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.

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 Foursquare⁴, 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

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—Data Mining

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

KDD ’14, August 24 - 27 2014, New York, NY, USA  
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.  
http://dx.doi.org/10.1145/2623330.2623645.

Keywords

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

1https://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, Mlt, 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, Mlt explores the meta path concept to generate useful features. Mlt 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”, Mlt 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 Mlt in predicting links for each network. Furthermore, Mlt 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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>information sources</td>
<td>multiple networks</td>
<td>multiple networks</td>
<td>multiple domains</td>
<td>multiple networks</td>
<td>multiple networks</td>
</tr>
<tr>
<td>source type</td>
<td>heterogeneous</td>
<td>heterogeneous</td>
<td>multi-view</td>
<td>heterogeneous</td>
<td>heterogeneous</td>
</tr>
<tr>
<td>source difference</td>
<td>solved</td>
<td>not solved</td>
<td>solved</td>
<td>not solved</td>
<td>not solved</td>
</tr>
<tr>
<td>predicted links</td>
<td>social links</td>
<td>heterogeneous links</td>
<td>n/a</td>
<td>social links</td>
<td>anchor links</td>
</tr>
<tr>
<td>settings</td>
<td>semi-supervised learning</td>
<td>supervised learning</td>
<td>supervised learning</td>
<td>supervised learning</td>
<td></td>
</tr>
<tr>
<td>knowledge to transfer</td>
<td>network structure</td>
<td>network structure</td>
<td>domain information</td>
<td>network structure</td>
<td></td>
</tr>
</tbody>
</table>

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 $\mathbb{G} = ((G^1, G^2, \ldots, G^n), (A^{1,2}, A^{1,3}, \ldots, A^{1,n}, A^{2,3}, \ldots, A^{n-1,n}))$, where $G^i, i \in \{1,2,\ldots,n\}$ is a heterogeneous social network and $A^{i,j} \neq \emptyset, i,j \in \{1,2,\ldots,n\}$ 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 an undirected anchor link between $G^i$ and $G^j$ iff $u^i \in U^i \cup U^j$, $v^j \in U^j \cup U^i$ and $u^i$ and $v^j$ are the accounts of the same user in $G^i$ and $G^j$ respectively.

Definition 4 (Anchor Users): User $u^i \in U^i$ is an anchor user in $G^i$ between $G^i$ and $G^j$ iff $\exists v^j \in U^j$ s.t. $(u^i, v^j) \in A^{i,j}$. The set of anchor users and non-anchor users is $U^i_{A,1} = \{u^i | u^i \in U^i \land \forall v^j \in U^j, (u^i, v^j) \notin A^{i,j}\}$.

Definition 5 (Non-Anchor Users): User $u^i \in U^i$ is an non-anchor user between $G^i$ and $G^j$ iff $u^i \notin U^i_{A,1}$. The set of non-anchor users in $G^i$ between $G^i$ and $G^j$ can be represented as $U^i_{A,1} = U^i - U^i_{A,1}$.

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 $(U^i_{A,1} = U^i) \land (U^j_{A,1} = U^j)$.

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 $(U^i_{A,1} \neq \emptyset) \lor (U^j_{A,1} \neq \emptyset)$.

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 $G^i, i \in \{1,2,\ldots,n\}$ 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 $P^i$, where $P^i = E^i_{u,u}$. Furthermore, a large set of unconnected user pairs are referred to as the unconnected links, $U^i$, and can be extracted from network $G^i$: $U^i = U^i \times U^i - P^i$. However, no information about links that will never be formed can be obtained from the network. In this paper, with $P^i$ and $U^i$, we formulate the link formation prediction as a PU link prediction problem.

Similarly, let $\{P^1, \ldots, P^n\}$, $\{U^1, \ldots, U^n\}$ and $\{L^1, \ldots, L^n\}$ be the sets of formed links, unconnected links, and links to be predicted of $G^1, G^2, \ldots, G^n$ respectively. With the formed and unconnected links of $G^1, G^2, \ldots, G^n$, 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 $V = U \cup P \cup L$ contains user, post and location nodes, $E = E_u \cup E_p \cup E_p, l$ contains the social links, write links and location links, then
$S_G = (T, R)$, $T = \{\text{User, Post, Location}\}$ and $R = \{\text{Social Link, Write Link, Location Link}\}$. 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 ex-
tensive 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 $T_i \in T, i \in \{1, 2, \ldots, k\}$ and $R_i \in R, i \in \{1, 2, \ldots, k-1\}$.

Meaning, depending on types of nodes and links that constitute it, $\Phi$ 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 $(T_1, \ldots , T_k$ are all the same $) \land (R_1, \ldots , R_{k-1}$ are all the same $)$, then $\Phi$ is a hom-
ogeneous 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 $T_1, \ldots, T_k$ and $R_1, \ldots, R_{k-1}$ 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 $\xrightarrow{\text{follow}}$ User, whose notation is
  \text{“$U \rightarrow U$” or $\Phi_0(U, U)$}.

- **ID 1.** Follower of Follower: User $\xrightarrow{\text{follow}}$ User $\xrightarrow{\text{follow}}$ User, whose notation is
  \text{“$U \rightarrow U \rightarrow U$” or $\Phi_1(U, U)$}.

- **ID 2.** Common Out Neighbor: User $\xrightarrow{\text{follow}}$ User $\xrightarrow{\text{follow}}$ User, whose notation is
  \text{“$U \rightarrow U \leftarrow U$” or $\Phi_2(U, U)$}.

- **ID 3.** Common In Neighbor: User $\xrightarrow{\text{follow}}$ User, whose notation is
  \text{“$U \leftarrow U \rightarrow U$” or $\Phi_3(U, U)$}.

**Heterogeneous Intra-Network Social Meta Path**

- **ID 4.** Common Words: User $\xrightarrow{\text{write}}$ Post $\xrightarrow{\text{contain}}$ User, whose notation is
  \text{“$U \rightarrow P \rightarrow W \leftarrow P \rightarrow U$” or $\Phi_4(U, U)$}.

**ID 5.** Common Timestamps: User $\xrightarrow{\text{write}}$ Post $\xrightarrow{\text{contain}}$ Time $\xrightarrow{\text{write}}$ User, whose notation is
  \text{“$U \rightarrow P \rightarrow T \leftarrow P \rightarrow U$” or $\Phi_5(U, U)$}.

**ID 6.** Common Location Checkins: User $\xrightarrow{\text{write}}$ Post $\xrightarrow{\text{attach}}$ Location $\xrightarrow{\text{attach}}$ User, whose notation is
  \text{“$U \rightarrow P \rightarrow L \leftarrow P \rightarrow U$” 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 $n \in T$ (or $l \in R$). Identity function $I(a, A) = \begin{cases} 1, & \text{if } a \in A \\ 0, & \text{otherwise}\end{cases}$
can check whether node/link $a$ is an in-
stance 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:

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

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

Features extracted based on $\Phi = \{\Phi_1, \ldots, \Phi_6\}$ are named
as the intra-network social meta path based social features.
($\Phi_0$ 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 $G^3$ 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 $0$. 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 $B^1 \rightarrow B^2 \rightarrow G^2 \rightarrow A^2 \rightarrow A^1$.

By following this idea, we propose to transfer useful in-
formation from aligned networks with the following anchor
meta path and the inter-network 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 $G^j$ if $T_1 = U^i$ and
$T_2 = U^j$ and $R_1 = A^{i-j}$. The notation of anchor meta
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 $\Psi = T_1 \rightarrow T_2 \rightarrow \cdots \rightarrow T_k$ is an inter-network meta path across $G^i$ and $G^j$ iff $\exists m \in \{1, 2, \cdots, k-1\}$, $T_m \rightarrow T_{m+1} = \Upsilon(U^i, U^j)$.

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:** $\Upsilon(U^i, U^j) \circ (\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j)) \circ \Upsilon(U^j, U^i)$, whose notation is $\Psi_1(U^i, U^j)$;
- **Category 2:** $\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j) \circ \Upsilon(U^i, U^j) \circ (\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j)) \circ \Upsilon(U^j, U^i)$, whose notation is $\Psi_2(U^i, U^j)$;
- **Category 3:** $\Upsilon(U^i, U^j) \circ (\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j)) \circ \Upsilon(U^j, U^i) \circ (\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j)) \circ \Upsilon(U^i, U^j)$, whose notation is $\Psi_3(U^i, U^j)$;
- **Category 4:** $\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j) \circ \Upsilon(U^i, U^j) \circ (\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j)) \circ \Upsilon(U^j, U^i) \circ (\Phi(U^i, U^j) \cup \Phi_0(U^i, U^j))$, whose notation is $\Psi_4(U^i, U^j)$.

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

Let $\Psi = \{\Psi_1, \Psi_2, \Psi_3, \Psi_4\}$. $\Psi$ is a comprehensive inter-network social meta path set and features extracted based on $\Psi$ 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 $\Psi_4(U^1, U^1)$; by following path $C^1 \rightarrow A^1 \rightarrow A^2 \rightarrow B^2 \rightarrow B^1$, we can go from non-anchor user $C^1$ to anchor user $B^1$, which is an instance of $\Psi_4(U^1, U^1)$; 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, $[x_\Phi^T, x_{\Psi}^T]^T$, from the multiple partially aligned heterogeneous networks.

Let $X_i \in [x_\Phi^T, x_{\Psi}^T]^T$ 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 $P(X_i = x, Y = y)$ 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 $[x_\Phi^T, x_{\Psi}^T]^T$ be the features of the top $K$ mi score selected from $[x_\Phi^T, x_{\Psi}^T]^T$. In the next subsection, we will use the selected feature vector $[x_\Phi^T, x_{\Psi}^T]^T$ 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 $P$ and unconnected links $U$. 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 $P$ and $U$ to be: $z(P) = +1$ and $z(U) = -1$.

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 $P$ and $U$ to be: $y(P) = +1$ but $y(U)$ can be either $+1$ or $-1$, as $U$ can contain both links to be formed and links that will never be formed.

By using $P$ and $U$ as the positive and negative training sets, we can build a link connection prediction model $M_c$, 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 $M_c$ 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 $l$: $p(z(l) = +1| x(l)) = [z_\Phi(l)^T, z_{\Psi}(l)^T]^T$. 

![Figure 2: Meta Path across Aligned Networks.](image)

![Figure 3: PU Link Prediction.](image)
Meanwhile, if we can obtain a set of links that “will never be formed”, i.e., “-1” links, from the network, which together with \( P \) (“+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 \( l \) : \( p(y(l) = +1|x(l)) \).

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’s connection probability and formation probability has been proved in existing works [5] to be:

\[
p(y(l) = +1|x(l)) \propto p(z(l) = +1|x(l)).
\]

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 \( \mathcal{M}_s \), built with \( P \) and \( U \) to classify links in \( U \) to extract the reliable negative link set.

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

\[
\mathcal{R}_N = \{ l \in U, p(z(l) = +1|x(l)) < \epsilon \}.
\]

Some heuristic methods have been proposed to set the optimal threshold \( \epsilon \), e.g., the spy technique proposed in [13]. As shown in Figure 3, we randomly select a subset of links in \( P \) as the spy, \( \mathcal{S}P \), whose proportion is controlled by \( s\% \). \( s\% = 15\% \) is used as the default sample rate in this paper. Sets \(( P - \mathcal{S}P )\) and \(( U \cup \mathcal{S}P )\) 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 \(( U \cup \mathcal{S}P )\), we can get their connection probabilities to be:

\[
p(z(l) = +1|x(l)), l \in (U \cup S P),
\]

and parameter \( \epsilon \) is set as the minimal connection probability of spy links in \( \mathcal{S}P \):

\[
\epsilon = \min \{ p(z(l) = +1|x(l)) \}.
\]

With the extracted reliable negative link set \( \mathcal{R}_N \), we can solve the PU link prediction problem with classification based link prediction methods, where \( \mathcal{P} \) and \( \mathcal{R}_N \) 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:

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

\[
= \arg \max_{\mathcal{Y}^i} p(y(\mathcal{L}^i) = \mathcal{Y}^i | [\hat{x}_s(\mathcal{L}^i)^T, \hat{x}_p(\mathcal{L}^i)^T]^T)
\]

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

---

### 3.3 Multi-Network Link Prediction Framework

Method M1i 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:

\[
\hat{\mathcal{Y}}^1, \hat{\mathcal{Y}}^2, \cdots, \hat{\mathcal{Y}}^n = \arg \max_{\mathcal{Y}^1, \mathcal{Y}^2, \cdots, \mathcal{Y}^n} p(y(\mathcal{L}^i) = \mathcal{Y}^i, y(\mathcal{L}^j) = \mathcal{Y}^j)
\]

\[
\cdots, y(\mathcal{L}^n) = \mathcal{Y}^n | G^1, G^2, \cdots, G^n)
\]

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]:

\[
\begin{aligned}
(\hat{\mathcal{Y}}^1)^{(r)} & = \arg \max_{\mathcal{Y}^1} p(y(\mathcal{L}^i) = \mathcal{Y}^1 | G^1 \cdots, G^n, \\
(\hat{\mathcal{Y}}^2)^{(r)} & = \arg \max_{\mathcal{Y}^2} p(y(\mathcal{L}^i) = \mathcal{Y}^2 | G^1 \cdots, G^n, (\hat{\mathcal{Y}}^1)^{(r-1)}, \cdots, (\hat{\mathcal{Y}}^n)^{(r-1)}) \\
& \cdots \cdots \cdots \\
(\hat{\mathcal{Y}}^n)^{(r)} & = \arg \max_{\mathcal{Y}^n} p(y(\mathcal{L}^i) = \mathcal{Y}^n | G^1 \cdots, G^n, (\hat{\mathcal{Y}}^1)^{(r)}, \cdots, (\hat{\mathcal{Y}}^{n-1})^{(r-1)})
\end{aligned}
\]

The structure of framework M1i 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, \( \hat{x}_s \), extracted from \( G^i \) and those extracted based on the inter-network social meta path, \( \hat{x}_p \), across \( G^1, G^2, \cdots, G^{i-1}, G^{i+1}, \cdots, G^n \) for links in \( P^i, U^i \) and \( \mathcal{L}^i \). Feature vectors \( \hat{x}_s(P^i), \hat{x}_s(U^i) \) and \( \hat{x}_p(P^i), \hat{x}_p(U^i) \) as well as the labels, \( y(P^i), y(U^i) \), of links in \( P^i \) and \( U^i \) are passed to the PU link prediction model \( \mathcal{M}^i \) and the meta path selection model \( \mathcal{M}^s \). 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

<table>
<thead>
<tr>
<th>property</th>
<th>Twitter</th>
<th>Foursquare</th>
</tr>
</thead>
<tbody>
<tr>
<td># node</td>
<td>5,223</td>
<td>5,392</td>
</tr>
<tr>
<td>Tweet/tip</td>
<td>9,490,707</td>
<td>48,756</td>
</tr>
<tr>
<td>Location</td>
<td>297,182</td>
<td>38,921</td>
</tr>
<tr>
<td>Friend/follow</td>
<td>164,920</td>
<td>76,972</td>
</tr>
<tr>
<td>Write</td>
<td>9,490,707</td>
<td>48,756</td>
</tr>
<tr>
<td>Locate</td>
<td>615,515</td>
<td>48,756</td>
</tr>
</tbody>
</table>

Figure 5: Effects of anchor link ratio $\rho_A$ 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.

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 $\Phi$ and $\Psi$ 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 $\Phi$ and $\Psi_1$ 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 on $\Psi_1$ 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 $\Phi$ 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 $\rho_A$ is set as 1.0.  

<table>
<thead>
<tr>
<th></th>
<th>Foursquare</th>
<th>Twitter</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>AUC</td>
<td></td>
</tr>
<tr>
<td>Mu</td>
<td>0.644 ± 0.01</td>
<td>0.665 ± 0.02</td>
<td>0.722 ± 0.13</td>
</tr>
<tr>
<td>Li</td>
<td>0.625 ± 0.01</td>
<td>0.637 ± 0.01</td>
<td>0.682 ± 0.07</td>
</tr>
</tbody>
</table>

For Foursquare, $\rho$ is applied to classify links in the test set, $L^T$, in which social meta path selection model $M_{ST}$ is applied as well. And the predicted formation probabilities of links in $L^T$ will also be used to update their weights in Twitter. Only the weights of potential social links in $L^F$ and $L^T$ will be updated in each iteration and this process continues until the predicted formation probabilities of links in $L^F$ and $L^T$ converge.

4.3 Experiment Results

To denote different degrees of network newness, in Table 3, we fix $\rho^F$ as 0.8 but changes $\rho^T$ 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 Mu is an integrated PU link prediction framework. The link prediction results in each part are evaluated by different metrics: Accuracy, AUC and F1. As shown in Table 3, Mu can outperform all other comparison methods consistently for $\rho^T \in \{0.1, 0.2, \cdots ,0.8\}$ in both Foursquare network and Twitter network.
Table 4: Performance comparison of different methods for inferring social and location links for Foursquare of different remaining information rates. The anchor sample rate \( \rho_A \) is set as 1.0.

<table>
<thead>
<tr>
<th>Network</th>
<th>Measure</th>
<th>Methods</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foursquare</td>
<td>AUC</td>
<td>Mli</td>
<td>0.862 ± 0.003</td>
<td>0.867 ± 0.004</td>
<td>0.873 ± 0.005</td>
<td>0.885 ± 0.007</td>
<td>0.891 ± 0.003</td>
<td>0.889 ± 0.004</td>
<td>0.914 ± 0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCAN</td>
<td>0.812 ± 0.007</td>
<td>0.812 ± 0.007</td>
<td>0.812 ± 0.005</td>
<td>0.812 ± 0.007</td>
<td>0.816 ± 0.001</td>
<td>0.815 ± 0.007</td>
<td>0.822 ± 0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAN</td>
<td>0.504 ± 0.007</td>
<td>0.512 ± 0.003</td>
<td>0.511 ± 0.005</td>
<td>0.516 ± 0.003</td>
<td>0.522 ± 0.004</td>
<td>0.530 ± 0.005</td>
<td>0.532 ± 0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>0.785 ± 0.003</td>
<td>0.786 ± 0.015</td>
<td>0.799 ± 0.004</td>
<td>0.794 ± 0.003</td>
<td>0.793 ± 0.004</td>
<td>0.789 ± 0.004</td>
<td>0.796 ± 0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LI</td>
<td>0.745 ± 0.011</td>
<td>0.762 ± 0.007</td>
<td>0.768 ± 0.007</td>
<td>0.772 ± 0.007</td>
<td>0.777 ± 0.008</td>
<td>0.783 ± 0.008</td>
<td>0.789 ± 0.006</td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>AUC</td>
<td>Mli</td>
<td>0.765 ± 0.004</td>
<td>0.774 ± 0.005</td>
<td>0.776 ± 0.006</td>
<td>0.784 ± 0.005</td>
<td>0.777 ± 0.004</td>
<td>0.785 ± 0.006</td>
<td>0.786 ± 0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCAN</td>
<td>0.749 ± 0.007</td>
<td>0.754 ± 0.006</td>
<td>0.754 ± 0.007</td>
<td>0.757 ± 0.006</td>
<td>0.758 ± 0.007</td>
<td>0.761 ± 0.008</td>
<td>0.763 ± 0.009</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAN</td>
<td>0.748 ± 0.003</td>
<td>0.748 ± 0.003</td>
<td>0.747 ± 0.003</td>
<td>0.748 ± 0.003</td>
<td>0.748 ± 0.003</td>
<td>0.748 ± 0.003</td>
<td>0.747 ± 0.003</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>0.692 ± 0.011</td>
<td>0.717 ± 0.008</td>
<td>0.725 ± 0.008</td>
<td>0.746 ± 0.008</td>
<td>0.741 ± 0.006</td>
<td>0.746 ± 0.004</td>
<td>0.757 ± 0.006</td>
<td></td>
</tr>
</tbody>
</table>

\( \{0.1, 0.2, \ldots, 0.8\} \). Similar to the results obtained in Table 3 where \( \rho_F \) 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, \( \rho_A \in [0, 1] \). To exclude other parameters’ interference, we fix \( \rho_A \) and \( \rho_F \) as 0.8 but change \( \rho_F \) with values in \( \{0.1, 0.2, \ldots, 1.0\} \) 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 \( \rho_A \) 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 \( \rho_A \) 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 \( \rho_F \), \( \rho_A \) and \( \rho_F \) 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 predictors 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-
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