Hubness-aware User Identity Linkage

Chaozhuo Li
Beihang University
cli@microsoft.com

Senzhang Wang
Central South University
szwang@csu.edu.cn

Feiran Huang
Jinan University
huangfr@jnu.edu.cn

Jie Xu*
Beijing Foreign Studies University
xujie1020@buaa.edu.cn

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

ABSTRACT

Nowadays, it is common for one natural person to join multiple social networks to enjoy different types of services. User identity linkage (UIL), which aims to link identical identities across different social platforms, has attracted increasing research interests recently. Most existing approaches focus on the sophisticated architecture engineering of the linkage model but ignore the challenge of hubness in the post-processing nearest neighbor search phase. Hubness appears as some identities in a social platform, called hubs, being extra-ordinary close to the identities in the other platform, which will degrade the alignment performance. Different from existing heuristic methods, in this paper we propose a hubness-aware user identity linkage model HAUIL to smoothly learn hubless linkage signals. A carefully-designed objective function is presented to explicitly mitigate the hubness information from the pre-learned linkage guidance. HAUIL can be easily adapted to most existing UIL models. Empirically, we evaluate HAUIL over multiple publicly available datasets, and the experimental results demonstrate its superiority.

CCS CONCEPTS

• **Information systems** → **Social networks**.

KEYWORDS

social identity linkage; social network analysis; deep learning

ACM Reference Format:
Chaozhuo Li, Senzhang Wang, Feiran Huang, Jie Xu, and Philip Yu. 2021. Hubness-aware User Identity Linkage. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482122

1 INTRODUCTION

Nowadays, users tend to simultaneously join a variety of social platforms to enjoy different types of services. When a user registers on a social platform, an identity is created to represent his unique personal figure. As an indispensable step in cross-platform social network mining, user identity linkage (UIL), which aims to link identities of one same natural person across different social networks, has attracted enormous attentions considering its significant research challenges and tremendous practical values.

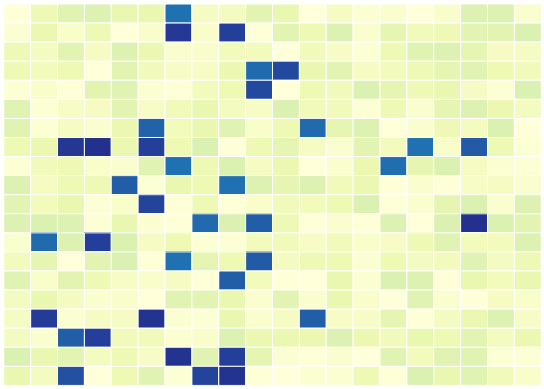


Figure 1: An illustrative example of hubness in the user identity linkage task.

Existing UIL models [6–8, 11, 12, 17, 20, 24] are mostly distance-based and usually aim to learn a desirable projection function to map user identities from the source platform to the target one. In the model training phase, the objective function is designed to minimize the distances between identities in the matched pairs and maximize the distances of unmatched ones. After that, in the inference phase, the ground distances between the projected source identity and the target candidates can be viewed as the linkage signals to retrieve the top-k nearest neighbors. Existing distance-based approaches mainly focus on the sophisticated architecture engineering of the projection function (e.g., adversarial learning [6, 7], graph neural network [1, 21] and translation models [23]). However, the post-processing nearest neighbor search in the inference phase is also critical to achieving the desirable linkage performance [1], which is seriously hindered by a phenomenon called hubness. How to alleviate the negative impact of hubness to the UIL task is rarely touched and still remains an open research problem.

Hubness is a tendency that a few target identities (hubs) appear unwontedly often among the k-nearest neighbor lists of the source identities [15]. Figure 1 intuitively illustrates the phenomenon of hubness in the UIL task. A SOTA UIL model SNNA is trained over a popular Facebook-Twitter dataset [7]. We randomly select 400

*Indicates Corresponding Author.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482122

Twitter identities and record the frequencies of these Twitter identities appearing in the top-5 lists of the Facebook identities. Each cell denotes a Twitter identity, and the darkness of color in a cell is proportion to the normalized occurrence times of the corresponding Twitter identities. A natural person usually registers very few identities in a social platform. Thus, an source identity may only link to one or few identities in the target platform. It would be more promising if the target identities uniformly appear in the top-k list of the source identities. However, this fine-grained one-to-few mapping is not explicitly enforced in most UIL models. As shown in Figure 1, several Twitter identities may appear much more frequent than others in the top-k nearest neighbor lists, which violates the reality of one-to-few mapping and will degrade the alignment performance. The underlying reason of hubness is the highly imbalanced distribution of the user activeness [1]. If an active user in the target network published a lot of microblogs and interacted frequently with others, then she has a larger chance to be close to many source identities due to the rich mutual information.

The hubness problem is rarely investigated by the existing UIL models. Chen et al. [1] propose a heuristic strategy named CGSS (Cross-Graph Similarity Scaling) as the distance measurement to mitigate hubness, which modifies the traditional cosine similarity with local scaling normalization. However, CGSS is a heuristic strategy without any mathematical foundation, which cannot explicitly demonstrate the motivations theoretically. Here we aim to alleviate the challenge of hubness from a totally different perspective with powerful mathematical derivations. To ensure the flexibility and generality, we assume the distances between projected source identities and target ones have been calculated by previous models. We argue that the ground distances cannot be directly utilized as the linkage guidance considering the severe challenge of hubness. Thus, new hubless linkage signals are required. This motivation leads to an optimization problem which smoothly converts the distances to the linkage probabilities. The learned probabilities should not only preserve the valuable linkage guidance in the pre-learned distances, but also can effectively mitigate the hubness. As our proposal focuses on the post-processing phase, it can be easily adapted to most existing UIL models, which demonstrates the generality of our proposal. The proposed HAUIL (**H**ubness-**A**ware **U**ser **I**dentity **L**inkage) model is thoroughly evaluated on five pairs of real-life datasets, and the experimental results demonstrate its superiority.

We summarize our main contributions as follows:

- We study the novel problem of hubness-aware UIL. Different from existing heuristic approaches, we introduce a flexible optimization-based strategy with strong mathematical foundation to smoothly mitigate the hubness.
- Extensively, we evaluate HAUIL on five groups of datasets. Experimental results demonstrate the superior performance of the proposed approach.

2 PROBLEM STATEMENT

We denote a source network $\mathbf{S} = \{s_1, s_2, \cdots, s_m\}$ with m identities and a target network $\mathbf{T} = \{t_1, t_2, \cdots t_n\}$ with n identities. We assume that a distance matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$ has already been calculated by the existing UIL models, in which $\mathbf{G}_{ij}$ denotes the ground distance between the target identity t_j and the source identity s_i. In this paper, we aim to learn a probability matrix $\mathbf{U} \in \mathbb{R}^{m \times n}$, in which $\mathbf{U}_{ij}$ denotes the probability of identity s_i and t_j belonging to the same natural person. Matrix $\mathbf{U}$ is expected to alleviate the challenge of hubness while preserving the valuable knowledge in the original distance matrix $\mathbf{G}$.

3 DISTANCE-BASED LINKAGE MODELS

Existing identity linkage models usually aim to learn a desirable projection function to minimize the distances between identities in the matched pairs and maximize the distances of unmatched ones [6, 8, 11, 12]. Given an aligned identity pair (s_i, t_p) and the unmatched pair (s_i, t_q), the objective function is defined as:

$$\min_{\mathcal{M}} \ \mathcal{L}_d = d(\mathcal{M}(s_i), t_p) - d(\mathcal{M}(s_i), t_q) \tag{1}$$

in which $\mathcal{M}$ is the projection function that maps the source identity into the target space. Function d measures the ground distance between two points, which is usually implemented as the Euclidean distance. With the learned projection function, we can achieve a distance matrix $\mathbf{G}$ which contains the learned distance-based linkage signals. However, these signals are seriously affected by hubness as discussed in the introduction section. Hence, we aim to smoothly mitigate the hubness from the learned distance matrix $\mathbf{G}$.

4 METHODOLOGY

We aim to learn a probability matrix $\mathbf{U}$ based on the learned distance matrix $\mathbf{G}$, which is expected to be distance preserving and hubless. Several works have been proposed to mitigate the challenge of hubness in different scenarios [2, 4, 5, 15]. Inspired by previous works [4, 5], we employ related techniques to remove hubness in the user identity linkage scenario. Next we will introduce the objective function from the perspectives of distance preserving and anti-hubness.

4.1 Distance Preserving

The distance matrix $\mathbf{G}$ contains crucial linkage signals learned by SOTA UIL models. Larger ground distances lead to the lower linkage probabilities. Based on this criterion, the objective function is presented as follows:

$$\begin{aligned} &\mathcal{L} = \min_{\mathbf{U}} \sum_{i,j} \mathbf{G}_{ij}\mathbf{U}_{ij} + \lambda \sum_{i,j} \mathbf{U}_{ij} \log \mathbf{U}_{ij}, \\ &s.t. \ \sum_{j} \mathbf{U}_{ij} = 1, \mathbf{U}\mathbf{U}^{\top} = \mathbf{I} \end{aligned} \tag{2}$$

The first term ($\min_{\mathbf{U}} \sum_{i,j} \mathbf{G}_{ij}\mathbf{U}_{ij}$) ensures that a larger distance $\mathbf{G}_{ij}$ leads to a lower linkage probability $\mathbf{U}_{ij}$, and vice versa. In addition, this objective function also minimizes the cost to transfer from the source space to the target space, which is equivalent to the Earth Mover's Distance (EMD) [13]. EMD estimates the minimum cost of turning the earth into the holes, which is calculated by the amount of dirt moved (i.e., $\mathbf{U}_{ij}$) times the moving distance (i.e., $\mathbf{G}_{ij}$), and has been proven as an effective measurement in UIL [8].

Loss $\mathcal{L}$ is regularized by $\lambda \sum_{i,j} \mathbf{U}_{ij} \log \mathbf{U}_{ij}$, which is the negative entropy of $\mathbf{U}$. This regularization term gives a measure of how uniform the probabilities are and a higher entropy tends to make the probabilities less extreme. Besides, negative entropy term has a clear bound which simplifies the process of finding the optimal

Table 1: Statistics of the datasets.

Dataset	Source Network	Target Network	$\#_{Matched}$
Twi.-Four.	Twitter (5,120)	Foursquare (5,313)	3,143
Twi.-Fli.	Twitter (6,005)	Flickr (4,403)	3,499
La.-My.	Lastfm (4,807)	Myspace (4,464)	1,777
Dou.-Wei.	Douban (15,151)	Weibo (28,646)	11,170
Twi.-Tum.	Twitter (127,736)	Tumblr (103,427)	30,746

solution. λ is the weight of regularization term. Given a source user i, the first constraint ensures that the summation of its matching probabilities on all target identities should be equal to 1. The second constraint forces the matrix $\mathbf{U}$ to be orthogonal as an orthogonal projection is theoretically appealing for its numerical stability.

4.2 Anti-hubness

The challenge of hubness is caused by some target identities being retrieved more frequently than others. Given a hub identity t_p and an anti-hub one t_q, we can get

$$\sum_{i=1}^{m} \mathbf{G}_{ip} < \sum_{i=1}^{m} \mathbf{G}_{iq} \tag{3}$$

as hub identities tend to be closer to other identities compared with anti-hub ones. A natural idea is to force all target identities being equally preferred to be retrieved [5]. Based on this motivation, the following constraint is presented:

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{U}_{ij} = \frac{1}{n} \tag{4}$$

in which the left expression denotes the preference of target identity t_j, namely on average how t_j is likely to be selected as the linkage candidate of a source identity. This constraint can force the preference to be uniformly distributed over all target identities. Overall, the final objective function is the combination of Formula 2 and Formula 4. This objective function can be easily optimized by the stochastic gradient descent algorithm.

5 EXPERIMENTS

5.1 Experimental Settings

We select five publicly available real-life datasets to evaluate the HAUIL model: Twitter-Foursquare [19], Twitter-Flickr [16], Lastfm-Myspace [22], Douban-Weibo [8] and Twitter-Tumblr [9]. The statistical information of the datasets is presented in Table 1. In order to thoroughly evaluate the generality of HAUIL, we select a set of SOTA UIL methods as the basic models to generate the distance matrix: MAH [14], COSNET [22], IONE [10], MEgo2Vec [18], ULink [12] and SNNA [7]. For the datasets without node attributes, we utilize Node2Vec [3] to generate the unsupervised node embeddings as representations. For the probability-based approaches, we use the normalized matching probabilities as the distance matrix. In the inference phase, following previous works [6, 7], we exploit the cosine similarity as the distance measurement. The dimension of node embeddings is set to 100, the negative entropy weight λ is set to 0.3 and the learning rate is set to 0.001. The parameters of baselines are carefully tuned on a small validation dataset following the guidelines in the original papers. Based on the previous works [7, 12], we select *Hit-Precision* score as the evaluation metric, which is formally defined as:

$$h(x) = \frac{k - (hit(x) - 1)}{k} \tag{5}$$

where $hit(x)$ is the rank position of the matched target user in the returned top-k candidate target identities. The *Hit-Precision* is calculated by the average on the scores of the matched identity pairs: $\frac{\sum_{i=0}^{i=m} h(x_i)}{m}$, in which m is the number of source identities in matched pairs.

5.2 Experimental Results

For each dataset, T_{tr} portion of aligned identity pairs are randomly selected as the training annotations, and 500 linked pairs are randomly selected as the test samples. T_{tr} increases from 0.1 to 0.5. We repeat this process 3 times and report the average *Hit-Precision* scores. Parameter k in the *Hit-Precision* measurement is set to 10. We select the recent CGSS [1] method as the hubness-aware baseline. Table 2 presents the experimental results. With the increase of training ratio T_{tr}, all models achieve better performance. Compared with traditional approaches, CGSS improves the performance by around 1%. Mitigating hubness with such a straight-forward method still outperforms baselines, which proves the importance of the studied hubless UIL task. Our proposal consistently achieves the best performance over all datasets under different settings. HAUIL beats the CGSS method by nearly 1.6% and outperforms traditional approaches by around 2.5%. Experimental results proves that our proposal can effectively alleviate the challenge of hubness, and thus achieves desirable linkage performance.

5.3 Ablation Study

The objective function contains a negative entropy regularizer and three constraints. In order to evaluate the importance of different components, we remove them from the objective function as four ablation models. Namely, HAUIL_1, HAUIL_2, HAUIL_3 and HAUIL_4 denote the models without negative entropy regularizer, non-negativity constraint, summation constraint and the anti-hubness constraint, respectively. Table 3 presents the linkage performance of ablation models on the five datasets. One can see that the performance of all ablation models is lower than the performance of vanilla HAUIL model, which demonstrates that all the components are indispensable to obtain the promising linkage results. Note that, without the anti-hubness component, HAUIL_4 achieves the worst performance, which also proves the importance of mitigating the hubness information.

5.4 Parameter Sensitivity Study

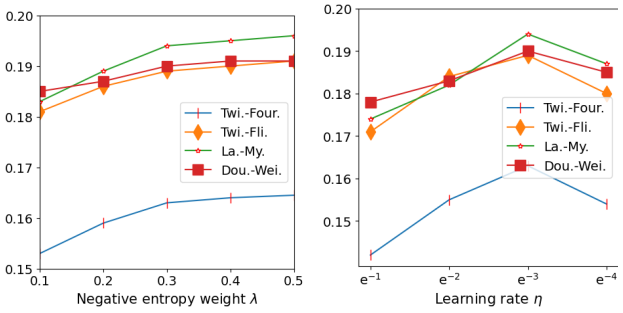
Here we study the performance sensitivity of HAUIL model on two core parameters: the negative entropy weight λ and the learning rate η. Training ratio T_{tr} is set to 10%. λ varies from 0.1 to 0.5, and η is set from e^{-1} to e^{-4}. $Hit-Precision$ scores under different settings on four dataset are reported. Figure 2 presents the experimental results. With the increase of λ, the performance over all the datasets first increases and then keeps steady or slightly

Table 2: User identity linkage performance with different training ratio T_r (*Hit-Precision* score).

	Twitter-Foursquare			Twitter-Flickr			Lastfm-Myspace			Douban-Weibo			Twitter-Tumblr		
T_{tr}	T_{tr}=0.1	T_{tr}=0.3	T_{tr}=0.5	T_{tr}=0.1	T_{tr}=0.3	T_{tr}=0.5	T_{tr}=0.1	T_{tr}=0.3	T_{tr}=0.5	T_{tr}=0.1	T_{tr}=0.3	T_{tr}=0.5	T_{tr}=0.1	T_{tr}=0.3	T_{tr}=0.5
MAH	0.102	0.162	0.192	0.126	0.177	0.206	0.147	0.232	0.288	0.136	0.171	0.213	0.238	0.281	0.310
MAH+CGSS	0.104	0.169	0.206	0.131	0.186	0.217	0.154	0.238	0.294	0.144	0.173	0.221	0.246	0.288	0.217
MAH+HAUIL	**0.112**	**0.188**	**0.219**	**0.139**	**0.204**	**0.224**	**0.167**	**0.253**	**0.309**	**0.153**	**0.189**	**0.239**	**0.254**	**0.298**	**0.327**
COSNET	0.116	0.190	0.217	0.132	0.189	0.216	0.152	0.234	0.282	0.139	0.171	0.210	0.256	0.296	0.319
COS.+CGSS	0.123	0.192	0.219	0.139	0.192	0.228	0.159	0.240	0.293	0.148	0.179	0.224	0.264	0.303	0.330
COS.+HAUIL	**0.133**	**0.197**	**0.226**	**0.152**	**0.204**	**0.245**	**0.170**	**0.252**	**0.303**	**0.154**	**0.187**	**0.229**	**0.275**	**0.316**	**0.341**
IONE	0.121	0.167	0.191	0.134	0.178	0.221	0.148	0.227	0.284	0.125	0.154	0.205	0.231	0.268	0.297
IONE+CGSS	0.126	0.172	0.214	0.140	0.185	0.225	0.157	0.236	0.291	0.129	0.158	0.207	0.239	0.276	0.306
IONE+HAUIL	**0.135**	**0.184**	**0.222**	**0.148**	**0.202**	**0.226**	**0.165**	**0.243**	**0.297**	**0.141**	**0.172**	**0.208**	**0.248**	**0.287**	**0.314**
MEgo2Vec	0.136	0.201	0.239	0.147	0.193	0.218	0.159	0.238	0.284	0.153	0.182	0.218	0.267	0.304	0.326
MEg.+CGSS	0.141	0.216	0.250	0.152	0.206	0.249	0.165	0.249	0.295	0.161	0.195	0.238	0.274	0.316	0.331
MEg.+HAUIL	**0.154**	**0.230**	**0.261**	**0.161**	**0.215**	**0.252**	**0.172**	**0.253**	**0.308**	**0.173**	**0.204**	**0.242**	**0.279**	**0.318**	**0.346**
ULink	0.138	0.178	0.209	0.155	0.207	0.235	0.155	0.236	0.283	0.176	0.207	0.246	0.274	0.313	0.336
ULink+CGSS	0.147	0.192	0.224	0.159	0.209	0.242	0.168	0.248	0.306	0.179	0.209	0.262	0.281	0.319	0.345
ULink+HAUIL	**0.159**	**0.200**	**0.231**	**0.167**	**0.213**	**0.258**	**0.173**	**0.252**	**0.310**	**0.187**	**0.216**	**0.269**	**0.289**	**0.326**	**0.357**
SNNA	0.147	0.215	0.253	0.174	0.220	0.242	0.176	0.260	0.304	0.181	0.215	0.250	0.286	0.328	0.349
SNNA+CGSS	0.155	0.227	0.259	0.181	0.231	0.257	0.187	0.273	0.317	0.184	0.220	0.254	0.293	0.337	0.358
SNNA+HAUIL	**0.163**	**0.239**	**0.272**	**0.189**	**0.350**	**0.267**	**0.194**	**0.275**	**0.328**	**0.190**	**0.228**	**0.267**	**0.307**	**0.346**	**0.373**

Table 3: Ablation study of HAUIL.

Method	Twi.-Four.	Wei.-Dou.	La.-My.	Dou.-Wei.	Twi.-Tum.
Baseline	0.147	0.174	0.176	0.181	0.286
HAUIL_1	0.151	0.179	0.182	0.183	0.292
HAUIL_2	0.155	0.181	0.185	0.187	0.295
HAUIL_3	0.158	0.184	0.189	0.186	0.298
HAUIL_4	0.149	0.177	0.179	0.182	0.290
HAUIL	**0.163**	**0.189**	**0.194**	**0.190**	**0.307**



(a) Negative entropy weight λ. (b) Learning rate η.

Figure 2: Parameter sensitivity analysis.

increases, which demonstrates appropriate negative entropy regularization may benefit the linkage performance. However, a larger λ will lead the training procedure to focus more on the negative entropy minimization task, which may interrupt and slow down the optimization speed to achieve optimal linkage solution. From the right sub-figure, one can see that with the increase of learning rate η, the performance over all the datasets first significantly increases and then dramatically drops. When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. A learning rate that is too small may never converge or get stuck on a sub-optimal solution and brings more time consuming. Thus, we have to carefully choose an appropriate value to balance the model efficiency and effectiveness.

6 CONCLUSION

In this paper, we study the problem of mitigating hubness information from the UIL task, which is a critical problem but rarely explored. Different from existing heuristic methods, in this paper we introduce a hubness-aware user identity linkage model HAUIL with mathematical foundations to smoothly learn hubless linkage signals. HAUIL has strong generality and can be easily adapted to the most existing UIL models. Our proposal is extensively evaluated over five real-life datasets. Experimental results demonstrate the superiority of our proposal.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61906074, 61906075, 61932011, 62172443), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011276), Guangdong Provincial Key R&D Plan (Grant No. 2019B1515120010, 202020022911500032).

REFERENCES

[1] Chaoqi Chen, Weiping Xie, Tingyang Xu, Yu Rong, Wenbing Huang, Xinghao Ding, Yue Huang, and Junzhou Huang. 2019. Unsupervised adversarial graph alignment with graph embedding. *arXiv preprint arXiv:1907.00544* (2019).

[2] Roman Feldbauer and Arthur Flexer. 2019. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge and Information Systems* 59, 1 (2019), 137–166.

[3] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[4] Jiaji Huang, Xingyu Cai, and Kenneth Church. 2020. Improving Bilingual Lexicon Induction for Low Frequency Words. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1310–1314.

[5] Jiaji Huang, Qiang Qiu, and Kenneth Church. 2019. Hubless nearest neighbor search for bilingual lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4072–4080.

[6] 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 *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 249–258.

[7] Chaozhuo Li, Senzhang Wang, Yukun Wang, Philip S. Yu, Yanbo Liang, Yun Liu, and Zhoujun Li. 2019. Adversarial Learning for Weakly-Supervised Social Network Alignment. In *AAAI*. 996–1003.

[8] 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.

[9] Bang Hui Lim, Dongyuan Lu, Tao Chen, and Min-Yen Kan. 2015. # mytweet via instagram: Exploring user behaviour across multiple social networks. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 113–120.

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

[11] 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.

[12] 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.

[13] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40, 2 (2000), 99–121.

[14] 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.

[15] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic. 2013. The role of hubness in clustering high-dimensional data. *IEEE transactions on knowledge and data engineering* 26, 3 (2013), 739–751.

[16] Ming Yan, Jitao Sang, Tao Mei, and Changsheng Xu. 2013. Friend transfer: Cold-start friend recommendation with cross-platform transfer learning of social knowledge. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

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

[18] Jing Zhang, Bo Chen, Xianming Wang, Hong Chen, Cuiping Li, Fengmei Jin, Guojie Song, and Yutao Zhang. 2018. Mego2vec: Embedding matched ego networks for user alignment across social networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 327–336.

[19] Jiawei Zhang and S Yu Philip. 2015. Integrated anchor and social link predictions across social networks. In *Twenty-fourth international joint conference on artificial intelligence*.

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

[21] Wen Zhang, Kai Shu, Huan Liu, and Yalin Wang. 2019. Graph neural networks for user identity linkage. *arXiv preprint arXiv:1903.02174* (2019).

[22] 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.

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

[24] Xiaoping Zhou, Xun Liang, Haiyan Zhang, and Yuefeng Ma. 2015. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE transactions on knowledge and data engineering* 28, 2 (2015), 411–424.