Transferring Heterogeneous Links across Location-Based Social Networks

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

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

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

ABSTRACT

Location-based social networks (LBSNs) are one kind of online social networks offering geographic services and have been attracting much attention in recent years. LBSNs usually have complex structures, involving heterogeneous nodes and links. Many recommendation services in LBSNs (e.g., friend and location recommendation) can be cast as link prediction problems (e.g., social link and location link prediction). Traditional link prediction researches on LBSNs mostly focus on predicting either social links or location links, assuming the prediction tasks of different types of links to be independent. However, in many real-world LBSNs, the prediction tasks for social links and location links are strongly correlated and mutually influential. Another key challenge in link prediction on LBSNs is the data sparsity problem (i.e., "new network" problem), which can be encountered when LBSNs branch into new geographic areas or social groups. Actually, nowadays, many users are involved in multiple networks simultaneously and users who just join one LBSN may have been using other LBSNs for a long time. In this paper, we study the problem of predicting multiple types of links simultaneously for a new LBSN across partially aligned LBSNs and propose a novel method TRAIL (TRAnsfer heterogeneous lInks across LBSNs). TRAIL can accumulate information for locations from online posts and extract heterogeneous features for both social links and location links. TRAIL can predict multiple types of links simultaneously. In addition, TRAIL can transfer information from other aligned networks to the new network to solve the problem of lacking information. Extensive experiments conducted on two real-world aligned LBSNs show that TRAIL can achieve very good performance and substantially outperform the baseline methods.

Categories and Subject Descriptors

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

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2559894.

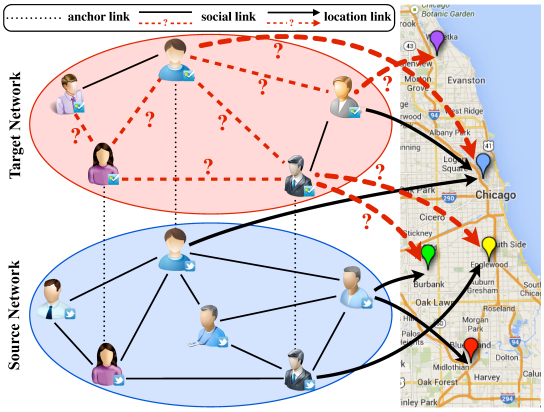


Figure 1: Example of collective link transferring across two aligned location-based social networks.

Keywords

Location-Based Social Networks, Link Prediction, Transfer Learning, Data Mining

1. INTRODUCTION

Location-based social networks (LBSNs) are one kind of online social networks that can provide geographic services, e.g., location check-ins and posting reviews, and have been attracting much attention in recent years [20, 16, 19, 5, 21]. LBSNs usually have very complex structures, including multiple kinds of nodes (e.g., users, locations, etc.) and different types of links among these nodes (e.g., social links among users and location links between users and locations). For example, Foursquare[1] is a mainstream LBSN. It involves millions of users and locations. Foursquare users can add friends, check in at different locations with cellphones, write reviews and share with others.

Link prediction, which aims at predicting whether two entities have certain relationships, has become a hot topic in recent years [3, 12, 7, 15, 18, 14, 9]. Many important services offered by LBSNs can be cast as link prediction problems. For example, friend recommendation involves predicting social links among users; location recommendation aims at predicting location links between users and locations. LBSNs can benefit a lot from the high-quality social link and location link prediction results. The reason is that well-established social ties can improve user's engagement in so-

[1]https://foursquare.com

Table 1: Summary of related problems.

Property	Transferring Heterogeneous Links across LBSNs	Collective Link Prediction across Multi-Domains [4]	Collaborative Recommendation for Networks[12, 7]	Predicting Social Links across Aligned Networks [22]	Inferring Anchor Links across Networks [11]
# networks	multiple	multiple	single	multiple	multiple
network type	heterogeneous	bipartite graph	heterogeneous	heterogeneous	heterogeneous
network aligned?	partially aligned	no	no	fully aligned	fully aligned
predicted links	multiple kinds (social and location links)	multiple kinds	multiple kinds	single kind (social link)	single kind (anchor link)
settings	transfer learning	transfer learning	unsupervised learning	transfer learning	transfer learning
knowledge to transfer	network structure through anchor links	network structure via task similarities	n/a	network structure through anchor links	network structure through anchor links

cial networks [13]. Meanwhile, in location-based social networks, high-quality predicted location links can enhance the value of the location services of the networks.

Conventional link prediction researches on LBSNs mostly focus on predicting either social links [16, 19] or location links [21, 5] and usually assume that the prediction tasks of different types of links to be independent. However, in many real-world LBSNs, the link prediction tasks for social links and location links are strongly correlated and mutually influential. For example, if two users are friends with each other, they are more likely to check-in at similar locations. Thus the performance of location recommendation can be significantly improved if we could make accurate friendship predictions. Similarly, if two users often check-in at similar locations, they are more likely to know each other and have friend links in real life. The performance of friend recommendation can be greatly improved if we could make accurate location-link predictions.

Another major challenge in link prediction for LBSNs is the information sparsity problem, where the linkage information within the network can be very sparse. Conventional link prediction methods usually assume that there are sufficient links within the network to compute features (e.g., common neighborhoods) between each pair of nodes. However, LBSNs often encounter "new network" problems when they branch into new geographic areas or social groups. For example, when a LBSN decides to extend services in a new geographic area (e.g., Foursquare's expansion into Chinese market), the linkage information within the area (both social links and location links) can be very sparse. Similarly, when a LBSN decides to promote in a new group of users (e.g., Facebook's expansion from college students to white collar), the linkage information within the social group can be largely missing. The constituent of a LBSN is not quite connected to the existing members, and it can be considered to be a "new network".

In order to solve the "new network" problems, we need to utilize additional information sources to facilitate the link prediction process. Actually, nowadays, many people are involved in multiple LBSNs to enjoy specific services offered by different ones. Users who just joined one LBSN may have been using other LBSNs for a long time. For example, in Figure 1, we have two LBSNs. The LBSN on the top is a new network, e.g., Foursquare, and the social links and location links in it are very sparse. However, some users have joined another LBSN that has already existed in the geographic area or social group for a long time with abundant linkage information. We refer such users as "anchors" and the link between the two accounts of the same user as an "anchor link". For simplicity, the new network is called the *target network* and the developed network is called the *source network*.

In this paper, we study the collective link prediction problem for a new network across aligned LBSNs and the links to be predicted include both social links and location links. This problem has not been studied before. Meanwhile, it is also very challenging to solve in the following aspects:

1. *Collective link prediction.* The first challenge mainly lies in the fact that social links and location links in LBSNs are correlated instead of being independent. The prediction tasks on social links and location links should be considered at the same time. Many existing works focus on predicting one single type of links in LBSNs [16, 19, 21, 5], which do not consider the correlation between different link prediction tasks.

2. *Lack of information in the target network.* The target network is a new network, information in which is quite rare. We need to overcome the information sparsity problem in the target network. Existing works on link prediction mainly focus on one single network [16, 19, 21, 5, 12, 7]. In real-world LBSNs, the anchor users between two networks can serve as a bridge to transfer information from one LBSN to the other LBSN. Such knowledge transfer can benefit the link prediction for both social links and location links.

A more detailed comparison with previous works is shown in Table 1.

In this paper, we propose a supervised collective linkage transferring method, TRAIL, to address the above challenges. TRAIL can accumulate auxiliary information for locations from online posts which have check-ins at them and can extract heterogeneous features for both social links and location links. TRAIL can predict social links and location links simultaneously. In addition, TRAIL can use both information in the target network and that transferred from the aligned source network at the same time.

2. PROBLEM FORMULATION

2.1 Location-Based Social Networks

A location-based social network (LBSN) can be modeled as a heterogeneous network $G = (V, E, W)$, where $V = \bigcup_i V_i$ is the union of different types of nodes and $V_i, i \in \{1, 2, \cdots, |V|\}$ is the set of nodes of the i_{th} type. $E = \bigcup_j E_j$ is the union of link sets among nodes in V and $E_j, j \in \{1, 2, \cdots, |E|\}$ is the set of links of the j_{th} type, $W : E \to \mathbb{R}$ denotes the weight of links in E.

Specially, for a LBSN, node set $V = \mathcal{U} \cup \mathcal{L} \cup \mathcal{T} \cup \mathcal{W}$ is the union of node sets of users, locations, time and words.

The link set $E = E_s \cup E_l \cup E_t \cup E_w$ is the union of link sets consisting of social relationships, location check-ins, active time and published words of users, W denotes the weight of links in E.

2.2 Aligned LBSNs

Based on the definition of heterogeneous network, the aligned LBSNs can be definied as $\mathcal{G} = (G_{set}, A_{set})$, where network set $G_{set} = \{G^1, G^2, \cdots, G^{|G_{set}|}\}$ is the set of $|G_{set}|$ different LBSNs, anchor link set $A = \{A^{1,2}, A^{1,3}, \cdots, A^{1,|G_{set}|}, A^{2,1}, \cdots, A^{|G_{set}|,|G_{set}|-1}\}$ contains the directed *anchor links* between pairwise networks in G_{set}. $A^{i,j} \subseteq \mathcal{U}^i \times \mathcal{U}^j$ is the set of directed *anchor links* from network G^i to G^j, where $\mathcal{U}^i$,$\mathcal{U}^j$ are the sets of users in network G^i and G^j. Link $(u_m^i, u_n^j) \in A^{i,j}$ is an *anchor link* between G^i and G^j iff. $(u_m^i \in \mathcal{U}^i) \wedge (u_n^j \in \mathcal{U}^j) \wedge (u_m^i$ and u_n^j are accounts owned by the same user in G^i and $G^j)$.

Given two aligned heterogeneous networks G^i and G^j, if all user accounts in one network are related to accounts in the other network by anchor links mutually, then G^i and G^j are *fully aligned*, in which case $|\mathcal{U}^i| = |\mathcal{U}^j| = |A^{i,j}|$ and the anchor links in $A^{i,j}$ have an inherent *one-to-one* property [2]. While, if some users in G^i do not have the corresponding accounts in G^j or some users in G^j do not have the corresponding accounts in G^i, then G^i and G^j are *partially aligned* and $|A^{i,j}| \leq \min\{|\mathcal{U}^i|, |\mathcal{U}^j|\}$.

2.3 Collective Link Prediction

Fully aligned social networks merely exist in the real world. In this paper, we are predicting multiple kinds of links for new networks across two *partially aligned* LBSNs. Let $\mathcal{G} = (\{G^t, G^s\}, \{A^{t,s}, A^{s,t}\})$ be the networks studied in this paper, where G^t is the target network, which is very new, and G^s is the aligned well-established source network, $A^{t,s}, A^{s,t}$ are the sets of directed *anchor links* between G^t and G^s. The set of users and locations in G^t are denoted as $\mathcal{U}^t$ and $\mathcal{L}^t$, while the sets of existing social links and location links in G^t are represented as E_s^t and E_l^t. What we want to predict are a subset of potential social links among users in G^t: $\boldsymbol{L}_s^t \subset (\mathcal{U}^t \times \mathcal{U}^t - E_s^t)$ and a subset of potential location links in G^t: $\boldsymbol{L}_l^t \subset (\mathcal{U}^t \times \mathcal{L}^t - E_l^t)$. In other words, we want to build a mapping: $f : \{\boldsymbol{L}_s^t, \boldsymbol{L}_l^t\} \to \{-1, 1\}$ to decide whether potential links in $\{\boldsymbol{L}_s^t, \boldsymbol{L}_l^t\}$ exist or not and a confidence score function $P : \{\boldsymbol{L}_s^t, \boldsymbol{L}_l^t\} \to [0, 1]$ denoting their existence probabilities.

3. PROPOSED METHODS

In this section, we will introduce the supervised collective link transferring method, TRAIL, in details.

3.1 Information Accumulation and Feature Extraction

TRAIL is based on a supervised setting, as a result, we need to extract features for both social links and location links using the heterogeneous information in the network. Before introducing the extracted features, we will introduce a method to accumulate information for locations at first.

3.1.1 Information Accumulation for Locations

Locations are represented as $(latitude, longitude)$ pairs in our problem, which possess no auxiliary information except location links with users in the network. As a result, we will confront problems of lacking auxiliary information when extracting heterogeneous features for location links. Actually, we notice users can publish online posts at the locations. And we propose to accumulate the text and timestamps information of the online posts checked in at a certain location as the auxiliary information possessed by that location.

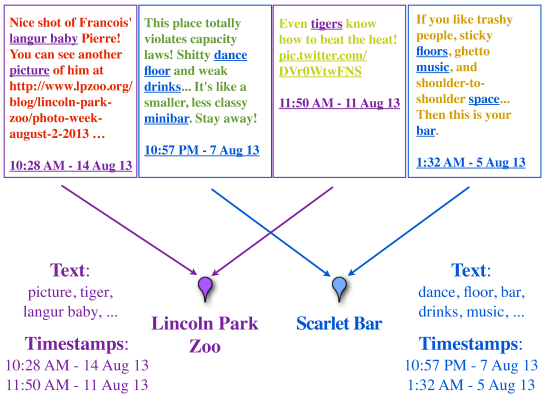


Figure 2: Example of information accumulation for locations from online posts.

From a statistical point of view, information from posts published at a certain location, including both timestamps and text contents, can reveal some properties of the location. For example, the timestamps of most posts published at nightlife sites are after 6:00 PM. While, those of posts published at restaurants serving brunch are during the daytime. Posts published at national parks can contain some phrases depicting the scenes, while posts published at basketball court may be mostly talking about games, teams and players. So, we can know more about the locations from the information accumulated from online posts.

For example, in Figure 2, we have two totally different locations: the Lincoln Park Zoo[2] and Scarlet Bar[3]. The Lincoln Park Zoo is the largest free zoo in Chicago and is open during 10:00 AM - 5:00 PM. The Scarlet Bar is one of the most famous bar in Chicago, where people can drink with friends, dance to enjoy their night life, and it is open during 8:00 PM - 2:00 AM.

We also have 4 online posts published by people at these two places in either Foursquare or Twitter. From the content of these posts, we find that people usually publish words about animals, pictures and the scene at the Lincoln Park Zoo. However, people who visit the Scarlet Bar mainly talk about the atmosphere in the bar, the drinks, the dance floor and the music there. So, users who frequently talk about animals in daily life can be interested in the Lincoln Park Zoo, while those who usually post words about the drinks may like the Scarlet Bar more. Meanwhile, we can also accumulate the timestamps of posts published at these two places. The timestamps of posts published at the Lincoln Park Zoo are mostly during the daytime, while those of posts published at the Scarlet Bar are at night. So, users who are usually active in the daytime can be more likely to visit the Lincoln Park Zoo, while people who are active during the night may prefer the Bar.

[2]http://www.lpzoo.org
[3]http://www.scarletbarchicago.com

Table 2: Features extracted from vector $\boldsymbol{x}$ and $\boldsymbol{y}$

Features	Descriptions
Extended Degree Count (EDC)	$\|\|\boldsymbol{x}\|\|_1, \|\|\boldsymbol{y}\|\|_1$
Extended Degree Ratio (EDR)	$\|\|\boldsymbol{x}\|\|_1/\|\|\boldsymbol{y}\|\|_1$
Extended Common Neighbour (ECN)	$\boldsymbol{x} \cdot \boldsymbol{y}$
Extended Jaccard's Coefficient (EJC)	$\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\|\boldsymbol{x}\|\|_1 \cdot \|\|\boldsymbol{y}\|\|_1}$
Extended Preferential Attachment (EPA)	$\|\|\boldsymbol{x}\|\|_1 \cdot \|\|\boldsymbol{y}\|\|_1$
Euclidean Distance (ED)	$(\sum_k (x_k - y_k)^2)^{1/2}$
Cosine Similarity (CS)	$\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\|\boldsymbol{x}\|\|_2 + \|\|\boldsymbol{y}\|\|_2}$

3.1.2 Heterogeneous Features

In this part, we will extract 4 different categories of features for both social links and location links from the heterogeneous information in the network networks, which include *social features*, *spatial distribution features*, *text usage features* and *temporal distribution features*. A summary of frequently used features is available in Table 2, where $||\boldsymbol{x}||_p = (\sum_{i=1}^{|\boldsymbol{x}|} |x_i|^p)^{1/p}$ is the L^p-norm of vector $\boldsymbol{x}$.

- **Features of Social Links:** For a certain social link (u_i, u_j), we can get their followers from the network: $\Gamma(u_i)$ and $\Gamma(u_j)$. Based on $\Gamma(u_i)$, we can construct the social link weight vector $\tilde{s}(u_i)$ for u_i, where $\tilde{s}(u_i) = (p_{1,i}, p_{2,i}, \cdots, p_{k,i}, \cdots, p_{n,i})^T, n = |\mathcal{U}|$ is the size of user set and $p_{k,i}$ is the weight of social link $(u_k, u_i), \forall u_k \in \mathcal{U}$: if $u_k \in (\mathcal{U} - \Gamma(u_i))$, $p_{k,i} = 0.0$; if $u_k \in \Gamma(u_i)$ and link (u_k, u_i) exists originally, then $p_{k,i} = 1.0$; otherwise, $p_{k,i}$ is the existence probability of link (u_k, u_i). Similarly, we can construct vector $\tilde{s}(u_j)$ for user u_j, which is of the same length as $\tilde{s}(u_i)$. From $\tilde{s}(u_i)$ and $\tilde{s}(u_j)$, we extract 7 different *social features* for social link (u_i, u_j), which are summarized in Table 2.
  In a similar way, for a certain social link (u_i, u_j), we can get the set of locations visited by user u_i and u_j: $\Phi(u_i)$ and $\Phi(u_j)$, from which we can obtain their location link weight vectors: $\tilde{l}(u_i)$ and $\tilde{l}(u_j)$. From the timestamps of posts published by users, we can obtain the users' active patterns. Each day is divided into 24 slots and the ratio of online posts published by user u in each hour is saved in a temporal distribution vector $\tilde{t}(u)$, whose length is 24. For social link (u_i, u_j), we can construct the temporal distribution vectors: $\tilde{t}(u_i)$ and $\tilde{t}(u_j)$ for u_i and u_j. In addition, we transform the words used by two users u_i and u_j into two text usage vectors: $\tilde{w}(u_i)$ and $\tilde{w}(u_j)$ weighted by TF-IDF, which are of the same length. From these vectors, we can extract the *spatial distribution features*, *temporal distribution features* and *text usage features* similar to the social link features summarized in Table 2 for social link (u_i, u_j).

- **Features of Location Links:** Similarly, we can obtain the set of users who have visited a location and regard them as the "neighbours" of that location. And for a location link (u_i, l_j), we can get the sets of neighbours of u_i and l_j: $\Gamma(u_i)$ and $\Psi(l_j)$, from which we can construct the social link weight vectors: $\tilde{s}(u_i)$ and $\tilde{s}(l_j)$. From the accumulated text and timestamps information of locations and the auxiliary information owned by users, we can also constract the temporal distribution vectors: $\tilde{t}(u_i)$ and $\tilde{t}(l_j)$ and the text usage vectors: $\tilde{w}(u_i)$ and $\tilde{w}(l_j)$ for location link (u_i, l_j). From these vectors, we can extract the *social features*, *temporal distribution features* and *text usage features* for location link (u_i, l_j).
  In addition, according to previous definitions, we can get the locations that user u has visited in the past: $\Phi(u)$ and the location link weight vector of u: $\tilde{l}(u)$ as well as the neighbors of a location l: $\Psi(l)$ and its social link weight vector: $\tilde{s}(l)$. For a certain location link (u_i, l_j), we extract 3 spatial distribution features from the network:

  (1) average weighted geographic distance between locations in $\Phi(u_i)$ and l_j

$$\frac{\sum_{l_k \in \Phi(u_i)} GeoD(l_k, l_j) \cdot \tilde{l}(u_i)_{l_k}}{||\tilde{l}(u_i)||_1 \cdot |\Phi(u_i)|}$$

  where, $GeoD(l_k, l_j)$ is the geographic distance of l_k and l_j and $\tilde{l}(u_i)_{l_k}$ is the weight of location link (u_i, l_k) saved in u_i's location link weight vector $\tilde{l}(u_i)$.

  (2) weighted number of users who have visited both locations in $\Phi(u_i)$ and l_j

$$\sum_{l_k \in \Phi(u_i)} \tilde{s}(l_k) \cdot \tilde{s}(l_j) \cdot \tilde{l}(u_i)_{l_k}$$

  (3) average weighted number of users who have visited both locations in $\Phi(u_i)$ and l_j

$$\frac{\sum_{l_k \in \Phi(u_i)} \tilde{s}(l_k) \cdot \tilde{s}(l_j) \cdot \tilde{l}(u_i)_{l_k}}{||\tilde{l}(u_i)||_1 \cdot \sum_{l_k \in \Phi(u_i)} ||\tilde{l}(s_k)||_1}$$

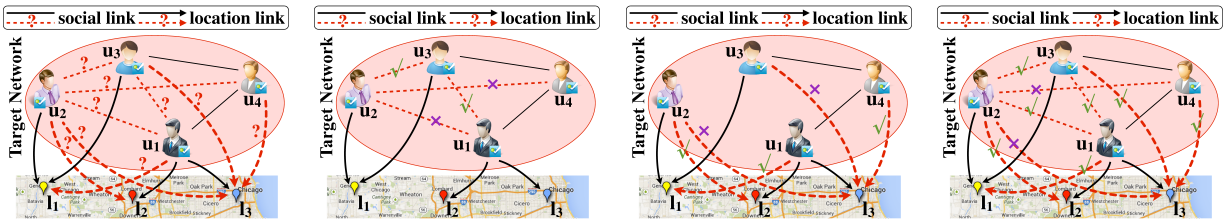
3.2 Collective Link Predictions

In this section, we will analyze and formulate the correlation between the social link prediction task and the location link prediction task.

3.2.1 Correlation Between Different Tasks

When predicting a link, the classifiers will give a score within range $[0, 1]$ to show its existence probability. Newly predicted social links will update the social link existence probability information in the network, which can affect other location link prediction tasks. For example, these updated social link existence probabilities can change the extended common neighbours of a location and a user. Similarly, the location link prediction task can also influence the social link prediction result.

For example, in Figure 3, we show an example of different link prediction methods. Figure 3(a) is the input aligned networks, in which there are 4 users and some existing social links $(u_3, u_4), (u_1, u_4)$ and location links (u_2, l_1), (u_3, l_1), (u_1, l_2), (u_1, l_3) as well as many other potential links to be predicted. Based on the information in the network, including social information (e.g., common neighbours), location information (e.g., co-checkins) and other auxiliary information, traditional link prediction methods can predict social links and locate links independently. Figure 3(b) shows the result of independent social link prediction result, in which social link (u_2, u_3) and (u_1, u_3) are predicted to be existent, while social link (u_1, u_2) and (u_2, u_4) are predicted to be nonexistent. Figure 3(c) shows the independent location link prediction result and in the result, location links $(u_2, l_2), (u_1, l_1), (u_4, l_3)$ are predicted to be existent, while (u_2, l_3) and (u_3, l_3) is predicted to be nonexistent.



(a) input network (b) social link prediction (c) location link prediction (d) collective link prediction

Figure 3: An example of different link prediction methods. (a) is the input network. (b)-(c) is independent social link and location link prediction result. (d) shows the collective link prediction result.

From the results in Figures 3(b) and 3(c), we can find some problematic phenomena. For example, user u_2 and u_1 are predicted to have visited locations l_1, l_2 and they are also predicted to share a common neighbour: u_3. Based on the result, it is highly likely that the potential social link (u_2, u_3) will be predicted to be existent. However, it is predicted to be nonexistent in Figure 3(b). Another example is that many neighbours of user u_3, both the originally existing u_4 and the newly predicted u_1 both have visited or are predicted to have visited l_3. By using Friend-based Collaborative Filtering (FCF) [21], u_3 is highly likely to be predicted to have visited l_3. However, the location link between u_3 and l_3 is predicted to be nonexistent in Figure 3(c).

If we consider the correlation between these two link prediction tasks and predict social links and location links simultaneously, the predicted results of social link (u_1, u_2) and location link (u_3, l_3) are highly likely to be predicted as existent. In Figure 3(d), we show a potential result of collective link prediction methods.

3.2.2 Collective Link Prediction

We formulate the sets of potential social links and potential location links to be predicted as $\boldsymbol{L}_s^t \subset (\mathcal{U}^t \times \mathcal{U}^t - E_s^t)$ and $\boldsymbol{L}_l^t \subset (\mathcal{U}^t \times \mathcal{L}^t - E_l^t)$ in the problem formulation section. For links $l_s^t \in \boldsymbol{L}_s^t$ and $l_l^t \in \boldsymbol{L}_l^t$, the supervised models built with the existing information in the network will give them the predicted labels: $y(l_s^t)$ and $y(l_l^t)$, as well as the existence probability scores: $P(y(l_s^t) = 1)$ and $P(y(l_l^t) = 1)$. Traditional methods predicting social links and location links independently aims at finding the set of labels achieving the maximum probability scores for each kind of links. In other words, let $\hat{\mathcal{Y}}_s^t \subset \{-1, 1\}^{|\boldsymbol{L}_s^t|}$, $\hat{\mathcal{Y}}_l^t \subset \{-1, 1\}^{|\boldsymbol{L}_l^t|}$ be the sets of optimal labels

$$\hat{\mathcal{Y}}_s^t = arg \max_{\mathcal{Y}_s^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t)$$

$$\hat{\mathcal{Y}}_l^t = arg \max_{\mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t)$$

where, $P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t)$ and $P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t)$ denote the probability scores achieved when links in $\boldsymbol{L}_s^t$ and $\boldsymbol{L}_l^t$ are assigned with labels in $\mathcal{Y}_s^t$ and $\mathcal{Y}_l^t$.

However, considering connections between these two link prediction tasks, the inferred social link or location link information should all be used in other link prediction tasks. The optimal selection of label sets $\hat{\mathcal{Y}}_s^t$ and $\hat{\mathcal{Y}}_l^t$ will be

$$\hat{\mathcal{Y}}_s^t, \hat{\mathcal{Y}}_l^t = arg \max_{\mathcal{Y}_s^t, \mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t | \mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t) \times P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t | \mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t)$$

3.3 Collective Linkage Transfer across LBSNs

3.3.1 Supervised Link Prediction

Traditional supervised link prediction methods by using one single network implicitly or explicitly assume that information in the target network itself is enough to build effective link prediction models. These methods use the extracted features of existing links in the traget network to train classifiers, which will be applied to predict other potential links. For example, we want to predict the existence probability of a social link (u_i^t, u_j^t) in the target network G^t, which is:

$$P(y(u_i^t, u_j^t) = 1 | G^t)$$

where $y(u_i^t, u_j^t)$ is the label of link (u_i^t, u_j^t). From G^t, we can extract a set of heterogeneous features for social link (u_i^t, u_j^t). Then

$$P(y(u_i^t, u_j^t) = 1 | G^t) = P(y(u_i^t, u_j^t) = 1 | \boldsymbol{x}(u_i^t, u_j^t))$$

where $\boldsymbol{x}(u_i^t, u_j^t) = [x(u_i^t, u_j^t)^1, x(u_i^t, u_j^t)^2, \cdots, x(u_i^t, u_j^t)^n]^T, n = |\boldsymbol{x}(u_i^t, u_j^t)|$ and $x(u_i^t, u_j^t)^k, k \in \{1, 2, \cdots, n\}$ is the k_{th} feature extracted from the target network for social link (u_i^t, u_j^t). Usually, feature $x(u_i^t, u_j^t)^k$ can be the summarized properties of social link (u_i^t, u_j^t), e.g., extended common neighbours.

Similarly, for a certain location link (u_i^t, l_j^t) in G^t, we can also use the extracted features for it from the target network, $\boldsymbol{x}(u_i^t, l_j^t)$, to predict its existence probability.

$$P(y(u_i^t, l_j^t) = 1 | G^t) = P(y(u_i^t, l_j^t) = 1 | \boldsymbol{x}(u_i^t, l_j^t))$$

If the target network is quite new, the features vectors extracted for both social links and location links can be very sparse, which can hardly build good link prediction models. Next, we will transfer information from the aligned source network to solve the problem.

3.3.2 Linkage Transfer across LBSNs

With the *anchor links*, we can locate users' corresponding accounts in the aligned source network, information in which can be transferred to the target network. Suppose, for instance, we want to predict a potential social link (u_i^t, u_j^t) by using information in both networks. By taking advantages of the *anchor links*, we can obtain the corresponding

Algorithm 1 TRAIL

Input: two aligned heterogeneous LBSNs, G^s, G^t.
existing social links and location links: E_s^t, E_l^t
anchor links between G^t and G^s: $A^{t,s}$, $A^{s,t}$
potential social links and location links: L_s^t, L_l^t
Output: the inferred labels and existence probabilities of links in L_s^t and L_l^t: $\hat{\mathcal{Y}}_s^t, \hat{\mathcal{P}}_s^t, \hat{\mathcal{Y}}_l^t, \hat{\mathcal{P}}_l^t$

```
construct training sets, test sets with E_s^t, E_l^t, L_s^t and L_l^t.
converge ← False
while converge is False do
    extract features x^t(E_s^t) and x^t(L_s^t) for social links in E_s^t and L_s^t from G^t.
    extract features x^s(E_s^t) and x^s(L_s^t) for social links in E_s^t and L_s^t from G^s by utilizing anchor links in A^{t,s}.
    C_s ← train([x^t(E_s^t)^T, x^s(E_s^t)^T, y^s(E_s^t)]^T, y^t(E_s^t))
    Ŷ_s^t, P̂_s^t ← C_s.classify([x^t(L_s^t)^T, x^s(L_s^t)^T, y^s(L_s^t)]^T)
    update G^t with Ŷ_s^t, P̂_s^t
    Accumulate information for locations
    extract features x^t(E_l^t) and x^t(L_l^t) for location links in E_l^t and L_l^t from G^t.
    extract features x^s(E_l^t) and x^s(L_l^t) for location links in E_l^t and L_l^t from G^s by utilizing anchor links in A^{t,s}.
    C_l ← train([x^t(E_l^t)^T, x^s(E_l^t)^T, y^s(E_l^t)]^T, y^t(E_l^t))
    Ŷ_l^t, P̂_l^t ← C_l.classify([x^t(L_l^t)^T, x^s(L_l^t)^T, y^s(L_l^t)]^T)
    update G^t with Ŷ_l^t, P̂_l^t
    if Ŷ_s^t, P̂_s^t, Ŷ_l^t, P̂_l^t all converge then
        converge ← True
    end if
end while
Return Ŷ_s^t, P̂_s^t, Ŷ_l^t, P̂_l^t
```

accounts of u_i^t and u_j^t in the aligned source network: u_i^s and u_j^s. If u_i^s and u_j^s both exist in G^s, then we will only transfer information related to the corresponding social link (u_i^s, u_j^s) in the aligned source network to the target network, which is represented as a feature vector extracted from G^s for link (u_i^s, u_j^s): $\boldsymbol{x}(u_i^s, u_j^s)$. We notice that the existence information of link (u_i^s, u_j^s) in the aligned source network, $y(u_i^s, u_j^s)$, is very useful, which is defined as *pseudo label* of link (u_i^t, u_j^t).

Definition 1 (Pseudo Label): Let (n_i^t, n_j^t) be a link in G^t, where n_i^t, n_j^t are nodes in it and they can be users, locations, etc., the corresponding link of (n_i^t, n_j^t) in the aligned source network G^s will be (n_i^s, n_j^s). The existence indicator of link (n_i^s, n_j^s) in G^s: $y(n_i^s, n_j^s)$ is defined as the *pseudo label* of link (n_i^t, n_j^t).

The *pseudo label* is used as an extra feature added to the extended feature vector, obtained by merging feature vectors extracted from G^t and G^s.

$$P(y(u_i^t, u_j^t) = 1 | G^t, G^s) = P\left(y(u_i^t, u_j^t) = 1 | \left[\boldsymbol{x}(u_i^t, u_j^t)^T, \boldsymbol{x}(u_i^s, u_j^s)^T, y(u_i^s, u_j^s)\right]^T\right)$$

Similarly, for a certain location link (u_i^t, l_j^t), we have

$$P(y(u_i^t, l_j^t) = 1 | G^t, G^s) = P\left(y(u_i^t, l_j^t) = 1 | \left[\boldsymbol{x}(u_i^t, l_j^t)^T, \boldsymbol{x}(u_i^s, l_j^s)^T, y(u_i^s, l_j^s)\right]^T\right)$$

Actually, we can also use *pseudo label* as the prediction result of link (n_i^t, n_j^t) in G^t and the method is called the NAIVE, which will be used as a baseline in our experiment.

Table 3: Properties of the Heterogeneous Social Networks

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

3.3.3 Collective Linkage Transfer across LBSNs

By using two aligned networks, the optimization equation will be revised as follows

$$\hat{\mathcal{Y}}_s^t, \hat{\mathcal{Y}}_l^t = arg \max_{\mathcal{Y}_s^t, \mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t) \times P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_l^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t)$$

For the given optimization equation, there are many different solutions. In this part, we will give an iterative method, TRAIL, to approach it, which can predict the social links and location links iteratively until convergence. Let τ be the τ_{th} iteration and the optimal label sets of social links and location links achieved in the τ_{th} iteration be $\hat{\mathcal{Y}}_s^{t(\tau)}$ and $\hat{\mathcal{Y}}_l^{t(\tau)}$, then

$$\hat{\mathcal{Y}}_s^{t(\tau)} = arg \max_{\mathcal{Y}_s^t} P(\mathbf{y}(\boldsymbol{L}_s^t) = \mathcal{Y}_s^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_s^t) = \hat{\mathcal{Y}}_s^{t(\tau-1)}, \mathbf{y}(\boldsymbol{L}_l^t) = \hat{\mathcal{Y}}_l^{t(\tau-1)})$$

$$\hat{\mathcal{Y}}_l^{t(\tau)} = arg \max_{\mathcal{Y}_l^t} P(\mathbf{y}(\boldsymbol{L}_l^t) = \mathcal{Y}_s^t | G^t, G^s, \mathbf{y}(\boldsymbol{L}_s^t) = \hat{\mathcal{Y}}_s^{t(\tau)}, \mathbf{y}(\boldsymbol{L}_l^t) = \hat{\mathcal{Y}}_l^{t(\tau-1)})$$

The pseudo code of TRAIL is available in Algorithm 1.

4. EXPERIMENTS

To testify the effectiveness of TRAIL in dealing with real-world aligned LBSNs, in this section, we will conduct extensive experiments on two real-world network datasets: Foursquare and Twitter.

4.1 Datasets Description

The networks used in this paper are: Foursquare, a famous LBSN, and Twitter, the hottest micro-blogging social network. Users play the key roles in both networks and they can follow/make friends with others, write posts online. Foursquare is constructed mainly around locations and can offer many location-related services, e.g., location check-ins and posting online reviews. Affected by the success of LBSNs, Twitter also starts to offer location-related services, e.g., online tweets can attach location check-ins. The locations in both of these social networks are represented as (latitude, longitude) pairs.

1. **Foursquare**: Users together with their online tips are crawled from Foursquare, whose numbers are 5,392 and 94,187 respectively. All these tips can attach location check-ins and the total number of locations crawled

from Foursquare is 38,921. The bidirectional friend link in Foursquare is decomposed into two unidirectional follow link and the original follow links are preserved. Detailed information about the Foursquare network is available in Table 3.

2. **Twitter**: Similarly, 5,223 users and all their online tweets are crawled from Twitter. The number of tweets crawled by us is 9,490,707, among which 615,515 tweets contain location check-ins and they account about 6.5% of all the tweets. The total number of locations obtained from the tweets is 297,182. The numbers of follow link among users, write link between users and tweets and the location link between tweets and locations are available in Table 3.

Both of these two networks are obtained from the webpages crawled with a shell script during the November, 2012. Meanwhile, if users display their Twitter accounts on their Foursquare homepages, we treat the connections between the accounts of Foursquare and Twitter as the anchor links between these two social networks. So, the anchor links are obtained by crawling the users' twitter IDs from their Foursquare homepages by using a shell script and the number of anchor links crawled is 3,388.

4.2 Experiment Settings

In this section, we will talk about the comparison methods, the evaluation measures and setups of the experiments in details.

4.2.1 Comparison Methods

To show that TRAIL can work well and outperform other state-of-art link prediction methods, we will compare TRAIL with many comparison methods, which can be divided into two categories: (1) supervised methods; (2) unsupervised methods.

- **TRAIL**: Method TRAIL is the link prediction method proposed in this paper. TRAIL is a supervised method and it can extract different categories of features for social links and location links from the heterogeneous networks. TRAIL can predict social links and location links simultaneously and can utilize both information in the target network and that transferred from the aligned source network.

- **Supervised Methods**: To show that predicting multiple kinds of links collectively can achieve better performance than predicting each kind of links independently, we compare TRAIL with SCAN (Supervised Cross Aligned Networks link prediction) [22], which is a supervised methods and can predict each kind of link independently across aligned networks. To demonstrate that using information in two aligned networks at the same time can achieve better performance than using one single network, we compare TRAIL with collective link prediction methods TRAILs (TRAnsfer heterogeneous lInks for LBSNs with Source network), TRAILt (TRAnsfer heterogeneous lInks for LBSNs with Target network) and compare SCAN with SCANs (Supervised Cross Aligned Link Prediction with Source network), SCANt (Supervised Cross Aligned Link Prediction with Target network). Methods TRAILt and SCANt utilize information in the target network only, while TRAILs and SCANs only use that transferred from the aligned source network.

- **Unsupervised Methods**: Some traditional unsupervised social link prediction methods are also used as the baseline methods to be compared with TRAIL, which include Common Neighbour (CN) [10]: $CN(x,y) = |\Gamma(x) \cap \Gamma(y)|$, Jaccard Coefficient ($JC$) [10]: $JC(x,y) = \frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|}$ and Adamic/Adar (AA) [1]: $AA(x,y) = |\Gamma(x)|\,|\Gamma(y)|$, where $\Gamma(x), \Gamma(y)$ are sets of neighbours of user x and y. A traditional unsupervised location recommendation method: FCF(Friend based Collaborative Filtering) [21]: $r_{i,j} = \frac{\sum_{u_k \in \Gamma(u_i)} r_{k,j} w_{i,k}}{\sum_{u_k \in \Gamma(u_i)} w_{i,k}}$ is used as the location link prediction baseline method, where $r_{i,j}$ is the rating of user u_i on location l_j and $w_{i,k}$ is the similarity of user u_i and u_k. Naive introduced before is used as a baseline method.

4.2.2 Evaluation Methods

To measure the effectiveness of these methods in predicting links, we will use two evaluation methods to assess their performance, which include AUC (Area Under ROC Curve) and *Accuracy*. Traditional unsupervised socal link and location link prediction methods *CN*, *JC*, *AA* and *FCF* can only output scores of potential links and their results are assessed by AUC only. Meanwhile, Naive can only produce labels of potential links and its performance is evaluated by *Accuracy* only. All other methods are evaluate by both AUC and *Accuracy*.

4.2.3 Setups

In the experiment, Foursquare is used as the target network and Twitter is used as the aligned source network. Existing social links and location links in Foursquare are used as the ground-truth.

We delete all the users' reposted tweets in Twitter about users' activities in Foursquare. Then, we group existing social links and location links in the target network as the positive social link set and positive location link set. Considering that users can visit a certain place multiple times which can , we delete all the duplicated location links and preserve on one copy. Sets of non-existent social links and location links collected from the target network are used as the negative social link set and negative location link set, which are of the same size as the positive sets. All these link sets are partitioned into two subsets by the 5-fold cross validation partitioned by links. To show that TRAIL can work well when the training pairs are quite limited, we use 1 fold as the training set and the remaining 4 folds as the test set. The target network studied in this paper is a new network and to simulate the different degrees of newness of it, we randomly sample a proportion of information in it to use under the control of parameter *remaining information rate* σ, which can include temporal activities, words used, locations visited etc. For example, if $\sigma = 0.1$, then the network is very new and only 10% of the information in original the network is available; if $\sigma = 0.8$, then the network is not that new as 80% of the information exists. To control the existence of *anchor links* between these two aligned social networks, we use another parameter: *anchor link sample rate* $\rho \in [0, 1.0]$ in the experiment. If $\rho = 0.0$, then these two networks are totally independent and have no anchor links between them; if $\rho = 1.0$, then these two networks are

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 is set as 1.0.

link	measure	methods	remaining information rates σ 0.1	0.2	0.3	0.4	0.5	0.6	0.7
social	AUC	TRAIL	**0.810±0.012**	**0.824±0.009**	**0.837±0.008**	**0.844±0.009**	**0.832±0.003**	**0.852±0.009**	**0.847±0.009**
		TRAILT	0.691±0.040	0.684±0.039	0.704±0.033	0.729±0.006	0.718±0.020	0.732±0.005	0.730±0.008
		TRAILS	0.572±0.007	0.578±0.007	0.580±0.004	0.575±0.012	0.580±0.011	0.583±0.009	0.578±0.009
		SCAN	0.772±0.050	0.788±0.004	0.811±0.009	0.830±0.005	0.809±0.004	0.825±0.008	0.824±0.012
		SCANT	0.524±0.023	0.559±0.008	0.559±0.017	0.554±0.044	0.630±0.008	0.599±0.007	0.627±0.004
		SCANS	0.583±0.005	0.579±0.003	0.583±0.010	0.562±0.005	0.579±0.004	0.585±0.003	0.584±0.003
		CN	0.494±0.002	0.500±0.015	0.504±0.006	0.496±0.012	0.495±0.018	0.491±0.015	0.489±0.018
		JC	0.497±0.003	0.503±0.004	0.501±0.002	0.502±0.010	0.496±0.008	0.496±0.019	0.492±0.008
		AA	0.494±0.002	0.499±0.014	0.501±0.006	0.494±0.012	0.492±0.018	0.489±0.015	0.493±0.022
	Accuracy	TRAIL	**0.855±0.002**	**0.849±0.004**	**0.850±0.008**	**0.854±0.005**	**0.850±0.003**	**0.851±0.001**	**0.852±0.004**
		TRAILT	0.622±0.046	0.627±0.036	0.655±0.022	0.676±0.009	0.674±0.019	0.677±0.004	0.679±0.008
		TRAILS	0.548±0.004	0.551±0.006	0.552±0.004	0.549±0.000	0.551±0.002	0.553±0.003	0.544±0.001
		SCAN	0.747±0.003	0.752±0.007	0.748±0.000	0.754±0.008	0.746±0.005	0.745±0.007	0.747±0.003
		SCANT	0.512±0.009	0.522±0.002	0.520±0.001	0.537±0.006	0.554±0.008	0.542±0.003	0.567±0.007
		SCANS	0.557±0.002	0.547±0.006	0.553±0.002	0.545±0.006	0.552±0.007	0.551±0.002	0.551±0.004
		NAIVE	0.525±0.014	0.526±0.006	0.525±0.008	0.526±0.007	0.525±0.013	0.525±0.009	0.525±0.013
location	AUC	TRAIL	**0.848±0.005**	**0.856±0.010**	**0.870±0.010**	**0.878±0.007**	**0.899±0.007**	**0.886±0.022**	**0.887±0.009**
		TRAILT	0.839±0.006	0.850±0.003	0.857±0.009	0.866±0.008	0.862±0.005	0.871±0.005	0.869±0.003
		TRAILS	0.631±0.003	0.632±0.002	0.631±0.001	0.634±0.001	0.634±0.002	0.634±0.002	0.635±0.001
		SCAN	0.712±0.010	0.757±0.002	0.758±0.009	0.770±0.005	0.775±0.005	0.784±0.004	0.792±0.003
		SCANT	0.676±0.009	0.711±0.005	0.730±0.005	0.749±0.003	0.756±0.001	0.763±0.005	0.769±0.003
		SCANS	0.633±0.003	0.633±0.003	0.633±0.001	0.636±0.001	0.637±0.000	0.633±0.001	0.634±0.001
		FCF	0.598±0.008	0.638±0.015	0.638±0.005	0.654±0.012	0.664±0.007	0.661±0.007	0.664±0.010
	Accuracy	TRAIL	**0.719±0.004**	**0.736±0.001**	**0.749±0.006**	**0.754±0.003**	**0.753±0.002**	**0.760±0.002**	**0.761±0.002**
		TRAILT	0.674±0.009	0.697±0.004	0.706±0.005	0.709±0.001	0.717±0.006	0.716±0.007	0.717±0.002
		TRAILS	0.536±0.003	0.527±0.001	0.537±0.005	0.553±0.003	0.560±0.002	0.565±0.000	0.566±0.001
		SCAN	0.658±0.000	0.670±0.002	0.682±0.001	0.697±0.003	0.699±0.003	0.723±0.003	0.723±0.007
		SCANT	0.610±0.001	0.623±0.001	0.631±0.001	0.647±0.001	0.653±0.002	0.671±0.003	0.676±0.002
		SCANS	0.536±0.025	0.531±0.008	0.535±0.002	0.547±0.004	0.557±0.004	0.565±0.001	0.566±0.001
		NAIVE	0.536±0.014	0.536±0.002	0.536±0.001	0.537±0.008	0.536±0.012	0.536±0.009	0.537±0.019

fully aligned; otherwise they are partially aligned. From the networks, different categories of features are extracted for each kind of pairs in the training set and test set. To solve the problem of lacking information in the target network, we transfer information from the aligned to the target network via the *anchor links*. The feature vectors obtained from both the target network and the aligned source network together with the *pseudo label* are merged into an expended feature vector to make use of information in both networks simultaneously. We train and classify social links and location links with iterative update until convergence or meet a certain maximum iteration number, which is set as 10 in our experiment.

4.3 Experiment Results

Experiment results are available in Table 4, which is under the setting that *anchor link sample rate* ρ is set as 1.0 and the *remaining information rate* σ changes from 0.1 to 0.8, and in Table 5, which is under the setting that *remaining information rate* σ is set as 1.0 and the *anchor link sample rate* ρ changes from 0.0 to 1.0 with an increasing step of 0.2. The results in these two tables can be divided into two parts: the first part is about the social links and the seciond part is about the location links, whose performance are evaluated by AUC and *Accuracy*.

In Table 4, compared with traditional unsupervised methods, like *FCF*, *CN*, *JC* and *AA*, supervised method TRAIL can substantially outperform them under the evaluation metric. For example, when $\sigma = 0.5$, the evaluation metric (AUC) of TRAIL is over 60% higher than that of *CN*, *JC*, *AA* and the evaluation metric (*Accuracy*) achieved by TRAIL is about 34% higher than that *FCF*. And compared with NAIVE, TRAIL can also perform far better, e.g., the evaluation metric (*Accuracy*) is over 40% higher than that of NAIVE when $\sigma = 0.5$ in Table. 4 By comparing TRAIL with SCAN, TRAILS with SCANS, TRAILT with SCANT in predicting both social links and location links, we can find that the methods predicting links collectively with iterative update can achieve better performance consistently than methods predicting each type of links independently. By comparing TRAIL with TRAILS and TRAILT, we can find that TRAIL using information in both the target network and the aligned source network can achieve better performance than using information in one single network, which can also be obtained by comparing SCAN with SCANS and SCANT. Similar results can be obtained in Table 5 as the *anchor link sample rate* ρ changes.

So, TRAIL can outperform all these state-of-art supervised baseline methods and traditional unsupervised methods for networks of different *remaining information rate* and different *anchor link sample rate* in Table 4 and Table 5, when the training cases is very limited under the evaluation of AUC and *Accuracy*.

In addition, the prediction result of method TRAIL can also converge very quickly. In Figure 4, we show the social and location link prediction results of TRAIL evaluated by *Accuracy* and AUC. Figures 4(a)- 4(d) are the results obtained by TRAIL when $\sigma = 0.5$ and $\rho = 1.0$. While, Figures 4(e)- 4(h) show the results obtained by TRAIL when $\sigma = 1.0$ and $\rho = 0.5$. We can find that all the results can converge quickly in less than 5 iterations.

5. RELATED WORK

Link prediction first proposed by D. Liben-Nowell et al. [14] has become a significant research topic in social network studies in recent years. M. A. Hasan et al. [9] are the first to study the link prediction problem as a supervised problem. However, their method is based on a homogeneous

Table 5: Performance comparison of different methods for inferring social and location links for Foursquare of different anchor link sample rates. The remaining informaiton rate σ is set as 1.0.

link	measure	methods	anchor link sample rates ρ 0.0	0.2	0.4	0.6	0.8	1.0
social	AUC	TRAIL	**0.712±0.004**	**0.733±0.019**	**0.761±0.017**	**0.782±0.007**	**0.821±0.012**	**0.855±0.008**
		TRAIL$_T$	0.712±0.012	0.711±0.007	0.711±0.012	0.711±0.010	0.712±0.014	0.712±0.005
		TRAIL$_S$	0.500±0.000	0.507±0.005	0.524±0.005	0.555±0.036	0.577±0.028	0.583±0.015
		SCAN	0.603±0.020	0.621±0.036	0.539±0.022	0.664±0.026	0.748±0.027	0.827±0.002
		SCAN$_T$	0.603±0.009	0.603±0.014	0.603±0.016	0.603±0.027	0.603±0.006	0.604±0.011
		SCAN$_S$	0.500±0.000	0.496±0.001	0.513±0.013	0.515±0.015	0.570±0.060	0.572±0.007
		CN	0.525±0.000	0.525±0.008	0.524±0.013	0.525±0.005	0.525±0.013	0.525±0.007
		JC	0.527±0.008	0.527±0.011	0.527±0.010	0.528±0.002	0.527±0.016	0.528±0.009
		AA	0.493±0.006	0.490±0.006	0.490±0.012	0.490±0.009	0.493±0.013	0.490±0.006
	Accuracy	TRAIL	**0.654±0.014**	**0.746±0.009**	**0.756±0.009**	**0.764±0.008**	**0.768±0.012**	**0.839±0.002**
		TRAIL$_T$	0.655±0.004	0.653±0.008	0.655±0.014	0.655±0.008	0.655±0.008	0.655±0.005
		TRAIL$_S$	0.500±0.000	0.501±0.003	0.535±0.009	0.529±0.006	0.535±0.004	0.545±0.014
		SCAN	0.554±0.028	0.567±0.009	0.563±0.007	0.605±0.014	0.656±0.011	0.748±0.012
		SCAN$_T$	0.553±0.002	0.553±0.004	0.553±0.003	0.554±0.002	0.553±0.001	0.553±0.003
		SCAN$_S$	0.500±0.000	0.498±0.003	0.515±0.008	0.529±0.003	0.536±0.003	0.541±0.005
		NAIVE	0.500±0.000	0.508±0.001	0.514±0.006	0.517±0.002	0.519±0.003	0.526±0.000
location	AUC	TRAIL	**0.871±0.020**	**0.876±0.011**	**0.891±0.006**	**0.881±0.028**	**0.916±0.016**	**0.925±0.007**
		TRAIL$_T$	0.871±0.015	0.872±0.004	0.872±0.013	0.872±0.003	0.872±0.017	0.872±0.014
		TRAIL$_S$	0.500±0.000	0.492±0.002	0.479±0.004	0.504±0.002	0.580±0.001	0.652±0.003
		SCAN	0.745±0.005	0.746±0.011	0.773±0.010	0.788±0.012	0.796±0.016	0.797±0.009
		SCAN$_T$	0.745±0.021	0.744±0.011	0.745±0.025	0.744±0.020	0.743±0.011	0.744±0.010
		SCAN$_S$	0.500±0.000	0.490±0.002	0.481±0.002	0.504±0.001	0.578±0.005	0.651±0.005
		FCF	0.682±0.006	0.683±0.002	0.682±0.007	0.683±0.002	0.683±0.006	0.682±0.003
	Accuracy	TRAIL	**0.734±0.008**	**0.754±0.005**	**0.765±0.006**	**0.775±0.003**	**0.789±0.008**	**0.797±0.010**
		TRAIL$_T$	0.735±0.002	0.734±0.007	0.734±0.007	0.734±0.006	0.735±0.004	0.735±0.004
		TRAIL$_S$	0.500±0.000	0.509±0.003	0.514±0.006	0.511±0.001	0.533±0.000	0.569±0.001
		SCAN	0.731±0.002	0.753±0.001	0.754±0.002	0.755±0.002	0.767±0.002	0.777±0.003
		SCAN$_T$	0.732±0.013	0.732±0.010	0.732±0.016	0.732±0.009	0.732±0.004	0.732±0.004
		SCAN$_S$	0.500±0.000	0.511±0.002	0.516±0.006	0.517±0.005	0.534±0.001	0.568±0.002
		NAIVE	0.500±0.000	0.509±0.001	0.517±0.001	0.517±0.005	0.525±0.010	0.536±0.004

network and many networks are heterogeneous nowadays. Y. Sun et al. [17] propose a meta path-based prediction model to predict co-author relationship in the heterogeneous bibliographic network.

As the Location-based social networks (LBSNs) are becoming more and more popular in recent years, many works have been done on such kind of social networks. M. Ye et al. [20] study the semantic annotation of locations in location-based social networks. Meanwhile, some works have also been done on predicting links for LBSNs. S. Scellato et al. [16] predict social links by using heterogeneous information in the network. D. Wang et al. [19] try to predict social links by utilizing the moving pattern of users. These works are all predicting social links and some other works focus on predicting location links. M. Ye et al. [21] study location recommendation problem by using friend-based collaborative filtering method. E. Cho et al. [5] regard the location recommendation problem as a supervised link prediction problem. Y. Zheng et al. propose to mine interesting locations and travel sequences from GPS trajectories in [23].

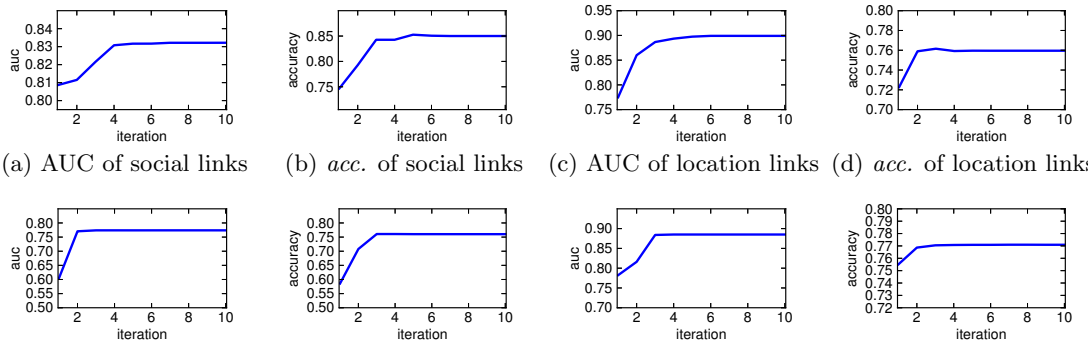
Most existing works focus on predicting one single type of link but some other works can predict multiple kinds of links simultaneously. I. Konstas et al. [12] propose to use collaborative filtering method to recommend multiple kinds of links for networks. While, F. Fouss et al. [7] use a traditional method: random walk to predict multiple kinds of links. B. Cao et al. [4] propose to predict links in different domains simultaneously with transfer learning. Some works propose to combine link prediction with other classification tasks. For example, M. Bilgic et al. [3] propose to do collective classification and link prediction for networks simultaneously.

All these works are based on one single network but many researchers start to shift their attention to multiple networks. Tang et al. [18] focus on inferring the type of links over multiple heterogeneous networks. Z. Lu et al. [15] propose to do supervised link prediction by using multiple information sources. Y. Dong et al. [6] propose to predict links across heterogeneous social networks. To deal with the differences in information distributions of multiple networks, G. Qi et al. [8] propose to use biased cross-network sampling to predict links across networks.

When studying multiple social networks, the first problem will be how to construct the bridges between networks to transfer information across them. X. Kong et al. [11] propose a method to infer the links between the accounts owned by the same users in different social networks and they are the first one to introduce the concepts of "anchor links" and "multiple aligned heterogeneous networks". J. Zhang et al. propose to predict social links for new users with information transferred from aligned source network through anchor links to solve the cold start problem in [22] and they are the first one to propose to transfer information across "aligned networks" through "anchor links".

6. CONCLUSION

In this paper, we study the collectively link prediction problem for new networks across aligned LBSNs and the links to be predicted in this paper include both social links and location links. We propose method TRAIL to deal with the challenges and solve the problem. TRAIL can accumulate information for locations and can extract different categories of features for both social links and location links from the networks. By taking advantage of the *anchor links*, TRAIL can utilize the information transferred from the aligned source network to ease the information sparsity problem. TRAIL can predict social links and location links by iterative updating the network with newly predicted re-



(a) AUC of social links (b) *acc.* of social links (c) AUC of location links (d) *acc.* of location links

(e) AUC of social links (f) *acc.* of social links (g) AUC of location links (h) *acc.* of location links

Figure 4: Social link and location link prediction results of each iteration under the evaluation of AUC and *Accuracy*. (a)-(d) are the results when $\sigma = 0.5$ and $\rho = 1.0$; (e)-(h) are the same results when $\sigma = 1.0$ and $\rho = 0.5$, where σ is the remaining information rate and ρ is the anchor link sample rate.

sults. Extensive experiments conducted on two real-world data sets demonstrate that TRAIL can achieve good prediction result for the target network of different degrees of newness and different anchor link sample rates.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, pages 211–230, 2001.

[2] A. E. Aladağ and C. Erten. Spinal: scalable protein interaction network alignment. Bioinformatics, pages 917–924, 2013.

[3] M. Bilgic, G. M. Namata, and L. Getoor. Combining collective classification and link prediction. In *ICDMW*, pages 381–386, 2007.

[4] B. Cao, N. Liu, and Q. Yang. Transfer learning for collective link prediction in multiple heterogenous domains. In *ICML*, pages 159–166, 2010.

[5] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[6] Y. Dong, J. Tang, S. Wu, J. Tian, N. Chawla, J. Rao, and H. Cao. Link prediction and recommendation across heterogeneous social networks. In *ICDM*, pages 181–190, 2012.

[7] 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*, pages 355–369, 2007.

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

[9] M. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM*, pages 71–83, 2006.

[10] M. A. Hasan and M. J. Zaki. A survey of link prediction in social networks. In *Social Network Data Analytics*, pages 243–275. Springer, 2011.

[11] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, pages 179–188, 2013.

[12] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR*, pages 195–202, 2009.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[14] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, pages 556–559, 2003.

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

[16] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD*, pages 1046–1054, 2011.

[17] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *ASONAM*, pages 121–128, 2011.

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

[19] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. Barabasi. Human mobility, social ties, and link prediction. In *KDD*, pages 1100–1108, 2011.

[20] M. Ye, D. Shou, W. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *KDD*, pages 520–528, 2011.

[21] M. Ye, P. Yin, and W. Lee. Location recommendation for location-based social networks. In *GIS*, pages 458–461, 2010.

[22] J. Zhang, X. Kong, and P. Yu. Predicting social links for new users across aligned heterogeneous social networks. In *ICDM*, pages 1289–1294, 2013.

[23] Y. Zheng, L. Zhang, X. Xie, and W. Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, pages 791–800, 2009.