# Integrated Anchor and Social Link Predictions across Social Networks

**Jiawei Zhang**[⋆] and **Philip S. Yu**[⋆†]
[⋆]University of Illinois at Chicago, IL, USA
[†]Institute for Data Science, Tsinghua University, Beijing, China.
jzhan9@uic.edu and psyu@cs.uic.edu

## Abstract

To enjoy more social network services, users nowadays are usually involved in multiple online social media sites at the same time. Across these social networks, users can be connected by both intra-network links (i.e., social links) and inter-network links (i.e., anchor links) simultaneously. 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. To solve the collective link identification problem, a unified link prediction framework, CLF (Collective Link Fusion) is proposed in this paper, which consists of two phases: step (1) collective link prediction of anchor and social links, 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.

## 1 Introduction

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" [Kong *et al.*, 2013] 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* [Kong *et al.*, 2013] 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. Across partially aligned social networks, multiple link prediction tasks exist, which can be conducted simultaneously. In this paper, we will study the *collective link identification problem*, which covers the following two different link prediction tasks at the same time:

• **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.

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 [Kong *et al.*, 2013].

The *collective link identification* problem studied in this paper is novel and conventional classification based link prediction models [Backstrom and Leskovec, 2011] cannot be applied to solve it directly due to the following challenges. Firstly, in traditional classification based methods [Backstrom and Leskovec, 2011; Hasan and Zaki, 2011], links in social networks are assigned with different labels according to their physical meanings, e.g., friends vs enemies [Wilcox and Stephen, 2012], trust vs distrust [Yao *et al.*, 2013], positive attitude vs negative attitude [Ye *et al.*, 2013], 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 [Zhang *et al.*, 2014a]. 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.

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 PU (Positive and Unlabeled [Liu *et al.*, 2003]) 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 Section 2, we will give the problem formulation. Methods will be introduced in Section 3. Extensive experiments are done in Section 4. Section 5 is about the related works. Finally, in Section 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 [Kong *et al.*, 2013], which are not introduced here due to the limited space. Different from [Kong *et al.*, 2013], the major assumptions about the aligned networks in this paper are: (1) no restrictions exist on the constraint of anchor links, which can be either *one-to-one* or *many-to-many*; (2) *partial alignment of networks*: fully aligned networks rarely exist in the real world and networks studied in this paper are partially aligned [Zhang *et al.*, 2014b].

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 [Kong *et al.*, 2013], 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 $G^s$.

### 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.

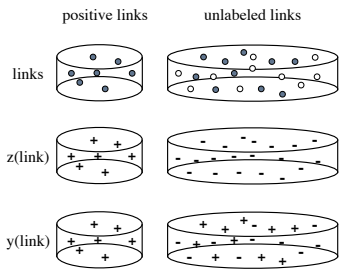



Figure 1: Example of connection states and labels of links in PU link prediction.

## 3 Proposed Methods

### 3.1 Preliminary

As introduced in [Zhang *et al.*, 2014b], 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 [Zhang *et al.*, 2014b]. 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 Figure 1, 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 [Liu *et al.*, 2003] to extract reliable negative links from the unconnected link set was proposed in [Zhang *et al.*, 2014b]. However, the correlation between links' *connection state* and *labels* is not clearly addressed in [Zhang *et al.*, 2014b], 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 [Kong *et al.*, 2013; Zhang *et al.*, 2013]) 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 Figure 1, 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

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

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

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

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 [Zhang *et al.*, 2014b]. However, [Zhang *et al.*, 2014b] fails to study the correlation between links' "*existence probability*" and "*formation probability*", which can be represented as follows:

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

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

**Definition 1** (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 [Elkan and Noto, 2008]. 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 have

**Bridging Probability Inference Equation**:

$$p(z = 1|y = 1) = \frac{1}{|\text{Pos}|} \sum_{link \in \text{Pos}} p(z = 1|y = 1)$$
$$= \frac{1}{|\text{Pos}|} \sum_{link \in \text{Pos}} p(z = 1|\mathbf{x}, y = 1),$$

where $p(z = 1|y = 1) = p(z = 1|\mathbf{x}, y = 1)$ can hold according to proof in previous parts. For links in $Pos$, we have $p(y = 1|\mathbf{x}) = 1$, $p(z = 1|\mathbf{x}, y = -1) = 0$ and $p(y = -1|\mathbf{x}) = 0$. So,

$$p(z = 1|y = 1) = \frac{1}{|\text{Pos}|} \sum_{link \in \text{Pos}} (p(z = 1|\mathbf{x}, y = 1)p(y = 1|\mathbf{x})$$
$$+ p(z = 1|\mathbf{x}, y = -1)p(y = -1|\mathbf{x}))$$
$$= \frac{1}{|\text{Pos}|} \sum_{link \in \text{Pos}} p(z = 1|\mathbf{x}).$$

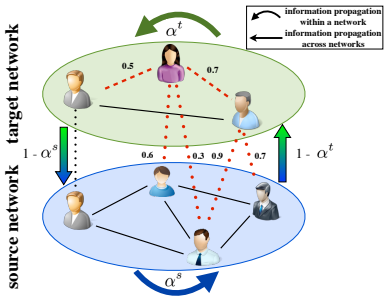



Figure 2: Collective Link Fusion across Networks.

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 [Zhang *et al.*, 2014b].

### 3.3 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 $G^t$.

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

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

where, $\mathbf{x}^T$ 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. 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 address such a problem, we will define the concept of "*probabilistic networks*" in the next section and introduce the collective random walk to transfer information for both anchor and non-anchor users across "*aligned probabilistic networks*".

### 3.4 Loose Co-Existence Transfer across Aligned Probabilistic Networks

As shown in Figure 2, 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 2** (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 [Fouss *et al.*, 2007; Tong *et al.*, 2006; Fujiwara *et al.*, 2012; Konstas *et al.*, 2009; Backstrom and Leskovec, 2011]. 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 $(\boldsymbol{p}^t)^{(\tau)} \in \mathbb{R}^{|U^t|}$ to store the probabilities of walking from a certain starting user to other users in the $G^t$ with $\tau$ steps. Entries of $(\boldsymbol{p}^t)^{(0)}$ are initialized with 0s except the entry corresponding to the starting user is initialized as 1. 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 1 step. Vector $\boldsymbol{p}^t$ can be updated with the following 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

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

where vector $\boldsymbol{p}^s$ is initialized with 0s, while initialization of $\boldsymbol{p}^t$ is identical to that in traditional single-network random walk. 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.

### 3.5 Collective Link Fusion

Furthermore, as illustrated in Figure 2, 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:

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

where $\alpha^t$ and $\alpha^s$ denote the weights of information within $G^t$ and $G^s$ respectively in updating the vectors. Careful choice of $\alpha^t$ and $\alpha^s$ can control the usage of information from other networks to avoid negative transfer problem effectively [Perkins and Salomon, 1992].

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 $\boldsymbol{p}^{(\tau)} = \left[\left(\left(\boldsymbol{p}^t\right)^{(\tau)}\right)^T, \left(\left(\boldsymbol{p}^s\right)^{(\tau)}\right)^T\right]^T$ 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.

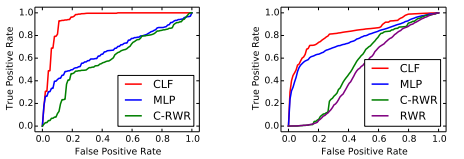
Table 1: Properties of the Aligned Social Networks

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

## 4 Experiments

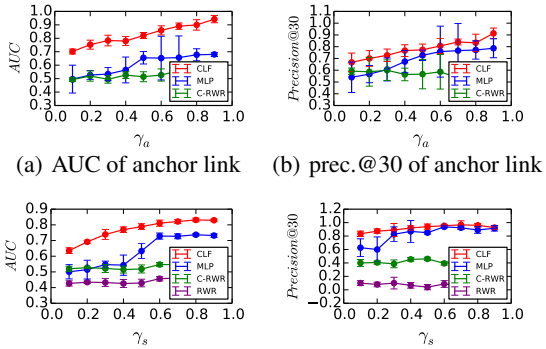
### 4.1 Data Description

Datasets used in this paper include Foursquare, a famous location-based online social networks, and Twitter, the hottest microblogging social network. A more detailed comparison of these two datasets is available in Table 1. The anchor link between Foursquare and Twitter is obtained by crawling users' Twitter accounts from their Foursquare homepages,




(a) ROC of anchor link (b) ROC of social link

Figure 3: ROC curve of 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

Figure 4: Anchor and social link prediction results.

whose number is $3,388$. 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. For more information about the datasets and the crawling method, please refer to [Zhang *et al.*, 2014b].

## 4.2 Experiment Setting

### Comparison Methods

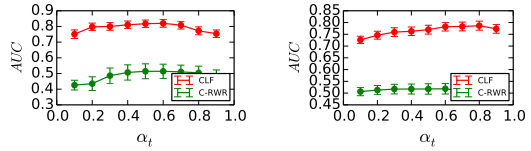
We compare CLF with many different baseline methods in predicting both social links and anchor links, in which SVM of linear kernel with optimal parameters is used as the base classifier. The comparison methods used in the experiment include:

• **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.

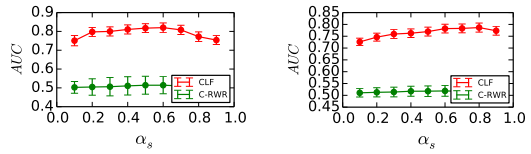
• **Multi-Network Link Prediction**: MLP extends the state-of-art PU link prediction method proposed in [Zhang *et al.*, 2014b] to infer the existence probabilities of both anchor and social links.

• **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.

• **Random Walk with Restart**: RWR (Random Walk with Restart) [Tong *et al.*, 2006] can calculate the "similarity" between any pairs of users within one network.




(a) AUC of anchor link (b) AUC of social link




(c) AUC of anchor link (d) AUC of social link

Figure 5: Analysis of parameters $\alpha^t$ and $\alpha^s$.

### Evaluation Methods

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

### 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 5 folds cross validation: 3 folds as the training set, 1 fold as the validation set and the remaining 1 fold as the test set. We randomly sampled a portion of links with percentage $\gamma_s$ ($\gamma_s$ varies from $0.1$ to $0.9$) from the positive social links in the 3 folds 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 $\gamma_s$ 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 $\gamma_a$ ($\gamma_a$ varies 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 probabilities of returning to the starting point is $c \in [0, 1.0]$. In the experiment, we set $\alpha^t$ and $\alpha^s$ as $0.6$ and $c$ is set as $0.1$, whose sensitivities will be analyzed in the following parts.

## 4.3 Experiment Result

In Figure 3, we show the ROC curve of the anchor and social link prediction results. In Figure 3(a), we set $\gamma_s = 0.5$ and $\gamma_a = 0.9$ and In Figure 3(b), 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.

In Figure 4, 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 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 Figure 4. When $\gamma_a(\gamma_s)$ is small, all the baseline methods can not work well, but CLF can still achieve good performance. Figures 4(a)- 4(b) show the result of anchor link prediction, in which $\gamma_s = 0.5$ and $\gamma_a$ changes from 0.1 to 0.9, and Figures 4(c)- 4(d) are the social link prediction result, where $\gamma_a = 0.5$ and $\gamma_s$ changes from 0.1 to 0.9.

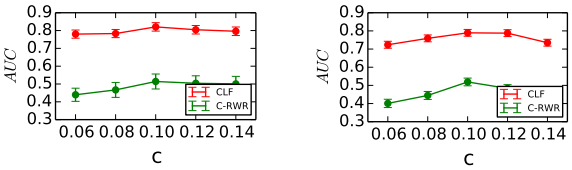
In Figure 4(a), we show the performance evaluated by AUC. The AUC of CLF is over over 40% better than MLP and over 50% better than C-RWR consistently in the whole changing range of $\gamma_a$. It demonstrates that the combination of MLP and C-RWR can lead to better results. In Figure 4(b), the performance of CLF is also better than both MLP and C-RWR under the evaluation of $Precision@30$. In Figure 4(c), 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 enhances 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 Figure 4(b), in Figure 4(d), CLF can beat all the baseline methods and perform very well when $\gamma_s$ is small. CLF can out-perform C-RWR shows that the multi-network link prediction step is essential and can work very well, while CLF can out-perform MLP demonstrates that the collective link fusion step can improve the prediction results of both anchor and social links.

In sum, CLF can out-perform all the baseline methods under the evaluation of both AUC and $Precision@30$ within the changing range of $\gamma_a$ and $\gamma_s$ in predicting both anchor and social links.

### 4.4 Parameter Analysis

CLF has three parameters in all, which are $c$, $\alpha_t$, $\alpha_s$. To analyze the effects of parameters in the experiment, we assign $\alpha_t$, $\alpha_s$ 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 Figures 5 - 6, where Figures 5(a) - 5(d) show the effects of parameter $\alpha^t$ and $\alpha^s$ and Figures 6(a) - 6(b) show the effects of parameter $c$.

In Figure 5(a) - 5(b), we only change $\alpha^t$ 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 $\alpha^t$ but CLF in Figure 5(b) 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. Figures 5(c) - 5(d) show the effects of $\alpha^s$. The performance of CLF and C-RWR is more stable compared with that in Figures 5(a) - 5(b), which shows that $\alpha^t$ has a much more significant effects than $\alpha^s$.




(a) AUC of anchor link (b) AUC of social link

Figure 6: Analysis of parameter $c$.

In Figures 6(a) - 6(b), we show the effects of parameter $c$ in the experiment where $\alpha^t$ and $\alpha^s$ are both set as 0.6. Performance of both CLF and C-RWR will varies as $c$ changes and they can achieve the best performance around $c = 0.1$.

## 5 Related Works

PU learning has been studied for several years and dozens of papers on this topic have been published. Liu et al. [Liu *et al.*, 2003] propose different settings to find the reliable negative instances in text classification. Zhao et al. propose to classify graphs with only positive and unlabeled examples in [Zhao *et al.*, 2011]. Zhang et al. are the first to propose the concept of PU link prediction in [Zhang *et al.*, 2014b] and study the PU social link prediction in multiple networks simultaneously. However, it does not address the collective prediction of anchor links and social links together, as 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 [Zhang *et al.*, 2014b].

Link prediction first proposed by Liben-Nowell et al. [Liben-Nowell and Kleinberg, 2003] 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 [Hasan *et al.*, 2006]. Meanwhile, Xiang et al. [Xiang *et al.*, 2010] develop an unsupervised model to estimate relationship strength. In heterogeneous social networks, multiple types of links can be predicted simultaneously. Namata propose a collective graph identification problem in [Namata *et al.*, 2011]. Some works labels links as positive and negative links according to their physical meanings, e.g., friendship vs. antagonism [Leskovec *et al.*, 2010], trust vs. distrust [Song and Meyer, 2014], and propose to predict these links in online social networks.

Entity identification across networks(communities) gets lots of attention in recent years. Sahraeian et al. [Sahraeian and Yoon, 2013] introduces a scalable algorithm to align proteins across large-scale PPI network. Zafarani et al. [Zafarani and Liu, 2009] propose to connect corresponding identities across communities. Iofciu et al. [Iofciu *et al.*, 2011] propose to identify common users across social tagging systems. Liu et al. [Liu *et al.*, 2013] propose an unsupervised to link users across communities. Kong et al. [Kong *et al.*, 2013] 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. [Zhang *et al.*, 2013; 2014a] propose 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. [Zhang *et al.*, 2014b] introduces an integrated PU link prediction framework to predict social links in multiple social networks concurrently.

## 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.

## 7 Acknowledgement

This work is supported in part by NSF through grants CNS-1115234 and OISE-1129076, Google Research Award, and the Pinnacle Lab at Singapore Management University.

## References

[Backstrom and Leskovec, 2011] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM*, 2011.

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

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

[Fujiwara *et al.*, 2012] Y. Fujiwara, M. Nakatsuji, M. Onizuka, and M. Kitsuregawa. Fast and exact top-k search for random walk with restart. *VLDB*, 2012.

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

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

[Iofciu *et al.*, 2011] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In *ICWSM*, 2011.

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

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

[Leskovec *et al.*, 2010] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, 2010.

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

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

[Liu *et al.*, 2013] J. Liu, F. Zhang, X. Song, Y. Song, C. Lin, and H. Hon. What's in a name?: An unsupervised approach to link users across communities. In *WSDM*, 2013.

[Namata *et al.*, 2011] G. Namata, S. Kok, and L. Getoor. Collective graph identification. In *KDD*, 2011.

[Perkins and Salomon, 1992] D. Perkins and G. Salomon. Transfer of learning. 1992.

[Sahraeian and Yoon, 2013] S. Sahraeian and B. Yoon. Smetana: Accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. *PLoS One*, 2013.

[Song and Meyer, 2014] D. Song and D. Meyer. A model of consistent node types in signed directed social networks. 2014.

[Tong *et al.*, 2006] H. Tong, C. Faloutsos, and J. Pan. Fast random walk with restart and its applications. In *ICDM*, 2006.

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

[Xiang *et al.*, 2010] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW*, 2010.

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

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

[Zafarani and Liu, 2009] R. Zafarani and H. Liu. Connecting corresponding identities across communities. In *ICWSM*, 2009.

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

[Zhang *et al.*, 2014a] J. Zhang, X. Kong, and P. Yu. Transferring heterogeneous links across location-based social networks. In *WSDM*, 2014.

[Zhang *et al.*, 2014b] J. Zhang, P. Yu, and Z. Zhou. Meta-path based multi-network collective link prediction. In *KDD*, 2014.

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