The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

A Community Sensing Approach for User Identity Linkage

Zexuan Wang Teruaki Hayashi Yukio Ohsawa

Department of Systems Innovation, School of Engineering, The University of Tokyo

User Identity Linkage aims to detect the same individual or entity across different Online Social Networks, which is a crucial step for information diffusion among isolated networks. While many pair-wise user linking methods have been proposed on this important topic, the community information naturally exists in the network is often discarded. In this paper, we proposed a novel embedding-based approach that considers both individual similarity and community similarity by jointly optimize them in a single loss function. Experiments on real dataset obtained from Foursquare and Twitter illustrate that proposed method outperforms other commonly used baselines that only consider the individual similarity.

1. Introduction

In recent years, Online Social Networks (OSNs) such as Twitter, Facebook and Foursquare tend to become the central platform of people's social life. Tons of contextual (e.g. tweets, photos) and network structure related (e.g. users' profiles, relations) data is created every day on these OSNs, which is an important resource for many valuable applications such as user behavior prediction, and cross-domain recommendation. All such applications require a crucial step called User Identity Linkage (UIL) [Shu 17], which aims to identify and link the same person/entity across different OSNs. These linkages are also called anchor links as they help align different networks under the common scene that users usually don't explicitly claim the ownership of their different accounts, and due to privacy protection rules, personal information is always restricted inside each isolated OSNs.

Abundant literature has been focusing on the UIL problem, and the majority of them fall into two categories: (1) Structure-based approaches: these approaches focus directly on the structural features of a social network, such as user names, following relationship and common neighbors between different users [Malhotra 12, Kong 13], while the problem of those approaches lies on the difficulty to find an optimal distance function between nodes to evaluate their similarity as networks are not presented in the Euclidean space [Zhang 18]. (2) Embedding-based approaches: network embedding is a new way of network representation that is able to encode the network in a continuous low-dimensional vector space while effectively preserving the network structure, for example, [Zhou 18] proposed a dual-learning embedding paradigm to improve the linking result.

However, existing methods haven't paid enough attention to the social communities naturally formed by people in the real world. Users who have limited profile information could be evaluated easier when they are located in interest groups together with their close neighbors. To better resolve the UIL problem, we proposed a novel method called Community Sensing User Identity Linkage (CSUIL), which takes advantage of both structural and embedded features of a network by designing a jointly learning model. It aids user mapping by driving some of users to the same communities they belong to, which enhances the method's accuracy and generalization ability. Experiment results on real-world dataset show feasibility of our method.

Contact: Zexuan Wang, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, wangzexuan@g.ecc.u-tokyo.ac.jp

2. Problem Definition

Definition 1 ***Social Network Graph*** *An unweighted and undirected network is denoted as* $G = \{V, E\}$*, where* V *is the set of nodes and each node represents a user,* E *is the set of edges reflecting connections between nodes.*

Definition 2 ***Node Embedding*** *In a given network* $G = \{V, E\}$*, node embedding (a sub-task of network embedding) learns a projection function* $\psi : V \mapsto \mathbb{R}^{|V|\times d}$*, where* $d \ll |V|$*. For each node* $v_i \in V$*,* $\boldsymbol{\psi}(v_i) \in \mathbb{R}^d$ *denotes its latent representation in the vector space.*

Definition 3 ***n-th order neighbors*** *The collection of all nodes which can be reached from the given root node* $v_r \in G$ *within exactly n hops, denoted as* $C_r = \{v_i | hop(v_i, v_r) = n\}$*.*

Definition 4 ***User Identity Linkage*** *Given two different networks,* $G^S = \{V^S, E^S\}$ *and* $G^T = \{V^T, E^T\}$*. The goal of User Identity Linkage (UIL) is to predict a pair-wise linkage between a user node* v_s *selected from the source network* G^S *and an unlabeled user node* v_s *in the target network* G^T*, which indicates the same user/entity (i.e.,* $v^s = v^t$*).*

3. Community Sensing User Identity Linkage

This proposed method consists of three main components: network embedding, community clustering and latent space mapping. A brief overview is shown in Figure 1, where blocks are the core elements in each phase, green lines indicate structural information flow directions and blue lines show how algorithms connect different phases.

3.1 Network Embedding

The quality of the latent representation of each node in both source and target network is important to the results

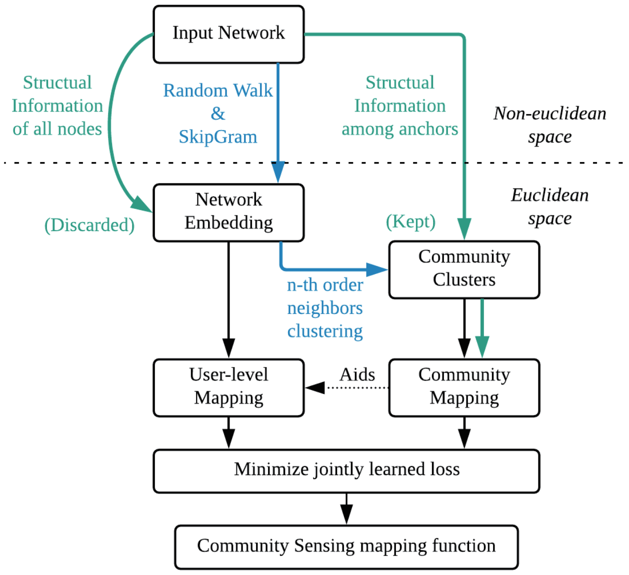


Figure 1: A brief overview of CSUIL

of the following clustering and mapping stages. Ideally, user nodes that have stronger connection, like sharing more common neighbors, or having shorter path between them should be closer to each other after they are projected into the latent space. To obtain the network embedding in good quality, an efficient model called DeepWalk [Perozzi 14] was adopted. DeepWalk mainly utilizes the truncated random walk and the SkipGram [Mikolov 13] model.

In particular, A random walk generator is first applied to the network, which will sample uniformly a random node $v_i \in G$ as the root of a random walk sequence W_{v_i}, then the generator samples uniformly from the neighbors of the last node visited until the maximum sequence length(l) is reached. The generated sequences could be thought of as short sentences, while the nodes within sequences are treated as words of a special kind of language. We could then obtain the embedding of nodes as a byproduct when updating the weight matrix in the derived SkipGram model, which aims to maximize the co-occurrence probability of nodes that appear within a window size w near the center v_j in the sequence W_{v_i}, that is to maximize the following log probability:

$$\max \quad \frac{1}{l}\sum_{i=1}^{l}\sum_{j=-w,j\neq 0}^{w} \log \Pr(v_{i+j}|v_i) \tag{1}$$

where $\Pr(v_{i+j}|v_i)$ is calculated with a hierarchical softmax function:

$$\Pr(v_{i+j}|v_i) = \frac{\exp(\boldsymbol{\psi}(v_{i+j})^{\mathrm{T}}\boldsymbol{\psi}(v_i))}{\sum_{m=1}^{l}\exp(\boldsymbol{\psi}(v_m)^{\mathrm{T}}\boldsymbol{\psi}(v_i))} \tag{2}$$

where $\boldsymbol{\psi}(v_i)$ is the embedding of node v_i we want to update at each training step and finally output to the next phase.

3.2 Community Clustering

In some supervised User Identity Linkage models such as PALE [Man 16], it only focuses on learning the user level, pair-wise matching patterns between source and target network. However, these methods failed to consider the social communities naturally formed by people in the real world. Some drawbacks may exist under such settings that users with very limited profile information could be hard to distinguish from others and the model may fall into over-fitting of local pair-wise features when trained with small amount of labeled data. More importantly, the knowledge contained in the structural relationship among anchor and non-anchor users in the original non-euclidean space is discarded after SkipGram is applied (shown by green lines in Figure 1) and later phases are not able to reuse such information.

Therefore, we made an assumption that compared to only considering the generated embedding or user-level similarity matching, the fact that which neighbors a user has in the original network, and which community a user belongs to could reveal more diffusible structural knowledge. Thus, we consider clustering the $n-th$ order neighbors of an anchor user to form their social community, the users in the same community have a closer relationship and higher similarity, which could be evaluated in some metrics including: the amount of common neighbors, or the minimum walk length between each other.

To utilize all the user information in a community, we reuse the structural information in the original network and derive a new embedding to represent this community by adopting the mean value of all community member embedding generated in Section 3.1 that are non-anchor nodes. The center that represents a certain community cluster C_i is denoted as μ_i:

$$\psi(\mu_i) = \frac{\psi(v_r) + \sum_{v' \in C_i}\psi(v')}{N+1} \tag{3}$$

where v_r is the root user, and N is the community size.

3.3 Latent Space Mapping

Let $\boldsymbol{z^s} = \psi(v^s)$ and $\boldsymbol{z^t} = \psi(v^t)$ be the node embedding generated in Section 3.1 and the final stage of CSUIL is Latent Space Mapping. In this phase, we try to find a mapping function from the source network to the target network Φ: $\mathbb{R}^{|V^s|\times d} \mapsto \mathbb{R}^{|V^t|\times d}$, that will minimize the distance between the predicted embedding $\Phi(\boldsymbol{z^s})$ and the true corresponding embedding $\boldsymbol{z^t}$ of $\boldsymbol{z^s}$ in the target network:

$$\min \quad \left\|\Phi\left(\boldsymbol{z^s}\right) - \boldsymbol{z^t}\right\|_F \tag{4}$$

We then train a novel two-inputs and two-outputs neural network model, which breaks down the whole task above into two simultaneously conducted parts: (1) minimize the distance between predicted and real user node (2) minimize the distance between predicted and real community center. The second sub-task will drive the mapping function to the direction that also exploits the relationship between community centers in both source and target networks to increase the generalization ability of the model on new unseen data.

Next, the design of the loss function could be one of the most critical parts of a machine learning model, a good loss function should reflect the error during training as well as the generalization error that guides parameters to optimize the model. Therefore, for the goal of above two sub-tasks, a new community sensing loss function is proposed as:

$$\begin{aligned} loss = (1-\gamma) & \sum_{(v^s, v^t) \in \{S,T\}} \left\| \Phi(\boldsymbol{z^s}; \theta) - \boldsymbol{z^t} \right\|_F \\ & + \gamma \sum_{\mu \in C} \left\| \Phi(\boldsymbol{\mu^s}; \theta) - \boldsymbol{\mu^t} \right\|_F \end{aligned} \tag{5}$$

where $\{S, T\}$ is the set of groundtruth anchor pairs, C is the set of community centers, F is the Frobenius norm, θ is the collection of all parameters in the model, and γ is the hyper-parameter of the weight coefficient of the community loss that could be co-optimized during the learning of the mapping function.

We finally employed a Multi-Layer Perceptron (MLP) model that does not require extensive feature selection or difficult parameter tuning to learn the optimized mapping function, while this model also has the flexibility of dealing with the non-linear relationships that may exist between the source and target network.

The whole algorithm design is shown in Algorithm 1.

Algorithm 1: CSUIL

Input: network $G(V, E)$, anchor nodes $\{S, T\}$, test nodes $\{S', T'\}$, community clustering parameter n, community loss parameter γ
Output: mapping function Φ, matching result list R
foreach *node* $v_i \in G$ **do**
  Generate the embedding of v_i as $\boldsymbol{z_i}$
end
foreach *anchor node pair* $\{s_i, t_i\}$ *in* $\{S, T\}$ **do**
  Reuse the original network structure information, cluster the n-order neighbors of s_i and t_i
  Derive the community center $\boldsymbol{\mu_i^s}$ and $\boldsymbol{\mu_i^t}$
end
Train the MLP model by jointly minimize the node mapping loss $\left\| \Phi(\boldsymbol{z^s}; \theta) - \Phi(\boldsymbol{z^t}) \right\|_F$ and community loss $\left\| \Phi(\boldsymbol{\mu^s}; \theta) - \Phi(\boldsymbol{\mu^t}) \right\|_F$
foreach *test node* $s_i' \in S'$ **do**
  Add the predicted t_i' to result list R
end
$Evaluate(R, T')$

4. Experiment

4.1 Data Preparation

A real-world social network dataset collected from Twitter and Foursquare [Zhang 15] is used in this experiment, which was released in [Liu 16]. All the sensitive personal information is removed under privacy concerns to form the final training and testing data. The ground truth of anchors is obtained by crawling users' Twitter accounts from their Foursquare homepage. Table 1 lists the statistics of this dataset.

Network	#Users	#Relations	#Anchors
Twitter	5,220	164,919	1,609
Foursquare	5,315	76,972	

Table 1: Statistics of Twitter-Foursquare Dataset

4.2 Evaluation Metrics

In this experiment, in a similar form to [Zhou 18], a metric called $Precision@k$ was adopted, which is defined as:

$$Precision@k = \frac{\sum_i^n TOP_k(\Phi(\boldsymbol{z_i^s}))}{N} \tag{6}$$

where $TOP_k(\Phi(\boldsymbol{z_i^s}))$ is a binary output function (0 or 1), for each predicted embedding $\Phi(\boldsymbol{z_i^s})$, it tells whether the positive match $\boldsymbol{z_i^t}$ exists in the $top-k$ list or not, and N is the number of all testing nodes. In the context of UIL, as $Precision@k$ is a metric of the true positive rate, it could be treated the same as $Recall@k$, and $F_1@k$.

4.3 Comparative Methods

We compare the proposed CSUIL with several existing embedding-based methods, and take them as the baseline of this task.

- **CSUIL**: the proposed method, it could explicitly exploit the individual as well as community features of a network, by jointly optimizing mapping functions that concentrate on user-level and community-level similarity respectively.
- **IONE**: Proposed in [Liu 16] and adopted as a baseline result, Input-Output Network Embedding (IONE) is a network embedding and partial network alignment method. It takes follower-ship and followee-ship as input and output contexts and generates all three representations together with the user node.
- **INE**: INE is a simplified version of IONE, which only consider node and input representation for matching.

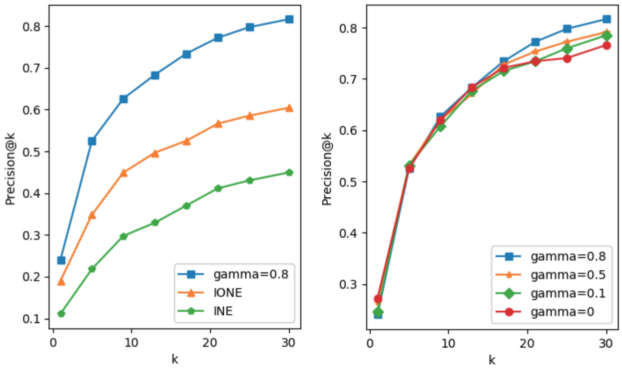
4.4 Results

The performance results are illustrated in Table 2 and Figure 2. In the experiment, during the community clustering phase, the cluster size is set to first-order neighbors for the simplicity. Then we examine the ability of the final model (with training rate=90%) on the link prediction task. For CSUIL, we report the result in different settings of precision metrics k and community loss weight coefficient γ. For INE and INOE, we report the result in the original paper's default setting.

	Precision@k							
γ	P@1	P@5	P@9	P@13	P@17	P@21	P@25	P@30
INE	0.1108	0.2184	0.2975	0.3291	0.3703	0.4114	0.4304	0.4494
IONE	0.1899	0.3481	0.4494	0.4968	0.5253	0.5665	0.5854	0.6044
0.8	**0.2405**	**0.5190**	**0.6203**	**0.6835**	**0.7342**	**0.7722**	**0.7975**	**0.8165**

Table 2: Performance comparison between baselines

From the experiment results, we could conclude that:



(a) Baselines (b) Precision on different γ

Figure 2: Link prediction precision results. X-axis is the different value of k, for the top-k list being evaluated; Y-axis is the precision result in percentage.

- Compared to the baseline model, INE and IONE, the best performance (when $\gamma = 0.8$, shown in Table 2 and Figure 2(a)) of our approach has an improvement from about 6% to 21 % at most in different settings of precision metrics, which shows the feasibility of this approach.
- Figure 2(b) also illustrates that by changing the setting of community loss weight coefficient γ, the ability of the model to sense more positive matching in a larger search space (higher k setting in precision), could be enhanced, which is an important improvement because many other papers only stress their performance at the $k = 30$ setting. However, adding too much weight to community loss may lead to a slight reduction of the ability to narrow the target to a finer scale (lower k in precision), compared with the $\gamma = 0$ setting.

5. Conclusion

In this paper, we aim to study the UIL problem by reusing the discarded knowledge in the original Online Social Network after network embedding. Not limited to anchor same users across networks, we would also like the community formed by close users to have a positive match across networks. This is because some users may have limited profile and it could be hard to distinguish them from others. However, in the context of a community, users share common features, and they will be driven to the correct direction where group of users with high similarity locates, even if community members are known little. This could also help to avoid overfitting the input data and increase the generalization ability of the method.

Therefore, we break down the main task into two simultaneously learned sub-tasks: User Mapping and Community Mapping, this is achieved by jointly optimizing the user loss and community loss in a single MLP model. Based on above theories, Community Sensing User Identity Linkage (CSUIL) is proposed. Results show that our approach outperforms current baseline models, and has the flexibility to adapt hyper-parameters for different needs or data input.

Acknowledgments

This work was funded by JSPS KAKENHI JP16H01836, JP16K12428, and industrial collaborators.

References

[Kong 13] Kong, X., Zhang, J., and Yu, P. S.: Inferring anchor links across multiple heterogeneous social networks, in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 179–188ACM (2013)

[Liu 16] Liu, L., Cheung, W. K., Li, X., and Liao, L.: Aligning Users across Social Networks Using Network Embedding., in *IJCAI*, pp. 1774–1780 (2016)

[Malhotra 12] Malhotra, A., Totti, L., Meira Jr, W., Kumaraguru, P., and Almeida, V.: Studying user footprints in different online social networks, in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 1065–1070IEEE Computer Society (2012)

[Man 16] Man, T., Shen, H., Liu, S., Jin, X., and Cheng, X.: Predict Anchor Links across Social Networks via an Embedding Approach., in *IJCAI*, Vol. 16, pp. 1823–1829 (2016)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[Perozzi 14] Perozzi, B., Al-Rfou, R., and Skiena, S.: Deepwalk: Online learning of social representations, in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710ACM (2014)

[Shu 17] Shu, K., Wang, S., Tang, J., Zafarani, R., and Liu, H.: User identity linkage across online social networks: A review, *Acm Sigkdd Explorations Newsletter*, Vol. 18, No. 2, pp. 5–17 (2017)

[Zhang 15] Zhang, J. and Philip, S. Y.: Integrated Anchor and Social Link Predictions across Social Networks., in *IJCAI*, pp. 2125–2132 (2015)

[Zhang 18] Zhang, J.: Social Network Fusion and Mining: A Survey, *CoRR*, Vol. abs/1804.09874, (2018)

[Zhou 18] Zhou, F., Liu, L., Zhang, K., Trajcevski, G., Wu, J., and Zhong, T.: DeepLink: A Deep Learning Approach for User Identity Linkage, in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 1313–1321IEEE (2018)