *intra-network social meta paths* of network G^i introduced in Subsection 3.1.1.

Let $\Psi = \{\Psi_1, \Psi_2, \Psi_3, \Psi_4\}$. Ψ is a comprehensive *inter-network social meta path* set and features extracted based on Ψ can transfer information for both anchor users and non-anchor users from other aligned networks. For example, in Figure 2, by following path "$B^1 \rightarrow B^2 \rightarrow A^2 \rightarrow A^1$", we can go from *anchor user* B^1 to *anchor user* A^1 and such path is an instance of $\Psi_1(U^1, U^1)$; by following path $C^1 \rightarrow A^1 \rightarrow A^2 \rightarrow D^2 \rightarrow D^1$, we can go from *non-anchor user* C^1 to *anchor user* D^1, which is an instance of $\Psi_2(U^1, U^1)$; in addition, by following path $C^1 \rightarrow A^1 \rightarrow A^2 \rightarrow B^2 \rightarrow B^1 \rightarrow E^1$, we can go from *non-anchor user* C^1 to *non-anchor user* E^1, which is an instance of $\Psi_4(U^1, U^1)$.

3.1.5 Social Meta Path Selection

As introduced in Section 1, information transferred from aligned networks is helpful for improving link prediction performance in a given network but can be misleading as well, which is called the *network difference problem*. To solve the *network difference problem*, we propose to rank and select top K features from the feature vector extracted based on the *intra-network* and *inter-network social meta paths*, $[\mathbf{x}_\Phi^T, \mathbf{x}_\Psi^T]^T$, from the multiple *partially aligned heterogeneous networks*.

Let variable X_i $[\mathbf{x}_\Phi^T, \mathbf{x}_\Psi^T]^T$ be a feature extracted based on a meta path in $\{\Phi, \Psi\}$ and variable Y be the *label*. $P(Y = y)$ denotes the *prior probability* that links in the training set having label y and $P(X_i = x)$ represents the *frequency* that feature X_i has value x. Information theory related measure *mutual information* (mi) is used as the ranking criteria:

$$mi(X_i) = \sum_x \sum_y P(X_i = x, Y = y) \log \frac{P(X_i = x, Y = y)}{P(X_i = x)P(Y = y)}$$

Let $[\bar{\mathbf{x}}_\Phi^T, \bar{\mathbf{x}}_\Psi^T]^T$ be the features of the top K *mi* score selected from $[\mathbf{x}_\Phi^T, \mathbf{x}_\Psi^T]^T$. In the next subsection, we will use the selected feature vector $[\bar{\mathbf{x}}_\Phi^T, \bar{\mathbf{x}}_\Psi^T]^T$ to build a novel PU link prediction model.

3.2 PU Link Prediction

In this subsection, we will propose a method to solve the *PU link prediction* problem in one single network.

As introduced in Section 2, from a given network, e.g., G, we can get two disjoint sets of links: connected (i.e., formed) links $\mathcal{P}$ and unconnected links $\mathcal{U}$. To differentiate these links, we define a new concept "*connection state*", z, in this paper to show whether a link is connected (i.e., formed) or unconnected in network G. For a given link l, if l is connected in the network, then $z(l) = +1$; otherwise, $z(l) = -1$. As a result, we can have the "*connection states*" of links in $\mathcal{P}$ and $\mathcal{U}$ to be: $z(\mathcal{P}) = +\mathbf{1}$ and $z(\mathcal{U}) = -\mathbf{1}$.

Besides the "*connection state*", links in the network can also have their own "*labels*", y, which can represent whether a link is to be formed or will never be formed in the network. For a given link l, if l has been formed or to be formed, then $y(l) = +1$; otherwise, $y(l) = -1$. Similarly, we can have the "*labels*" of links in $\mathcal{P}$ and $\mathcal{U}$ to be: $y(\mathcal{P}) = +\mathbf{1}$ but $y(\mathcal{U})$ can be either $+1$ or -1, as $\mathcal{U}$ can contain both links to be formed and links that will never be formed.

By using $\mathcal{P}$ and $\mathcal{U}$ as the positive and negative training sets, we can build a *link connection prediction model* $\mathcal{M}_c$, which can be applied to predict whether a link exists in the original network, i.e., the *connection state* of a link. Let l be a link to be predicted, by applying $\mathcal{M}_c$ to classify l, we can get the *connection probability* of l to be:

Definition 14: (Connection Probability): The probability that link l's *connection states* is predicted to be *connected* (i.e., $z(l) = +1$) is formally defined as the *connection probability* of link l: $p(z(l) = +1|\mathbf{x}(l))$, where $\mathbf{x}(l) = [\bar{\mathbf{x}}_\Phi(l)^T, \bar{\mathbf{x}}_\Psi(l)^T]^T$.

Meanwhile, if we can obtain a set of links that "will never be formed", i.e., "-1" links, from the network, which together with $\mathcal{P}$ ("+1" links) can be used to build a *link formation prediction model*, $\mathcal{M}_f$, which can be used to get the *formation probability* of l to be:

Definition 15: (Formation Probability): The probability that link l's *label* is predicted to be *formed or will be formed* (i.e., $y(l) = +1$) is formally defined as the *formation probability* of link l: $p(y(l) = +1|\mathbf{x}(l))$.

However, from the network, we have no information about "links that will never be formed" (i.e., "-1" links). As a result, the *formation probabilities* of potential links that we aim to obtain as proposed in Section 2 can be very challenging to calculate. Meanwhile, the correlation between link l's *connection probability* and *formation probability* has been proved in existing works [5] to be:

$$p(y(l) = +1|\mathbf{x}(l)) \quad p(z(l) = +1|\mathbf{x}(l)).$$

In other words, for links whose *connection probabilities* are low, their *formation probabilities* will be relatively low as well. This rule can be utilized to extract links which can be more likely to be the reliable "-1" links from the network. We propose to apply the the *link connection prediction model* $\mathcal{M}_c$ built with $\mathcal{P}$ and $\mathcal{U}$ to classify links in $\mathcal{U}$ to extract the *reliable negative link set*.

Definition 16: (Reliable Negative Link Set): The *reliable negative links* in the *unconnected link* set $\mathcal{U}$ are those whose *connection probabilities* predicted by the *link connection prediction model*, $\mathcal{M}_c$, are lower than threshold ϵ $[0, 1]$:

$$\mathcal{RN} = \{l|l \quad \mathcal{U}, p(z(l) = +1|\mathbf{x}(l)) < \epsilon\}.$$

Some Heuristic methods have been proposed to set the optimal threshold ϵ, e.g., the *spy technique* proposed in [13]. As shown in Figure 3, we randomly selected a subset of links in $\mathcal{P}$ as the spy, $\mathcal{SP}$, whose proportion is controlled by $s\%$. $s\% = 15\%$ is used as the default sample rate in this paper. Sets $(\mathcal{P} - \mathcal{SP})$ and $(\mathcal{U} \quad \mathcal{SP})$ are used as positive and negative training sets to the *spy prediction* model, $\mathcal{M}_s$. By applying $\mathcal{M}_s$ to classify links in $(\mathcal{U} \quad \mathcal{SP})$, we can get their *connection probabilities* to be:

$$p(z(l) = +1|\mathbf{x}(l)), l \quad (\mathcal{U} \quad \mathcal{SP}),$$

and parameter ϵ is set as the minimal *connection probability* of spy links in $\mathcal{SP}$:

$$\epsilon = \min_{l \quad \mathcal{SP}} p(z(l) = +1|\mathbf{x}(l)).$$

With the extracted *reliable negative link set* $\mathcal{RN}$, we can solve the *PU link prediction* problem with *classification based link prediction methods*, where $\mathcal{P}$ and $\mathcal{RN}$ are used as the positive and negative training sets respectively. Meanwhile, when applying the built model to predict links in $\mathcal{L}^i$, the optimal labels, $\hat{\mathcal{Y}}^i$, of $\mathcal{L}^i$, should be those which can maximize the following *formation probabilities*:

$$\hat{\mathcal{Y}}^i = \arg\max_{\mathcal{Y}^i} p(y(\mathcal{L}^i) = \mathcal{Y}^i|G^1, G^2, \cdots, G^k)$$
$$= \arg\max_{\mathcal{Y}^i} p(y(\mathcal{L}^i) = \mathcal{Y}^i| \left[\bar{\mathbf{x}}_\Phi(\mathcal{L}^i)^T, \bar{\mathbf{x}}_\Psi(\mathcal{L}^i)^T\right]^T)$$

where $y(\mathcal{L}^i) = \mathcal{Y}^i$ represents that links in $\mathcal{L}^i$ have labels $\mathcal{Y}^i$.

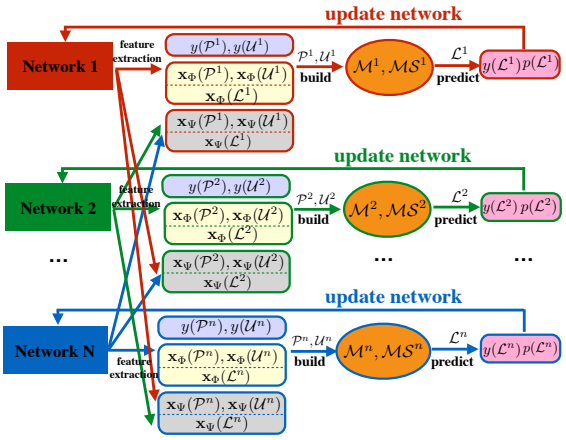


Figure 4: Multi-PU Link Prediction Framework.

3.3 Multi-Network Link Prediction Framework

Method MLI proposed in this paper is a general link prediction framework and can be applied to predict social links in n *partially aligned networks* simultaneously. When it comes to n partially aligned network formulated in Section 2, the optimal labels of potential links $\{\mathcal{L}^1, \mathcal{L}^2, \cdots, \mathcal{L}^n\}$ of networks $G^1, G^2, \cdots, G^n$ will be:

$$\hat{\mathcal{Y}}^1, \hat{\mathcal{Y}}^2, \cdots, \hat{\mathcal{Y}}^n = \arg\max_{\mathcal{Y}^1, \mathcal{Y}^2, \cdots, \mathcal{Y}^n} p(y(\mathcal{L}^1) = \mathcal{Y}^1, y(\mathcal{L}^2) = \mathcal{Y}^2, \cdots, y(\mathcal{L}^n) = \mathcal{Y}^k|G^1, G^2, \cdots, G^n)$$

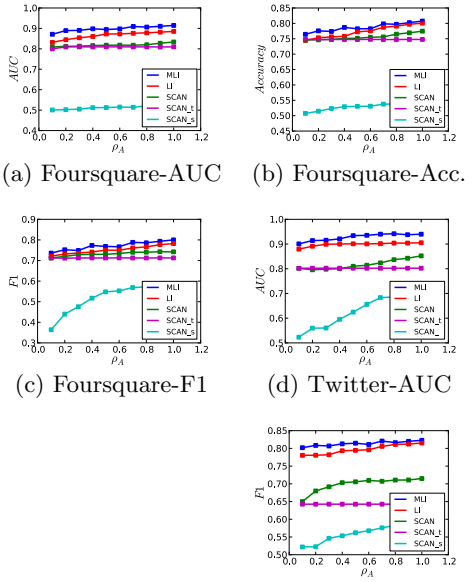
The above target function is very complex to solve and, in this paper, we propose to obtain the solution by updating one variable, e.g., $\mathcal{Y}^1$, and fix other variables, e.g., $\mathcal{Y}^2, \cdots, \mathcal{Y}^n$, alternatively with the following equation [26]:

$$\begin{cases} (\hat{\mathcal{Y}}^1)^{(\tau)} & = \arg\max_{\mathcal{Y}^1} p(y(\mathcal{L}^1) = \mathcal{Y}^1|G^1, G^2, \cdots, G^n, (\hat{\mathcal{Y}}^2)^{(\tau-1)}, (\hat{\mathcal{Y}}^3)^{(\tau-1)}, \cdots, (\hat{\mathcal{Y}}^n)^{(\tau-1)}) \\ (\hat{\mathcal{Y}}^2)^{(\tau)} & = \arg\max_{\mathcal{Y}^2} p(y(\mathcal{L}^2) = \mathcal{Y}^2|G^1, G^2, \cdots, G^n, (\hat{\mathcal{Y}}^1)^{(\tau)}, (\hat{\mathcal{Y}}^3)^{(\tau-1)}, \cdots, (\hat{\mathcal{Y}}^n)^{(\tau-1)}) \\ & \cdots\cdots \\ (\hat{\mathcal{Y}}^n)^{(\tau)} & = \arg\max_{\mathcal{Y}^n} p(y(\mathcal{L}^n) = \mathcal{Y}^n|G^1, G^2, \cdots, G^n, (\hat{\mathcal{Y}}^1)^{(\tau)}, (\hat{\mathcal{Y}}^2)^{(\tau)}, \cdots, (\hat{\mathcal{Y}}^{(n-1)})^{(\tau)}) \end{cases}$$

The structure of framework MLI is shown in Figure 4. When predicting social links in network G^i, we can extract features based on the *intra-network social meta path*, $\mathbf{x}_\Phi$, extracted from G^i and those extracted based on the *inter-network social meta path*, $\mathbf{x}_\Psi$, across G^1, G^2, $\cdots$, G^{i-1}, G^{i+1}, $\cdots$, G^n for links in $\mathcal{P}^i$, $\mathcal{U}^i$ and $\mathcal{L}^i$. Feature vectors $\mathbf{x}_\Phi(\mathcal{P})$, $\mathbf{x}_\Phi(\mathcal{U})$ and $\mathbf{x}_\Psi(\mathcal{P})$, $\mathbf{x}_\Psi(\mathcal{U})$ as well as the labels, $y(\mathcal{P})$, $y(\mathcal{U})$, of links in $\mathcal{P}$ and $\mathcal{U}$ are passed to the PU link prediction model $\mathcal{M}^i$ and the meta path selection model $\mathcal{MS}^i$. The formation probabilities of links in $\mathcal{L}^i$ predicted by model $\mathcal{M}^i$ will be used to update the network by replace the weights of $\mathcal{L}^i$ with the newly predicted formation probabilities. The initial weights of these potential links in $\mathcal{L}^i$ are set as 0 (i.e., the *formation probability* of links mentioned in Definition 11). After finishing these steps on G^i, we will move to conduct similar operations on G^{i+1}. We iteratively predict links in G^1 to G^n alternatively in a sequence until the results in all of these networks converge.

Table 2: Properties of the Heterogeneous Networks

	property	network: Twitter	network: Foursquare
# node	user	5,223	5,392
	tweet/tip	9,490,707	48,756
	location	297,182	38,921
# link	friend/follow	164,920	76,972
	write	9,490,707	48,756
	locate	615,515	48,756



(a) Foursquare-AUC (b) Foursquare-Acc.

(c) Foursquare-F1 (d) Twitter-AUC

(e) Twitter-Acc. (f) Twitter-F1

Figure 5: Effects of anchor link ratio ρ_A on prediction results in different networks evaluated by different metrics.

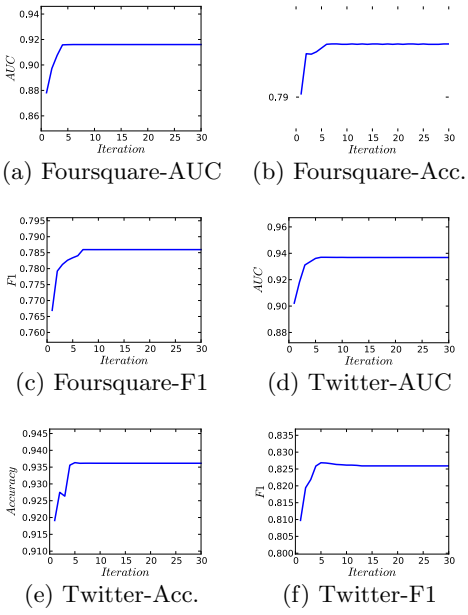
4. EXPERIMENTS

To demonstrate the effectiveness of MLI in dealing with real-world multiple *partially aligned heterogeneous networks*, we will conduct extensive experiments in this section. This section includes 3 parts: (1) dataset description; (2) experiment setting; (3) experiment results.

4.1 Datasets

The datasets used in this paper are Foursquare and Twitter, both of which are famous heterogeneous online social networks. These two networks were crawled during November of 2012 [10, 25, 26]. The structures of both Foursquare and Twitter have been introduced in Section 2. Statistical information about these two datasets is available in Table 2:

- **Foursquare**: 5, 392 users, 48, 756 tips and 38, 921 locations are crawled from Foursquare. The social links among the crawled users is 76, 972 and each user has about 14 friends in Foursquare.
- **Twitter**: 5, 223 users together with their tweets are crawled from Twitter, whose number is 9, 490, 707. Among these 5, 223 users, there exist 164, 920 follow links. Among all these tweets, about 615, 515 have location checkins, accounting for about 6.48% of all the tweets.



(a) Foursquare-AUC (b) Foursquare-Acc.

(c) Foursquare-F1 (d) Twitter-AUC

(e) Twitter-Acc. (f) Twitter-F1

Figure 6: Convergence analysis in different networks under the evaluation of different metrics.

4.2 Experiment Setting

4.2.1 *Comparison Methods*

To show the advantages of MLI, we compare MLI with many other baseline methods, which include:

- MLI: Method MLI is the multi-network link prediction framework proposed in this paper, which can predict social links in multiple online social networks simultaneously. The features used by MLI are extracted based on the meta paths selected from Φ and Ψ across aligned networks.
- LI: Method LI (Link Identifier) is identical to MLI except that LI predict the formation of social links in each network independently.
- SCAN: Method SCAN (Cross Aligned Network link prediction) proposed in [25, 26] is similar to MLI except that (1) SCAN predicts social links in each network independently; (2) features used by SCAN are those extracted based on meta paths Φ andΨ_1 without meta path selection.
- SCAN-S: Method SCAN-S (SCAN with Source Network) proposed in [25, 26] is identical to SCAN except that the features used by SCAN-S are those extracted based onΨ_1 without meta path selection.
- SCAN-T: Method SCAN-T (SCAN with Target Network) proposed in [25, 26]) is identical to SCAN except that the features used by SCAN-S are those extracted based on Φ without meta path selection.

4.2.2 *Evaluation Metrics*

The social links in both Foursquare and Twitter are used as the ground truth to evaluate the prediction results. SVM

Table 3: Performance comparison of different methods for inferring social and location links for Foursquare of different remaining information rates. The anchor link sample rate ρ_A is set as 1.0.

network	measure	methods	Remaining information rates ρ_F of Foursquare. 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Foursquare	AUC	MLI	**0.677±0.023**	**0.776±0.011**	**0.844±0.008**	**0.887±0.005**	**0.906±0.003**	**0.912±0.005**	**0.912±0.003**	**0.916±0.004**
		LI	0.573±0.019	0.68±0.023	0.806±0.01	0.853±0.004	0.866±0.003	0.874±0.007	0.881±0.003	0.878±0.005
		SCAN	0.549±0.009	0.56±0.009	0.662±0.03	0.745±0.009	0.786±0.014	0.804±0.01	0.812±0.005	0.82±0.004
		SCAN-T	0.5±0.083	0.503±0.007	0.613±0.012	0.739±0.008	0.764±0.013	0.787±0.007	0.8±0.006	0.81±0.007
		SCAN-S	0.524±0.013	0.524±0.017	0.524±0.012	0.524±0.005	0.524±0.002	0.524±0.01	0.524±0.003	0.524±0.005
	Accuracy	MLI	**0.632±0.01**	**0.692±0.007**	**0.755±0.005**	**0.769±0.004**	**0.779±0.002**	**0.798±0.006**	**0.799±0.004**	**0.797±0.005**
		LI	0.568±0.013	0.624±0.053	0.699±0.004	0.722±0.006	0.761±0.01	0.782±0.01	0.789±0.005	0.791±0.006
		SCAN	0.558±0.007	0.6±0.006	0.683±0.071	0.714±0.009	0.721±0.007	0.736±0.007	0.75±0.008	0.765±0.009
		SCAN-T	0.491±0.019	0.568±0.004	0.65±0.008	0.685±0.007	0.714±0.007	0.727±0.009	0.736±0.012	0.747±0.003
		SCAN-S	0.548±0.011	0.548±0.055	0.548±0.007	0.548±0.008	0.548±0.007	0.548±0.01	0.548±0.003	0.548±0.006
	F1	MLI	**0.644±0.01**	**0.695±0.022**	**0.722±0.013**	**0.742±0.005**	**0.761±0.005**	**0.789±0.006**	**0.783±0.005**	**0.786±0.006**
		LI	0.63±0.017	0.635±0.015	0.66±0.007	0.684±0.01	0.715±0.016	0.753±0.014	0.764±0.007	0.766±0.009
		SCAN	0.6±0.02	0.609±0.006	0.614±0.031	0.632±0.018	0.645±0.018	0.676±0.016	0.701±0.01	0.726±0.013
		SCAN-T	0.534±0.196	0.559±0.004	0.565±0.016	0.584±0.011	0.645±0.011	0.674±0.016	0.696±0.019	0.712±0.01
		SCAN-S	0.56±0.016	0.56±0.041	0.56±0.015	0.56±0.015	0.56±0.013	0.56±0.013	0.56±0.005	0.56±0.01
Twitter	AUC	MLI	**0.884±0.004**	**0.891±0.003**	**0.915±0.003**	**0.917±0.003**	**0.923±0.002**	**0.929±0.003**	**0.927±0.003**	**0.937±0.003**
		LI	0.841±0.003	0.847±0.002	0.852±0.003	0.862±0.002	0.873±0.002	0.884±0.003	0.894±0.003	0.904±0.003
		SCAN	0.801±0.003	0.814±0.002	0.819±0.003	0.817±0.002	0.819±0.002	0.823±0.003	0.831±0.002	0.837±0.003
		SCAN-T	0.802±0.002	0.802±0.002	0.802±0.002	0.802±0.002	0.802±0.002	0.802±0.002	0.802±0.002	0.802±0.002
		SCAN-S	0.508±0.002	0.543±0.002	0.584±0.003	0.631±0.001	0.653±0.002	0.666±0.003	0.673±0.003	0.686±0.003
	Accuracy	MLI	**0.92±0.003**	**0.927±0.002**	**0.927±0.003**	**0.929±0.004**	**0.93±0.003**	**0.932±0.003**	**0.936±0.003**	**0.936±0.004**
		LI	0.899±0.004	0.904±0.004	0.908±0.004	0.913±0.002	0.916±0.003	0.918±0.003	0.918±0.003	0.92±0.004
		SCAN	0.831±0.005	0.835±0.003	0.837±0.006	0.842±0.001	0.844±0.002	0.848±0.004	0.848±0.002	0.849±0.004
		SCAN-T	0.827±0.003	0.827±0.003	0.827±0.003	0.827±0.003	0.827±0.003	0.827±0.003	0.827±0.003	0.827±0.003
		SCAN-S	0.568±0.004	0.577±0.003	0.585±0.002	0.587±0.002	0.591±0.003	0.594±0.003	0.596±0.003	0.598±0.004
	F1	MLI	**0.804±0.002**	**0.808±0.002**	**0.809±0.003**	**0.811±0.003**	**0.812±0.003**	**0.818±0.003**	**0.826±0.003**	**0.826±0.004**
		LI	0.776±0.005	0.785±0.005	0.792±0.005	0.8±0.003	0.804±0.003	0.808±0.003	0.809±0.003	0.811±0.004
		SCAN	0.682±0.006	0.686±0.004	0.69±0.006	0.699±0.001	0.703±0.003	0.707±0.004	0.709±0.002	0.711±0.005
		SCAN-T	0.683±0.003	0.683±0.003	0.683±0.003	0.683±0.003	0.683±0.003	0.683±0.003	0.683±0.003	0.683±0.003
		SCAN-S	0.53±0.006	0.546±0.006	0.559±0.004	0.564±0.004	0.571±0.004	0.575±0.004	0.581±0.004	0.583±0.005

[2] with linear kernel and optimal parameters is used as the base classifier of all comparison methods. *Accuracy*, *AUC* and *F1* score are used as the evaluation metrics in this paper.

4.2.3 Experiment Setups

All the existing links in Foursquare are used as the positive link set and a proportion of unconnected links among users except the positive links are sampled as the negative link set, which is of the same size as the positive link set. Both the positive and negative link sets are divided into 5 folds: 4 folds as the training set and 1 fold as the test set. To represent different degrees of newness (available information users have in the networks), a fraction of information, which include posts, location checkins, temporal records, is randomly sampled from Foursquare as the available information under the control of parameter ρ^F $[0, 1]$ and the remaining information are deleted. Meanwhile, ρ^F proportion of the positive links are randomly sampled as the final positive link set and the remaining $(1 - \rho^F)$ proportion of positive links are mixed with the negative links to form the final unlabeled link set: $\mathcal{U}^F$. A subset of links in $\mathcal{U}^F$ are extracted as the reliable negative link set, $\mathcal{RN}^F$, with the spy technique. In a similar way, we can obtain the positive and reliable negative links in Twitter to be $\mathcal{P}^T$ and $\mathcal{RN}^T$, controlled by the sampling parameter ρ^T $[0, 1]$.

Supervised models, $\mathcal{M}^F$, built with $\mathcal{P}^F$ and $\mathcal{RN}^F$, are applied to classify links in the test set, $\mathcal{L}^F$. Depending on the specific methods, features used to build the models can be different and *social meta path* selection model $\mathcal{MS}^F$ is applied to select the most useful *social meta path* based features (K is set as 7 in the experiment) to build model $\mathcal{M}^F$. The predicted formation probabilities of links in $\mathcal{L}^F$ will be used to update their link weights in Foursquare. Based on the updated aligned networks, supervised models, $\mathcal{M}^T$, built with $\mathcal{P}^T$ and $\mathcal{RN}^T$, will be applied to classify links in the test set, $\mathcal{L}^T$, in which *social meta path* selection model $\mathcal{MS}^T$ is applied as well. And the predicted formation probabilities of links in $\mathcal{L}^T$ will also used to update their weights in Twitter. Only the weights of potential social links in $\mathcal{L}^F$ and $\mathcal{L}^T$ will be updated in each iteration and this process continues until the predicted formation probabilities of links in $\mathcal{L}^T$ and $\mathcal{L}^F$ converge.

4.3 Experiment Results

To denote different degrees of network newness, in Table 3, we fix ρ^T as 0.8 but changes ρ^F within $\{0.1, 0.2, \cdots, 0.8\}$. Table 3 has two parts: the upper part is the link prediction results in Foursquare and the lower part is that in Twitter, as MLI is an integrated PU link prediction framework. The link prediction results in each part are evaluated by different metrics: *AUC*, *Accuracy* and *F1*. As shown in Table 3, MLI can outperform all other comparison methods consistently for ρ^F $\{0.1, 0.2, \cdots, 0.8\}$ in both Foursquare network and Twitter network. For example, in Foursquare when $\rho^F = 0.5$, the *AUC* achieved by MLI is about 5% better than LI, 15% better than SCAN, 19% better than SCAN-T and 73% better than SCAN-S; the *Accuracy* achieved by MLI is about 2.3% better than LI, 8% better than SCAN, 9.1% higher than SCAN-T and over 40% higher than SCAN-S; the *F1* of MLI is 6.4% higher than LI, 18% higher than SCAN and SCAN-T and 36% higher than SCAN-S. When $\rho^F = 0.5$, the link prediction results of MLI in Twitter are also much better than all other baseline methods. For instances, in Twitter the *AUC* of MLI is 0.923 ± 0.002, which is about 6% better than LI, over 13% better than SCAN, SCAN-T and over 40% better than SCAN-S. Similar results can be obtained when evaluated by *Accuracy* and *F1*.

In Table 4, we fix $\rho^F = 0.8$ but change ρ^T with values in

Table 4: Performance comparison of different methods for inferring social and location links for Foursquare of different remaining information rates. The anchor link sample rate ρ_A is set as 1.0.

network	measure	methods	Remaining information rates ρ_T of Twitter 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Foursquare	AUC	MLI	**0.862±0.003**	**0.867±0.004**	**0.87±0.003**	**0.873±0.005**	**0.885±0.003**	**0.891±0.003**	**0.895±0.004**	**0.916±0.004**
		LI	0.831±0.005	0.834±0.004	0.846±0.004	0.853±0.005	0.855±0.005	0.867±0.004	0.868±0.005	0.87±0.005
		SCAN	0.81±0.007	0.81±0.008	0.812±0.005	0.817±0.007	0.816±0.01	0.815±0.007	0.822±0.006	0.82±0.004
		SCAN-T	0.81±0.007	0.81±0.007	0.81±0.007	0.81±0.007	0.809±0.007	0.809±0.007	0.81±0.007	0.81±0.007
		SCAN-S	0.504±0.007	0.51±0.003	0.511±0.005	0.516±0.005	0.522±0.004	0.53±0.005	0.53±0.004	0.53±0.005
	Accuracy	MLI	**0.78±0.003**	**0.786±0.005**	**0.789±0.004**	**0.794±0.005**	**0.793±0.004**	**0.789±0.004**	**0.796±0.005**	**0.797±0.005**
		LI	0.745±0.011	0.762±0.005	0.768±0.007	0.772±0.007	0.777±0.008	0.783±0.008	0.789±0.006	0.791±0.006
		SCAN	0.749±0.007	0.754±0.006	0.754±0.007	0.757±0.006	0.758±0.007	0.761±0.008	0.763±0.009	0.765±0.009
		SCAN-T	0.748±0.003	0.748±0.003	0.747±0.003	0.748±0.003	0.748±0.003	0.748±0.003	0.748±0.003	0.747±0.003
		SCAN-S	0.692±0.011	0.717±0.008	0.725±0.008	0.746±0.008	0.741±0.006	0.746±0.004	0.75±0.007	0.758±0.006
	F1	MLI	**0.768±0.004**	**0.774±0.005**	**0.778±0.006**	**0.784±0.006**	**0.785±0.005**	**0.777±0.004**	**0.785±0.006**	**0.786±0.006**
		LI	0.721±0.02	0.734±0.01	0.734±0.012	0.736±0.012	0.744±0.012	0.755±0.011	0.764±0.01	0.766±0.009
		SCAN	0.717±0.01	0.718±0.007	0.714±0.009	0.715±0.009	0.718±0.011	0.72±0.012	0.721±0.013	0.726±0.013
		SCAN-T	0.713±0.01	0.712±0.01	0.712±0.01	0.713±0.01	0.713±0.01	0.712±0.01	0.713±0.01	0.712±0.01
		SCAN-S	0.509±0.02	0.514±0.014	0.524±0.014	0.529±0.013	0.54±0.009	0.542±0.007	0.559±0.012	0.559±0.01
Twitter	AUC	MLI	**0.837±0.004**	**0.858±0.004**	**0.905±0.005**	**0.926±0.003**	**0.924±0.002**	**0.932±0.003**	**0.934±0.002**	**0.937±0.003**
		LI	0.772±0.009	0.829±0.008	0.871±0.009	0.887±0.002	0.887±0.002	0.897±0.003	0.899±0.003	0.904±0.003
		SCAN	0.706±0.008	0.771±0.012	0.799±0.009	0.817±0.002	0.819±0.002	0.829±0.003	0.83±0.003	0.834±0.003
		SCAN-T	0.555±0.133	0.678±0.006	0.753±0.044	0.754±0.019	0.764±0.014	0.781±0.004	0.794±0.003	0.802±0.002
		SCAN-S	0.687±0.008	0.687±0.002	0.687±0.005	0.687±0.002	0.687±0.002	0.687±0.004	0.687±0.003	0.687±0.003
	Accuracy	MLI	**0.821±0.005**	**0.864±0.001**	**0.892±0.008**	**0.914±0.004**	**0.925±0.002**	**0.926±0.004**	**0.936±0.002**	**0.936±0.004**
		LI	0.706±0.002	0.834±0.011	0.877±0.003	0.898±0.005	0.912±0.001	0.92±0.004	0.924±0.002	0.92±0.004
		SCAN	0.594±0.006	0.716±0.009	0.781±0.005	0.801±0.003	0.823±0.002	0.831±0.004	0.842±0.002	0.849±0.004
		SCAN-T	0.547±0.062	0.645±0.038	0.723±0.048	0.786±0.004	0.8±0.002	0.815±0.005	0.824±0.002	0.827±0.003
		SCAN-S	0.59±0.009	0.59±0.007	0.59±0.004	0.59±0.004	0.59±0.002	0.59±0.004	0.59±0.003	0.59±0.004
	F1	MLI	**0.713±0.009**	**0.762±0.005**	**0.791±0.006**	**0.81±0.004**	**0.81±0.002**	**0.819±0.004**	**0.821±0.002**	**0.826±0.004**
		LI	0.651±0.006	0.671±0.023	0.749±0.014	0.779±0.007	0.801±0.003	0.813±0.005	0.818±0.003	0.811±0.004
		SCAN	0.6±0.017	0.633±0.023	0.657±0.013	0.684±0.004	0.703±0.004	0.714±0.005	0.716±0.002	0.711±0.005
		SCAN-T	0.552±0.113	0.574±0.016	0.604±0.031	0.618±0.003	0.63±0.001	0.641±0.004	0.67±0.002	0.686±0.003
		SCAN-S	0.575±0.025	0.575±0.016	0.575±0.005	0.575±0.006	0.575±0.004	0.575±0.004	0.575±0.003	0.575±0.005

$\{0.1, 0.2, \cdots, 0.8\}$. Similar to the results obtained in Table 3 where ρ_F varies, MLI can beat all other methods in both Twitter and Foursquare when the degree of newness of the Twitter network changes.

MLI can perform better than LI in both Foursquare and Twitter, which shows that predicting social links in multiple networks simultaneously in MLI framework can do enhance the results in both networks; the fact that LI can beat SCAN shows that features extracted based on cross network meta paths can do transfer useful information for both anchor and non-anchor users; SCAN works better than both SCAN-T and SCAN-S denotes that link prediction with information in two networks simultaneously is better than that with information in one single network.

4.4 Parameter Analysis

An important parameter that can affect the performance of all these methods is the rate of anchor links existing across networks. In this part, we will analyze the effects of the anchor link rate, ρ_A $[0, 1.0]$. To exclude other parameters' interference, we fix ρ^F and ρ^T as 0.8 but change ρ^A with values in $\{0.1, 0.2, \cdots, 1.0\}$ and study the link prediction results in both Foursquare and Twitter under the evaluation of *AUC*, *Accuracy* and *F1*. The results are shown in Figure 5.

As shown in Figure 5, where Figures 5(a)-5(c) are the link prediction results in Foursquare and the Figures 5(d)-5(f) are those in Twitter, almost all the methods can perform better as ρ^A increases, except SCAN-T as it only utilizes information in the target network only. It shows that with more anchor links, MLI, LI, SCAN and SCAN-S can transfer much more information from other aligned source networks to the target network to enhance the results. In addition, MLI can work better than LI consistently as ρ^A varies, which can show the effectiveness of MLI in dealing with networks with different ratios of anchor links

4.5 Convergence Analysis

MLI need to predict the links in all the aligned networks alternatively and iteratively until convergence. In this part, we will analyze whether MLI can converge as this process continues. We show the link prediction results achieved by MLI in both Foursquare and Twitter under the evaluation of *AUC*, *Accuracy* and *F1* when ρ^F, ρ^T and ρ^A are all set as 0.8 in Figure 6. Figures 6(a)-6(c) are the results in Foursquare network from iteration 1 to iteration 30 and Figures 6(d)-6(f) are those in Twitter network. As shown in these figures, results achieved by MLI can converge in less than 10 iterations in both Foursquare and Twitter evaluated by all these three metrics.

5. RELATED WORK

Link prediction in online social networks first proposed by D. Liben-Nowell *et al.* [12] has been a hot research topic in recent years and many different methods have been proposed. D. Liben-Nowell *et al.* [12] propose many unsupervised link predicators to predict the social connections among users. M. Hasan *et al.* [8] propose to predict links by using supervised learning methods. An extensive survey of other link prediction methods is available in [9, 7].

Meanwhile, some works have also been done on predicting multiple kinds of links simultaneously. I. Konstas *et al.* [11] propose to recommend multiple kinds of links with collaborative filtering methods. F. Fouss *et al.* [6] propose to use a traditional model, random walk, to predict multiple kinds of links simultaneously in networks. M. Bilgic *et al.* [1] propose an approach to address two problems by interleaving object classification and link prediction in a collective algo-

rithm. P. Domingos *et al.* [4] propose a unifying framework, Markov Logic, for collective classification problems in social networks.

PU learning techniques have been proposed for many years and have been widely used in many different areas. B. Liu *et al.* [13] propose many different settings to obtain the reliable negative instance set from unlabeled instances in text mining tasks. Y. Zhao *et al.* [28] propose to apply PU learning techniques to graph mining area.

Nowadays, the researchers' focus start shifting to study multiple aligned heterogeneous online social networks simultaneously. X. Kong *et al.* [10] are the first to propose the concept of "multiple aligned heterogeneous social networks" and "anchor links" . They propose a two-phase method to predict the anchor links across networks. J. Zhang *et al.* [25] propose to transfer useful information across aligned networks to help predict social links for new users and they are the first to study social link prediction across aligned networks. J. Zhang *et al.* [26] propose to predict multiple kinds of links for new networks with information transferred across partially aligned networks and they are the first to study collective link prediction across partially aligned networks. J. Zhang *et al.* also gives a survey about link prediction problems and methods across social networks in [27].

6. CONCLUSION

In this paper, we have studied the *multi-network link prediction* problems across partially aligned networks. An effective general link prediction framework, Mli, has been proposed to solve the problem. Heterogeneous features can be extracted from the network based on both *intra-network* and *inter-network social meta paths*. Useful features are selected and transferred to aligned networks to enhance the prediction results mutually. Extensive experiments conducted on two real-world *aligned networks* demonstrate that Mli can work very well in predicting social links in multiple *partially aligned networks* simultaneously.

7. ACKNOWLEDGEMENT

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 the Pinnacle Lab at Singapore Management University, NSFC (61333014, 61321491) and 111 Program (B14020).

8. REFERENCES

[1] M. Bilgic, G. Namata, and L. Getoor. Combining collective classification and link prediction. In *ICDMW*, 2007.

[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] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, 2011.

[4] P. Domingos and M. Richardson. Markov logic: A unifying framework for statistical relational learning. In *ICML Workshop*, 2004.

[5] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *KDD*, 2008.

[6] F. Fouss, A. Pirotte, J. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *TKDE*, 2007.

[7] L. Getoor and C. P. Diehl. Link mining: A survey. *SIGKDD Explorations Newsletter*, 7(2), 2005.

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

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

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

[11] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR*, 2009.

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

[13] B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu. Building text classifiers using positive and unlabeled examples. In *ICDM*, 2003.

[14] Z. Lu, B. Savas, W. Tang, and I. Dhillon. Supervised link prediction using multiple sources. In *ICDM*, 2010.

[15] S. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 2010.

[16] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD*, 2011.

[17] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 2011.

[18] B. Tan, E. Xiang, Q. Yang, and E. Zhong. Multi-transfer: Transfer learning with multiple views and multiple sources. In *SDM*, 2013.

[19] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM*, 2012.

[20] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. Barabasi. Human mobility, social ties, and link prediction. In *KDD*, 2011.

[21] K. Wilcox and A. T. Stephen. Are close friends the enemy? online social networks, self-esteem, and self-control. Journal of Consumer Research, 2012.

[22] Y. Yao, H. Tong, X. Yan, F. Xu, and J. Lu. Matri: a multi-aspect and transitive trust inference model. In *WWW*, 2013.

[23] J. Ye, H. Cheng, Z. Zhu, and M. Chen. Predicting positive and negative links in signed social networks by transfer learning. In *WWW*, 2013.

[24] M. Ye, P. Yin, and W. Lee. Location recommendation for location-based social networks. In *GIS*, 2010.

[25] J. Zhang, X. Kong, and P. Yu. Predicting social links for new users across aligned heterogeneous social networks. In *ICDM*, 2013.

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

[27] J. Zhang and P. Yu. Link prediction across heterogeneous social networks: A survey. 2014.

[28] Y. Zhao, X. Kong, and P. Yu. Positive and unlabeled learning for graph classification. In *ICDM*, 2011.