TransLink: User Identity Linkage across Heterogeneous Social Networks via Translating Embeddings

Jingya Zhou
School of Computer Science and Technology
Soochow University
Suzhou, P.R. China
jy_zhou@suda.edu.cn

Jianxi Fan
School of Computer Science and Technology
Soochow University
Suzhou, P.R. China
jxfan@suda.edu.cn

***Abstract*—Nowadays people tend to create accounts with multiple social networks (SNs) to enjoy a variety of social network services. User identity linkage (UIL) aims to identify those multiple accounts belonging to a same person. UIL is of great importance to user behavior understanding and prediction, information dissemination, viral marketing, wellness diagnosis, etc. Most of existing solutions typically rely on the embedding of either user's attributes or behaviors into a latent vector space, and establish anchor link based on vector distance. However, these efforts still face challenges originated from the heterogeneity of SNs, incompleteness of user information and lack of enough known anchor links. In this paper, we investigate the UIL problem by presenting a translation-modeling approach – TransLink. It jointly embeds both users and interactive behaviors of various SNs into a unified low-dimensional representation space according to a set of known anchor links. More specifically, we primarily study three typical SNs, e.g, Twitter, Foursquare and Instagram. Before embedding, we abstract schemas of three SNs and extract interaction metapaths for each SN. By doing this we can efficiently address the first two challenges. Furthermore, iterative linkage can ensure linkage performance by using a very small set of known anchor links. Experiment results on two real-world datasets demonstrate the superiority of TransLink over the state-of-the-art approaches.**

***Index Terms*—User identity linkage, Social networks, Network alignment, Translating embeddings**

I. Introduction

The prevalence of social networks (SNs) have spawned a variety of social network services and significantly enrich people's daily life. For example, Twitter, Foursquare and Instagram help users to generate and share texts, locations and pictures online respectively. Each SN can only partially reflect people's daily life, in order to reflect the whole social life, so we have to fuse multiple SNs. In fact today's SNs explicitly coexist but most of them fuse implicitly, because they usually are maintained by different providers and there exist many potential anchor users across varied SNs. Here anchor user refers to individual who has more than one account with varied SNs. It is reported that 69% of online active users have more than three SN accounts[1], where 53% of Twitter users are using Instagram as well[2], so anchor users become the key to SN fusion. Obtaining anchor users across multiple SNs will help to study user behavior understanding and prediction [1]–[3], information dissemination [4], viral marketing [5], [6], wellness diagnosis [7], etc. However, few users voluntarily declare their multiple SN accounts. User identity Linkage (UIL) [8] aims to identify latent anchor users across multiple SNs by connecting anchor links between their accounts.

In recent years, many studies have focused on the UIL problem. Existing linkage approaches can be generally divided into three main categories: *a) Attribute-based approaches* [9]–[13], which infer latent anchor links among accounts by calculating their distances of attributes, such as username, location, avatar, etc. *b) User generated content (UGC) based approaches* [14]–[16], which capture special characteristics of user identities by extracting features of UGC, such as interest, writing style, trajectory, etc. *c) Network-based approaches* [17]–[20], in which the connectivity of network topology structures is explored to measure the similarities of neighborhood and network features. Network representation learning techniques are frequently used to make the features preserving the original network structure. Though many efforts have been devoted to solve the UIL problem, the following challenges make it still hard to be addressed.

First, most of existing approaches focus on anchor user detection among similar types of SNs, e.g., Facebook and Twitter are blogging sites that provide text-dominated blogging services. Both SNs can easily share content semantic space without additional domain knowledge. As for heterogeneous SNs, they suffer from very big differences spanned from network structure, user behavior to user information (including attributes and UGC). For example, a Twitter user posts tweets while a Foursquare user check-ins at a specific location. Network heterogeneity makes the problem more complicated and challenging.

Second, user information in multiple SNs are sparse and

[1]http://www.pewinternet.org/fact-sheet/social-media/

[2]http://www.marketingcharts.com/online/majority-of-twitter-users-also-use-instagram-38941/

incomplete, where UGC are highly unstructured. For example, some of user's attributes may be missing on some SNs, while some of other's generated content on one SN may not have counterpart on other SNs.

Third, the existing supervised and semi-supervised approaches achieve better linkage accuracy than unsupervised approaches, but they suffer from lack of sufficient number of known anchor links. Though human-involved approach can provide a set of known anchor links with high accuracy, it is time-consuming and cannot be applied to large-scale networks.

As we know, birds of a feather flock together. Identifying an anchor user not only depends on her attributes, but also on her neighbors and interactive behaviors to each neighbor. Most existing approaches neglect the semantic information of interactive behaviors. In this paper, inspired by translation-based techniques, e.g. TransE [21], CANE [22], we propose a novel approach TransLink to address UIL with the favor of translating embeddings which are used to encode both users and interactive behaviors of multiple SNs into a unified low-dimensional vector space. More specifically, TransLink consists of the following three components:

1) Extractions of network schemas and interaction metapaths. To capture network characteristics of heterogenous SNs, we abstract network schema for each SN according to its service execution logic. Based on the schema, we extract interaction metapaths from interactive behaviors for the next embeddings.

2) Translating embeddings. We extend translation-based techniques to learn both user and interactive behavior embeddings. Afterwards, we learn to map the above embeddings of multiple SNs into a unified semantic space with the favor of a small set of known anchor links.

3) Iterative identity linkage. We iteratively connect anchor links among counterpart accounts, and update the translating embeddings by appending those high-confident anchor links increasingly found by our approach into the set of known anchor links. Therefore, TransLink can ensure linkage performance by using a very small set of known anchor links.

The main contributions of this paper are summarized as follows:

- We firstly investigate the UIL problem by leveraging translating embeddings which interpret interaction relationship between users based on translation concept. Motivated by the concept, we design two types of network embeddings to realize a unified representation space across heterogeneous SNs.
- We propose a reliable iterative identity linkage mechanism in TransLink. It can maximally improve the precision and recall rate by iteratively appending the newly connected anchor links into the training set. TransLink is specifically suitable to address the UIL problem encountered across low aligned heterogeneous networks.
- We conduct extensive experiments on two types of real-world datasets which correspond to two scenarios: highly aligned networks and low aligned networks respectively. The results indicate that TransLink can achieve better performance compared to the state-of-the-art approaches.

The remainder of this paper is organized as follows. Section II summarizes the related work. Section III formalizes the UIL problem. Section IV elaborates the details of TransLink. The experimental evaluation is done in Section V, and finally this paper is concluded in Section VI.

II. Related work

Dozens of studies have been devoted to address the UIL problem. Depending on the information used to find the potential anchor links, we group existing work into the following three main categories.

a) Attribute-based approaches assume that anchor user's accounts have close similarities on their attributes, and exploit attribute information for detecting potential anchor links. Liu et al. [9] proposed an unsupervised approach to determine the similarity between usernames, and use it as the criterion for identifying anchor links. A user may register varied names on multiple SNs. Zafarani et al. [10] found that users often exhibit certain behavioral patterns when selecting usernames, and proposed a supervised approach to learn patterns and identify anchor users. However, different people may use the same username. It is hard to achieve satisfactory linkage precision only by usernames. Zhang et al. [11] improved accuracy by presenting a naive bayes classifier where username, location, language and avatar are taken into account. Mu et al. [13] focused on user's intrinsic structure by projecting user accounts into latent user space based on user attributes. It can achieve an acceptable precision only if user's attribute information is complete.

b) UGC-based approaches insist that UGC contains more potential valuable information for user identification. Liu et al. [14] studied writing style enhanced UIL. The writing style refers to personalized wording and emoticon adoption extracted from UGC such as tweets and comments. But writing-style based approach only works for text-dominated SNs. Nie et al. [15] proposed to identify users based on their core interests which can be obtained by topic extraction from temporal and post information. Interest is a coarse grained criterion and we cannot distinguish different users precisely according to their interests. Riederer et al. [16] adopted a trajectory-based framework which extracts trajectory from a series of timestamped location data and captures the unique footprints of users' activities. This framework cannot be used for anchor user linkage across heterogenous SNs, e.g., user linkage between text-dominated SNs and location-based social networks (LBSNs).

c) Network-based approaches explore user identification based on network information which can reflect user's social characteristics to a great extent. Zhang et al. [17] utilized neighborhood-based features directly to capture the match degree of two accounts. Due to the successes of network representation learning in many applications, many recent work [18]–[20], [23] begin to extract latent network features by means of network representation learning techniques. For

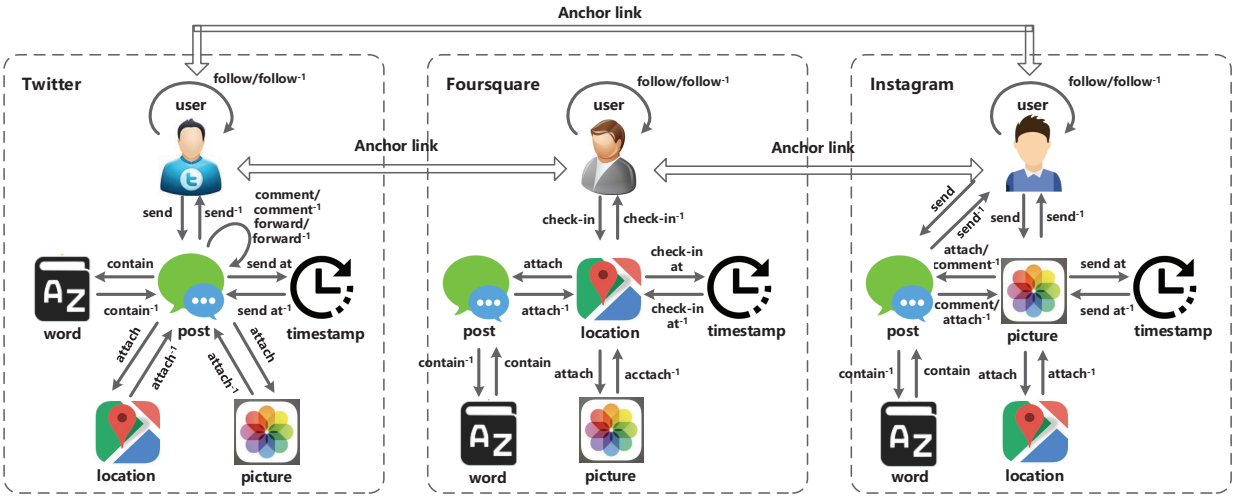


Fig. 1. An illustration of schemas of three aligned SNs.

example, Liu et al. [19] leveraged a network representation learning technique for learning both follower-ship and followee-ship of individual users, and embedded them into a common space for further identification. Zhang et al. [23] transformed the partial co-alignment problem into a joint optimization problem by incorporating both attributes and network features. These approaches typically embed users via learning separately in different SNs. But the features learned from different sources have varied feature spaces, and they may not be suitable for being concatenated as input features. Tan et al. [18] combined two SNs together as an entire network and map it to a hypergraph. Instead of using pairwise neighborhood relationship they rely on multi-neighborhood relationship for the purpose of learning more useful latent network features. Zhou et al. [20] encoded users into vector representation based on DeepWalk [24] to capture local and global network structures, and presented a supervised linkage learning method to learn the mapping function between two SNs. Deep neural networks can help to enhance the learning results, but they neglect the valuable semantic information hidden on the edge, which is of critical importance for identification. Translation-based techniques have strong abilities to represent vertices and edges given a network graph. They model relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. For instance, as a widely-used method for knowledge representation learning, TransE [21] can project both entities and relations into a continuous low-dimensional vector space. Different from the knowledge graph which has a well pre-defined relation categories, the interactions between users in SNs exhibit rich and variant meanings. For example, the following behaviors to a user in social media may be caused by varied reasons. CANE [22] combines text vector representation and structure vector representation together, and utilizes text information to explain the relationships between vertices. Similar to [14], [16], CANE only works for homogeneous networks. Moreover, existing work use the same type of graph to represent a social network. The social graph only consists of user set and edge set, cannot reflet the heterogeneity of different SNs. Our approach distinguishes heterogeneity of different SNs by extracting network schemas which contributes to appropriate embeddings on different SNs. Hence TransLink can capture more knowledge about interactive behaviors from multiple SNs. In addition, we perform joint embeddings based on a small set of known anchor links to unify semantic spaces.

III. UIL Problem Formulation

In this section, we introduce preliminary concepts used in this paper and define the UIL problem. Let $\{\mathrm{G}_1,\mathrm{G}_2,...,\mathrm{G}_m\}$ denote the set of SNs available for alignment, where G_i is the social network graph and is defined as follows:

Definition 1. (Social Network Graph) *The social network is represented by a directed graph $G = (V, E)$, where V is the set of vertices and $E = V \times V$ is the set of edges.*

Different from social network graph defined in prior work, where a vertex corresponds to a user, here a vertex can also represent other entity such as blog, location, picture, etc. The types of entities depend on the specific social network (we elaborate this in the next section). User set U is the subset of V, i.e., $U \subset V$, and the set of connections between users is also the subset of E.

Without loss of generality, we primarily focus on UIL between two heterogeneous SNs, while the settings of two SNs can be easily generalized to multiple networks.

Definition 2. (User Identity Linkage) *Given two SNs G_i and G_j, the task of user identity linkage is to find out every pair of users $u_a^{(i)} \in U_i$ and $u_x^{(j)} \in U_j$, such that they belong to a same real person, i.e., $u_a^{(i)} = u_x^{(j)}$.*

IV. TransLink Framework

Facing plenty of heterogenous SNs, in this paper, we specifically choose Twitter, Foursquare and Istagram as underlying SN platforms for user identity linkage. The reason is twofold: a) they typically represent three distinct SNs with different

network structures, different user behaviors and even different user information; b) three SNs are currently most famous platforms regarding micro-blogging, location check-in and picture-sharing respectively. In this section, we will elaborate three parts of TransLink.

A. Extractions of Network Schemas and interaction metapaths

To better capture the heterogeneities among three SNs, we borrow the concept of network schema [25] to model the basic structure of each network, and its definition is described as follows:

Definition 3. (Network Schema) *The schema of a social network G is a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of G's vertex types and $\mathcal{E}$ is the set of G's edge types.*

Fig .1 illustrates schemas of three aligned SNs, and each of them consists of various types of vertices. For example, Twitter users send posts which mainly consist of a group of words, while sometimes posts may attach non-text information such as locations and pictures. If a user u follows another user v, u can comment or forward v's posts, and each posting record has a timestamp. The vertex set of Twitter schema contains six types of vertices including user, post, timestamp, word, location and picture. While the edge set of Twitter schema contains 14 types of edges including $follow/follow^{-1}$, $send/send^{-1}$, $contain/contain^{-1}$, $send\ at/send\ at^{-1}$, $attach/attach^{-1}$, $comment/comment^{-1}$ and $forward/forward^{-1}$, where $follow^{-1}$ is the inverse relationship of $follow$, e.g., user u $follows$ v can also be regarded as that user v is followed by u. Foursquare is a famous location-based SN, which allows users to record trajectory information via checking-in at locations. The contents generated by a Foursquare user are locations, and sometimes may attach posts and pictures. By contrast, an Instagram user often records her life by sending pictures, while posts and locations are also allowed to be attached. Three SNs' schemas have the same vertex set, i.e., three SNs have the same types of vertices. But the edge sets of three schemas are different due to different SN service execution logics.

After obtaining network schema, the interaction between any pair of users can be represented by a path connecting them. To categorize all possible paths in three heterogeneous networks, we define the concept of interaction metapath as follows:

Definition 4. (Interaction Metapath) *Given a schema $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of social network $G = (V, E)$, the interaction metapath p_{ab} between two users u_a and u_b consists of a sequence of vertices and edges such that $p_{ab} = u_a \xrightarrow{r_1} v_1 \xrightarrow{r_2} v_2 \xrightarrow{r_3} ... \xrightarrow{r_k} u_b$, where $u_a, u_b \in V$, $r_1, ..., r_k \in \mathcal{E}$, k is the path length.*

As shown in Fig. 2, based on the network schemas studied above, we extract interaction metapaths to cover nearly all possible interaction instances in three heterogenous SNs. For example, there are three types of interactive behavior in Twitter: a) user u_b follows u_a (equivalent to u_a follow^{-1} u_b); b) u_b comments u_a's post; c) u_b forwards u_a's post. Each of them corresponds to an interaction metapath. If u_a and u_b are Foursquare users, besides of followship, they may interact

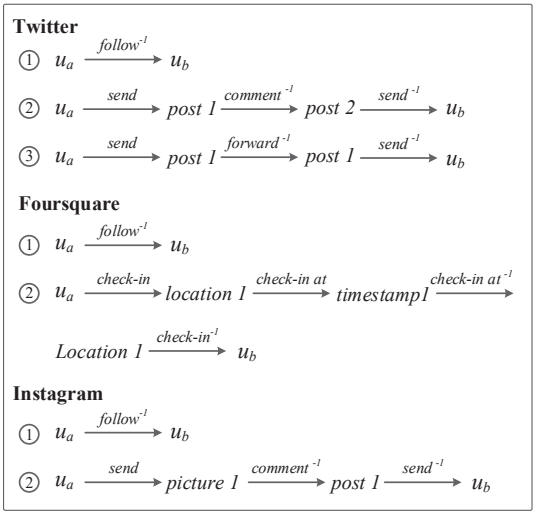


Fig. 2. Three sets of interaction metapaths extracted from three SNs.

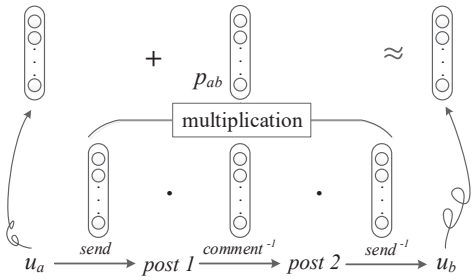


Fig. 3. A simple example to illustrate the metapath embedding.

by coexisting at the same place. While the most frequent interactive behaviors in Instagram are following others and commenting on their pictures.

B. Translating Embeddings

As an effective method for feature extraction, embedding projects network vertices into a low-dimensional latent space where each vertex is encoded as a vector representing its own feature. As for SNs, there are valuable semantic information hidden on the edge, and the features extracted via vertex embedding is limited. Different from traditional embeddings, translation-based techniques can embed both vertices and edges simultaneously. Translation means the movement that changes the position of a vector in representation space. In a typical translation-based framework, triple (head vertex, relation edge, tail vertex) acts as the basic unit to represent the relationship among three components. More specifically, a triple (h, e, t) implies that tail vertex t should be close to head vertex h plus edge e in representation space, i,e., $h + e \approx t$. We use interaction triple (u_a, p_{ab}, u_b) to describe the interactive fact between users u_a and u_b, where p_{ab} is the interaction metapath between them. Note that we do not strictly distinguish vertex/edge/metapath from their vector representations in this paper.

In order to elaborate clearly, we redefine social network graph by adding the set T of triples, i.e., $G = (V, E, T)$, and

divide translating embeddings into the following two steps:

1) Intra-network Embeddings: The purpose of intra-network embeddings is to project users and interaction meta-paths within one SN into a low-dimensional latent space such that every triple meets the aforementioned approximate equation. For a triple (u_a, p_{ab}, u_b), we define an energy function as $E(u_a, p_{ab}, u_b) = \|u_a + p_{ab} - u_b\|$ denoting energy of translation via metapath p_{ab}. Then the energy function of translations from u_a to u_b is defined as follow:

$$E(u_a, \mathcal{P}_{ab}, u_b) = \frac{1}{\Gamma} \sum_{p_{ab} \in \mathcal{P}_{ab}} R(p_{ab}|u_a, u_b) E(u_a, p_{ab}, u_b), \tag{1}$$

where $\mathcal{P}_{ab}$ is the set of metapaths from u_a to u_b, $R(p_{ab}|u_a, u_b)$ indicates the reliability of metapath p_{ab} given user pair (u_a, u_b), and Γ is the normalization factor, i.e., $\Gamma = \sum_{p_{ab} \in \mathcal{P}_{ab}} R(p_{ab}|u_a, u_b)$.

The semantic meaning of a metapath primarily relies on its involved edges. Thus it is reasonable for us to build metapath embeddings via semantic composition of edge embeddings. As illustrated in Fig. 3, the embedding of metapath p_{ab} is composed by embeddings of $send$, $comment^{-1}$ and $send^{-1}$. The composition here is defined as the multiplication operation for its simplicity and efficiency, so the metapath embedding is defined as

$$p_{ab} = r_1 \cdot r_2 \cdot \ldots \cdot r_k. \tag{2}$$

We define a margin-based score function as follows, and use it as the training objective to guide embeddings.

$$\begin{aligned} S_{intra} = \sum_{T \in \{T^{(i)} | i \in [1,m]\}} \sum_{(u_a, r, u_b) \in T} (L(u_a, r, u_b) \\ + \tfrac{1}{\Gamma} \sum_{p_{ab} \in \mathcal{P}_{ab}} R(p_{ab}|u_a, u_b)\ L(p_{ab}, r)), \end{aligned} \tag{3}$$

where $L(u_a, r, u_b)$ and $L(p_{ab}, r)$ are two margin-based loss functions with respect to triples (u_a, r, u_b) and (p_{ab}, r) respectively. Their definitions are given by

$$\begin{cases} L(u_a, r, u_b) = \sum\limits_{(u'_a, r', u'_b) \in T^-} \max\{0, \\ \qquad \gamma + E(u_a, r, u_b) - E(u'_a, r', u'_b)\}, \\ L(p_{ab}, r) = \sum\limits_{(u_a, r', u_b) \in T^-} \max\{0, \\ \qquad \gamma + E(p_{ab}, r) - E(p'_{ab}, r')\}, \end{cases} \tag{4}$$

where T^{-1} is the negative sample set of T, and can be obtained by replacing one of three components in a triple:

$$\begin{aligned} T^- = \{(u'_a, r, u_b) | u'_a \in V\} \cup \{(u_a, r, u'_b) | u'_b \in V\} \\ \cup \{(u_a, r', u_b) | r' \in E\}, (u_a, r, u_b) \in T. \end{aligned} \tag{5}$$

2) Inter-network Embeddings: The first step are conducted separately in heterogeneous SNs, so the embeddings belong to different spaces. For identity linkage, these embeddings should be merged into a unified representation space, which is the purpose of inter-network embeddings. We conduct inter-network embeddings supervised by a set of known anchor links. Given two users $u_a^{(i)} \in U_i, u_x^{(j)} \in U_j$, if they actually belong to the same person, there must be an anchor link

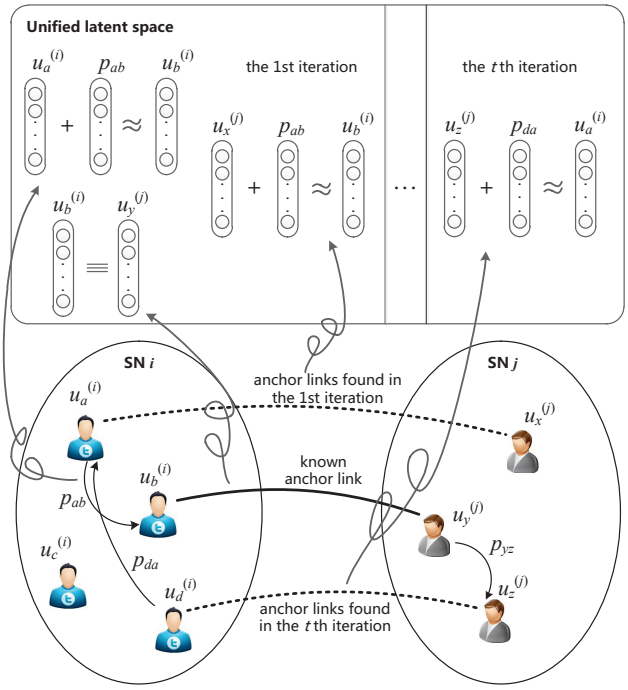


Fig. 4. An illustration of iterative identity linkage between two heterogeneous SNs.

$r_{ax}^{(ij)}$ between them, such that $u_a^{(i)} + r_{ax}^{(ij)} \approx u_x^{(j)}$. The energy function of inter-network embeddings is thus defined as:

$$E(u_a^{(i)}, u_x^{(j)}) = \left\| u_a^{(i)} + r_{ax}^{(ij)} - u_x^{(j)} \right\|. \tag{6}$$

The score function is accordingly defined as the sum of energy functions over the set $\mathbb{Z}$ of known anchor links.

$$S_{inter} = \sum_{r_{ax}^{(ij)} \in \mathbb{Z}} \lambda E(u_a{}^{(i)}, u_x{}^{(j)}), \tag{7}$$

where λ is a weighted factor.

C. Iterative Identity Linkage

Based on the aforementioned embeddings, we continue to conduct identity linkage in the unified latent space. The similarity between users is measured by the energy function defined in Eq. (6). As illustrated in Fig. 4, for a non-anchor user $u_a{}^{(i)} \in U_i$, our objective is to find the nearest non-anchor user $u_x{}^{(j)} \in U_j$ who can minimize the energy, i.e., $\arg\min_{u_x{}^{(j)}} E(u_a{}^{(i)}, u_x{}^{(j)})$. More specifically, given an anchor link connecting anchor users $u_b{}^{(i)} \in U_i$ and $u_y{}^{(j)} \in U_j$, if we have $u_x{}^{(j)} + p_{ab} \approx u_b{}^{(i)}$, then $u_a{}^{(i)}$ and $u_x{}^{(j)}$ must be anchor users and should be connected by an anchor link. Here we define a threshold ψ as a hyper-parameter for similarity judgement. If $E(u_a{}^{(i)}, u_x{}^{(j)}) < \psi$, it will be confident that $u_x{}^{(j)} + p_{ab} \approx u_b{}^{(i)}$ is very likely to hold, otherwise we will not think they are anchor users.

As the given labeled information, a large set $\mathbb{Z}$ of known anchor links is of critical importance for improving UIL performance. We notice that the newly connected anchor links can be appended into set $\mathbb{Z}$ to update inter-network

embeddings and find more anchor links in an iterative manner. Nevertheless, there are inevitable errors in identity linkage, UIL may suffer from error propagation once connecting wrong anchor links. To address this drawback, we propose a reliable iterative identity linkage to maximize the number of anchor links and meanwhile the linkage accuracy does not decline. Specifically, let the newly connected anchor links be iteratively appended into a predefined set $\mathbb{Z}'$, and define a reliability score for each new link (e.g., connecting ${u_a}^{(i)}, {u_x}^{(j)}$) as follows:

$$R({u_a}^{(i)}, {u_x}^{(j)}) = \phi\left(\alpha(\psi - E({u_a}^{(i)}, {u_x}^{(j)}))\right), \quad (8)$$

where $\phi(\cdot)$ is a sigmoid function, α is a hyper-parameter, $\alpha > 0$. Then we can obtain the score of iterative identity linkage:

$$R(i,j) = \sum_{({u_a}^{(i)}, {u_x}^{(j)}) \in \mathbb{Z}'} R({u_a}^{(i)}, {u_x}^{(j)}) \Big(\Phi({u_a}^{(i)}, {u_x}^{(j)}) + \Phi({u_x}^{(j)}, {u_a}^{(i)})\Big), \quad (9)$$

where $\Phi({u_a}^{(i)}, {u_x}^{(j)})$ indicates the loss function of anchor link, and it is calculated by adding losses $L'(\cdot)$ on triples, so we have

$$\begin{aligned} \Phi({u_a}^{(i)}, {u_x}^{(j)}) &= \sum_{({u_a}^{(i)}, r, u_b)} L'({u_x}^{(j)}, r, u_b) \\ &\quad + \sum_{(u_a, r, {u_a}^{(i)})} L'(u_a, r, {u_x}^{(j)}), \\ where\, L'(u_a, r, u_b) &= L(u_a, r, u_b) \\ &\quad + \tfrac{1}{\Gamma} \sum_{p_{ab} \in \mathcal{P}_{ab}} R(p_{ab}|u_a, u_b) L(p_{ab}, r). \end{aligned} \quad (10)$$

We calculate the reliability score R_{ij} at the end of each iteration. If the score exceeds a predefined threshold θ, it implies that they may be many wrong anchor links connected in the current iteration. Then the iterative identity linkage procedure must rollback to the previous iteration and update $\mathbb{Z}'$ till the reliability score does not exceed threshold, i.e., $R_{ij} \leq \theta$.

The proposed approach TransLink is illustrated by Algorithm 1, which firstly extracts network schemas and interaction metapaths from two SNs (lines 1-2). After that, intra-network embeddings project each SN's users and interaction metapaths into a latent space (lines 3-4), while inter-network embeddings joint these embeddings into a unified space with the favor of known anchor links (line 7). To accurately find more anchor links, both inter-network embeddings and user identity linkage are conducted iteratively. In each iteration the newly connected anchor links will be appended into the current anchor link set $\mathbb{A}$ as long as wrong links are tiny (lines 13-16). Finally, the algorithm terminates when no new anchor links were found (line 17). For SNs G_i and G_j, the number of anchor links would not exceed the size of smaller user set, i.e., $\min(|U_i|, |U_j|)$. Even though there is only one anchor link found in each iteration, our algorithm is bound to converge after $\min(|U_i|, |U_j|) - |\mathbb{Z}|$ iterations at most. In real world, SNs are usually partially aligned and a batch of anchor links would be found in each iteration, so our algorithm would converge even faster.

Algorithm 1: *TransLink*

Input: SNs G_i and G_j, the set $\mathbb{Z}$ of known anchor links, weight and parameters λ, α, ψ, θ;
Output: the set $\mathbb{Z}'$ of anchor links found;

1 Extract network schemas $\mathcal{G}_i$ and $\mathcal{G}_j$ from G_i and G_j;
2 Extract interaction metapaths based on schemas;
3 **for** *each SN* **do**
4 Embed its users and interaction metapaths by minimizing S_{intra} defined in Eq. (3);
5 $\mathbb{A} \leftarrow \mathbb{Z}$, $\mathbb{Z}' \leftarrow \emptyset$;
6 **do**
7 Joint network embeddings supervised by $\mathbb{A}$;
8 $\mathbb{C} \leftarrow \emptyset$;
9 **for** *each pair of non-anchor users* $({u_a}^{(i)}, {u_x}^{(j)})$, ${u_a}^{(i)} \in V_i, {u_x}^{(j)} \in V_j$ **do**
10 **if** $E({u_a}^{(i)}, {u_x}^{(j)}) < \psi$ **then**
11 $\mathbb{Z}' \leftarrow \cup\{r_{ax}^{ij}\}$;
12 $\mathbb{C} \leftarrow \mathbb{C} \cup \{r_{ax}^{ij}\}$;
13 **if** $R(i,j) > \theta$ **then**
14 $\mathbb{Z}' \leftarrow \mathbb{Z}' \backslash \mathbb{C}$;
15 **Continue**;
16 $\mathbb{A} \leftarrow \mathbb{A} \cup \mathbb{Z}'$
17 **while** $\mathbb{C} \neq \emptyset$;
18 **return** $\mathbb{Z}'$;

D. Discussions

As we pointed out in Section IV-B, metapath embeddings should perform semantic composition of edge embeddings. Note that multiplication is not the only choice and there are other composition methods. For example, recurrent neural network (RNN) is also recently used for semantic composition [26]. To incorporate RNN-based composition in our framework, we need to build a matrix $\mathbf{M}$ such that

$$c_l = \Theta(\mathbf{M}[c_{l-1}; r_l]), \quad (11)$$

where $\Theta(\cdot)$ is a non-linearity or identical function, and $[x; y]$ indicates the concatenation of two vectors. Starting from $c_1 = r_1$ we can obtain the path embedding $p_{ab} = c_k$ by recursively performing RNN along the path.

V. Experiments

In this section, we first introduce the datasets and experiment settings. After that we show the experimental results together with analysis on the performance comparison.

A. Datasets

Till now, there are no agreed benchmark datasets for UIL tasks [8]. We use two real-world datasets in our following experiments. One is labeled as **TF** which is crawled from Twitter and Foursquare, and is provided by Zhang et al. [27]. **TF** contains 5,223 Twitter users and 5,392 Foursquare users, and they share 3,388 anchor users, so the proportion of ground truth is high (about 62%∼64%). Another dataset is labeled as

TABLE I
STATISTICS OF SOCIAL NETWORK DATASETS

Dataset		Social network		
		Twitter	Foursquare	#anchor
TF	#user	5,223	5,392	3,388
	#edge	164,920	76,972	
		Twitter	Instagram	#anchor
TI	#user	217,432	282,476	23,512
	#edge	9,837,937	20,548,254	

TI which is crawled by ourselves from Twitter and Instagram. We firstly crawled 217,432 users in Twitter, and chose 10 users who claim to have Instagram accounts. Then we let 10 users as seeds to crawl more users. To obtain the ground truth, we use a shell script to match anchor users who : a) claim they have accounts in both networks; b) share their Instagram links on Twitter. Finally we found a set of 23,512 anchor users as the ground truth, which is only a small fraction of users (about 8%∼11%). Two datasets just stand for two different types of alignment: highly aligned networks and low aligned networks. The basic statistics of datasets are shown in Table I.

B. Experiment Setttings

1) Experiment Setup: In implementation, we use stochastic gradient descent (SGD) as our optimizer to train embeddings. As for hyper-parameters, we set $\alpha = 1$, $\psi = 3, \theta = 10$. To keep comparison fair, the vector dimension of all approach are identical, i.e., $d = 96$. The learning rate is set to be 0.001 by following the optimal settings in [21]. The ground truth anchor links are divided into two sets GT_1 and GT_2: GT_1 is used for training while GT_2 is combined with non-anchor users for testing. We use the proportion of GT_1 to define the training ratio, i.e., $tr = |GT_1| : |GT_1 \cup GT_2|$ and set $tr = 0.8$ by default. We repeat each experiment for 10 times and report the mean performance.

2) Comparison Approaches: We evaluate TransLink by comparing to several state-of-the-art approaches described as follows:

- *RNN-TransLink*. It is a variant of TransLink which utilizes RNN-based composition instead of multiplication for metapath embeddings.
- *Input Output Network Embedding (IONE)* [19]. It learns a network embedding with the follower-ship/followee-ship of each user, and uses input/output context vectors to preserve the proximity and makes alignment accordingly.
- *DeepLink* [20]. It learns to encode network vertices into vector representation to capture local and global network structures, and use a policy gradient method to update the linkage.
- *K Nearest Neighbors (KNN)*. It is a simple unsupervised method based on K nearest neighbors. Here we utilize common neighbors as user features to calculate the K nearest neighbors. K is set to be the optimal value to make KNN achieving as good performance as possible, i.e., $K = 5$ and 12 for datasets **TF** and **TI** respectively.

3) Evaluation Metrics: To evaluate the performance, we mainly focus on three metrics: precision@n, recall@n and mean-rank. The precision@n indicates the fraction of true positive anchor links among the anchor links returned, where @n is used to represent the anchor link r is true positive as long as r belongs to top-n list. Similarly, recall@n indicates the fraction of true positive anchor links over the ground truth. Mean-rank is defined as the average ranks of all true positive anchor links and a lower value indicates a better performance.

C. Results

Comparison of TransLink and Its Variant. In the first group of experiments, we make a comparison between TransLink and its variant. The results are illustrated in Table II, where we notice that TransLink performs slightly better than RNN-TransLink on both datasets except for recall@1 on dataset **TF**. For highly aligned networks, RNN-TransLink gains a weak advantage on recall@1 by sacrificing precision@1 and mean-rank. In general, we conclude that two approaches perform similarly and the composition operation in metapath embeddings has a limited impact on the result of identity linkage.

TABLE II
RESULTS OF TRANSLINK AND ITS VARIANT ON TWO DATASETS

	precision@1		recall@1		mean-rank	
	TF	**TI**	**TF**	**TI**	**TF**	**TI**
TransLink	**0.3625**	**0.2136**	0.5127	**0.4253**	**15.72**	**102.7**
RNN-TransLink	0.3598	0.2083	**0.5164**	0.4118	16.28	103.1

Convergence Analysis. TransLink attempts to improve UIL performance by conducting inter-network embeddings and identity linkage iteratively. In this group of experiments, we evaluate the convergence performance by tuning the training ratio. As illustrated in Fig. 5, the number of iterations indicates a decreasing trend along with the increasing training ratio. Because increasing training ratio implies more known anchor links would be used for training and accordingly more new anchor links would be found with high confidence. Another reason is that the latent anchor links waiting to be found in testing set is decreasing. When $tr > 0.3$, the number of iterations almost no longer changes with training ratio. It is because that TransLink can quickly find a great proportion of ground truth anchor links via iterations as long as the size of training set reaches a certain scale. Interestingly we still notice that TransLink converges slower on dataset **TI**. As far as **TI** concerned, the limited knowledge learned from training set is probably to connect many wrong anchor links and then TransLink must rollback to avoid error propagation. Rollback operations increase the number of iterations.

Comparison Among Multiple Approaches. In this group of experiments, we compare TransLink with the state-of-the-art approaches based on three metrics. Fig. 6 shows the comparison results on precision@n when n varies among {1, 5, 10, 20, 50, 100}. Compared to the accurate matching (i.e., $n = 1$), precision@n can be improved greatly by using the set of top-n candidates for matching. Our approach outperforms

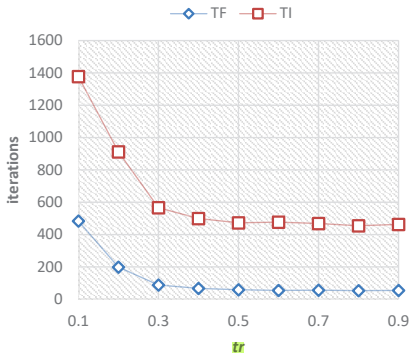
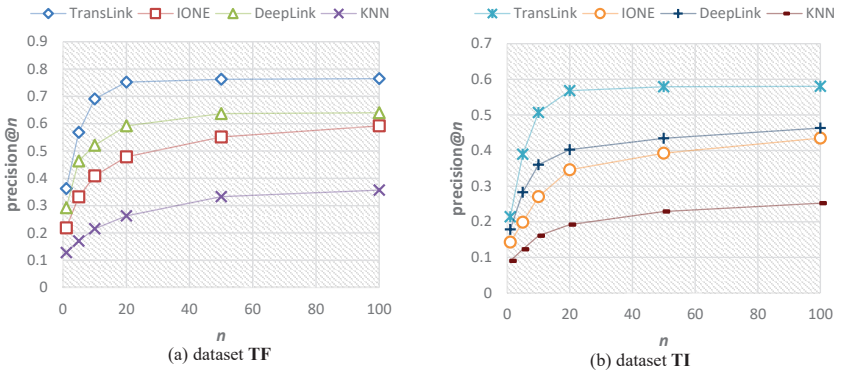
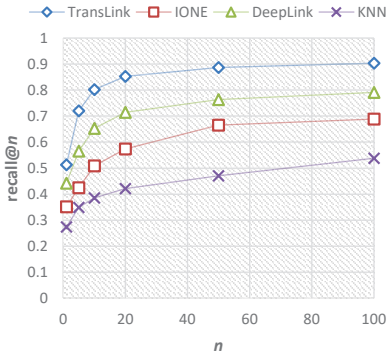


Fig. 5. The iteration changes along with the increasing training ratio.

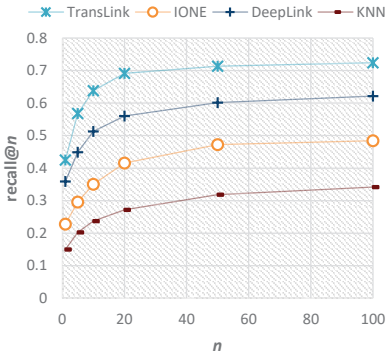


(a) dataset **TF** (b) dataset **TI**

Fig. 6. The performance comparison based on precision@n.



(a) dataset **TF**



(b) dataset **TI**

Fig. 7. The performance comparison based on recall@n.

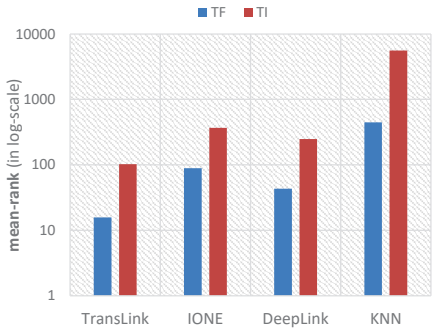


Fig. 8. The performance comparison based on mean-rank.

TABLE III
THE PERFORMANCE COMPARISON UNDER DIFFERENT TRAINING RATIOS

		precision@100				recall@100				mean-rank			
	tr	0.1	0.3	0.5	0.9	0.1	0.3	0.5	0.9	0.1	0.3	0.5	0.9
TF	TransLink	**0.2380**	**0.7103**	**0.7467**	**0.7653**	**0.3117**	**0.8233**	**0.8668**	**0.9031**	**156.3**	**37.15**	**22.58**	**15.72**
	IONE	0.1765	0.3660	0.4906	0.5921	0.2088	0.4254	0.5835	0.6889	532.7	287.2	186.14	88.66
	DeepLink	0.2107	0.4784	0.5721	0.6403	0.2490	0.5798	0.7029	0.7914	426.4	232.1	103.05	43.42
TI	TransLink	**0.1632**	**0.4897**	**0.5667**	**0.5812**	**0.2230**	**0.5816**	**0.6931**	**0.7243**	**1,025**	**312.3**	**171.8**	**102.7**
	IONE	0.1128	0.2187	0.3323	0.4345	0.1862	0.3014	0.3735	0.4845	4,327	2,108	788.7	376.3
	DeepLink	0.1386	0.2939	0.3891	0.4637	0.2021	0.4220	0.5331	0.6218	2,896	1,097	482.1	251.6

the others and improves precision@100 to 76.5%. Moreover, the improvement would become very small after n grows to a certain value. For example, the improvement of TansLink tends to be zero after $n > 20$, which implies TransLink can use less candidates for precision improvement.

Fig. 7 reports the changes of recall@n under different values of n, from which we find a similar trend as precision@n has. KNN does not have training procedure and directly utilizes feature vectors of different SNs to do identity linkage, so its performance fall behind others on both datasets. IONE learns abundant context vectors via training and achieves much better performance than KNN. Compared to IONE and KNN, DeepLink iterates millions of dual-learning processes to update vertex features, so it can search relatively more anchor links. Different from traditional embeddings that primarily focus on vertex embedding, TransLink encodes both users and interaction metapaths into a unified latent space by leveraging translating embeddings. The anchor links are connected based on the representations of users and their interactions, which can achieve a better UIL. As for mean-rank, Fig. 8 shows our approach obtains the lowest mean-rank on both datasets, which accordingly interprets the highest performance on precision@n and recall@n.

Besides of n, we also make comparisons under different training ratios and report the results in Table III. KNN needs no training, so we remove its performance from the table. As can be seen in the table, the percentage of known anchor links used for training has a significant impact upon three metrics of approaches. For each approach, training with more known anchor links will contribute to learning more valuable knowledge and improve precision@100 and recall@100 accordingly. Among these approaches, TransLink achieves the best performance,

and its observed precision@100 and recall@100 on **TF** reach 76.53% and 90.31%. Even for IONE which performs worst, its performance on **TF** achieves 59.21% and 68.89%. In real world, most of SNs are generally partially aligned. Besides, the actual ground truth of aligned networks is very difficult to obtain, and the already obtained ground truth is a tiny fraction of large-scale networks. It is difficult for current approaches to address the above issue. It is particularly worth mentioning that TransLink explores to reach an acceptable performance by means of incremental iterations that can automatically find more anchor links with high confidence to compensate for the small size of training set.

VI. Conclusion

In this paper, we investigated the UIL problem by leveraging translating embedding techniques and proposed a novel UIL approach, called TransLink. TransLink is specifically designed for heterogeneous SNs which have considerable differences on many aspects including network structure, user behavior and user information. We extracted network schemas and interaction metapaths to capture the heterogeneities of SNs, and embedded both users and their interactions of different SNs into a unified latent space. Meanwhile TransLink utilized a reliable iterative identity linkage to ensure the linkage performance with only a small set of known anchor links. The extensive experiments on two different types of real-world datasets demonstate that TransLink outperforms the state-of-the-art approaches.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No. 61502328, No. 61572337, No. 61672370), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1701173B), Open Project Program of Jiangsu Provincial Key Laboratory for Computer Information Processing Technology (No. KJS1740) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] A. Farseev, L. Nie, M. Akbari, and T. Chua, "Harvesting multiple sources for user profile learning: a big data study," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR)*, 2015, pp. 235–242.

[2] X. Song, L. Nie, L. Zhang, M. Akbari, and T. Chua, "Multiple social network learning and its application in volunteerism tendency prediction," in *Proceedings of the 38th ACM SIGIR*, 2015, pp. 213–222.

[3] X. Wu, Z. Hu, X. Fu, L. Fu, X. Wang, and S. Lu, "Social network deanonymization with overlapping communities: Analysis, algorithm and experiments," in *Proceedings of the 37th IEEE International Conference on Computer Communications (INFOCOM)*, 2018.

[4] Y. T. Wen, P. Lei, W. Peng, and X. Zhou, "Exploring social influence on location-based social networks," in *Proceedings of IEEE International Conference on Data Mining (ICDM)*, 2014, pp. 1043–1048.

[5] J. Li, Z. Cai, M. Yan, and Y. Li, "Using crowdsourced data in location-based social networks to explore influence maximization," in *Proceedings of the 35th IEEE International Conference on Computer Communications (INFOCOM)*, 2016.

[6] J. Zhou, J. Fan, J. Wang, X. Wang, and L. Li, "Cost-efficient viral marketing in online social networks," *World Wide Web, accept*, 2018.

[7] A. Farseev and T. Chua, "Tweetfit: Fusing multiple social media and sensor data for wellness profile learning," in *Proceedings of the 31st Conference on Artificial Intelligence (AAAI)*, 2017, pp. 95–101.

[8] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *SIGKDD Explorations*, vol. 18, no. 2, pp. 5–17, 2016.

[9] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "Whats in a name? an unsupervised approach to link users across communities," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*, 2013, pp. 495–504.

[10] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2013, pp. 41–49.

[11] H. Zhang, M. Kan, Y. Liu, and S. Ma, "Online social network profile linkage," in *Proceedings of 10th AIRS Conference on Information Retrieval Technology*, 2014, pp. 197–208.

[12] Y. Shen and H. Jin, "Controllable information sharing for user accounts linkage across multiple online social networks," in *Proceedings of the 23rd ACM on Conference on Information and Knowledge Management (CIKM)*, 2014, pp. 381–390.

[13] X. Mu, F. Zhu, E. Lim, J. Xiao, J. Wang, and Z. Zhou, "User identity linkage by latent user space modelling," in *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2016, pp. 1775–1784.

[14] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "HYDRA: large-scale social identity linkage via heterogeneous behavior modeling," in *Proceedings of International Conference on Management of Data (SIGMOD)*, 2014, pp. 51–62.

[15] Y. Nie, Y. Jia, S. Li, X. Zhu, A. Li, and B. Zhou, "Identifying users across social networks based on dynamic core interests," *Neurocomputing*, vol. 210, pp. 107–115, 2016.

[16] C. J. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016, pp. 707–719.

[17] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, "COSNET: connecting heterogeneous social networks with local and global consistency," in *Proceedings of the 21th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2015, pp. 1485–1494.

[18] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen, "Mapping users across networks by manifold alignment on hypergraph," in *Proceedings of the 28th Conference on Artificial Intelligence (AAAI)*, 2014.

[19] L. Liu, W. K. Cheung, X. Li, and L. Liao, "Aligning users across social networks using network embedding," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[20] F. Zhou, L. Liu, K. Zhang, G. Trajcevski, J. Wu, and T. Zhong, "Deeplink: A deep learning approach for user identity linkage," in *Proceedings of the 37th IEEE International Conference on Computer Communications (INFOCOM)*, 2018.

[21] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proceedings of 27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013, pp. 2787–2795.

[22] C. Tu, H. Liu, Z. Liu, and M. Sun, "CANE: context-aware network embedding for relation modeling," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 1722–1731.

[23] J. Zhang and P. S. Yu, "PCT: partial co-alignment of social networks," in *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016, pp. 749–759.

[24] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: online learning of social representations," in *Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2014, pp. 701–710.

[25] Y. Sun and J. Han, *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers, 2012.

[26] A. Neelakantan, B. Roth, and A. McCallum, "Compositional vector space models for knowledge base completion," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015, pp. 156–166.

[27] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM)*, 2014, pp. 303–312.