Knowl Inf Syst
https://doi.org/10.1007/s10115-018-1210-1

REGULAR PAPER

Integrated anchor and social link predictions across multiple social networks

Qianyi Zhan[1,4] · **Jiawei Zhang**[2] · **Philip S. Yu**[3]

Received: 26 October 2016 / Revised: 24 January 2018 / Accepted: 6 May 2018

Abstract In recent years, various online social networks offering specific services have gained great popularity and success. To enjoy more online social services, some users can be involved in multiple social networks simultaneously. A challenging problem in social network studies is to identify the common users across networks to gain better understanding of user behavior. This is referred to as the anchor link prediction problem. Meanwhile, across these partially aligned social networks, users can be connected by different kinds of links, e.g., social links among users in one single network and anchor links between accounts of the shared users in different networks. Many different link prediction methods have been proposed so far to predict each type of links separately. In this paper, we want to predict the formation of social links among users in the target network as well as anchor links aligning the target network with other external social networks. The problem is formally defined as the "collective link identification" problem. Predicting the formation of links in social networks with traditional link prediction methods, e.g., classification-based methods, can be very challenging. The reason is that, from the network, we can only obtain the formed links (i.e., positive links) but no information about the links that will never be formed (i.e., negative links). To solve the collective link identification problem, a unified link prediction frame-

A preliminary version of this work appeared in: Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI '15), 2015.

✉ Qianyi Zhan
zhanqianyi@gmail.com

Jiawei Zhang
jzhang@cs.fsu.edu

Philip S. Yu
psyu@cs.uic.edu

1 School of Digital Media, Jiangnan University, Wuxi, China

2 IFM Lab, Department of Computer Science, Florida State University, Tallahassee, FL, USA

3 University of Illinois at Chicago, Chicago, IL, USA

4 Jiangsu Key Laboratory of Media Design and Software Technology, Wuxi, China

work, collective link fusion (CLF) is proposed in this paper, which consists of two phases: step (1) collective link prediction of anchor and social links with positive and unlabeled learning techniques, and step (2) propagation of predicted links across the partially aligned "probabilistic networks" with collective random walk. Extensive experiments conducted on two real-world partially aligned networks demonstrate that CLF can perform very well in predicting social and anchor links concurrently.

Keywords Link prediction · Transfer learning · PU learning · Data mining

1 Introduction

In recent years, there has been a surge of interest in studying multiple online social networks simultaneously [14,34,35,39]. In part, this interest is driven by the burgeoning growth of various online social networks. Meanwhile, to enjoy more social network services, users nowadays are usually involved in multiple online social networks at the same time, e.g., Foursquare, Facebook and Twitter. These shared users of different online social networks are defined as the "anchor users" [14] as they can act like "anchors" aligning the networks they participate in, while the remaining unshared users are called the "non-anchor users". Across partially aligned online social networks, users are connected by various kinds of links: (1) intra-network links, i.e., the *social links* among users within networks; and (2) inter-network links, i.e., the *anchor links* [14] connecting the accounts of the anchor users across different networks.

Predicting the formation of links in online social networks has been a hot research topic in recent years and many different kinds of link prediction methods have been proposed so far, e.g., classification-based methods [2,9,34,35,39] built with links in the networks, where "formed links" and links which will "never be formed" can be labeled as "positive links" and "negative links" respectively. Actually, when predicting the formation of links in social networks, we can only have the formed links (i.e., positive links) but no information on links that will never be formed (i.e., negative links). Predicting the formation of links merely with the existing formed links (i.e., positive links) is formally defined as the *link formation prediction* problems. Many important services across aligned social networks can be cast as link formation prediction tasks, e.g., anchor user identification [14] across networks can be converted into an inter-network anchor link formation prediction task, friend recommendation [34,35] within one network can be regarded as an intra-network social link formation prediction task.

Furthermore, as discovered in [35], multiple link prediction tasks in the same networks may be strongly correlated and can be done concurrently to improve the prediction results through mutual re-enhancement. Formally, we define the process of simultaneously predicting multiple kinds of links among users across multiple partially aligned networks merely with positive examples as the *collective link identification problem*. In this paper, the *collective link identification* problem covers the following two different link formation prediction tasks simultaneously:

- *Social Link Formation Prediction* discover social links to be formed among users in the future in a network that we target on.
- *Anchor Link Formation Prediction* uncover the hidden anchor links connecting accounts of anchor users between the target network and other aligned social networks.

Figure 1 gives an example of the *collective identification* problem. As illustrated in Fig. 1a, we have two partially aligned social networks, both of which contain some users. The network

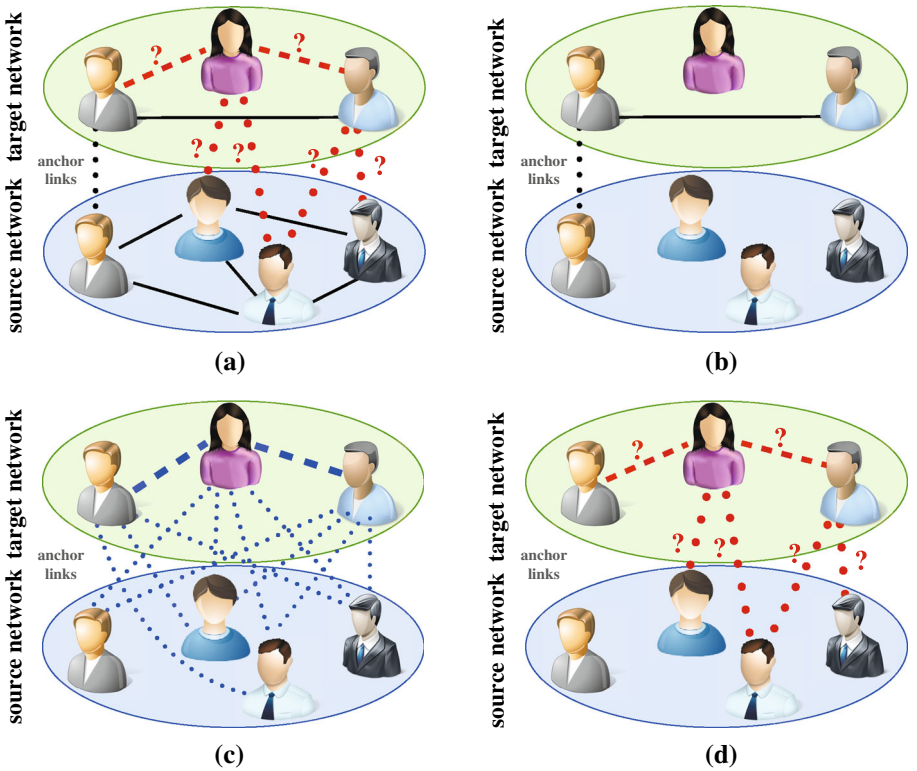


Fig. 1 Collective link identification across partially aligned networks. **a** Input networks, **b** existing anchor and social links, **c** unconnected anchor and social links, **d** anchor and social links to be predicted (color figure online)

at the top, which is new and has very few social links, is defined as the *target network*, while the network at the bottom, which is well developed and contains many social links, is defined as the *source network*. Some common users of the target and source networks are connected by anchor links. In Fig. 1b, black dotted links across the networks are the existing anchor links and the black solid lines in the target network are the existing social links. All the possible anchor and social links among users in the target network and those across different networks in the target network except the black ones can be viewed as the unconnected potential anchor/social links, which are the blue dotted/dashed lines shown in Fig. 1c. Furthermore, the red dotted/dashed lines are the anchor/social links to be predicted, which are shown in Fig. 1d.

These two link formation prediction tasks covered in the *collective link identification* problem are both of great importance for online social networks, especially when the target network is very new and social connections among users in it are sparse: (1) *anchor link formation prediction* can add more inter-network connections between different networks, which is a crucial prerequisite for many cross-network applications, e.g., friend recommendation and information diffusion across social networks, (2) *social link formation prediction* can add more intra-network social connections among users in the target network, which is helpful for inter-network anchor link identification [14].

To support such claims, we also conduct some statistical investigations on the information distributions of two real-world partially aligned social networks: Foursquare and Twitter,

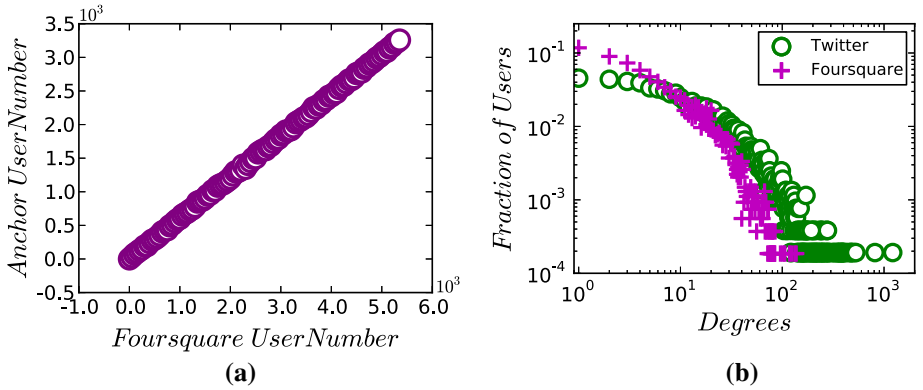


Fig. 2 Some statistical information about two partially aligned networks: Foursquare and Twitter. **a** The anchor user number in Foursquare. **b** Degree distribution of users in both Foursquare and Twitter

where results are shown in Fig. 2. As shown in Fig. 2a, for a given random sample of Foursquare users, about 60% of whom are also involved in Twitter (i.e., anchor users). However, the remaining 40% Foursquare users' aligned accounts in Twitter are unknown and discovering the hidden anchor links for these 40% non-anchor users can be critical to improve the quality of services for these users. Fig. 2b shows the degree distributions (i.e., number of social links) of users in both Twitter and Foursquare, which both follow the power law. In addition, as the data used in the experiment, we crawled 5223 users' information in Twitter, and they posted 9,490,707 tweets, checked-in 297,182 locations and generated 164,296 social links. It means each user in Twitter has 1817.1 tweets, 56.9 locations and 31.5 social degree averagely. While according to the experiment data, each user in Foursquare only has 9.04 tweets, 7.2 locations and 14.3 social degree averagely. It shows that social information of users in Foursquare is much sparser than that in Twitter, and from Fig. 2b, we can get the same observation too. Therefore, information transferred from Twitter can be helpful for the social link prediction task in Foursquare [34,35].

The *collective link identification* problem studied in this paper is novel and conventional classification-based link prediction models [2] cannot be applied to solve it directly due to the following challenges. Firstly, in traditional classification-based methods [2,9], links in social networks are assigned with different labels according to their physical meanings, e.g., friends vs enemies [27], trust vs distrust [30], positive attitude vs negative attitude [31], etc. However, when predicting the formation of links in social networks, we can only have the formed links (i.e., positive links) but no information about links that will never be formed (i.e., negative links). Secondly, traditional classification-based link prediction models are based on the assumption that information in the target network is sufficient to build effective models. This assumption will be seriously violated when the network is new, available information in which would be very sparse [35]. Furthermore, traditional classification-based link prediction models mostly focus on predicting one single type of links without considering the correlation between different link prediction tasks. In Table 1, we show the comparisons of the *collective link identification* problem with some correlated problems in many aspects, and more detailed description is available in Sect. 5.

Despite its importance and novelty, the *collective link identification* problem is very challenging to solve due to the following reasons:

Table 1 Summary of related problems

Property	Collective link prediction	[39]	[22]	[35]	[34]
# Networks	Multiple	Multiple	Single	Multiple	Multiple
Network type	Heterogeneous	Heterogeneous	Homogeneous	Heterogeneous	Heterogeneous
Network alignment	Partially	Partially	No	Partially	Fully
PU learning	Yes	Yes	No	No	No
Link fusion	Across networks	Across networks	n/a	n/a	n/a
Predicted links	Social links in target network + anchor links across networks	Social links in target network	Social links	Social links in target network	Social links in target network

- *Lack of Negative Labeled Links* Though there are plenty of missing links, it is difficult to tell which links will never been formed (i.e. negative labeled links) and which are not formed temporarily. A reasonable way to address this problem is using only positive links to do prediction. However few work about it has been done before, and supervised link prediction methods merely with positive links is still an open problem to this context so far.
- *Lacking Social Features for Anchor Links* Information of users who form anchor links is located in two different networks and is disjoint actually. Existing social features of links defined for single-network setting cannot be applied to anchor links across multiple networks directly.
- *Partial Alignment of Networks* Networks studied in this paper are partially aligned and the new network that we target on contains both *anchor users* and *non-anchor users*. Few works have been done to transfer information for non-anchor users in the new target network yet.
- *Collective Link Prediction* The *collective link identification* problem studied in this paper covers two tasks simultaneously. Analysis and utilization of the correlation between these two tasks to help improve the prediction results mutually is critical.

To solve these challenges, a two-phase link prediction framework, CLF, is proposed in this paper. In the first step, CLF predicts anchor and social links independently by (1) formulating the link formation problem with positive links as a Positive and Unlabeled (PU [19]) learning problem, and (2) transferring information for social links formed by anchor users from other source networks to the target network via existing anchor links. In the second step, CLF propagates information across the partially aligned "probabilistic networks" constructed with the prediction results of the first step. With *collective random walk*, CLF can (1) transfer information for both anchor users and non-anchor users, (2) fuse newly predicted results of both anchor and social links for mutual enhancement, and (3) control the proportion of information diffused across networks.

This paper is organized as follows. In Sect. 2, we will give the problem formulation. Methods will be introduced in Sect. 3. Extensive experiments are done in Sect. 4. Section 5 is about the related works. Finally, in Sect. 6, we will conclude the paper.

2 Problem formulation

2.1 Partially aligned heterogeneous networks

In this paper, we will follow the definitions of "anchor users", "anchor links", etc., proposed in [14]. Different from [14], the major assumptions about the aligned networks in this paper is *partial alignment of networks*: fully aligned networks rarely exist in the real world and networks studied in this paper are partially aligned [39].

Aligned social networks first introduced in [14] are defined as $\mathcal{G} = ((G^1, G^2, \ldots, G^n), (A^{1,2}, A^{1,3}, \ldots, A^{1,n}, A^{2,3}, \ldots, A^{(n-1),n}))$, where $G^i = (\mathcal{V}^i, \mathcal{E}^i), i \in \{1, 2, \ldots, n\}$ is a heterogeneous network containing multiple kinds of nodes and complex links and $A^{i,j}$ is the set of undirected anchor links between G^i and G^j. If users in G^i and G^j are all connected by anchor links in $A^{i,j}$, then networks G^i and G^j are fully aligned. Otherwise, G^i and G^j are partially aligned.

Link (u, v) is an anchor link between network G^i and G^j iff $(u \in U^i) \wedge (v \in U^j) \wedge$ (u and v are the accounts of the same user), where U^i and U^j are the user sets of G^i and G^j, respectively. If all users in one network (e.g., G^i) are connected by anchor links with users in another network (e.g., G^j) and all users in G^j are connected by anchor links with users in G^i as well, then networks G^i and G^j are fully aligned. Otherwise, G^i and G^j are partially aligned.

The partially aligned heterogeneous social networks studied in this paper are Foursquare and Twitter, which are used as the target and source networks respectively. According to the definition of aligned heterogeneous networks in [14], networks studied in this paper can be formulated as $\mathcal{G} = ((G^t, G^s), (A^{t,s}))$, where G^t, G^s are the target network and source network, respectively, and $A^{t,s}$ is the set of undirected anchor links between G^t and G^s.

Users in both Foursquare and Twitter can write posts online, which can contain text information, timestamps and attached location check-ins. As a result, G^t and G^s can both be formulated as $G = (\{U \cup L \cup W \cup T\}, \{E_{u,u} \cup E_{u,l} \cup E_{u,w} \cup E_{u,t}\})$ where U, L, W and T are the sets of user, location, word and timestamp nodes in the network and $E_{u,u}$, $E_{u,l}$, $E_{u,w}$ and $E_{u,t}$ are the sets of links among users and those between locations, words, timestamps and users.

2.2 Integrated PU link prediction problem

The *collective link identification problem* studied in this paper includes the simultaneous inference of both anchor links between G^t and G^s and social links in G^t merely with the positive links. Across aligned networks, in addition to the positive links, we can identify lots of unconnected links as well. For example, let $E^t_{u,u}$ and U^t be the sets of existing links and users in G^t, we can represent the existing and unconnected social links to be $E^t_{u,u}$ and $U^t \times U^t - E^t_{u,u}$ respectively. If these unconnected links are viewed as "unlabeled links", then the link formation prediction problem with positive and unlabeled links can be formally defined as *PU link prediction problems*. In this paper, we formulate the *collective link identification* problem as the *integrated PU link prediction* problem, which covers the (1) *PU anchor link prediction*; (2) *PU social link prediction* simultaneously.

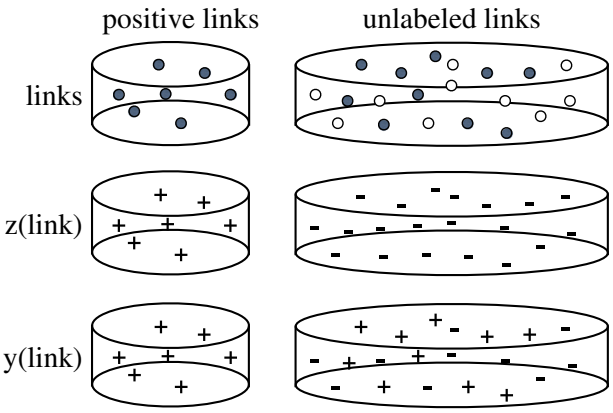


Fig. 3 Example of connection states and labels of links in PU link prediction

3 Proposed methods

3.1 Preliminary

As introduced in [39], from networks, we can extract both existing and unconnected links. To differentiate these links, a term named "*connection state*": $z \in \{-1, +1\}$ was introduced in [39]. If a certain link (u, v) is an existing link in the network, then $z(u, v) = +1$; if (u, v) is an unconnected link, then $z(u, v) = -1$. Meanwhile, besides the "*connection state*", all the links can also have their own *labels*, $y \in \{-1, +1\}$, e.g., friends vs enemies, trust vs distrust, formed vs will never be formed, etc. In this paper, if link (u, v) has been/will be formed, then $y(u, v) = +1$; if (u, v) will never be formed, then $y(u, v) = -1$. As shown in Fig. 3, for all existing links in the network, their connection states z and labels y are all $+1$, while the connection states z of all initially unconnected links are -1 but the labels y of these unconnected links can be either $+1$ or -1, as the unconnected links include both links to be formed and links that will never be formed. These unconnected links are referred to as the unlabeled links in the PU link prediction.

A PU social link prediction model applying spy technique [19] to extract reliable negative links from the unconnected link set was proposed in [39]. However, the correlation between links' *connection state* and *labels* is not clearly addressed in [39], which will be analyzed and derived in details in this paper. A new PU link prediction method based on the analysis and derivations will be introduced in the next subsection, which can be applied to infer both anchor and social links across multiple partially aligned networks.

3.2 Link formation probability inference

For each anchor/social link, a set of features (e.g., the features proposed in [14,34]) can be extracted from the networks, e.g., the feature vector extracted for certain anchor/social link (u, v) can be represented as $\mathbf{x}(u, v)$. As a result, each anchor/social link (u, v) in the networks can be denoted as a tuple $\langle \boldsymbol{x}(u, v), y(u, v), z(u, v)\rangle$. Let $p(\boldsymbol{x}, y, z)$ be the joint distribution of $\boldsymbol{x}$, y and z. As shown in Fig. 3, all the existing links ($z = 1$) are positive links ($y = 1$):

$$p(y = 1|\boldsymbol{x}, z = 1) = p(y = 1|z = 1) = 1.0.$$

A basic assumption about PU link prediction is that *the existing positive links are randomly sampled from the whole positive link set*, which means that for two arbitrary positive links (u_1, v_1) and (u_2, v_2) we have

$$\begin{aligned} p(z(u_1, v_1) = 1 | \boldsymbol{x}(u_1, v_1), y(u_1, v_1) = 1) \\ = p(z(u_2, v_2) = 1 | \boldsymbol{x}(u_2, v_2), y(u_2, v_2) = 1). \end{aligned}$$

In other words, the conditional distribution $p(z = 1|\boldsymbol{x}, y = 1)$ is independent of variable $\boldsymbol{x}$, i.e.,

$$\begin{aligned} p(z = 1|y = 1) = & \sum_{\text{link} \in \mathcal{G}} p(z = 1|\boldsymbol{x}(\text{link}), y = 1) p(\boldsymbol{x}(\text{link})|y = 1) \\ = & p(z = 1|\boldsymbol{x}, y = 1) \cdot \sum_{\text{link} \in \mathcal{G}} p(\boldsymbol{x}(\text{link})|y = 1) \\ = & p(z = 1|\boldsymbol{x}, y = 1). \end{aligned}$$

Meanwhile, the probabilities that link l is predicted to be "existing" ($z = +1$) and "formed" ($y = +1$) can be defined as the "*existence probability*" (i.e., $p(z = 1|\boldsymbol{x})$) and "*formation probability*" (i.e., $p(y = 1|\boldsymbol{x})$), respectively, as introduced in [39].

Definition 1 (*Existence Probability*) Probability that a link originally exists in the networks is formally defined as the *existence probability* of the link, $p(z = 1|\boldsymbol{x})$.

Definition 2 (*Formation Probability*) Probability that a link will be formed is formally defined as the *formation probability* of the link, $p(y = 1|\boldsymbol{x})$.

However, [39] fails to study the correlation between links' "*existence probability*" and "*formation probability*", which can be represented as follows:

$$\begin{aligned} p(z = 1|\boldsymbol{x}) &= p(z = 1|\boldsymbol{x}) \cdot p(y = 1|\boldsymbol{x}, z = 1) = p(y = 1, z = 1|\boldsymbol{x}) \\ &= p(y = 1|\boldsymbol{x}) \cdot p(z = 1|\boldsymbol{x}, y = 1) \\ &= p(y = 1|\boldsymbol{x}) \cdot p(z = 1|y = 1). \end{aligned}$$

As a result, links' *formation probabilities* can be inferred from their *existence probabilities* if we know $p(z = 1|y = 1)$ in advance.

Definition 3 (*Bridging Probability*) $p(z = 1|y = 1)$ is formally defined as the *bridging probability* between the existence probability and the formation probability.

The bridging probability can be inferred with the binary classification models built with the existing ($z = +1$) and unconnected ($z = -1$) links [4]. We split all the existing and unconnected links into "training set" and "validation set" via cross-validation. Classification models built based on the training set can be applied to the validation set. Let Pos be the subset of links that are positive in the validation set. We use the *method of moments* to estimate the bridging probability, and here is the inference equation:

$$\begin{aligned} & p(z = 1|y = 1) \\ & = \frac{1}{|\text{Pos}|} \sum_{\text{link} \in \text{Pos}} p(z = 1|y = 1) && \text{moment estimator} \\ & = \frac{1}{|\text{Pos}|} \sum_{\text{link} \in \text{Pos}} p(z = 1|\mathbf{x}, y = 1), && p(z = 1|y = 1) = p(z = 1|\mathbf{x}, y = 1) \end{aligned}$$

Algorithm 1 PU Link Prediction (MLP)

Input: heterogeneous social networks: G
sets of positive and unlabeled: P, U
anchor link validation and test sets: V and T
Output: existence probabilities of links in V and T: $p(y(V) = \mathbf{1}|\mathbf{x}(V))$ and $p(y(T) = \mathbf{1}|\mathbf{x}(T))$
1: extract feature vectors, $\mathbf{x}$, for links in P, U, V and T
2: assign links in P, U with labels $+\mathbf{1}$, $-\mathbf{1}$ respectively
3: $SVM.train([\mathbf{x}(P), \mathbf{x}(U)], [1, \cdots, 1, -1 \cdots, -1]^T)$
4: $p(z(V) = \mathbf{1}|\mathbf{x}(V)) = SVM.classify(\mathbf{x}(V))$
5: calculate $p(z = 1|y = 1)$ with the **Bridging Probability Inference Equation**
6: $p(z(T) = \mathbf{1}|\mathbf{x}(T)) = c.classify(\mathbf{x}(T))$
7: $p(y(T) = \mathbf{1}|\mathbf{x}(T)) = \frac{p(z(T)=\mathbf{1}|\mathbf{x}(T))}{p(z=1|y=1)}$
8: **return** $p(y(V) = \mathbf{1}|\mathbf{x}(V))$ and $p(y(T) = \mathbf{1}|\mathbf{x}(T))$

$$
\begin{aligned}
&= \frac{1}{|\text{Pos}|} \sum_{\text{link}\in\text{Pos}} (p(z = 1|\mathbf{x}, y = 1) \cdot 1 + 0 \cdot 0 \\
&= \frac{1}{|\text{Pos}|} \sum_{\text{link}\in\text{Pos}} (p(z = 1|\mathbf{x}, y = 1)p(y = 1|\mathbf{x}) \\
&\qquad + p(z = 1|\mathbf{x}, y = -1)p(y = -1|\mathbf{x})) \qquad \text{link} \in \text{Pos} \\
&= \frac{1}{|\text{Pos}|} \sum_{\text{link}\in\text{Pos}} p(z = 1|\mathbf{x}). \qquad \text{law of total probability}
\end{aligned}
$$

As a result, the average existence probabilities of links in Pos works as an estimator of the bridging probability, which clearly clarifies the correlation between link's existence probability and formation probability. Based on the inferred bridging probability $p(z = 1|y = 1)$, we can predict the formation probabilities of anchor and social links based on their existence probabilities, which is totally different from the spy technique introduced in [39].

The pseudo code of the *PU Link Prediction* (MLP) method is available in Algorithm 1.

3.3 Inter-network social features for anchor links

Social information of users who form certain anchor links is located in different networks and is disjoint as a result. Consider a certain anchor link (u, v) between G^t and G^s, for example, we can get the neighbors of u and v from G^t and G^s, which are $\Gamma(u)$ and $\Gamma(v)$ respectively. User u and v are in two different networks and their neighbors $\Gamma(u)$ and $\Gamma(v)$ are in different networks as well, i.e., $\Gamma(u) \cap \Gamma(v) = \emptyset$. In this case, traditional social features, like "Common Neighbor" [9], "Jaccard's Coefficient" [9] and "Adamic/Adar" [1], will not work. To solve the problem, we use the extended definition of these three social features proposed in [14] instead, which are named as the *inter-network social features* for anchor links in this paper.

- *Extended Common Neighbor*, which denotes the number shared "neighbors" of u and v: $ECN(u, v) = \left|\Gamma(u) \bigcap_{A^{t,s}} \Gamma(v)\right|$, and $\left|\Gamma(u) \bigcap_{A^{t,s}} \Gamma(v)\right| = \left|\{(u', v') \in A^{t,s}, u' \in \Gamma(u), v' \in \Gamma(v)\}\right|$.
- *Extended Jaccard's Coefficient*: $EJC(u, v)$ takes the size of $\Gamma(u)$ and $\Gamma(v)$ into account, considering that $ECN(u, v)$ can be very large because u and v both have lots of neighbors rather than the strong correlation between them. Let $\left|\Gamma(u) \bigcup_{A^{t,s}} \Gamma(v)\right| = |\Gamma(u)| + |\Gamma(v)| - \left|\Gamma(u) \bigcap_{A^{t,s}} \Gamma(v)\right|$, $EJC(u, v)$ can be represented as $EJC(u, v) = \frac{\left|\Gamma(u) \bigcap_{A^{t,s}} \Gamma(v)\right|}{\left|\Gamma(u) \bigcup_{A^{t,s}} \Gamma(v)\right|}$.

- *Extended Adamic/Adar*, which gives different "common neighbors" different weights: $EAA(u, v) = \sum_{\forall(u',v')\in\Gamma(u)\bigcap_{A^{t,s}}\Gamma(v)} \log^{-1}\left(\frac{|\Gamma(u')|+|\Gamma(v')|}{2}\right)$.

From the definitions of these *inter-network social features*, we can find that social links in both G^t and G^s are essential for the social feature extraction in PU anchor link prediction.

3.4 Strict co-existence transfer across networks

To solve the information sparsity problem in the new target network, we propose to transfer information from source networks via the anchor links with the *strict co-existence (of anchor links) transfer* method.

Given a certain social link (u^t, v^t) in G^t, we can extract features for (u^t, v^t), which are represent as vector, $\boldsymbol{x}(u^t, v^t)$. Meanwhile, we notice that by utilizing the anchor links $A^{t,s}$, we can locate the corresponding accounts of user u^t and v^t in G^s, which are u^s and v^s, respectively (if both u^t and v^t are anchor users). The dense feature vector $\boldsymbol{x}(u^s, v^s)$ together with its label $y(u^s, v^s)$ extracted for social link (u^s, v^s) from the more established G^s is correlated with (u^t, v^t) and can be transferred to G^t. A detailed description of the extracted features for social links is available in the Appendix.

Furthermore, we notice that the existence information $y(u^s, v^s)$ of link (u^s, v^s) is also very important and can be transferred to G^t to help improve the result. With full consideration about the network differences, features from different networks are assigned with different weights, which can be decided by classifiers, e.g., SVM [3], when training the model. With the information in G^t and that transferred from G^s, we can get the formation probability of link (u^t, v^t) to be

$$p\left(y(u^t, v^t) = 1 | \left[\boldsymbol{x}(u^t, v^t)^{\mathrm{T}}, \boldsymbol{x}(u^s, v^s)^{\mathrm{T}}, y(u^s, v^s)\right]^{\mathrm{T}}\right),$$

where, $\mathbf{x}^{\mathrm{T}}$ denotes the transpose of vector/matrix $\mathbf{x}$.

According to the above descriptions, the *strict co-existence transfer* method can transfer information for social links formed by anchor users effectively, but it can only work for social links formed by anchor users in G^t, as it needs anchor links to help locate users' corresponding accounts in G^s. However, in real-world partially aligned networks, many users are non-anchor users, in which case, strict co-existence transfer method will not work very well. To overcome the shortcomings of strict co-existence transfer method, we will propose a *loose co-existence transfer* method and introduce the collective random walk to transfer information for both anchor and non-anchor users across "*aligned probabilistic networks*" by relaxing the strict co-existence requirements of anchor links.

3.5 Loose co-existence transfer across aligned probabilistic networks

As shown in Fig. 4, collective anchor and social link prediction can add many *uncertain* anchor links and social links across networks (i.e., the red dotted/dashed lines), whose weights are represented as their "*formation probabilities*".

Definition 4 (*Aligned Probabilistic Networks*) The original partially aligned social added with the newly predicted anchor and social links are formally defined as the *aligned probabilistic networks*, where weights of the originally existing links are 1 and those of newly added ones are their inferred formation probabilities.

Traditional random walk approach has been shown to be effective in computing the similarities between nodes and propagate information within one single network [2,5,6,15,26].

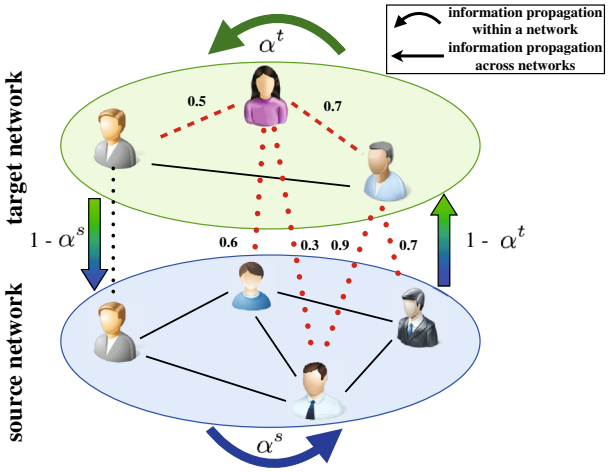


Fig. 4 Collective link fusion across networks (color figure online)

Algorithm 2 Collective Link Fusion (C-RWR)

Input: sets of positive anchor and social links: P_a, P_s
parameters: α^t, α^s and c
probabilities: $p_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a))$, $p_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a))$, $p_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s))$, $p_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s))$
Output: formation confidence: $\tilde{p}_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a))$, $\tilde{p}_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a))$, $\tilde{p}_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s))$, $\tilde{p}_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s))$
1: Initialize matrices $\bar{\boldsymbol{W}}^t$, $\bar{\boldsymbol{W}}^{t,s}$, $\bar{\boldsymbol{W}}^s$ and $\bar{\boldsymbol{W}}^{s,t}$ with existing links in P_a, P_s and probabilities in $p_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a))$, $p_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a))$, $p_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s))$, $p_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s))$
2: $\boldsymbol{W} = \begin{bmatrix} \alpha^t \bar{\boldsymbol{W}}^t & (1-\alpha^s)\bar{\boldsymbol{W}}^{st} \\ (1-\alpha^t)\bar{\boldsymbol{W}}^{ts} & \alpha^s \bar{\boldsymbol{W}}^s \end{bmatrix}$
3: **while** ∃ potential anchor/social link of user u **do**
4: Initialize vector $\boldsymbol{q}$ to be $\mathbf{0}$.
5: $\boldsymbol{q}[u] = 1$
6: $\boldsymbol{p} = c\,[\boldsymbol{I} - (1-c)\boldsymbol{W}]^{-1}\,\boldsymbol{q}$
/*update $p^e(V_a)$, $p^e(T_a)$, $p^e(V_s)$, $p^e(T_s)$ with $\boldsymbol{p}$*/
7: **for** anchor link $(u, v) \in V_a, T_a$ **do**
8: $p_a(u, v) = \boldsymbol{p}[v]$
9: **end for**
10: **for** social link $(u, v) \in V_s, T_s$ **do**
11: $p_s(u, v) = \boldsymbol{p}[v]$
12: **end for**
13: **end while**
14: **return** $\tilde{p}_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a))$, $\tilde{p}_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a))$, $\tilde{p}_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s))$, $\tilde{p}_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s))$

Based on the social links in the "*probabilistic target network*" (i.e., G^t), we can construct the adjacency matrix $\boldsymbol{W}^t \in \mathbb{R}^{|U^t|\times|U^t|}$ of the network, where $\boldsymbol{W}^t_{j,i}$ denotes weight of link (u_i, v_j), $u_i, v_j \in U^t$. We use vector $\left(\boldsymbol{p}^t\right)^{(\tau)} \in \mathbb{R}^{|U^t|}$ to store the probabilities of walking from a certain starting user to other users in the G^t with τ steps. Let $\bar{\boldsymbol{W}}^t = \boldsymbol{W}^t \boldsymbol{D}^{-1}$ be the column-normalized adjacency matrix of $\boldsymbol{W}^t$, where $\boldsymbol{D}_{i,i} = \sum_j \boldsymbol{W}^t_{j,i}$ and $\bar{\boldsymbol{W}}^t_{j,i}$ denotes the probability of walking from u_i to u_j in one step. Vector $\boldsymbol{p}^t$ can be updated with the following

Algorithm 3 Collective PU Link Fusion (CLF)

Input: two partially aligned heterogeneous social networks: $\mathcal{G} = ((G^t, G^s), (A^{t,s}))$
sets of positive and unlabeled anchor links: P_a, U_a
anchor link validation and test sets: V_a and T_a
sets of positive and unlabeled social links: P_s, U_s
social link validation and test sets: V_s and T_s
parameters: α^t, α^s and c

Output: formation confidence of links in V_a, T_a, V_s and T_s: $\tilde{p}_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a))$, $\tilde{p}_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a))$, $\tilde{p}_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s))$, $\tilde{p}_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s))$

Step 1: PU anchor link prediction across networks

1: $\begin{pmatrix} p_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a)) \\ p_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a)) \end{pmatrix} = \text{MLP}(\mathcal{G}, P_a, U_a, V_a, T_a)$

Step 2: PU social link prediction across networks

2: $\begin{pmatrix} p_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s)) \\ p_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s)) \end{pmatrix} = \text{MLP}(\mathcal{G}, P_s, U_s, V_s, T_s)$

Step 3: collective link fusion across networks

3: $\begin{pmatrix} \tilde{p}_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a)), \\ \tilde{p}_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a)), \\ \tilde{p}_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s)), \\ \tilde{p}_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s)) \end{pmatrix} = \text{C} - \text{RWR} \begin{pmatrix} P_a, P_s, \alpha^t, \alpha^s, c, \\ p_a(y(V_a) = \mathbf{1}|\mathbf{x}(V_a)), \\ p_a(y(T_a) = \mathbf{1}|\mathbf{x}(T_a)), \\ p_s(y(V_s) = \mathbf{1}|\mathbf{x}(V_s)), \\ p_s(y(T_s) = \mathbf{1}|\mathbf{x}(T_s)) \end{pmatrix}$

equation until convergence:

$$\left(\boldsymbol{p}^t\right)^{(\tau+1)} = \bar{\boldsymbol{W}}^t \left(\boldsymbol{p}^t\right)^{(\tau)}.$$

Values in vector $\boldsymbol{p}$ at convergence denote the "*formation confidence*" scores of social links between the starting user and other users within the target network G^t.

Furthermore, the newly added uncertain anchor link attached to non-anchor users can provide the opportunity to propagate information form G^s for non-anchor user in the new target network G^t. We propose to extend the traditional random walk to aligned social networks. Similar to $\mathbf{W}^t$, we define $\bar{\boldsymbol{W}}^{ts}$ and $\bar{\boldsymbol{W}}^{st}$ to be the column-normalized adjacency matrices from G^t to G^s and from G^s to G^t respectively. With $\bar{\boldsymbol{W}}^{ts}$ and $\bar{\boldsymbol{W}}^{st}$, we can define the updating equations of inter-network random walks from G^t to G^s and that from G^s back to G^t to be

$$\left(\boldsymbol{p}^s\right)^{(\tau+1)} = \bar{\boldsymbol{W}}^{ts} \left(\boldsymbol{p}^t\right)^{(\tau)},$$
$$\left(\boldsymbol{p}^t\right)^{(\tau+1)} = \bar{\boldsymbol{W}}^{st} \left(\boldsymbol{p}^s\right)^{(\tau+1)}.$$

Vector $\boldsymbol{p}^t$ obtained at convergence denotes the "*formation confidence*" scores of social links between the starting user and other users within the target network G^t, while vector $\boldsymbol{p}^s$ obtained at convergence denotes the "*formation confidence*" scores of anchor links between the starting user and other users in the source network G^s. Intra-network random walk together with inter-network random walk are defined as *collective random walk* in this paper formally.

Different from *strict co-existence transfer*, the inter-network random walk across aligned probabilistic networks relaxes the requirements of anchor links and is named as the *loose co-existence transfer* in this paper.

3.6 Collective link fusion

Furthermore, as illustrated in Fig. 4, newly predicted information of both anchor and social links can propagate within G^t and G^s as well as propagating across G^t and G^s. This process

of fusing predicted information of anchor and social links across partially aligned networks is formally defined as the *collective link fusion* (CLF) in this paper. By integrating the intra-network random walks in G^t and G^s as well as the inter-network random walks from G^t to G^s and from G^s and G^t (i.e., the collective random walk), we can obtain the updating equations of CLF across the aligned probabilistic networks:

$$\begin{cases} (\boldsymbol{p}^s)^{(\tau+1)} = \alpha^s \bar{\boldsymbol{W}}^s (\boldsymbol{p}^s)^{(\tau)} + (1-\alpha^s)\bar{\boldsymbol{W}}^{ts} \left(\boldsymbol{p}^t\right)^{(\tau)}, \\ \left(\boldsymbol{p}^t\right)^{(\tau+1)} = \alpha^t \bar{\boldsymbol{W}}^t \left(\boldsymbol{p}^t\right)^{(\tau)} + (1-\alpha^t)\bar{\boldsymbol{W}}^{st} (\boldsymbol{p}^s)^{(\tau)}, \end{cases}$$

where α^t and α^s denote the weights of information within G^t and G^s, respectively, in updating the vectors. Careful choice of α^t and α^s can control the usage of information from other networks to avoid negative transfer problem effectively [23].

If the walkers are allowed to return to the starting point, then the integrated updating equation will be

$$\boldsymbol{p}^{(\tau+1)} = (1-c)\boldsymbol{W}\boldsymbol{p}^{(\tau)} + c\boldsymbol{q},$$

where $\boldsymbol{W} = \begin{bmatrix} \alpha^t \bar{\boldsymbol{W}}^t & (1-\alpha^t)\bar{\boldsymbol{W}}^{st} \\ (1-\alpha^s)\bar{\boldsymbol{W}}^{ts} & \alpha^s \bar{\boldsymbol{W}}^s \end{bmatrix}$ constant c denotes the probability of returning to the starting point, vector $\boldsymbol{p}^{(\tau)} = \left[\left(\left(\boldsymbol{p}^t\right)^{(\tau)}\right)^{\mathrm{T}}, \left(\left(\boldsymbol{p}^s\right)^{(\tau)}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$ stores the probabilities of walking from the starting user to users in both G^t and G^s and vector $\boldsymbol{q} \in \{0,1\}^{|U^t|+|U^s|}$ is filled with 0 except the cell corresponding to the starting user, which is set as 1. Keep updating $\boldsymbol{p}$ until convergence, we can get

$$\boldsymbol{p} = c\left[\boldsymbol{I} - (1-c)\boldsymbol{W}\right]^{-1}\boldsymbol{q},$$

where matrix $\boldsymbol{I} \in \{0,1\}^{(|U^t|+|U^s|)^2}$ is an identity matrix. Entries in vector $\boldsymbol{p}$ at convergence store the "*formation confidence*" scores of potential anchor and social links connecting the starting user with other users in G^s and G^t, respectively.

MLP together with C-RWR will form the CLF framework, whose pseudo code is given in Algorithm 3. As shown in Algorithm 3, in CLF framework, we conduct PU link prediction of *anchor links* and *social links* at first, whose results are passed to C-RWR. In C-RWR, we construct the *aligned probabilistic networks* and use *random walk* to propagate information across *probabilistic networks*. Vector $\boldsymbol{p}$ at convergence contains the "*formation confidence scores*" of potential anchor and social links connecting the starting user and other users. These scores will be used as the prediction scores of these links returned to C-RWR (i.e., the *formation confidence* returned in Algorithm 2) as well as the final output of CLF framework.

PU link prediction and RWR (random walk with restart) technique comprise the CLF method. The PU learning for one bit matrix completion needs $O(n^2)$ time [10] and the time complexity of second part RWR in calculating inverse and pseudoinverse is approximately $O(n^3)$ [21], where n is the total number of nodes in the target network and source network.

4 Experiments

4.1 Data description

Datasets used in this paper include Foursquare, a famous location-based online social networks, and Twitter, the hottest microblogging social network. The anchor link between

Table 2 Properties of the aligned social networks

Property	Network	
	Twitter	Foursquare
# Node		
User	5223	5392
Tweet/tip	9,490,707	48,756
Location	297,182	38,921
# Link		
Friend/follow	164,920	76,972
Write	9,490,707	48,756
Locate	615,515	48,756

Foursquare and Twitter is obtained by crawling users' Twitter accounts from their Foursquare homepages, whose number is 3388. A more detailed comparison of these two datasets is available in Table 2 [39]:

- *Foursquare* 5392 users, 48,756 tips and 38,921 locations are crawled from Foursquare. These users generated 76,972 social links which means each user has about 14 friends in Foursquare on average.
- *Twitter* 5223 users who generated 9,490,707 tweets are crawled from Twitter. Among all these users, there exist 164,920 follow links.

In the experiment, social links in Foursquare and anchor links between Foursquare and Twitter are used as the ground truth to evaluate the performance of CLF and other baseline methods.

4.2 Experiment setting

4.2.1 Comparison methods

We compare CLF with many different baseline methods in predicting both social links and anchor links, in which SVM of linear kernel with optimal parameters is used as the base classifier. The comparison methods used in the experiment include:

- *Collective Link Fusion* CLF proposed in this paper include multiple phases: (1) collective multi-network link prediction; (2) collective link fusion across partially aligned probabilistic networks. CLF can utilize the extended definition of social features for anchor links in the PU anchor link prediction task, transfer information from the source network for social links formed by both "anchor users" and "non-anchor users" in PU social link prediction in the target network and can fuse the prediction results of both anchor and social links with cross-network random walk.
- *Multi-Network Link Prediction* MLP extends the state-of-art PU link prediction method proposed in [39] to infer the existence probabilities of both anchor and social links independently with positive and unlabeled links.
- *Collective Random Walk* C-RWR is the second step of CLF and can propagate information of both anchor and social links across networks. When C-RWR is used as a baseline method, only the existing anchor and social links are used in constructing the adjacency matrices. C-RWR can transfer information for "non-anchor users" with loose co-existence transfer and can fuse the results of both anchor and social link prediction across partially aligned networks.

- *Random Walk with Restart* RWR (Random Walk with Restart) [26] can calculate the "similarity" between any pairs of users within one network.

4.2.2 Evaluation methods

Considering that CLF, MLP, C-RWR and RWR can only output scores of both anchor links and social links, we will use AUC and Precision@30 to evaluate their performance.

4.2.3 Experiment setups

We use all existing social links in G^t as the sets of positive social links in the experiment. Then, we randomly sampled a set of non-existent social links as the negative social link set from G^t, which is of the same size as the set of positive social link. These links are partitioned into 3 parts with fivefold cross-validation: threefold as the training set, onefold as the validation set and the remaining onefold as the test set. We randomly sampled a portion of links with percentage γ_s (γ_s varies from 0.1 to 0.9) from the positive social links in the threefold as the final positive training set. The remaining $(1 - \gamma_s)$ positive social links are mixed with the negative training links. Classifiers built with the γ_s sampled positive and mixed social links (negative links and the remaining $(1 - \gamma_s)$ positive social links) are applied to classify social links in the validation set and test set. Existence probabilities obtained on the positive social links in the validation set are used to approximate the bridging probability, $p(z = 1|y = 1)$, which will be used to get the final formation probabilities of social links in both the validation set and the test set. In a similar way, we can get the formation probabilities of anchor links in the validation set and test set. The parameter used to control the percentage of positive anchor links used to train models is γ_a (γ_a varies from 0.1 to 0.9). When predicting social links, we set γ_a as 0.5 and vary γ_s from 0.1 to 0.9; meanwhile, when predicting anchor links, we set γ_s as 0.5, but vary γ_a from 0.1 to 0.9

Based on the multi-network link prediction result, we further propagate the predicted information across networks. The probabilities of propagating within G^t and G^s instead of crossing the networks are $\alpha_t, \alpha_s \in [0, 1.0]$. The probability of returning to the starting point is $c \in [0, 1.0]$. In the experiment, we set α^t and α^s as 0.6 and c is set as 0.1, whose sensitivities will be analyzed in the following parts.

4.3 Experiment result

In Fig. 5, we show the ROC curve of the anchor and social link prediction results. In Fig. 5a, we set $\gamma_s = 0.5$ and $\gamma_a = 0.9$ and in Fig. 5b, we set $\gamma_a = 0.5$ and $\gamma_s = 0.9$. We can find that the area under the ROC curve of CLF is the largest among all the baseline methods.

Figure 6a, b show the precision of top {200, 400, 600, . . . , 2000} predictions of anchor and top {500, 1000, 1500, . . . , 5000} predictions of social link respectively. Figures illustrate that the increase in prediction number leads to the decrease of precision for all methods. However, CLF can outperform others when evaluating by precision of all top predictions. Moreover, the decline of CLF is the smallest. For example, when predicting anchor links, the precision of CLF drops down from 0.88 at top 200 to 0.79 at top 2000, which the decrease is 10.23%. While this number of MLP is 22.54% and C-RWR's is 29.43%. It demonstrates that CLF method has a good and consistent performance. The result of social link prediction leads to the similar conclusion with the anchor link prediction.

In Fig. 7, we show the experiment results (mean $\pm$ std) of both anchor links and social links of different method under the evaluation of AUC and $Precision@30$ over all links

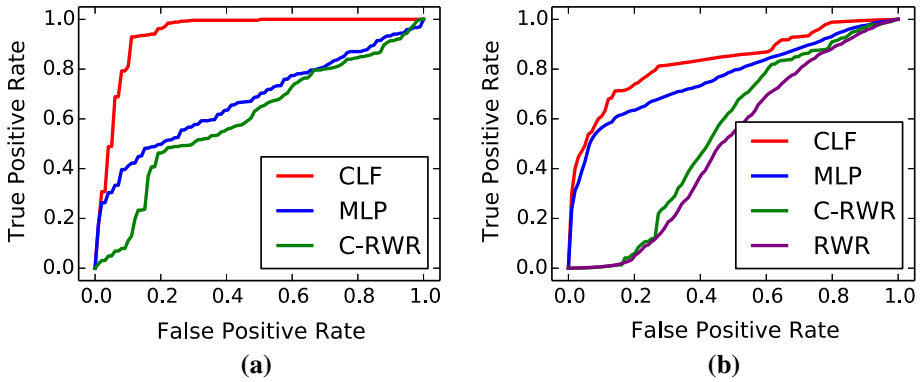


Fig. 5 ROC curve of link prediction results. **a** ROC of anchor link, **b** ROC of social link

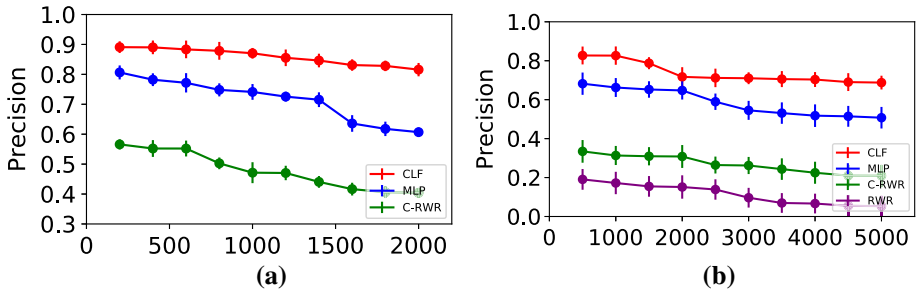


Fig. 6 Precision of link prediction results. **a** Precision of anchor link, **b** precision of social link

of all users, where $\gamma_a(\gamma_s)$ changes from 0.1 to 0.9. The performance of most methods will increase as $\gamma_a(\gamma_s)$ increases in Fig. 7. When $\gamma_a(\gamma_s)$ is small, all the baseline methods cannot work well, but CLF can still achieve good performance. Figure 7a, b shows the result of anchor link prediction, in which $\gamma_s = 0.5$ and γ_a changes from 0.1 to 0.9, and Fig. 7c, d shows the social link prediction result, where $\gamma_a = 0.5$ and γ_s change from 0.1 to 0.9.

In Fig. 7a, we show the performance evaluated by AUC. The AUC of CLF is over 40% better than MLP and over 50% better than C-RWR consistently in the whole changing range of γ_a. It demonstrates that the combination of MLP and C-RWR can lead to better results. In Fig. 7b, the performance of CLF is also better than both MLP and C-RWR under the evaluation of $Precision@30$. In Fig. 7c, we show the social link prediction result under the evaluation of AUC. CLF can perform well in predicting social links and outperform all other baseline methods with a big advantage. Method C-RWR, which propagate information of existing links across networks, can perform better than RWR, which shows that *loose co-existence transfer* for "non-anchor users" can indeed improve the result. However, CLF using the probabilistic network will further enhance the performance over C-RWR. This shows the importance of the first step on using the multi-network link prediction to build the probabilistic network. Similar to the result in Fig. 7b, d, CLF can beat all the baseline methods and perform very well when γ_s is small. CLF can outperform C-RWR shows that the multi-network link prediction step is essential and can work very well, while CLF can outperform MLP demonstrates that the collective link fusion step can improve the prediction results of both anchor and social links.

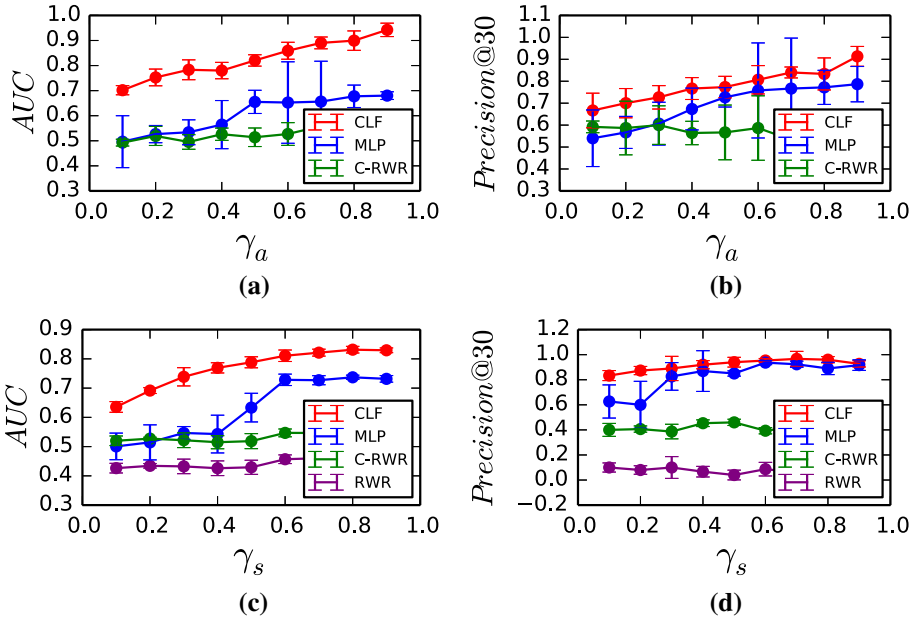


Fig. 7 Anchor and social link prediction results. **a** AUC of anchor link, **b** prec.@30 of anchor link, **c** AUC of social link, **d** prec.@30 of social link

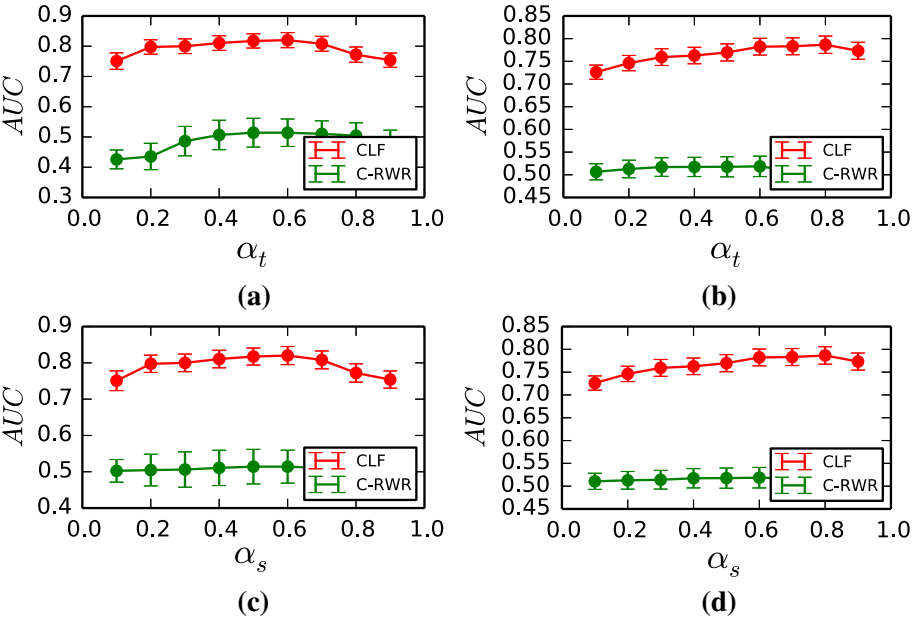


Fig. 8 Analysis of parameters α^t and α^s. **a** AUC of anchor link, **b** AUC of social link, **c** AUC of anchor link, **d** AUC of social link

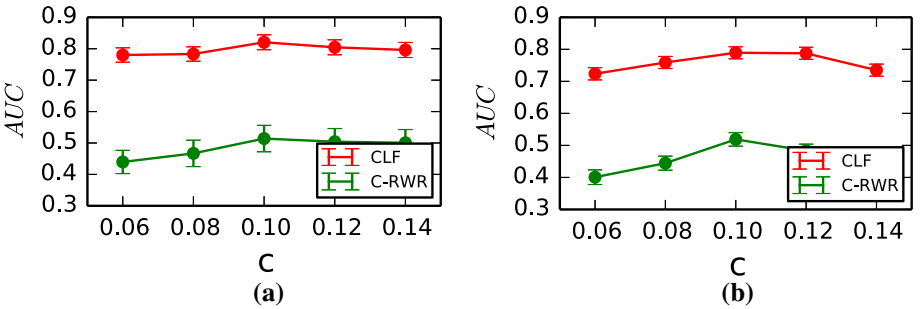


Fig. 9 Analysis of parameter c. **a** AUC of anchor link, **b** AUC of social link

In sum, CLF can outperform all the baseline methods under the evaluation of both AUC and $Precision@30$ within the changing range of γ_a and γ_s in predicting both anchor and social links.

4.4 Parameter analysis

CLF has three parameters in all, which are c, α_t, α_s. To analyze the effects of parameters in the experiment, we assign α_t, α_s with values in [0.1, 0.9], and assign parameter c with values in {0.06, 0.08, 0.10, 0.12, 0.14} to compare the performance of CLF and C-RWR under the evaluation of AUC. The results are available in Figs. 8 and 9, where Fig. 8a, d show the effects of parameter α^t and α^s and Fig. 9a, b show the effects of parameter c.

In Fig. 8a, b, we only change α^t with values in [0.1, 0.9] and fix all other parameters. Both CLF and C-RWR can perform very stable within the changing range of α^t but CLF in Fig. 8b has an visible increasing trend when $\alpha^t \in [0.1, 0.6]$ and stay stable when $\alpha^t \in [0.6, 0.8]$ and drops at 0.9. Fig. 8c, d show the effects of α^s. The performance of CLF and C-RWR is more stable compared with that in Fig. 8a, b, which shows that α^t has a much more significant effects than α^s.

In Fig. 9a, b, we show the effects of parameter c in the experiment where α^t and α^s are both set as 0.6. Performance of both CLF and C-RWR will vary as c changes, and they can achieve the best performance around $c = 0.1$.

4.5 Case study

We show a case study to demonstrate that the two-phase method CLF can work well in predicting both anchor and social links. As illustrated in Fig. 10, we have five real-world users who own both Foursquare and Twitter accounts. Originally, the social connections among users in Foursquare are identical to those in Twitter. As shown in Fig. 10a, to test the effectiveness of CLF in predicting anchor and social links, all the social links in Foursquare except the social link between "Nathan Levinson" and "Michelle Jacobson" and all the anchor links between Foursquare and Twitter except those between "Nathan Levinson" and "Nathan Levinson", "Andrew Nystrom" and "Andrew Nystrom" are deleted. Many social and anchor links in Fig. 10a are to be predicted. Figure 10b shows the prediction result of PU link prediction method, MLP. We can find that $precision@3$ scores achieved by MLP in predicting both anchor links and social links are 33.3%, as only the anchor link between "Michelle Jacobson" and "Michelle Jacobson" and the social link between "Nathan Levinson" and "Andrew Nystrom" are correctly predicted among the *top-3* link prediction

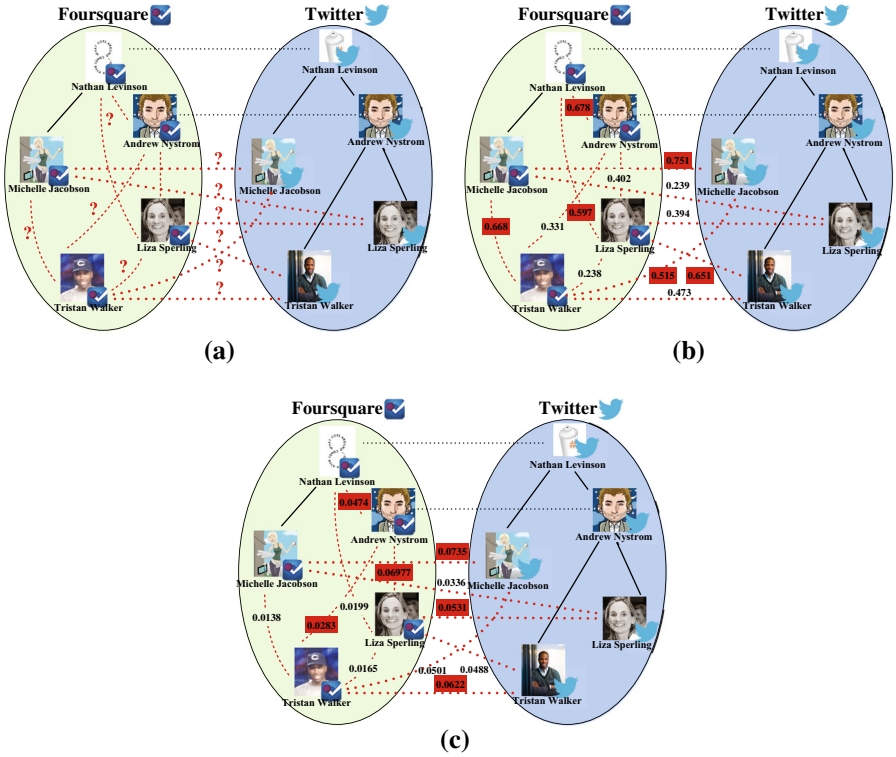


Fig. 10 Case study. **a** The input network, **b** the result of MLP, **c** the result of CLF

results. Based on the results in Fig. 10b, we further propagate predicted information across networks, which is the second step of CLF. The results are shown in Fig. 10c, in which the scores of both anchor links and social links have been updated. In the updated results, $precision@3$ scores of CLF in predicting anchor links and social links are both 100%, respectively, and links exists in the real world can get the highest scores among all the links to be predicted.

5 Related works

PU learning has been studied for several years and dozens of papers on this topic have been published. Liu et al. [19] proposed different settings to find the reliable negative instances in text classification. Zhao et al. proposed to classify graphs with only positive and unlabeled examples in [40]. Zhang et al. were the first to propose the concept of PU link prediction in [39] and studied the PU social link prediction in multiple networks simultaneously. However, [39] does not address the collective prediction of anchor links and social links together, which we have studied in this paper. A new PU link prediction method is introduced in this paper, which is totally different from the spy technique used in [39].

Link prediction first proposed by Liben-Nowell et al. [18] has been a hot research topic in recent years. Predicting the labels of links with supervised models is formulated as a *supervised link prediction problem* [8]. Meanwhile, Xiang et al. [29] developed an unsupervised

model to estimate relationship strength. In heterogeneous social networks, multiple types of links can be predicted simultaneously. Namata proposed a collective graph identification problem in [22]. Some works label links as positive and negative links according to their physical meanings, e.g., friendship vs. antagonism [17], trust vs. distrust [25], and propose to predict these links in online social networks. Hwang et al. [11] proposed a heterogeneous label propagation algorithm for disease gene discovery. Xi et al. proposed a unified link analysis framework for multi-type interrelated data objects in [28]. Extensive surveys of link prediction problems are available in [7,9,16].

Entity identification across networks(communities) gets lots of attention in recent years. Sahraeian et al. [24] introduces a scalable algorithm to align proteins across large-scale PPI network. Zafarani et al. [32] proposed to connect corresponding identities across communities. Iofciu et al. [12] proposed to identify common users across social tagging systems. Liu et al. [20] proposed an unsupervised to link users across communities. Kong et al. [14] notice that users are involved in multiple social networks nowadays and propose to infer the links between accounts of the anchor users. Zhang et al. [34,35] proposed transfer links across networks to predict links for new users and new networks respectively. Furthermore, links in multiple partially aligned social networks can be strongly correlated and Zhang et al. [39] introduced an integrated PU link prediction framework to predict social links in multiple social networks concurrently.

The work in this paper has made a great progress when comparing with our prior works. Multiple aligned heterogeneous networks, first studied by Kong et al. [14], have become a hot research topic in recent years. Kong et al. [14,36] were the first to propose the concepts of "anchor link", "aligned heterogeneous networks" and studied the link prediction problem across aligned networks. But in [14], two networks are fully aligned. Zhang et al. [34,35,39] were the first to study link prediction problem for new users, using information transferred from other aligned source networks via anchor links. These works extended the problem setting from fully aligned networks to partially aligned networks, which were much closer to the real situation. Zhang et al. [35] can predict both social links and location links, while Zhang et al. [39] can predict the formation of social links in multiple partially aligned networks simultaneously. However, unlike this paper, links to be predicted in prior works are limited in social links and other heterogeneous links, none of them can predict anchor links across networks.

Besides link prediction, other topic and applications on multiple social networks also attract researchers interest. Zhang and Jin et al. [13,37,38] also proposed to study the community detection problems across aligned networks, where information from all these aligned networks can be transferred to prune and refine the community structures of each network mutually. In addition, Zhan et al. introduced the cross-aligned-network information diffusion problem in [33], where multiple diffusion channels were extracted based on various types of intra and inter-network meta paths.

6 Conclusion

In this paper, we study the *collective link identification* problem merely with formed links (i.e., positive links) in the networks. By using unconnected links in networks as the unlabeled links, we propose a two-phase method, CLF, to infer the anchor and social links simultaneously. Extensive experiments conducted on two real-world partially aligned networks, Foursquare and Twitter, demonstrate that CLF can address the challenges of *collective link identification* very well and achieve good results in predicting both anchor and social links.

Acknowledgements This work is supported by the Fundamental Research Funds for the Central Universities under grant JUSRP11852. This work was partially supported by Florida State University Council on Research and Creativity (CRC) via the Project ID 041776. This work is also supported in part by NSF through Grants IIS-1526499, IIS-1763325, CNS-1626432 and NSFC 61672313. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

Appendix

Social features of anchor links have been introduced in previous part, in this part, we will introduce the *social features* of social links and *spatial distribution features*, *temporal distribution features* and *text usage features* of both anchor links and social links.

6.1. Social features

See Table 3.

Table 3 Social features defined for social link (u, v)

Features	Descriptions
Degree count	$d_{\text{in}}(u), d_{\text{in}}(v), d_{\text{out}}(u), d_{\text{out}}(v)$
Degree ratio	$d_{\text{in}}(u)/d_{\text{in}}(v), d_{\text{out}}(u)/d_{\text{out}}(v)$
Common neighbor	$\lvert\Gamma(u) \cap \Gamma(v)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\cup\Gamma(v)\rvert}$
Preferential attachment	$\lvert\Gamma(u)\rvert \cdot \lvert\Gamma(v)\rvert$
Adamic/Adar	$\sum_{\forall w\in(\Gamma(u)\cap\Gamma(v))} \frac{1}{\log\lvert\Gamma(w)\rvert}$

$\Gamma(u)$ is the set of neighbors of user u.

In addition to social information, we also extract features from users' location check-ins. For a certain anchor/social link (u, v), we can get the locations that u and v have been to $\Phi(u)$ and $\Phi(v)$, respectively. Since each user can visit a location many times, we construct vector $l(u)$ and $l(v)$ for u and v, respectively, each cell in which record the times that u and v visit a certain location in $\Phi(u) \cup \Phi(v)$.

6.2. Spatial distribution features

See Table 4.

Table 4 Spatial distribution features for link (u, v)

Features	Descriptions
Location count (LC)	$\lvert\Phi(u)\rvert, \lvert\Phi(v)\rvert$
Extended CN (ECN)	$\lvert\Phi(u) \cap \Phi(v)\rvert$
Extended JC (EJC)	$\frac{\lvert\Phi(u)\rvert\cap\Phi(v)}{\lvert\Phi(u)\cup\Phi(v)\rvert}$
Extended AA (EAA)	$\lvert\Phi(u)\rvert \cdot \lvert\Phi(v)\rvert$

Similarly, we can get the set of locations that u has visited from the networks, $\Phi(u)$. For a certain anchor/social link (u, v), we can extract the spatial distribution features for it with those summarized in Table 3 except the "Adamic/Adar" measure based on $\Phi(u)$ and $\Phi(u)$.

6.3. Temporal distribution features

See Table 5.

Table 5 Other frequently features for link (u, v)

Features	Descriptions
Inner product (IP)	$\mathbf{x}(u) \cdot \mathbf{x}(v)$
Euclidean distance (ED)	$(\sum_k (\mathbf{x}(u)_k - \mathbf{x}(v)_k)^2)^{1/2}$
Cosine similarity (CS)	$\frac{\mathbf{x}(u) \cdot \mathbf{x}(v)}{\|\mathbf{x}(u)\| \cdot \|\mathbf{x}(v)\|}$

Users' temporal activity information is also used to extract features for link (u, v). Each day is divided into 24 h slots, and the number of online posts published at certain hours is stored in vector $\mathbf{x}(u)$ and $\mathbf{x}(v)$, from which we can extract $IP(\mathbf{x}(u), \mathbf{x}(v))$, $ED(\mathbf{x}(u), \mathbf{x}(v))$ and $CS(\mathbf{x}(u), \mathbf{x}(v))$ summarized in Table 5 as the temporal distribution features of link (u, v).

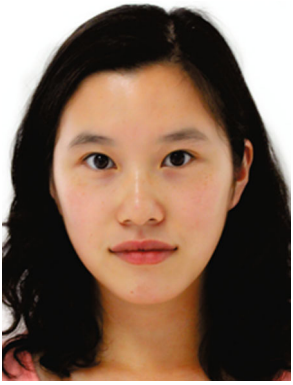
6.4. Text usage features

For a certain link (u, v), we can get the words that u and v have used in the past and group them as two bag-of-words vectors, $\mathbf{x}(u)$ and $\mathbf{x}(v)$, weighted by TF-IDF. From $\mathbf{x}(u)$ and $\mathbf{x}(v)$, we also extract $IP(\mathbf{x}(u), \mathbf{x}(v))$, $ED(\mathbf{x}(u), \mathbf{x}(v))$ and $CS(\mathbf{x}(u), \mathbf{x}(v))$ summarized in Table 5 as the text usage features of link (u, v).

References

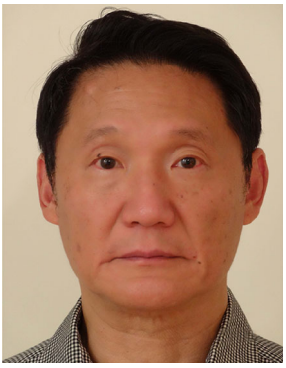
1. Adamic L, Adar E (2001) Friends and neighbors on the web. Soc Netw 25:211–230
2. Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: WSDM
3. Chang C-C, Lin C-J (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. In: KDD
5. Fouss F, Pirotte A, Renders J, Saerens M (2007) Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. TKDE 19:355–369
6. Fujiwara Y, Nakatsuji M, Onizuka M, Kitsuregawa M (2012) Fast and exact top-k search for random walk with restart. VLDB 55:442–453
7. Getoor L, Diehl CP (2005) Link mining: a survey. SIGKDD Explor Newslett 7:3–12
8. Hasan M, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In: SDM
9. Hasan M, Zaki MJ (2011) A survey of link prediction in social networks. In: Aggarwal CC (ed) Social network data analytics. Springer, New York
10. Hsieh C-J, Natarajan N, Dhillon IS (2015) PU learning for matrix completion. In: ICML, pp 2445–2453
11. Hwang T, Kuang R (2010) A heterogeneous label propagation algorithm for disease gene discovery. In: SDM
12. Iofciu T, Fankhauser P, Abel F, Bischoff K (2011) Identifying users across social tagging systems. In: ICWSM
13. Jin S, Zhang J, Yu P, Yang S, Li A (2014) Synergistic partitioning in multiple large scale social networks. In: IEEE BigData

14. Kong X, Zhang J, Yu P (2013) Inferring anchor links across multiple heterogeneous social networks. In: CIKM
15. Konstas I, Stathopoulos V, Jose JM (2009) On social networks and collaborative recommendation. In: SIGIR
16. Lü L, Zhou T (2011) Link prediction in complex networks: a survey. Phys A Stat Mech Its Appl 390:1150–1170
17. Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. In: WWW
18. Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: CIKM
19. Liu B, Dai Y, Li X, Lee W, Yu P (2003) Building text classifiers using positive and unlabeled examples. In: ICDM
20. Liu J, Zhang F, Song X, Song Y, Lin C, Hon H (2013) What's in a name? An unsupervised approach to link users across communities. In: WSDM
21. Lü L, Zhou T (2011) Link prediction in complex networks: a survey. Phys A Stat Mech Its Appl 390(6):1150–1170
22. Namata G, Kok S, Getoor L (2011) Collective graph identification. In: KDD
23. Perkins D, Salomon G (1992) Transfer of learning Pergamon Press, Oxford, England
24. Sahraeian S, Yoon B (2013) Smetana: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. PLoS ONE 8:e67995
25. Song D, Meyer D (2014) A model of consistent node types in signed directed social networks. In: ASONAM '14 Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE Press, Piscataway, NJ, USA, pp 72–80
26. Tong H, Faloutsos C, Pan J (2006) Fast random walk with restart and its applications. In: ICDM
27. Wilcox K, Stephen AT (2012) Are close friends the enemy? Online social networks, self-esteem, and self-control. J Consum Res 40:90–103
28. Xi W, Zhang B, Chen Z, Lu Y, Yan S, Ma W, Fox E (2004) Link fusion: a unified link analysis framework for multi-type interrelated data objects. In: WWW
29. Xiang R, Neville J, Rogati M (2010) Modeling relationship strength in online social networks. In: WWW
30. Yao Y, Tong H, Yan X, Xu F, Lu J (2013) Matri: a multi-aspect and transitive trust inference model. In: WWW
31. Ye J, Cheng H, Zhu Z, Chen M (2013) Predicting positive and negative links in signed social networks by transfer learning. In: WWW
32. Zafarani R, Liu H (2009) Connecting corresponding identities across communities. In: ICWSM
33. Zhan Q, Wang S, Zhang J, Yu P, Xie J (2015) Influence maximization across partially aligned heterogenous social networks. In: PAKDD
34. Zhang J, Kong X, Yu P (2013) Predicting social links for new users across aligned heterogeneous social networks. In: ICDM
35. Zhang J, Kong X, Yu P (2014) Transferring heterogeneous links across location-based social networks. In: WSDM
36. Zhang J, Shao W, Wang S, Kong X, Yu P (2015) Pna: Partial network alignment with generic stable matching. In: IEEE IRI
37. Zhang J, Yu P (2015) Community detection for emerging networks. In: SDM
38. Zhang J, Yu P (2015) Mcd: Mutual clustering across multiple heterogeneous networks. In: IEEE BigData Congress
39. Zhang J, Yu P, Zhou Z (2014) Meta-path based multi-network collective link prediction. In: KDD
40. Zhao Y, Kong X, Yu P (2011) Positive and unlabeled learning for graph classification. In: ICDM



Qianyi Zhan received her B.S. degree in computer science from Nanjing Normal University, and the Ph.D. degree in computer science from the Nanjing University. She is an assistant professor in the School of Digital Media at Jiangnan University. Her principal research interests focus on data mining and social network analysis. She has published more than 10 referred papers in journals and major conferences.

Jiawei Zhang received the B.S. degree in computer science from Nanjing University, and the Ph.D. degree in computer science from the University of Illinois at Chicago. He is an assistant professor in the Department of Computer Science at Florida State University. His principal research interest is on data mining, deep learning and machine learning, with a focus on developing general methodologies for information fusion and mining, which will be applicable to a diverse set of applications. He has published more than 50 referred papers in books, journals, and major conferences.



Philips S. Yu received the B.S. Degree in E.E. from National Taiwan University, the M.S. and Ph.D. degrees in E.E. from Stanford University, and the M.B.A. degree from New York University. He is a Distinguished Professor in Computer Science at the University of Illinois at Chicago and also holds the Wexler Chair in Information Technology. Before joining UIC, Dr. Yu was with IBM, where he was manager of the Software Tools and Techniques department at the Watson Research Center. His research interest is on big data, including data mining, data stream, database and privacy. He has published more than 1000 papers in refereed journals and conferences. He holds or has applied for more than 300 US patents. Dr. Yu is a Fellow of the ACM and the IEEE. Dr. Yu is the recipient of ACM SIGKDD 2016 Innovation Award for his influential research and scientific contributions on mining, fusion and anonymization of big data, the IEEE Computer Society's 2013 Technical Achievement Award for pioneering and fundamentally innovative contributions to the scalable indexing, querying, searching, mining and anonymization of big data, and the Research Contributions Award from IEEE Intl. Conference on Data Mining (ICDM) in 2003 for his pioneering contributions to the field of data mining. He also received the ICDM 2013 10-year Highest-Impact Paper Award, and the EDBT Test of Time Award (2014). He was the Editor-in-Chiefs of ACM Transactions on Knowledge Discovery from Data (2011–2017) and IEEE Transactions on Knowledge and Data Engineering (2001–2004).