Partially Shared Adversarial Learning For Semi-supervised Multi-platform User Identity Linkage

Chaozhuo Li
Beihang University
lichaozhuo@buaa.edu.cn

Senzhang Wang
Nanjing University of Aeronautics and Astronautics
szwang@nuaa.edu.cn

Hao Wang
Southwest Jiaotong University
hwang@my.swjtu.edu.cn

Yanbo Liang
Facebook
yliang@apache.org

Philip S. Yu
University of Illinois at Chicago
psyu@uic.edu

Zhoujun Li*
Beihang University
lizj@buaa.edu.cn

Wei Wang
Ping An Bank Co., Ltd.
ww.cs.tj@gmail.com

ABSTRACT

With the increasing popularity and diversity of social media, users tend to join multiple social platforms to enjoy different types of services. User identity linkage, which aims to link identical identities across different social platforms, has attracted increasing research attentions recently. Existing methods usually focus on pairwise identity linkage between two platforms, which cannot piece up the information from multi-sources to depict the intrinsic figures of social users. In this paper, we propose a novel adversarial learning based framework MSUIL with partially shared generators to perform Semi-supervised User Identity Linkage across Multiple social networks. The isomorphism across multiple platforms is captured as the complementary to link identities. The insight is that we aim to learn the desirable projection functions (generators) to not only minimize the distance between the distributions of user identities in arbitrary pairs of platforms, but also incorporate the available annotations as the learning guidance. The projection functions of different platform pairs share partial parameters, which ensures MSUIL can capture the interdependencies among multiple platforms and improves the model efficiency. Empirically, we evaluate our proposal over multiple datasets. The experimental results demonstrate the superiority of the proposed MSUIL model.

ACM Reference Format:
Chaozhuo Li, Senzhang Wang, Hao Wang, Yanbo Liang, Philip S. Yu, Zhoujun Li, and Wei Wang. 2019. Partially Shared Adversarial Learning For Semi-supervised Multi-platform User Identity Linkage. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357904

1 INTRODUCTION

Nowadays, social networks are becoming increasingly prevalent and diversity. Users tend to register in multiple social platforms for different purposes, such as people use LinkedIn for job seeking and join Twitter to share the precious life moments. A natural person may create an identity in each platform to present his/her unique social figure, which is usually associated with the posted texts, social relations and user behaviors in the corresponding platform. User identity linkage (UIL), which aims to link the identities of same natural person across different social platforms, has attracted more attentions considering its tremendous research challenges and practical values. UIL contributes to depicting the intrinsic characteristics of users by fusing information from multiple social sources, which benefits many industry applications, such as social recommendation [21], information diffusing prediction [25, 30, 33] and network dynamics analysis [26, 31].

Most existing works focus on the pairwise identity linkage between two platforms [4, 5, 11, 12, 14, 16–18, 22, 24, 32, 35, 40]. They usually first encapsulate the contents (e.g., attributes or the behavior patterns) of a social identity into a feature vector, and then incorporate annotations to learn a matching function to decide whether two identities from different social networks belong to a same natural person. However, in real life users tend to join more than two social networks, and directly applying the pairwise models in the multi-platform scenario cannot achieve desirable performance. Given n platforms, there exist $\frac{n\times(n-1)}{2}$ possible combinations of two platforms. It means that $O(n^2)$ independent pairwise models need to be trained from scratch, which leads to the low efficiency and resource waste. In addition, the social multiplicity is the underlying nature of social data, and ignoring the interdependencies among multiple platforms may lead to the undesirable linkage performance.

For the multi-platform identity linkage task, most existing methods all supervised models, which need a large number of annotations to learn a classifier or a ranking model [15, 20, 29, 39]. ULink [15] first maps the user identities from different platforms into a shared latent space, and then tries to minimize the distances between the linked user identities and maximize the distances between the user identities belonging to different people. However, as social platforms are independent from each other, it is extremely

[*] Corresponding author.

CIKM '19, November 3–7, 2019, Beijing, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357904

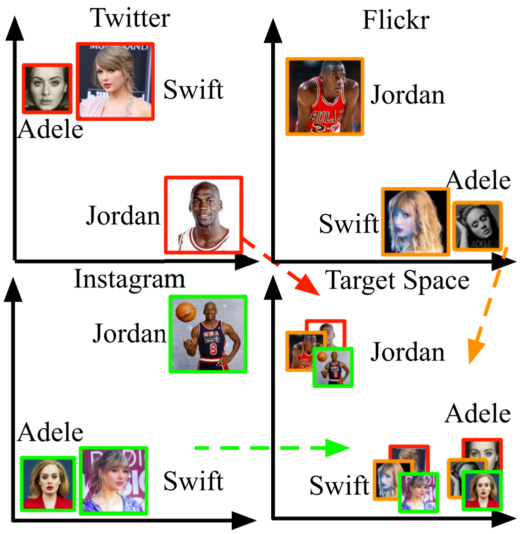


Figure 1: An illustration of the isomorphism across multiple social platforms. The dotted lines with different colors refer to the projection functions from different platforms to the target space.

time consuming and requires high labor costs to collect enough linked social identities as annotations.

In this paper, we study the novel problem of semi-supervised identity linkage across multiple platforms, which incorporates the unlabeled instances to reduce the reliance on the annotations. Following previous works [8, 12, 15], we assume the social identities in different platforms are the projections of natural people constrained by the features provided by the platforms. As shown in Figure 1, Taylor Swift, Adele are famous singers while Michael Jordan is a well-known basketball player. The social identities of Swift and Adele are much closer in all three platforms (Twitter, Flickr and Instagram) as they share common interests and tend to post music related tweets and follow other musicians. Thus, the distributions of three people in different platforms present similar shapes, which is also known as the isomorphism across social networks [13]. If we can project the identity distributions in the three social spaces into a target space (e.g., Instagram) through some operations (e.g., rotation), in which the distances between the projected distributions are minimized as shown in the bottom right corner of Figure 1, the identities belong to the same natural person will be grouped together. Similar idea has been adopted by previous work [9], but it focuses on the pairwise identity linkage and cannot fully exploit the interdependencies in the multi-platform setting.

The studied problem is challenging due to the following three reasons. First, the number of input platforms is unfixed in the multi-platform scenario. It is challenging to design a flexible framework which can be easily extended to suit a variable number of social platforms. Second, it is intractable to minimize the distances between multiple distributions simultaneously while keeping the model effectiveness. Previous work [9] has studied the distribution minimization technique in the pairwise scenario, but it cannot be directly applied on the multi-platform task due to the low efficiency. Third, incorporating multiple social sources can provide rich mutual information, which also increases the uncertainty. It would be better if few annotations can be utilized to provide extra guidances in the sample level. As the distances between distributions are minimized on the distribution level in an unsupervised manner, it is difficult to incorporate few annotations as indicators considering the totally different purposes and scenarios.

In this paper, we propose a multi-platform semi-supervised identity linkage model MSUIL to address the mentioned challenges. MSUIL aims to learn desirable projection functions to minimize the distribution distances between arbitrary pairs of social platforms. Specifically, for each input platform P_i, MSUIL contains an encoder E_i, a decoder O_i and a discriminator D_i. Encoder E_i maps the feature vectors of identities in platform P_i into a shared latent space, and decoder O_i projects the mapped vectors from the latent space into the social space P_i as the generated samples. The combination of an encoders and a decoder can be viewed as a projection function (generator) between two input social platforms. For example, encoder E_i and decoder O_j form up the projection function G_{ij} to project the identities from source platform P_i to the target platform P_j. Given n input platforms, we only need to train n encoders and n decoders to obtain n^2 projection functions, which reduces the model complexity from $O(n^2)$ to $O(n)$. Thus, MSUIL can be viewed as a partially shared adversarial learning framework. Discriminator D_i aims to distinguish the real instances in P_i from the samples generated by the decoders. Under the adversarial learning framework, the discriminator essentially estimates the approximate Wasserstein distance between the real distribution and the generated distribution. Through the competition with discriminators, the encoders and decoders will be updated to minimize the estimated Wasserstein distance. When handling a new pair of platforms, the distribution closeness information of previous pairs will be incorporated through the shared parameters to link identities, and thus the minimization of distances among multiple distributions can be latently achieved. We also design another annotation guided loss function to incorporate few annotations. The proposed MSUIL model is evaluated on two partially aligned data collections (three social networks and three academic coauthor networks), and the results demonstrate the superiority of MSUIL.

We summarize our main contributions as follows.

- We study the novel problem of semi-supervised user identity linkage under the multi-platform setting, and further propose to capture the isomorphism across multiple social networks as the guidance.
- We propose a novel adversarial learning based model MSUIL with partially shared generators to efficiently minimize the distances between multiple social distributions, and MSUIL can also incorporate few annotations as the indicators.
- Extensively, we evaluate the proposed model on multiple social networks. Experimental results demonstrate the superior performance of MSUIL.

The rest of this paper is organized as follows. Section 2 gives a brief review on related works. Then we formally define the studied problem in section 3. In section 4 we introduce the details of MSUIL model. Experimental results are shown and discussed in section 5. Finally, we give concluding remarks in section 6.

2 RELATED WORK

User identity linkage (UIL), also known as the social network alignment, aims to link the social identities belong to a same nature person across different social platforms. Existing UIL approaches can be roughly categorized into supervised, semi-supervised and unsupervised methods. Most existing methods are supervised models, which learn a ranking model or a binary classification model to distinguish the linked social identities from the unlinked ones [4, 12, 14–18, 24, 35, 38]. Zhang et al. [35] explored a probabilistic approach that utilized a domain-specific prior knowledge to perform user identity linkage. Man et al. [12] proposed an embedding based UIL model, which employed network embedding techniques with awareness of observed anchor links as supervised information to capture the major and specific structural regularities. ULink [15] first mapped the user identities from different platforms into a shared latent space, and then tried to minimize the distances between the linked user identities and maximize the distances between the user identities belonging to different people. Nie et al. [16] proposed a dynamic core interests mapping model, which jointly incorporated the social topology of users and the posted texts. Zhang et al. [38] introduced the popular graph convolutional network (GCN) to provide quality node embeddings to link identities. However, as social platforms are independent from each other, it requires high labor costs to collect enough linked social identities as annotations. Thus, some unsupervised models are proposed to link identities based on the unlabeled data [6, 7, 9, 10, 19, 27]. Liu et al. [10] first generated a set of labeled samples based on the rareness of screen names, and then trained a binary classier according to the generated samples. Lacoste-Julien et al. [7] proposed a greedy unsupervised learning based UIL model to estimate the heuristic string similarities between the texts of social identities. Li et al. [40] performed the UIL task in the distribution level and proposed an adversarial learning based method and a matrix factorization based method. Xie et al. [27] proposed a factoid embedding based model to learn the latent representations and link two user identities from different social platforms. Although unsupervised approaches can eliminate the reliance on the annotations, they usually suffer from comparatively low performance as no annotations are incorporated.

Recently several semi-supervised methods are proposed to utilize the unlabeled samples along with a few annotations to link user identities [5, 11, 22, 32, 34, 39, 40]. As semi-supervised models can both exploit the available annotations and incorporate the unlabeled samples to capture the shape of the underlying data distribution, which are more promising to link social identities effectively and efficiently. Zhong et al. [40] proposed a co-training UIL framework, which manipulated two independent models (the attribute-based model and the relationship-based model) and made them reinforce each other iteratively. Liu et al. [11] embedded the follower-ship/followee-ship into the learned embeddings and incorporated a few annotations to learn a semi-supervised classifier. Existing semi-supervised methods are designed for the pairwise identity linkage, which cannot be directly applied on the multi-platform scenario. Thus, in this paper we aim to propose a novel model to link user identities across multiple platforms in a semi-supervised manner, which is expected to incorporate the unsupervised isomorphism information along with few available annotations.

3 PROBLEM DEFINITION

In this section we formally define the studied problem. Denote the set of social media platforms as $\mathcal{P} = \{P_1, P_2, \cdots, P_n\}$. Each social platform is defined as $P_k = (V_k, F_k)$, in which V_k contains the user identities in this platform, and $F_k \in \mathbb{R}^{|V_k| \times d}$ contains the features of identities. Each row in F_k presents the d-dimensional feature vector of the corresponding social identity. Following the previous work [15], the dimensions of feature vectors in different platforms are set to the same, which can be easily satisfied by the popular network embedding methods. The studied multi-platform user identity linkage problem can be formally defined as follows:

Definition 3.1. **Multi-platform User Identity Linkage (MUIL).** Given two arbitrary social platforms P_i and P_j from the set $\mathcal{P}$, we aim to locate a set of user identity pairs $U_{ij} = \{(v_i, v_j) | v_i \in V_i, v_j \in V_j)\}$, in which v_i and v_j belong to a same natural person. The interdependencies between multiple social platforms are expected to be incorporated.

As mentioned in the introduction section, the studied problem can be transformed into the learning of a set of projection functions, which is formally defined as follows:

Definition 3.2. **Projection function learning for MUIL.** Given an arbitrary pair of input platforms P_i and P_j, we aim to learn a desirable projection function G_{ij} to project the source distribution $\mathbb{P}^I$ into the target space, which should minimize the distance between the projected distribution $G_{ij}(\mathbb{P}^I)$ and the target distribution $\mathbb{P}^J$. In addition, a set of few annotations $M_{ij} = \{(v_i, v_j) | v_i \in V_i, v_j \in V_j\}$ should also be incorporated. For a matched identity pair (v_i, v_j) in M_{ij}, G_{ij} should minimize the distance between the projected source point $G_{ij}(v_i)$ and the target point v_j.

As the pairwise methods need to train $O(n^2)$ independent models, in this paper we aim to reduce the model complexity to $O(n)$ by learning partially shared projection functions[1]. After the model training process, given a source identity v_i, his/her linked candidates can be selected based on the distances between the projected point $G_{ij}(v_i)$ and the identities v_j in target social network. A smaller distance means the two identities has a larger chance to be the same natural person.

4 METHODOLOGY

In this section, we will present the details of MSUIL, which is a partially shared adversarial learning based model to link identities across multiple social platforms. In each training step of MSUIL, a pair of social platforms P_i and P_j are randomly selected as the input. The social distribution $\mathbb{P}^I$ of platform P_i is viewed as the source distribution and $\mathbb{P}^J$ of platform P_j is viewed as the target distribution. We aim to learn a desirable projection function G_{ij} to minimize the distance between the generated distribution $\tilde{\mathbb{P}}^J = G_{ij}(\mathbb{P}^I)$ and the real target distribution $\mathbb{P}^J$. Previous work [9] introduces the

[1] In this paper we use "projection function" and "generator" interchangeably.

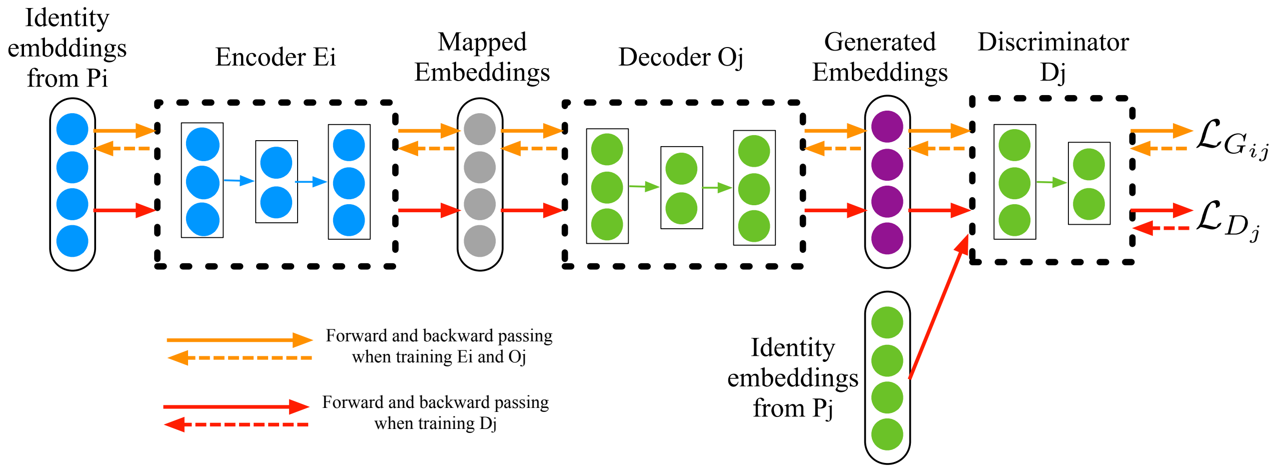

Identity embddings from Pi
Encoder Ei
Mapped Embeddings
Decoder Oj
Generated Embeddings
Discriminator Dj
$\mathcal{L}_{G_{ij}}$
$\mathcal{L}_{D_j}$
Forward and backward passing when training Ei and Oj
Forward and backward passing when training Dj
Identity embeddings from Pj

Figure 2: The details of the MSUIL model. Generator G_{ij} is composed of the encoder E_i and decoder O_j.

Wasserstein distance to measure the closeness between two distributions, which can be minimized under an adversarial learning framework. However, previous work [9] focuses on the pairwise identity linkage and its generator directly projects the source distribution into the target space, which needs to train $O(n^2)$ generators $\{G_{ij} | 1 \leqslant i, j \leqslant n\}$ in the multi-platform scenario. Besides, the training processes of generators are independent from each other, which cannot capture the interdependencies among multiple platforms.

Different from the traditional adversarial learning framework exploited in the previous works, inspired by [2], here we propose the novel partially shared generators for the studied problem. Instead of viewing the generator as a unit, the generator G_{ij} in MSUIL is composed by the encoder E_i of the platform P_i and the decoder O_j of the platform P_j as shown in Figure 2. Encoder E_i maps the source distribution $\mathbb{P}^I$ into a shared latent space as the mapped distribution $\hat{\mathbb{P}}^I$, while decoder O_j projects the mapped distribution into the target space as the generated distribution $\tilde{\mathbb{P}}^J$. The encoders and decoders are shared when handling different pairs of input platforms (e.g., the encoder E_i is also used in generator G_{ik} to link identities across platform P_i and P_k). Thus, we can construct n^2 generators with only n encoders and n decoders, which reduces the model complexity from $O(n^2)$ to $O(n)$.

Next we will introduce the details of the major components, the objective functions and the optimization algorithm of MSUIL model. Several important issues will also be discussed. Encoders and decoders will be presented in subsection 4.1 and 4.2, respectively. The details of discriminators will be demonstrated in the subsection 4.3 along with the objective functions, as the discriminator can be viewed as a result of objective formula derivation.

4.1 Encoder

Encoder maps the inputs from source space into a shared latent space. The latent space can be understood as an intermediate platform, which provides the possibility of reusing the encoders/decoders across different pairs of inputs. Following the previous works [9, 15], encoder is implemented as a linear transformation. We also tried non-linear functions using neural networks with non-linear activations, but they do not work well. This may be because the non-linear functions seriously alters the input distribution and further destroys the isomorphism. Previous works [9, 15] also point out that enforcing the linear transformations to be orthogonal can achieve higher performance, and thus the encoders are expected to perform the orthogonal linear transformation. Given an input identity feature vector $x_i \in \mathbb{R}^{d \times 1}$ from the source space, encoder E_i performs the following step:

$$\hat{x}_i = E_i(x_i) = \mathcal{M}_i \times x_i \tag{1}$$

in which $\hat{x}_i \in \mathbb{R}^{d \times 1}$ is the mapped latent embedding and $\mathcal{M}_i \in \mathbb{R}^{d \times d}$ is the linear transformation matrix of encoder E_i.

4.2 Decoder

Decoder projects the mapped latent embeddings to the target space. The combination of encoder E_i and decoder O_j can be viewed as the generator G_{ij}, which projects the identities from source space to the target space. Similar to the encoder, decoder also performs the linear transformation. When the source platform is same to the target one, decoder is expected to reconstruct the original feature vectors from the latent ones as: $O_j(E_j(x)) = x$. As the transformation matrix $\mathcal{M}_j$ in encoder E_j is assumed to be orthogonal, the transformation matrix of decoder O_j can be the transpose of $\mathcal{M}_j$ as $\mathcal{M}_j^{-1} = \mathcal{M}_j^{\intercal}$. The calculation of decoder O_j can be formulated as the follows:

$$\tilde{x}_j = O_j(\hat{x}_i) = \mathcal{M}_j^{\intercal} \times \hat{x}_i \tag{2}$$

in which $\mathcal{M}_j^{\intercal} \in \mathbb{R}^{d \times d}$ is the transformation matrix in the decoder O_j. $\tilde{x}_j \in \mathbb{R}^{d \times 1}$ is the generated embedding, which can be viewed as the projected point of the source identity in the target space.

4.3 Objective Function

Given a pair of platforms P_i and P_j, the objective of MSUIL can be formally defined as follows:

$$\min_{G_{ij}} \mathrm{WD}(\mathbb{P}^J, \mathbb{P}^{G_{ij}(I)}) = \inf_{\gamma \in \Gamma(\mathbb{P}^J, \mathbb{P}^{G_{ij}(I)})} \mathbb{E}_{(x_j, G_{ij}(x_i)) \sim \gamma}[dis(x_j, G_{ij}(x_i))]$$

The right part is the definition of the Wasserstein distance, which measures the minimum cost of transporting mass in converting a data distribution to another data distribution. $\Gamma(\mathbb{P}^J, \mathbb{P}^{G_{ij}(I)})$ is the joint probability distribution $\gamma(x_j, G_{ij}(x_i))$ with marginals $\mathbb{P}^{G_{ij}(I)}$ and $\mathbb{P}^J$. dis calculates the distance between two points, which is set to the Euclidean distance. Wasserstein distance aims to locate a desirable joint distribution Γ to reach the expectation infimum, but it is difficult to try all the possible joint distributions to compute the $\inf_{\gamma \in \Gamma(\mathbb{P}^J, \mathbb{P}^{G_{ij}(I)})}$ [37]. Fortunately, the Wasserstein distance can be transformed into a much simpler version based on the Kantorovich-Rubinstein duality [23] as follows:

$$\mathrm{WD} = \frac{1}{K} \sup_{\|f_j\|_L \leqslant K} \mathbb{E}_{x_j \sim \mathbb{P}^J} f_j(x_j) - \mathbb{E}_{G_{ij}(x_i) \sim \mathbb{P}^{G_{ij}(I)}} f_j(G_{ij}(x_i)). \tag{3}$$

Function f_j should satisfy the K-Lipschitz continuous restriction, which means $|f_j(x_1) - f_j(x_2)| \leqslant K|x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$ and $K \geqslant 0$ is the Lipschitz constant. Formula (3) aims to learn a desirable K-Lipschitz function to achieve the supremum. Considering the neural networks own powerful approximation abilities, here we select a multi-layer feed forward network to find a desirable function f_j. Thus, Formula (3) can be converted into the following objective function:

$$\max_{\theta_j} \; L_{f_j} = \mathbb{E}_{x_j \sim \mathbb{P}^J} f_j(x_j) - \mathbb{E}_{G_{ij}(x_i) \sim \mathbb{P}^{G_{ij}(I)}} f_j(G_{ij}(x_i)) \tag{4}$$

θ_j is the parameter set of the feed forward network f_j. The loss of f_j aims to maximize the output probabilities of the true samples, and minimize the probabilities of the generated samples, which performs the functions of a discriminator. Thus, the function f_j can be viewed as a discriminator D_j to distinguish the original target samples from the generated ones, and its loss can be formulated as:

$$\max_{\theta_j} \; L_{D_j} = \mathbb{E}_{x_j \sim \mathbb{P}^J} D_j(x_j) - \mathbb{E}_{G_{ij}(x_i) \sim \mathbb{P}^{G_{ij}(I)}} D_j(G_{ij}(x_i)) \tag{5}$$

Intuitively, discriminator D_j identifies how likely a given vector is from the platform P_j, which essentially estimates the approximate Wasserstein distance between distributions $\mathbb{P}^{G_{ij}(I)}$ and $\mathbb{P}^J$. In addition, in order to satisfy the K-Lipschitz restriction, the clipping trick [1] is adopted, which clamps the weights θ_j to a small window $[-c, c]$ after every gradient updating.

Next we will present the objective function of generators. As a partially shared generator, G_{ij} can be formulated as:

$$\tilde{x}_j = G_{ij}(x_i) = O_j(E_i(x_i)) = \mathcal{M}_j^{\mathsf{T}} \times \mathcal{M}_i \times x_i \tag{6}$$

The generator G_{ij} aims to minimize the approximate Wasserstein distance estimated by D_j. From Formula (5), one can see that G_{ij} only exists in the second term at the right of the equals sign, and thus we can maximize the second term to achieve a smaller Wasserstein distance as follows:

$$\max_{G_{ij}} \; L_{G_{ij}} = \mathbb{E}_{G_{ij}(x_i) \sim \mathbb{P}^{G_{ij}(I)}} D_j(G_{ij}(x_i)) \tag{7}$$

Intuitively, the objective of generator G_{ij} aims to confuse D_j, and thus the generators are learned in a way that D_j cannot differentiate the generated vectors from the real vectors in P_j. In essence, this objective minimizes the estimated Wasserstein distance between the original distribution $\mathbb{P}^J$ and the projected distribution $\mathbb{P}^{G_{ij}(I)}$.

Here we also aim to incorporate few annotations to provide extra guidances into the learning of generators. Given an arbitrary pair of platforms, we have a set of source identities and their matched target identities denoted as M_{ij}. For a matched identity pair (x_i, x_j), we aim to minimize the distance between the projected source node $G_{ij}(x_i)$ and the target node x_j:

Algorithm 1 Training algorithm of MSUIL.

```
Require: the input platforms $\mathcal{P}$, the batch size $b$, the number of discriminator training $n_k$ and the annotation guided weight $\lambda_c$.
 1: repeat
 2:     ▷ Training steps of discriminators $D$
 3:     for $u = 0 \rightarrow n_k$ do
 4:         $\mathrm{loss}_d = 0$                                        ▷ Discriminator loss
 5:         for $P_i \in \mathcal{P}$ do
 6:             randomly select a platform $P_j \in \mathcal{P}$
 7:             randomly sample a batch of source identities $x_i \sim P_i$
 8:             randomly sample a batch of target identities $x_j \sim P_j$
 9:             $\hat{x}_i = E_i(x_i) = \mathcal{M}_i \times x_i$                 ▷ Encoder $E_i$
10:             $\tilde{x}_j = O_j(\hat{x}_i) = \mathcal{M}_j^{\mathsf{T}} \times \hat{x}_i$                ▷ Decoder $O_j$
11:             $y_j = D_j(x_j)$                                  ▷ Discriminator $D_j$
12:             $\tilde{y}_j = D_j(\tilde{x}_j)$                                  ▷ Discriminator $D_j$
13:             $\mathrm{loss}_d$ += $(\tilde{y}_j - y_j)$
14:         end for
15:         update parameters in all discriminators $D$ to minimize $\mathrm{loss}_d$.
16:         clip the parameters in all discriminators.
17:     end for
18:     ▷ Training steps of encodes $E$ and decoders $O$
19:     $\mathrm{loss}_g = 0$                                             ▷ Generator loss
20:     $\mathrm{loss}_c = 0$                                       ▷ Annotation guided loss
21:     for $P_i \in \mathcal{P}$ do
22:         randomly select a platform $P_j \in \mathcal{P}$
23:         randomly sample a batch of source identities $x_i \sim P_i$
24:         $\hat{x}_i = E_i(x_i) = \mathcal{M}_i \times x_i$                     ▷ Encoder $E_i$
25:         $\tilde{x}_j = O_j(\hat{x}_i) = \mathcal{M}_j^{\mathsf{T}} \times \hat{x}_i$                    ▷ Decoder $O_j$
26:         $\tilde{y}_j = D_j(\tilde{x}_j)$                                      ▷ Discriminator $D_j$
27:         $\mathrm{loss}_g$ -= $\tilde{y}_j$
28:         $\mathrm{loss}_c$ += $\lambda_c \cdot \sum_{(x_i, x_j) \in M_{ij}} d(G_{ij}(x_i), x_j)$
29:     end for
30:     update all encoders and decoders to minimize $\mathrm{loss}_g$ and $\mathrm{loss}_c$.
31:     orthogonalize transformation matrices in encoders
32: until convergence
```

$$\min_{G_{ij}} \; L_{C_{ij}} = \frac{\lambda_c}{|M_{ij}|} \sum_{(x_i, x_j) \in M_{ij}} d(G_{ij}(x_i), x_j) \tag{8}$$

λ_c is a hyper-parameter to control the weight of loss L_C.

4.4 Orthogonalization

As discussed in subsection 4.1, we expect the transformation matrices $\mathcal{M}$ to be orthogonal. Thus, in each training iteration, the following step is processed to ensure the transformation matrices $\mathcal{M}$ are (approximately) orthogonal [3]:

$$\forall P_i \in \mathcal{P} : \mathcal{M}_i = (1 + \beta)\mathcal{M}_i - \beta \mathcal{M}_i \mathcal{M}_i^{\mathsf{T}} \mathcal{M}_i \tag{9}$$

in which β is the retraction parameter and is fixed to 0.001.

4.5 Optimization Algorithm

In this subsection we briefly introduce the optimization algorithm of the proposed MSUIL model as shown in Algorithm 1. The training steps of discriminators are presented in line 2 to 17, which are also shown as the red lines in Figure 2. In line 6, the target platform P_j is selected by uniform sampling from the set $\mathcal{P}$, which can be same to the source platform P_i. In line 7 and 8, the sampling probability of a social identity is proportional to his/her topology importance (e.g., the degree of social relations). The loss of discriminator in line 13 is same to the Formula 5, which will be updated in each training iteration. In line 16, the parameters of discriminators are clipped to satisfy the K-Lipschitz restriction. Line 18 to 31 present the training steps of the encoders and decoders, which are shown as the orange lines in Figure 2. Loss $loss_g$ is calculated in line 26 following the Formula (7), and the annotation guided loss $loss_c$ is calculated in line 27 according to the Formula (8). In the end of each training iteration, the transformation matrices $\mathcal{M}$ will be updated to orthogonal following Formula (9). In each training iteration, we record the average of generator losses.

4.6 Discussion

In this subsection, we will discuss several important issues.

Flexibility Analysis As mentioned in the introduction section, one challenge of the multi-platform UIL task is the flexibility, which requires the proposed model can be easily extended to suit a variable number of social platforms. As the components (encoder, decoder and discriminator) of each platform are independent in MSUIL, thus our proposal can be easily extended by adding or removing the corresponding components. For example, given a new platform, we only need to add its encoder, decoder and discriminator into the adversarial learning pipeline, and update them according to the Algorithm 1 without significant modifications, which proves the high flexibility of MSUIL.

Training Stability Analysis The adversarial learning framework is also famous for its instable training procedure. Traditional adversarial learning models usually minimize the Jensen-Shannon (JS) distance between the generated distribution and the original distribution. However, the gradient of the JS distance will be zero when the two distributions have no overlap, which leads to the vanishing gradient issue [1]. In this paper, we select the Wasserstein distance as the measurement, which can provide more smoother gradients [1]. Besides, Wasserstein distance can still provide correct gradients when two distributions have overlap area. Hence MSUIL can present more stable training process.

Multi-platform Interdependency Capturing Analysis As the generators are partially shared among different pairs of platforms, MSUIL can capture the interdependencies among multiple platforms. For example, give the pair of P_i and P_j as the input, encoder E_i is trained to preserve the interdependency between these two platforms. After that, given the pair of P_i and P_k as input, the encoder E_i is also trained to capture the correlations between these two platforms. After the model training, encoder E_i can preserve the interdependencies between P_i and all other platforms. Considering the distances between arbitrary pairs of distributions are minimized iteratively, MSUIL can achieve a local optimal of the minimization of the sum of distances between multiple distributions.

5 EXPERIMENT

In this section, first we will introduce the experimental settings, which include the datasets, baseline methods and the parameter settings. Then the experimental results of different methods are reported and analyzed. UIL models are evaluated on two scenarios: pairwise identity linkage and the multi-platform linkage. The sensitivity study is also performed on several core parameters.

5.1 Datasets

In order to thoroughly evaluate the performance of MSUIL, we introduce two data collections, which include three social networks and three academic coauthor networks. The datasets are crawled and formatted by our corporation coauthors.

- **Social Networks**: This data collection includes three popular social platforms: Weibo, Douban and Zhihu. Weibo (Sina Weibo) (*www.weibo.com*) is one of the most influential social platforms in China, which can be viewed as the hybrid of Twitter and Facebook. Zhihu (*www.zhihu.com*) is a popular community question answering platform in China, which is similar to the Quora website. Douban (*www.douban.com*) is also a popular social network in China, where users can publish their comments on the books, movies and musics. Douban users can present the corresponding Weibo and Zhihu accounts in their home pages, and thus we can obtain the linked identities as the ground truth. In addition, following the previous work [36], a set of unlinked identities are randomly selected from each platform to compose the partially aligned dataset. Finally we can obtain 2,536 Weibo identities, 2,246 Zhihu identities and 3,723 Douban identities. The number of matched identities in each pair of platforms is shown in Table 1a. The following relationships between identities along with the attributes of identities are crawled to construct the social feature spaces.
- **Coauthor Networks**: DBLP (*http://dblp.uni-trier.de*) provides open bibliographic information on major computer science journals and proceedings, whose dataset can be publicly available[2]. The authors of published papers in three years (2015, 2016 and 2017) and the corresponding coauthor relationships are collected to form three academic coauthor networks. In each network, Yoshua Bengio is selected as the center node, and then we construct a coauthor subnetwork by locating the coauthors who can be reached within three steps from the center node. We aim to identify whether two nodes in different coauthor networks belong to the same researcher. DBLP assigns each researcher an unique identity, which can be viewed as the ground truth. The statistics of this data collection are shown in Table 1b.

We also present the counts of linked identities across three networks in each data collection in Table 1c. It requires the three identities from different networks belong to a same person, which will be used in the multi-platform linkage task. Next, we will briefly introduce the data preprocessing process. The feature space of a input network is constructed to preserve both the network topology information and the node attributes. For example, for the DBLP networks, we collect the titles and abstracts of the published papers as the attributes of the authors. These text information published by a single user is represented by a $tf\text{-}idf$ feature vector. Then

[2] http://dblp.uni-trier.de/xml/

Table 1: Statistics of the datasets. The numbers in the brackets are the counts of nodes. The numbers in the table are the counts of matched identity pairs across networks.

(a) The count of pairwise linked identities in social network dataset.

Network	Weibo(2,536)	Zhihu(2,246)	Douban(3,723)
Weibo(2,536)	NA	1,556	1,739
Zhihu(2,246)	1,556	NA	1,644
Douban(3,723)	1,739	1,644	NA

(b) The count of pairwise linked identities in coauthor dataset.

Network	DBLP15(3,881)	DBLP16(5,989)	DBLP17(7,073)
DBLP15(3,881)	NA	1,852	1,492
DBLP16(5,989)	1,852	NA	2,570
DBLP17(7,073)	1,492	2,570	NA

(c) The count of linked identities across three networks.

	Social networks	Coauthor networks
Count of linked identites	642	856

TADW [28], a popular node attribute preserving network embedding model is introduced to embed the text information and the coauthor relationships into the distributed embeddings to form the feature spaces of DBLPs, The feature spaces of other datasets are also constructed in the similar manner, but differ in the user attributes (ages, genders and the interest tags for Zhihu users, the personal tags and joined groups for Douban users and interest tags, posted microblogs for Weibo users). We select the degree count of a node as its weight when sampling the training batches as shown in line 7 and 8 of Algorithm 1. The feature spaces of different platform are learned independently, which ensures the generality of the proposed MSUIL model.

5.2 Experimental Setting

Baseline Methods We select the following state-of-the-art baseline methods to make a comparison, which include the supervised and semi-supervised approaches. Besides, the baselines also contain the pairwise UIL models and multi-platform UIL models to extensively evaluate the proposed model.

- **MAH** [22] is a semi-supervised pairwise identity linkage model, which first constructs a hypergraph and then a novel subspace learning algorithm is applied on the hypergraph to match identities.
- **IONE** [11] is an embedding based pairwise UIL model, which solves both the network embedding problem and the user alignment problem simultaneously under a unified optimization framework.
- **CoLink** [40] is also a pairwise UIL model in the weakly-supervised manner, which employs a co-training algorithm and manipulates two independent models, the attribute-based model and the relationship-based model, and makes them reinforce each other iteratively.
- **COSNET** [39] is a supervised multi-platform UIL method, which introduces an energy based model to capture both local and global consistencies among multiple networks.
- **ULink** [15] is a supervised multi-platform model to link identities in a latent user space by minimizing the distances between the linked user identities and maximizing the distances between the identities belonging to different people.

Parameter Setup For the proposed MSUIL model, the dimension of identity feature vectors d is set 128. The transformation matrices $\mathcal{M}$ in the encoders and decoders are randomly initialized as orthogonal matrices. The number of neural cells in the hidden layer of discriminator is set to 256 and the clipping parameter c is set to 0.01. For the parameters in Algorithm 1, the training batch size b is set to 128, the number of discriminator training in each iteration n_k is set to 5 and the annotation guided weight λ_c is set to 0.3. The parameters of baselines are set according to the original papers.

Evaluation Metric Following the previous work [15], we select *Hit-Precision*, a popular evaluation metric is to compare the top-k candidates for the identity linkage:

$$h(x) = \frac{k - (hit(x) - 1)}{k} \tag{10}$$

where $hit(x)$ is the rank position of correctly linked user in the returned list of the top-k candidate target identities. Then the *Hit-Precision* is calculated by the average on the scores of the correctly matched identity pairs: $\frac{\sum_{i=0}^{i=m} h(x_i)}{m}$, in which m is the number of source identities in the matched pairs. For the multi-platform identity linkage task, the average $Hit-precision$ scores will be reported.

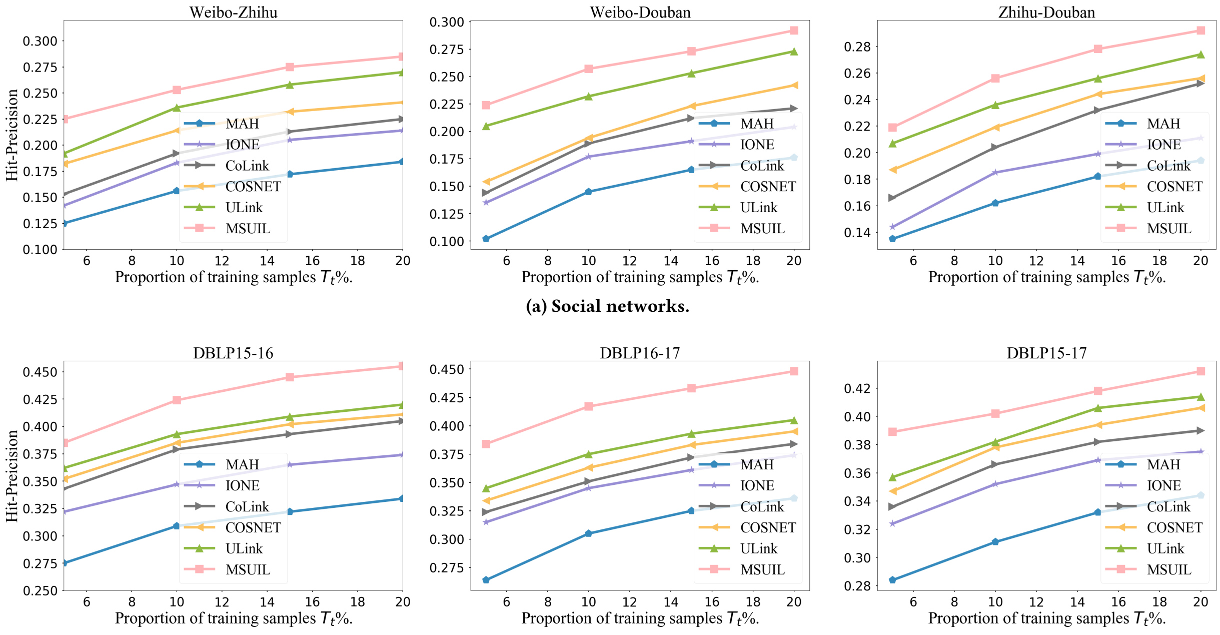
5.3 Evaluation on Pairwise Identity Linkage

In this subsection we will evaluate the performance of different methods on the pairwise identity linkage task. For each pair of networks, we randomly select T_t portion of the linked identities as the available annotations. Here we set the training ratio T_t to 10%. Besides, N_t linked identities are also randomly selected as the test set, and N_t is fixed to 200. The number of returned top candidates k is increased from 3 to 10.

Quantitative Analysis Table 2 shows the results of different approaches on the pairwise linkage task. The pairwise UIL models (MAH, IONE, CoLink) achieve comparatively undesirable performance, as they can only exploit the data from two social networks and ignore the available useful information from the multi-platforms. Compared with the pairwise methods, the multi-platform UIL models (COSNET, ULink) achieve better performance, which demonstrates that the interdependencies across multiple social networks contribute to improving the identity linkage performance. One can clearly see that the proposed MSUIL model achieves the best performance among all the methods. Compared with the best baseline (ULink), MSUIL improves the performance by nearly 2% on the social networks and 2.3% on the coauthor networks. By incorporating the isomorphism across social networks as the complementary, MSUIL can reduce the reliance on the annotations and thus better perform the pairwise identity linkage. In addition, MSUIL consistently outperforms baseline methods with different settings of k, which further proves the effectiveness of our proposal.

Table 2: Comparison with the baseline methods on the pairwise identity linkage task (*Hit-Precision* score).

	Weibo-Zhihu			Weibo-Douban			Zhihu-Douban			DBLP15-16			DBLP16-17			DBLP15-17		
k	k=3	k=5	k=10	k=3	k=5	k=10	k=3	k=5	k=10	k=3	k=5	k=10	k=3	k=5	k=10	k=3	k=5	k=10
MAH	0.124	0.156	0.181	0.119	0.145	0.173	0.134	0.162	0.197	0.277	0.309	0.354	0.275	0.305	0.356	0.267	0.311	0.363
IONE	0.154	0.183	0.211	0.132	0.177	0.204	0.159	0.185	0.221	0.302	0.347	0.397	0.308	0.345	0.396	0.310	0.352	0.377
CoLink	0.162	0.192	0.224	0.142	0.189	0.212	0.167	0.204	0.233	0.322	0.379	0.414	0.310	0.351	0.400	0.317	0.366	0.395
COSNET	0.173	0.214	0.231	0.166	0.194	0.225	0.172	0.219	0.246	0.346	0.385	0.433	0.332	0.363	0.421	0.339	0.378	0.414
ULink	0.195	0.236	0.257	0.193	0.232	0.253	0.197	0.236	0.271	0.358	0.393	0.442	0.346	0.375	0.433	0.347	0.382	0.421
MSUIL	**0.215**	**0.253**	**0.264**	**0.219**	**0.267**	**0.273**	**0.213**	**0.256**	**0.298**	**0.377**	**0.424**	**0.457**	**0.373**	**0.417**	**0.443**	**0.365**	**0.402**	**0.452**



(a) Social networks.

(b) Academic coauthor networks.

Figure 3: Pairwise identity linkage performance on two data collections w.r.t the proportion of training samples T_t.

We also study the influence of the training ratio T_t on the pairwise identity linkage performance. The training ratio T_t is increased from 5% to 15% and the parameter k is fixed to 5. Figure 3 shows the *Hit-Precision* scores of identity linkage models given different training ratios. From the results, one can see that with the increase of training ratio T_t, the performance of all methods present similar uptrend curves. Given 5% more training samples, the scores of MSUIL are further improved by nearly 3% on average. MSUIL beats the best pairwise UIL method CoLink by 7% and the best multi-platform UIL baseline ULink by nearly 3%. Under a unified adversarial learning framework, MSUIL can effectively incorporate both the supervised annotations and the unsupervised distribution similarity information, which ensures its superiority.

Abliation Study Here we perform the abliation study on the MSUIL model to evaluate its performance in handling different numbers of input networks. For each data collection, MSUIL takes all the three networks as the inputs, and we further propose a pairwise model MSUIL$_t$ which only considers two social networks as inputs. MSUIL$_t$ can be easily obtained by removing an adversarial learning pipeline (the corresponding encode, decoder and discriminator) from the model training framework. MSUIL$_t$ still can be trained according to the Algorithm 1 without any modifications. The training ratio T_r is set to 10% and k is fixed to 5. Figure 4 shows the results of MSUIL and MSUIL$_t$ on two data collections. From the Table 2 and Figure 4, one can see that the pairwise model MSUIL$_t$ outperforms the pairwise baselines (MAH, IONE and CoLink). MSUIL$_t$ beats the best pairwise model CoLink by around 2% on average, which demonstrates the distribution closenesses across the social networks are the powerful unsupervised indicators to link identities. Compared with MSUIL, MSUIL$_t$ performs less well as it can only incorporate the data from two social networks to make decisions, while MSUIL can effectively capture the interdependencies among multiple social networks to better link identities. In addition, the abliation study also proves the flexibility of MSUIL model as it can be easily extended to handle a variable number of input networks without significant modifications.

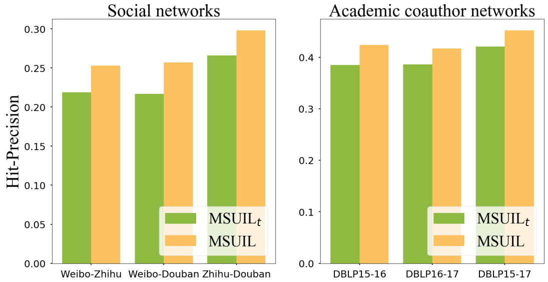


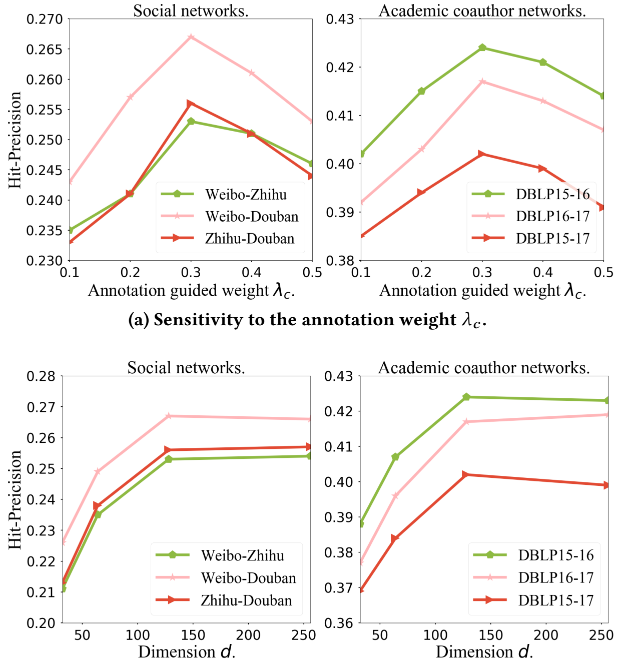
Figure 4: The results of ablation study on the pairwise linkage task. MSUIL$_t$ views only two networks as the inputs, while MSUIL incorporates all the three networks.

Parameter Sensitivity Study Here we study the performance sensitivities of MSUIL on two core hyper-parameters: the annotation guided weight λ_c and the feature dimension d. Weight λ_c is increased from 0.1 to 0.5, and the dimension d varies from 32 to 256. The sensitivity study is performed on both data collections, and the results are shown in Figure 5. From the Figure 5a, one can see that with the increase of the annotation guided weight λ_c, the model performance first increases and then decreases. It shows that incorporating appropriate annotation guided loss can improve the identity linkage performance. However, when λ_c is too large, the annotation guided loss may overwhelm the adversarial loss and further lead to the undesirable performance. From the Figure 5b, One can see that, with the increase of the dimension d, the *Hit-Precision* scores first increase and then keep stable. It demonstrates that a larger embedding dimension can provide stronger representation ability when d is comparatively small, which can generate quality node features to improve the linkage performance. But when d is too large, the model will encounter the performance bottleneck. In addition, a larger d will significantly increase the number of model parameters because the transformation matrices $\mathcal{M}$ to be learned have d^2 trainable parameters, which will slow down the model training speed.

5.4 Multi-Platform Identity Linkage

In this subsection, we evaluate the performance of MSUIL to link identities across three social networks in a data collection. The set of natural people M_m who have identities on all three networks in a data collection is viewed as the ground truth. We assume three identities x_i, x_j and x_k from different platforms belong to a same natural person. Given the identity x_i in the platform P_i, we can achieve two sets of top-k candidates from other two platforms by UIL models. If the returned candidates contain the two target identities (x_j and x_k), the average *Hit-Precision* scores are calculated, otherwise the score is set to 0 as the model fails to link identities across three platforms. The training ratio T_t is set to 10%.

Table 3 shows the experimental results of multi-platform identity linkage. Compared with the pairwise identity linkage results shown in Table 2, one can clearly see the performance of all methods drops significantly, which proves the multi-platform identity task is more challenging. Given three platforms P_i, P_j and P_k, the multi-platform identity task has $|V_i| \times |V_j| \times |V_k|$ candidate combinations, while the pairwise linkage task only has $|V_i| \times |V_j|$ ($|V_j| \times |V_k|$ or $|V_i| \times |V_k|$) candidates, where $|V|$ is the count of identities in the corresponding platform. The search space of multi-platform identity task is much larger than the one of pairwise linkage, which leads to the performance decline. From the results in Table 3, one can also see that the proposed MSUIL model still achieves the best *Hit-Precision* scores despite the performance drop. MSUIL beats the best baseline ULink by 2.0% on the social networks and 4.2% on the academic coauthor networks, which proves MSUIL owns the comparatively powerful identity linkage capacity in the multi-platform scenario.



(a) Sensitivity to the annotation weight λ_c.

(b) Sensitivity to the dimension d.

Figure 5: The parameter sensitivity of MSUIL on the pairwise linkage task.

Table 3: Comparison with the baseline methods on the multi-platform identity linkage task (*Hit-Precision* score).

	Weibo-Zhihu-Douban			DBLP15-16-17		
k	k=3	k=5	k=10	k=3	k=5	k=10
MAH	0.073	0.084	0.126	0.127	0.147	0.175
IONE	0.084	0.097	0.143	0.142	0.168	0.193
CoLink	0.092	0.106	0.157	0.153	0.192	0.216
COSNET	0.104	0.117	0.163	0.164	0.205	0.235
ULink	0.116	0.118	0.171	0.169	0.223	0.247
MSUIL	**0.132**	**0.148**	**0.186**	**0.214**	**0.257**	**0.276**

6 CONCLUSION

In this paper, we study the novel problem of multi-platform social identity linkage in the semi-supervised manner. The insight is that we capture the isomorphism across multiple platforms as the complementary to link identities in the distribution level, which reduces the reliance on the annotations. The studied problem is transformed to the learning of a set of projection functions to minimize the Wasserstein distance between the distributions of user identities in arbitrary pairs of platforms. Few annotations also can be incorporated to provide extra guidances to the learning of projection functions. We propose a novel adversarial learning based model MSUIL with partially shared projection functions to capture the interdependencies across multiple platforms. With the shared generators, the model complexity is reduced from $O(n^2)$ to $O(n)$. Under an unified adversarial learning framework, MSUIL can effectively incorporate both the supervised annotations and the unsupervised distribution similarity information. We evaluate the proposed model on two data collections include three social networks and three academic coauthor networks. The experimental results demonstrate MSUIL outperforms the state-of-the-art baselines on both pairwise identity linkage task and multi-platform identity linkage task.

ACKNOWLEDGMENTS

This work was supported the Natural Science Foundation of China (Nos. U1636211,61672081,61370126,61602237), Beijing Advanced Innovation Center for Imaging Technology (No.BAICIT-2016001), National Key R&D Program of China (No.2016QY04W0802 and 2017YFB0802203), Natural Science Foundation of Jiangsu Province (No. BK20171420) and Natural Science Foundation of Guangdong Province (No. 2017A030313334). This work was also supported in part by NSF under grants III-1526499, III-1763325, III-1909323, SaTC-1930941, and CNS-1626432.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv* (2017), 1–32.

[2] Xilun Chen and Claire Cardie. 2018. Unsupervised Multilingual Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 261–270.

[3] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 854–863.

[4] Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. 2011. Identifying Users Across Social Tagging Systems. In *ICWSM*.

[5] Nitish Korula and Silvio Lattanzi. 2014. An efficient reconciliation algorithm for social networks. *VLDB* (2014), 377–388.

[6] Sebastian Labitzke, Irina Taranu, and Hannes Hartenstein. 2011. What your friends tell others about you: Low cost linkability of social network profiles. In *Social Network Mining and Analysis*. 1065–1070.

[7] Simon Lacoste-Julien, Konstantina Palla, Alex Davies, Gjergji Kasneci, Thore Graepel, and Zoubin Ghahramani. 2013. Sigma: Simple greedy matching for aligning large knowledge bases. In *KDD*. ACM, 572–580.

[8] Chaozhuo Li, Senzhang Wang, Yukun Wang, Philip Yu, Yanbo Liang, Yun Liu, and Zhoujun Li. 2019. Adversarial Learning for Weakly-Supervised Social Network Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 996–1003.

[9] Chaozhuo Li, Senzhang Wang, Philip S Yu, Lei Zheng, Xiaoming Zhang, Zhoujun Li, and Yanbo Liang. 2018. Distribution Distance Minimization for Unsupervised User Identity Linkage. In *CIKM*. ACM, 447–456.

[10] Jing Liu, Fan Zhang, Xinying Song, Young-In Song, Chin-Yew Lin, and Hsiao-Wuen Hon. 2013. What's in a name?: an unsupervised approach to link users across communities. In *WSDM*. ACM, 495–504.

[11] Li Liu, William K Cheung, Xin Li, and Lejian Liao. 2016. Aligning Users across Social Networks Using Network Embedding.. In *IJCAI*. 1774–1780.

[12] Tong Man, Huawei Shen, Shenghua Liu, Xiaolong Jin, and Xueqi Cheng. 2016. Predict Anchor Links across Social Networks via an Embedding Approach. In *IJCAI*. 1823–1829.

[13] Brendan D McKay et al. 1981. Practical graph isomorphism. (1981).

[14] Marti Motoyama and George Varghese. 2009. I seek you: searching and matching individuals in social networks. In *WSDM*. ACM, 67–75.

[15] Xin Mu, Feida Zhu, Ee-Peng Lim, Jing Xiao, Jianzong Wang, and Zhi-Hua Zhou. 2016. User identity linkage by latent user space modelling. In *KDD*. ACM, 1775–1784.

[16] Yuanping Nie, Yan Jia, Shudong Li, Xiang Zhu, Aiping Li, and Bin Zhou. 2016. Identifying users across social networks based on dynamic core interests. *Neurocomputing* (2016), 107–115.

[17] Olga Peled, Michael Fire, Lior Rokach, and Yuval Elovici. 2013. Entity matching in online social networks. In *SocialCom*. IEEE, 339–344.

[18] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. 2011. How unique and traceable are usernames?. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 1–17.

[19] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. 2016. Linking users across domains with location data: Theory and validation. In *WWW*. 707–719.

[20] Vishal Sharma and Curtis Dyreson. 2018. LINKSOCIAL: Linking User Profiles Across Multiple Social Media Platforms. In *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 260–267.

[21] Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. User identity linkage across online social networks: A review. *ACM SIGKDD Explorations Newsletter* (2017), 5–17.

[22] Shulong Tan, Ziyu Guan, Deng Cai, Xuzhen Qin, Jiajun Bu, and Chun Chen. 2014. Mapping Users across Networks by Manifold Alignment on Hypergraph. In *AAAI*. 159–165.

[23] Cédric Villani. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

[24] Jan Vosecky, Dan Hong, and Vincent Y Shen. 2009. User identification across multiple social networks. In *Networked Digital Technologies*. IEEE, 360–365.

[25] Senzhang Wang, Xia Hu, Philip S Yu, and Zhoujun Li. 2014. MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades. In *KDD*. ACM, 1246–1255.

[26] Senzhang Wang, Zhao Yan, Xia Hu, S Yu Philip, and Zhoujun Li. 2015. Burst Time Prediction in Cascades.. In *AAAI*. 325–331.

[27] Wei Xie, Xin Mu, Roy Ka-Wei Lee, Feida Zhu, and Ee-Peng Lim. 2018. Unsupervised User Identity Linkage via Factoid Embedding. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1338–1343.

[28] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. 2015. Network Representation Learning with Rich Text Information. In *IJCAI*. 2111–2117.

[29] Yang Yang, De-Chuan Zhan, Yi-Feng Wu, and Yuan Jiang. 2018. Multi-network User Identification via Graph-Aware Embedding. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 209–221.

[30] Reza Zafarani and Huan Liu. 2014. Users joining multiple sites: distributions and patterns. In *ICWSM*.

[31] Reza Zafarani and Huan Liu. 2016. Users joining multiple sites: Friendship and popularity variations across sites. *Information Fusion* (2016), 83–89.

[32] Reza Zafarani, Lei Tang, and Huan Liu. 2015. User identification across social media. *TKDD* (2015), 16.

[33] Qianyi Zhan, Jiawei Zhang, Senzhang Wang, S Yu Philip, and Junyuan Xie. 2015. Influence maximization across partially aligned heterogenous social networks. In *PAKDD*. Springer, 58–69.

[34] Baichuan Zhang, Tanay Kumar Saha, and Mohammad Al Hasan. [n. d.]. Name disambiguation from link data in a collaboration graph. In *ASONAM*. 81–84.

[35] Haochen Zhang, Min-Yen Kan, Yiqun Liu, and Shaoping Ma. 2014. Online social network profile linkage. In *Asia Information Retrieval Symposium*. Springer, 197–208.

[36] Jiawei Zhang, Philip S Yu, and Zhi-Hua Zhou. 2014. Meta-path based multi-network collective link prediction. In *KDD*. ACM, 1286–1295.

[37] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *EMNLP*. 1934–1945.

[38] Wen Zhang, Kai Shu, Huan Liu, and Yalin Wang. 2019. Graph Neural Networks for User Identity Linkage. *arXiv preprint arXiv:1903.02174* (2019).

[39] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S Yu. 2015. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *KDD*. ACM, 1485–1494.

[40] Zexuan Zhong, Yong Cao, Mu Guo, and Zaiqing Nie. 2018. CoLink: An Unsupervised Framework for User Identity Linkage. In *AAAI*. 3379–3385.