Unsupervised User Identity Linkage via Graph Neural Networks

Fan Zhou*, Zijing Wen*, Ting Zhong*‡, Goce Trajcevski†, Xovee Xu* and Leyuan Liu*
*University of Electronic Science and Technology of China, China
†Department of Electrical and Computer Engineering, Iowa State University, USA
‡ Ting Zhong is the corresponding author: zhongting@uestc.edu.cn

***Abstract*—User identity linkage (UIL) aims to link identical users engaging in multiple social networks. It has received considerable attention in both academia and industry due to its profound implications for multiple applications. Although existing approaches have achieved promising progress in UIL using various graph learning methods, they usually require a large number of labeled anchor nodes which, however, are difficult to obtain in real-world social platforms due to privacy issues. We introduce a novel UIL model NWUIL (Network Wasserstein learning for UIL) to identify anchor users across social networks in a fully unsupervised manner. Instead of point vector embedding of nodes as in previous methods, NWUIL captures node distribution in Wasserstein space with graph neural networks. We also propose to reformulate the UIL task as an optimal network transport problem, and then introduce an unsupervised mapping process based on the network Wasserstein distance for UIL. In this way, our method not only improves the anchor node aligning accuracy but also alleviates the issues caused by insufficient labeled anchor nodes. We conduct extensive experiments using real-world datasets, and demonstrate that NWUIL significantly outperforms existing unsupervised baselines while showing competitive performance as some state-of-the-art supervised approaches.**

***Index Terms*—user identity linkage, unsupervised learning, social networks, optimal transport, graph neural networks**

I. Introduction

Nowadays, many people are involved in various online social networks to enjoy different services – chatting with friends using WhatsApp/Skype/WeChat; sharing daily activities and photos on Facebook/Instagram, and publishing or forwarding commentaries in Twitter/Weibo; searching scientific papers in arXiv, Google Scholar, and DBLP; etc. Discovering the account belonging to the same identity across different social networks, also known as user identity linkage (UIL) or network alignment in the literature, is an important task for many academic and commercial applications such as cross-site recommendation and advertising, authorship analysis, user profiling (to name but a few), and has received significant research interest in recent years [1]–[6].

Existing Approaches. Existing UIL methods can be generally grouped into few broad categories. Many of the current approaches [3]–[7] link cross-domain users using a number of labeled anchor nodes and their attributes (e.g., username, profiles, writing style, etc.). For example, DeepLink [4] is a graph representation learning based method which embeds the heterogeneous networks using graph embedding techniques and then learns a cross-domain mapping function with reinforcement learning for UIL. TransLink [5] jointly embeds users and their behaviors in various networks into a unified low-dimensional space with a set of known anchor links, combined with the extracted interaction metapaths of each network, to link users in a supervised manner. Recently, *d*NAME [6] tries to interpret the UIL results using robust statistics through investigating the importance of each labeled training anchor nodes on the linking performance. These approaches have achieved significant progress on UIL performance and explore network topology and user profile information for UIL, where graph representation methods such as node embedding methods (e.g., DeepWalk [8]) and graph neural networks (e.g., GCN [9]) are usually leveraged for user representation learning, and utilize the labeled anchor nodes to train a matching model for linking unknown anchor nodes. However, it is prohibitively expensive and even impossible to have enough labeled anchor nodes for training these models, mainly due to the privacy concerns and reluctance of social network platforms to share their account information.

Another line of works [10]–[13] attempt to tackle the insufficient labeled data issue and solve the UIL problem in a fully unsupervised manner. For example, SiGMa [10] jointly learns the structural information and attributes of users to measure the similarities of social identities, and greedily matches users. The unsupervised method UUIL [11] minimizes the distance distribution between user identities in different social networks. Factoid embedding [12] distinguishes each user from others by embedding the profile attributes, content types and social relations associated with users into a common embedding space. REGAL [14] is an unsupervised network alignment framework based on low-rank matrix approximation for speeding up calculation, which simultaneously embeds multiple networks into a latent space and leverages a nearest neighbor search method for node alignments rather than performing all pairwise comparisons. Though these methods are anchor-nodes free, they do not perform well compared to supervised methods and usually require additional user attribute information. Such information, however, is not always available - e.g., due to missing attributes, arbitrary or masked personal information, and casual writing styles of online users. Therefore, existing unsupervised methods are sensitive to

This is the author version of the paper presented at IEEE GLOBECOM '20
doi: 10.1109/GLOBECOM42002.2020.9322311

attribute noise that is prevalent in real-world social network platforms.

Present work. In this paper, we propose a novel UIL method based on graph neural networks (GNN) [15] and optimal transportation theory [16] for addressing the social user identity linkage problem. Our solution NWUIL (Network Wasserstein learning for UIL) leverages GNNs to learn the topological structures of individual social networks and embeds the users into a low-dimensional space preserving both node positions and representation variance. Unlike previous approaches that rely on deterministic node representation for UIL, our method is capable of examining the uncertainty of node embeddings while alleviating the confounding matching issue caused by deterministic embedding and linking. Subsequently, each network is represented as a mixture of Gaussian and the UIL task is accordingly reformulated as an optimal transportation problem by minimizing the cost of moving the mass in the source network into the target one. This enables NWUIL to learn the matching of anchor node pairs in an unsupervised manner. The main contributions of this work can be summarized as follows:

- We present a method to embed each node in network as a Gaussian distribution, which not only preserves the topology information w.r.t. individual social networks, but also captures the uncertainties of node representation. This enables our model to solve the confusing matching problem by exploring all possible anchor nodes while accounting for the uncertainty of matching results.
- We propose a novel unsupervised network matching model for addressing UIL problem. To our knowledge, NWUIL is among the first attempts to solve UIL with optimal transportation, which enables it to alleviate the issues caused by insufficient training anchor nodes. By minimizing the Wasserstein distance between two networks, our method is capable of optimizing the node distribution embedding to significantly improve the alignment accuracy.
- To evaluate the effectiveness of the proposed NWUIL method, we conducted extensive experiments on real-world datasets. The results show that our approach achieves state-of-the-art performance compared to unsupervised UIL method while showing superiority over weakly-supervised methods.

II. Preliminaries

Let $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ denote an unweighted and undirected graph representing a social network, where $\mathcal{U}$ is the set of nodes, and $\mathcal{E}$ is the set of edges connecting the nodes. $e_{i,j} \in \mathcal{E}$ indicates the existence of a relationship between users/nodes u_i and u_j. Given two networks $\mathcal{G}^s = (\mathcal{U}, \mathcal{E}^s)$ and $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}^t)$, we call a pair (u, v) ($u \in \mathcal{U}$ and $v \in \mathcal{V}$) *anchor* nodes, if they belong to the same identity.

Definition 1. ***(User Identity Linkage (UIL))*** *UIL problem is defined as identifying all the anchor node pairs among two social networks. In general, the objective of network alignment methods is to learn a mapping Φ such that two $\mathcal{G}^s$ and $\mathcal{G}^t$ are aligned by maximizing the similarity of all anchor user pairs.*

Let $(\mathbf{A}, \mathbf{X})$ be the matrices representing a network $\mathcal{G}$, where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is a symmetric adjacency matrix ($A_{ij} = 0$ if $(i, j) \notin \mathcal{E}$ and $A_{ij} = 1$ otherwise) and $\mathbf{X} \in \mathbb{R}^{N \times D}$ is a feature matrix assembling attribute information of each node, whose i^{th} row $\mathbf{x}_i$ denotes the feature vector of u_i. We embed each network into a Wasserstein space using the Gaussian-based network embedding method [17], [18], which learns the distributions of nodes while preserving network structure and uncertainty properties of node representation.

Definition 2. ***(Node Distribution Embedding)*** *Each node u_i is embedded as a Gaussian distribution $\mathbb{H}_i = \mathcal{N}(\mu_i, \Sigma_i)$ in the Wasserstein space, where mean μ_i preserves the position information of the nodes and variance Σ_i captures the uncertainty of the node representation.*

A function Φ projects the node vectors/distributions from $\mathcal{G}^s$ to $\mathcal{G}^t$, if for each $u_i \in \mathcal{G}^s$, the mapping function can be defined as $\Phi(u_i) = \Phi(\mathbb{H}_i) = \mathcal{N}(\Phi(\mu_i), \Phi(\Sigma_i)))$. For convenience, $\Phi(u_i)$ and $\Phi(\mathbb{H}_i)$ are used interchangeably in this paper.

Definition 3. ***(Wasserstein distance)*** *The p^{th} Wasserstein distance [19] between two probability distributions can be formally defined as follows:*

$$W_p(\mathbb{H}_i, \mathbb{H}_j)^p = \inf_{\gamma \in \Pi(\mathbb{H}_i, \mathbb{H}_j)} \mathbb{E}_{(u_i, u_j) \sim \gamma}[d(u_i, u_j)^p], \qquad (1)$$

where $\mathbb{E}[\cdot]$ is the expected value of $d(u_i, u_j)^p$, d is a distance function, and $\Pi(\mathbb{H}_i, \mathbb{H}_j)$ is the set of all joint distributions of the random variables with marginals $\mathbb{H}_i$ and $\mathbb{H}_j$.

Wasserstein distance, also called Earth Mover's Distance (EMD) [19], is appealing for measuring the similarity between the distributions of nodes, as it satisfies both the symmetry and the triangle inequality properties [18].

III. Methodology: NWUIL

We now present the details of NWUIL model which consists of two main components, i.e., network representation learning with graph neural networks and unsupervised networks matching with network Wasserstein distance.

A. GNN for Graph Representation Learning

Recently, graph neural networks such as GCN [15], GAT [20] and GIN [21], have enabled a generalization of deep learning techniques to graph structured data. This spurs significant progress on various graph-based learning tasks – e.g., link prediction, clustering, node classification, etc. In NWUIL, we employ an L-layer GCN [15] to exploit the topology structures of the networks using following layer-wise aggregation rule:

$$\mathbf{h}_i^{(l)} = \sigma \left\{ \text{MEAN} \left(\mathbf{h}_i^{(l-1)} + \sum_{u_j \in \text{N}_i} \mathbf{h}_j^{(l-1)} \right) \cdot \mathbf{W}_h \right\}, \qquad (2)$$

where $\mathbf{h}_i^{(l)}$ is the feature vector of a node u_i at the l-th layer, which is initialized as $\mathbf{h}_i^0 = \mathbf{x_i}$; σ denotes the non-linear activation (e.g., ReLU here); $\mathbf{W}_h$ is a learnable weight matrix; and MEAN is the element-wise mean pooling.

Instead of representing nodes as deterministic points in the latent space as in previous works [4]–[6], we learn a Gaussian distribution for each node:

$$\widehat{\mathbf{h}}_i = \mathrm{ReLU}\left(\mathbf{h}_i^{(L)}\mathbf{W}_1 + \mathbf{b}_1\right), \tag{3}$$

$$\mu_i = \widehat{\mathbf{h}}_i\mathbf{W}_2 + \mathbf{b}_2, \tag{4}$$

$$\Sigma_i = \mathrm{ReLU}\left(\widehat{\mathbf{h}}_i\mathbf{W}_3 + \mathbf{b}_3\right), \tag{5}$$

where $\mathbf{h}_i^{(L)}$ is the feature vector obtained at the last (L-th) layer; $\mathbf{W}_1 \in \mathbb{R}^{d\times S}$, $\mathbf{W}_2 \in \mathbb{R}^{S\times d}$, $\mathbf{W}_3 \in \mathbb{R}^{S\times d}$, $\mathbf{b}_1 \in \mathbb{R}^S$, $\mathbf{b}_2 \in \mathbb{R}^d$, and $\mathbf{b}_3 \in \mathbb{R}^d$ are learnable parameters, while mean and variance capture the positional information and the uncertainty of each node, respectively. That is, we use the Gaussian distribution $\mathcal{N}(\mu_i, \Sigma_i)$ as the representation of nodes, which is also referred to $\mathbb{H}_i$ for simplicity.

However, the general-formed Wasserstein distance (cf. Eq. (1)) is computationally expensive. Fortunately, the calculation process can be accelerated by 2[nd] Wasserstein distance (W_2), which can be computed as:

$$W_2(\mathbb{H}_i, \mathbb{H}_j)^2 = \|\mu_i - \mu_j\|_2^2 + \left\|\Sigma_i^{1/2} - \Sigma_j^{1/2}\right\|_F^2, \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

We can now leverage the learned node representation to learn a mapping function using the annotated anchor nodes, which is the mainstream in existing supervised UIL approaches [3]–[7]. However, it requires sufficient anchor nodes to train such a mapping function so as to distinguish the anchor nodes from the confounding factors, e.g., the neighbors and similar structural nodes, which is also known as matching confusions [6].

B. Unsupervised Matching with Optimal Transport

Network representation as a GMM: Given two partially aligned networks $\mathcal{G}^s = (\mathcal{V}^s, \mathcal{E}^s, \mathbb{P}^s)$ and $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t, \mathbb{P}^t)$, where $\mathbb{P}^s$ and $\mathbb{P}^t$ are the distributions of two networks in Wasserstein space reflecting the probability of each node appearing in observed interactions. In above network embedding, we have encoded each node in a network as a Gaussian distribution. Therefore, the whole network can be represented by a Gaussian Mixture Model (GMM):

$$\mathbb{P}^s(u) = \sum_{i=1}^{N} \pi_i^s \mathcal{N}(\mu_i^s, \Sigma_i^s), \qquad \text{s.t.} \sum_{i=1}^{N} \pi_i^s = 1 \tag{7}$$

$$\mathbb{P}^t(v) = \sum_{i=1}^{M} \pi_i^t \mathcal{N}(\mu_i^t, \Sigma_i^t), \qquad \text{s.t.} \sum_{i=1}^{M} \pi_i^t = 1 \tag{8}$$

where π_i^s and π_i^t indicate the component weights of each node distribution $\mathcal{N}(\mu_i^s, \Sigma_i^s)$ and $\mathcal{N}(\mu_i^t, \Sigma_i^t)$, respectively. For simplicity, we assume *a-priori* uniform distribution over component weights, i.e., $\pi_i^s = 1/N$ and $\pi_i^t = 1/M$.

UIL as an OT Problem: We now introduce an unsupervised model based on *optimal transport* (OT) [19]. In particular, we consider all the nodes in a network as a whole and align them from the distribution perspective. Specifically, UIL seeks to match $\mathbb{P}^t(v)$ with $\mathbb{P}^s(u)$ while minimizing a specific cost function $c(u, v)$.

UIL as an OT Plan Optimization: UIL can be alternatively formulated as optimizing over OT plans γ through minimizing the OT distance:

$$d_{OT}(\mathbb{P}^s(u)||\mathbb{P}^t(v)) := \inf_{\gamma} \int_{\mathcal{U}\times\mathcal{V}} c(u, v)d\gamma(u, v), \tag{9}$$

where marginals $\int_v \gamma(u, v)dv = \mathbb{P}^s(u)$ and $\int_u \gamma(u, v)du = \mathbb{P}^t(v)$. The optimal solution γ_* of the problem above is the optimal network alignment plan. We have:

Proposition 1. *The minimization problem (cf. Eq.* (9)*) can be solved via:*

$$d_{OT}(\mathbb{P}^s||\mathbb{P}^t) := \sup_{\phi,\psi\in\mathcal{I}} \left[\int_{\mathcal{U}} \phi(u)\mathbb{P}^s(u)du + \int_{\mathcal{V}} \psi(v)\mathbb{P}^t(v)dv\right]$$
$$\textit{s.t. } \mathcal{I} = \{\phi, \psi : \mathbb{R} \to \mathbb{R} | \phi(u) + \psi(v) \le c(u, v)\} \tag{10}$$

where $\mathcal{I}$ is the set of functions with respect to the cost function $c(u, v)$; $\phi(u)$ and $\psi(v)$ is optimal Kantorovich potential pair.

Our goal now becomes to learn the mapping function $\phi(u)$ and $\psi(v)$ on $\mathcal{G}^s$ and $\mathcal{G}^t$ for linking the nodes in two networks. The proof of Eq. (10) can be found in Appendix.

Because the real datasets in UIL task are discrete structured, we define a coupling matrix $\mathbf{\Gamma} \in \mathbb{R}_+^{N\times M}$ between the source space and target space, whose marginals recover $\mathbb{P}^s(u)$ and $\mathbb{P}^t(v)$, and $\mathbf{\Gamma}_{i,j}$ describes the amount of mass flowing from the mass found at u_i toward v_j in the formalism of discrete measures. Formally, Kantorovich's formulation seeks $\mathbf{\Gamma}$ in the transportation polytope:

$$\Pi(\pi^s, \pi^t) := (\mathbf{\Gamma} \in \mathbb{R}_+^{N\times M} | \sum_j \mathbf{\Gamma}_{ij} = \pi^s, \sum_i \mathbf{\Gamma}_{ij} = \pi^t). \tag{11}$$

Therefore, the discrete d_{OT} can be defined as follows where we aim to find an appropriate transportation plan $\mathbf{\Gamma}$ for minimizing the d_{OT}:

$$d_{OT}(\mathbb{P}^s||\mathbb{P}^t) := \min_{\mathbf{\Gamma}\in\Pi(\pi^s,\pi^t)} \langle \mathbf{C}, \mathbf{\Gamma}\rangle := \sum_{i,j} \mathbf{C}_{ij}\mathbf{\Gamma}_{ij}, \tag{12}$$

where $\mathbf{C} \in \mathbb{R}^{N\times M}$ is an element-wise cost matrix. Due to the representations of nodes obtained in our method are Gaussian distribution embeddings, the cost matrix can be defined as $\mathbf{C}_{ij} = \|\mu_i^s - \mu_i^t\|_2^2 + \left\|{\Sigma_i^s}^{1/2} - {\Sigma_i^t}^{1/2}\right\|_F^2$, which indicates the distance between a node pair (u_i, v_j) with its Gaussian distribution embeddings $\mathcal{N}(\mu_i^s, \Sigma_i^s)$ and $\mathcal{N}(\mu_j^t, \Sigma_j^t)$.

Solving Eq. (12) is a linear optimization problem with the complexity of $O(N^3 \log N)$. We add an entropy penalization $H(\mathbf{\Gamma})$ [22] to make the optimization more efficient, which can also lead to a better empirical results:

$$d_{OT}(\mathbb{P}^s||\mathbb{P}^t) := \min_{\mathbf{\Gamma}\in\Pi(\pi^s,\pi^t)} \langle \mathbf{C}, \mathbf{\Gamma}\rangle - \lambda H(\mathbf{\Gamma}), \tag{13}$$

TABLE I
STATISTICS OF DATASETS.

Dataset	# nodes	# edges	# anchor nodes
Foursquare	5,120	76,972	3,148
Twitter	5,313	164,920	
Last.fm	2,138	4,259	1,561
MySpace	2,117	3,798	

which is a strictly convex optimization problem and can be solved efficiently by Sinkhorn-Knopp algorithm [23]. The solution can be represented as the form of $\mathbf{\Gamma}^* = \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$, with the Gibbs kernel $\mathbf{K} = e^{-\frac{\mathbf{C}}{\lambda}} \in \mathbb{R}_+^{N\times M}$ associated to $\mathbf{C}$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}_+^N \times \mathbb{R}_+^M$ can be computed as [23]:

$$\mathbf{u} = \pi^s \oslash \mathbf{K}\mathbf{v}, \quad \mathbf{v} = \pi^t \oslash \mathbf{K}^\top \mathbf{u}, \tag{14}$$

where $\oslash$ denotes component-wise division. Note that in the implementation of NWUIL, we use Wasserstein distance W_2 (cf. Eq. (6)) to compute the optimal transportation distance d_{OT} between $\mathbb{P}^s$ and $\mathbb{P}^t$.

IV. EXPERIMENTS

A. Datasets

We use two benchmark UIL datasets for evaluating all the methods in our experiments:

• **Foursquare-Twitter** (F-T) [24]: this dataset consists of 3,148 anchor nodes across Foursquare and Twitter who have been identified as the same identity.

• **Last.fm-MySpace** (L-M) (cf. http://aminer.org/cosnet): this dataset consists of 1,561 anchor nodes who have registered accounts in both Last.fm and MySpace.

The details of the two datasets are shown in Table I.

B. Baselines & Metrics

Comparison approaches. We compare our NWUIL model with the following state-of-the-art unsupervised and supervised UIL baselines. For DeepLink and dNAME, we use 10% of anchor nodes for training the models.

• *Factoid Embedding (FE)* [12]: generates user attributes and embeds different attributes into respective embedding spaces.

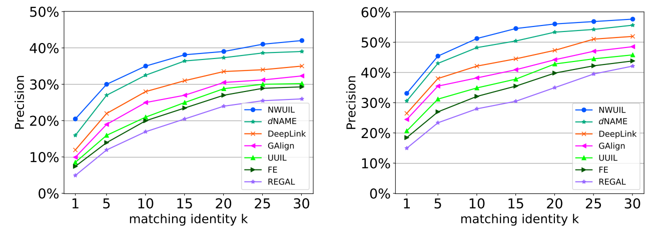
• *REGAL* [14]: uses matrix factorization, structural similarity, and attributes agreement between nodes in disjoint graphs.

• *UUIL* [11]: takes users as a whole and conduct UIL from the user space distribution level. It introduces a generative adversarial network to match the anchor users.

• *GAlign* [25]: exploits the multi-order nature of GCN for linking anchor nodes. It incorporates a data perturbation to make matching adaptive to noises and consistency violations.

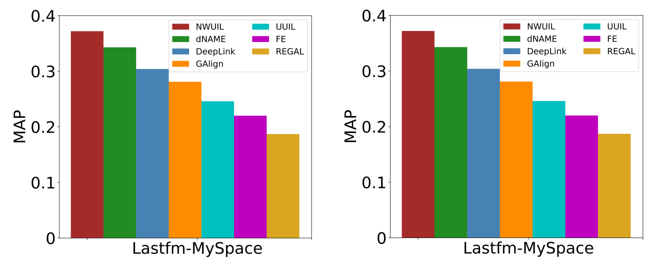
• *DeepLink* [4]: is an end-to-end approach in a supervised manner, which embeds network nodes as vector representation to capture local and global structural information of networks and links anchor nodes in a dual-learning way.

• *dNAME* [6]: is based on an adversarial matching technique, which embeds nodes in a disentangled and faithful manner.



(a) Precision vs. k (F-T). (b) Precision vs. k (L-M).

Fig. 1. Accuracy comparison among methods.



(a) Foursquare-Twitter. (b) Last.fm-Myspace.

Fig. 2. MAP results on two datasets.

Evaluation protocols. We use two commonly used metric – *Precision@k* and Mean Average Precision (MAP) for evaluating the model performance. While *Precision@k* indicates whether the positive matching identities exist in the predicted top-k ($k \le n$) anchor node list, MAP pays more attention to the ranking performance of the lining results.

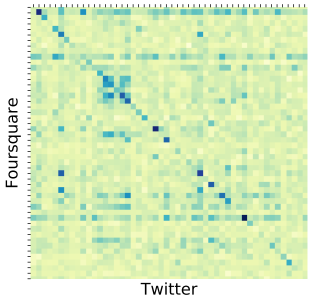
C. Experimental Results

UIL performance. We first systematically evaluate various methods on the user identity linking accuracy. The UIL precision of all methods in two datasets are illustrated in Fig. 1, from which we have following three observations:

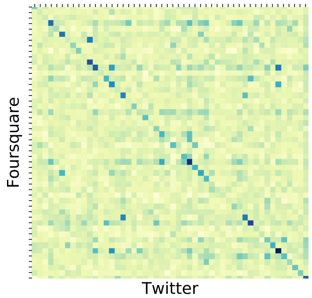
(1) NWUIL consistently outperforms the baselines. Compared to GAlign – often the best approach among *unsupervised* baselines, NWUIL achieves 6% and 11% higher accuracy on the Foursquare-Twitter dataset and the Last.fm-Myspace dataset, respectively. This result indicates that our Wasserstein distance based approach is more effective than previous unsupervised models for UIL, due to the incorporation of node representation uncertainty rather than deterministic point vector embedding in previous works.

(2) Previous supervised UIL methods (e.g., DeepLink and dNAME) relying on graph/node embedding techniques may not perform well without sufficient annotated anchor nodes. In another word, they require enough anchor nodes to boost their performance on UIL – otherwise, their UIL results are very skewed due to the underfitting issue inherent in training the mapping functions. In contrast, our method overcomes this issue by treating the network embedding as a whole and leverages the Wasserstein distance to compare the node similarity without explicitly training the mapping functions.

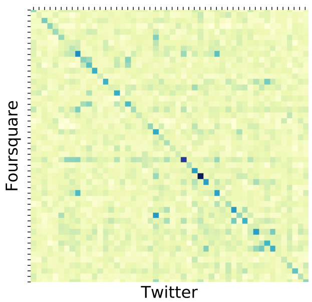
(3) NWUIL is more sensitive to the ranking performance and therefore achieves higher MAP results compared to other



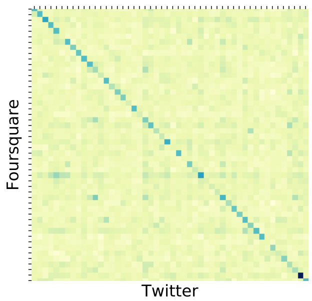
(a) GAlign Testing.



(b) DeepLink Testing.

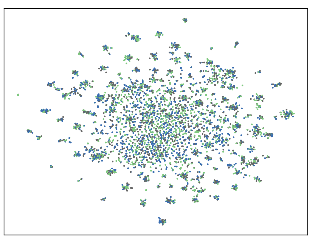


(c) *d*NAME Testing.

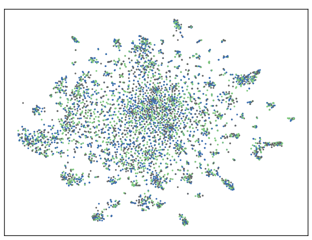


(d) NWUIL Testing.

Fig. 3. Visualization of the UIL performance (Foursquare-Twitter).



(a) Foursquare.



(b) Twitter.

Fig. 4. Latent space visualization on different datasets.

models, as illustrated in Fig. 2. This result indicates that NWUIL outperforms other baselines in disentangling the latent embedding of nodes, especially for those neighbored ones. Because the neighbor nodes are usually embedded very close using deterministic node embedding techniques, and therefore are difficult to be distinguished when linking the anchor nodes – a.k.a. "matching confusions" [6] and is a major bottleneck of improving the UIL performance. In this spirit, our NWUIL provides a new perspective of overcoming the matching confusion issue by exploring the node representation distribution and measuring network similarity using optimal transport distance.

Qualitative analysis. We further investigate the qualitative UIL results of our model against three baselines. The heat maps in Fig. 3 show the anchor node linking performance for GAlign, DeepLink, *d*NAME and NWUIL, where the diagonal dots correspond to the similarity between two user identities across different networks, i.e., the darker the dot, the better UIL performance. Apparently, NWUIL achieves better results (darker along the diagonals) on linking the anchor nodes. Note that here DeepLink and *d*NAME are boosted with 10% anchor nodes.

Visualization of the latent space. Finally, we plot the learned latent space of NWUIL in Fig. 4 using t-SNE to map the high-dimensional vector into the 2D space. Note that for better visualization we omit the distribution of node representation. We can observe a clear clustering effect of the node embedding of all nodes in the respective latent space. This phenomenon is desirable for later anchor node linking since the disentangled node embedding would benefits the model to distinguish the anchor nodes from the confounding nodes (e.g., neighbors and topological structure similar nodes).

V. Related work

Existing UIL approaches can be broadly divided into the supervised, semi-supervised and unsupervised methods. Most of the earlier approaches [3], [26], [27] are supervised, aiming to learn a binary classifier to identify the unknown node pairs by learning the network similarity with the known anchor nodes. For example, PALE [26] is a supervised model, which tries to capture the network structures and learns a mapping function for linking anchor nodes across networks. ULink [3] first introduces the concept of "latent user space" for linking user identity across different social platforms. DeepLink [4] addresses the UIL problem using reinforcement learning which treats the identity linkage as a dual learning process. *d*NAME [6] presents an adversarial learning method which disentangles the network embedding when linking individual anchor nodes. In practice, however, it is extremely expensive to obtain enough linked cross social identities as annotations to train these supervised models.

Another set of methods [10]–[13], [28] addresses the UIL without using the labeled anchor nodes. UUIL [11] is an unsupervised method from user space distribution level, focusing on minimizing the distance distribution between user identities in different social networks. CoLink [28] proposes to link identical users with co-training, which independently models user attributes and relationships, and makes them reinforce each other in an unsupervised way. Xie et al. [12] introduce a factoid embedding based model aiming at coping with different profile attributes, content types and network links of different social networks. Though these methods are anchor-nodes free, they do not perform well and usually require additional user attribute information for distinguishing the person from others.

Compared to these deterministic UIL solutions, our NWUIL is a stochastic model that can exploit the structural embedding uncertainty and optimize the UIL procedure with assured quality due to the optimal transportation theory. As demonstrated by the empirical results, our model not only shows superior performance against the baselines, but also provides interpretable UIL results.

VI. Conclusions

We presented NWUIL – a novel approach for cross-domain identical user linkage. Our user linkage algorithm addresses the biased inference problem by leveraging the graph neural networks for probabilistic node embedding and network Wasserstein distance for anchor node matching. NWUIL is a transportation based unsupervised UIL approach capable of alleviating the matching confusion problem inherent in existing cross-network UIL methods. Empirical results conducted on real-world datasets show that our method not only significantly improves the linking accuracy over existing unsupervised UIL methods but also outperforms the weakly-supervised models. As our ongoing work, we are focusing on two main aspects: taking into account various attributes associated with users (e.g., profiles and posting contents) to further improve the performance of NWUIL, and investigating the influence of topology consistency and constraints on UIL results.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No.61602097), Sichuan Science and Technology Program No.2020YFG0053 and NSF grants CNS 1646107.

Appendix

Proof of Proposition 1

Proof. In order to prove Proposition 1 (Eq.(10)), we reformulate the constrained optimization in the primal problem as an unconstrained one. Consider the following optimizations:

$$\sup_{\phi}\left[\int_{\mathcal{U}'}\phi\left(u'\right)\mathbb{P}^{s}\left(u'\right)du'-\int_{\mathcal{U}\times\mathcal{V}}\phi(u)d\gamma(u,v)\right], \quad (15)$$

$$\sup_{\psi}\left[\int_{\mathcal{V}'}\psi\left(v'\right)\mathbb{P}^{t}\left(v'\right)dv'-\int_{\mathcal{U}\times\mathcal{V}}\psi(v)d\gamma(u,v)\right], \quad (16)$$

where $\int_{\mathcal{U}'}\phi(u')\mathbb{P}^s(u')du'$ and $\int_{\mathcal{V}'}\phi(v')\mathbb{P}^t(v')dv'$ are the expectations of ϕ and ψ, respectively. In the distribution $\mathbb{P}^s$, $\int_{\mathcal{U}\times\mathcal{V}}\phi(u)d\gamma(u,v)$ is the expectation of ϕ under the marginal distribution $\int_v\gamma(u,v)dv$. Obviously, formula (15) and formula (16) are both zero if the marginal constraint over the distributions of $\mathbb{P}^s(u)$ and $\mathbb{P}^t(v)$ are met due to $\int_v\gamma(u,v)dv=\mathbb{P}^s(u)$ and $\int_u\gamma(u,v)du=\mathbb{P}^t(v)$. We then modify the formula of the optimal transport cost in Eq.(9) to incorporate the constraints as follows:

$$d_{OT}(\mathbb{P}^s||\mathbb{P}^t):=\inf_{\gamma}\left[\int_{\mathcal{U}\times\mathcal{V}}c(u,v)d\gamma(u,v)+\sup_{\psi,\phi}\mathcal{L}\right], \quad (17)$$

where

$$\mathcal{L}=\int_{\mathcal{U}\times\mathcal{V}}\left[\int_{\mathcal{U}'}\phi\left(u'\right)\mathbb{P}^{s}\left(u'\right)du'-\phi(u)\right]d\gamma(u,v)$$
$$+\int_{\mathcal{U}\times\mathcal{V}}\left[\int_{\mathcal{V}'}\psi\left(\mathcal{V}'\right)\mathbb{P}^{t}\left(v'\right)dv'-\psi(v)\right]d\gamma(u,v). \quad (18)$$

According to Sion's minimax theorem [29], we can reformulate Eq.(17) as

$$d_{OT}(\mathbb{P}^s||\mathbb{P}^t):=\sup_{\phi,\psi}\inf_{\gamma}\left[\int_{\mathcal{U}\times\mathcal{V}}c(u,v)d\gamma(u,v)+\mathcal{L}\right]$$
$$:=\sup_{\phi,\psi}\left\{\inf_{\gamma}\left[\int_{\mathcal{U}\times\mathcal{V}}[c(u,v)-(\phi(u)+\psi(v))]d\gamma(u,v)\right]\right.$$
$$\left.+\int_{\mathcal{U}'}\phi\left(u'\right)\mathbb{P}^{t}\left(u'\right)du'+\int_{\mathcal{V}'}\psi\left(v'\right)\mathbb{P}^{t}(v')dv'\right\}, \quad (19)$$

where term of $\inf_\gamma$ is constrained by $\phi(u)+\psi(v)\leq c(u,v)$ – otherwise the cost will become arbitrarily large. □

References

[1] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "Hydra: large-scale social identity linkage via heterogeneous behavior modeling," in *SIGMOD*, 2014.
[2] M. Jiang, P. Cui, N. J. Yuan, and X. Xie, "Little is much: Bridging cross-platform behaviors through overlapped crowds," in *AAAI*, 2016.
[3] X. Mu, F. Zhu, E. P. Lim, J. Xiao, J. Wang, and Z. H. Zhou, "User identity linkage by latent user space modelling," in *SIGKDD*, 2016.
[4] F. Zhou, L. Liu, K. Zhang, G. Trajcevski, J. Wu, and T. Zhong, "Deeplink: A deep learning approach for user identity linkage," in *INFOCOM*, 2018.
[5] J. Zhou and J. Fan, "Translink: User identity linkage across heterogeneous social networks via translating embeddings," in *INFOCOM*. IEEE, 2019, pp. 2116–2124.
[6] F. Zhou, Z. Wen, G. Trajcevski, K. Zhang, T. Zhong, and F. Liu, "Disentangled network alignment with matching explainability," in *INFOCOM*. IEEE, 2019, pp. 1360–1368.
[7] L. Liu, W. K. Cheung, X. Li, and L. Liao, "Aligning users across social networks using network embedding," in *IJCAI*, 2016.
[8] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *SIGKDD*, 2014.
[9] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arxiv preprint*, 2016.
[10] S. Lacoste-Julien, K. Palla, A. Davies, G. Kasneci, T. Graepel, and Z. Ghahramani, "Sigma: Simple greedy matching for aligning large knowledge bases," in *SIGKDD*. ACM, 2013, pp. 572–580.
[11] C. Li, S. Wang, P. S. Yu, L. Zheng, X. Zhang, Z. Li, and Y. Liang, "Distribution distance minimization for unsupervised user identity linkage," in *CIKM*. ACM, 2018, pp. 447–456.
[12] W. Xie, X. Mu, R. K.-W. Lee, F. Zhu, and E.-P. Lim, "Unsupervised user identity linkage via factoid embedding," in *ICDM*, 2018.
[13] G. Li, L. Sun, Z. Zhang, P. Ji, S. Su, and P. S. Yu, "Mc2: Unsupervised multiple social network alignment," in *BigData*, 2019.
[14] M. Heimann, H. Shen, T. Safavi, and D. Koutra, "Regal: Representation learning-based graph alignment," in *CIKM*. ACM, 2018, pp. 117–126.
[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.
[16] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.
[17] A. Bojchevski and S. Günnemann, "Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking," in *ICLR*, 2018.
[18] D. Zhu, P. Cui, D. Wang, and W. Zhu, "Deep variational network embedding in wasserstein space," in *SIGKDD*, 2018.
[19] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 43–59, 2017.
[20] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *ICLR*, 2018.
[21] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *ICLR*, 2018.
[22] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *NIPS*, 2013, pp. 2292–2300.
[23] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-wasserstein averaging of kernel and distance matrices," in *ICML*, 2016, pp. 2664–2672.
[24] J. Zhang and P. S. Yu, "Pct: Partial co-alignment of social networks," in *WWW*, 2016.

[25] H. T. Trung, T. Van Vinh, N. T. Tam, H. Yin, M. Weidlich, and N. Q. V. Hung, "Adaptive network alignment with unsupervised and multi-order convolutional networks," in *ICDE*, 2020.
[26] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach." in *IJCAI*, vol. 16, 2016, pp. 1823–1829.
[27] Y. Qiao, Y. Wu, F. Duo, W. Lin, and J. Yang, "Siamese neural networks for user identity linkage through web browsing," *TNNLS*, 2019.
[28] Z. Zhong, Y. Cao, M. Guo, and Z. Nie, "Colink: An unsupervised framework for user identity linkage," in *AAA*, 2018.
[29] M. Sion *et al.*, "On general minimax theorems." *Pacific Journal of mathematics*, vol. 8, no. 1, pp. 171–176, 1958.