User Identity Linkage across Online Social Networks: A Review

Kai Shu[†], Suhang Wang[†], Jiliang Tang[‡], Reza Zafarani [♮], and Huan Liu[†]
[†]Computer Science & Engineering, Arizona State University, Tempe, AZ, USA
[‡]Computer Science & Engineering, Michigan State University, East Lansing, MI, USA
[♮]Computer Science & Engineering, Syracuse University, Syracuse, NY, USA
[†]{kai.shu,suhang.wang,huan.liu}@asu.edu,
[‡]tangjili@msu.edu, [♮] rzafaran@syr.edu

ABSTRACT

The increasing popularity and diversity of social media sites has encouraged more and more people to participate on multiple online social networks to enjoy their services. Each user may create a *user identity*, which can includes profile, content, or network information, to represent his or her unique public figure in every social network. Thus, a fundamental question arises – *can we link user identities across online social networks?* User identity linkage across online social networks is an emerging task in social media and has attracted increasing attention in recent years. Advancements in user identity linkage could potentially impact various domains such as recommendation and link prediction. Due to the unique characteristics of social network data, this problem faces tremendous challenges. To tackle these challenges, recent approaches generally consist of (1) extracting features and (2) constructing predictive models from a variety of perspectives. In this paper, we review key achievements of user identity linkage across online social networks including state-of-the-art algorithms, evaluation metrics, and representative datasets. We also discuss related research areas, open problems, and future research directions for user identity linkage across online social networks.

1. INTRODUCTION

In recent years, users have been introduced to many online social networks such as Twitter, Instagram, or LinkedIn. Due to diverse functionalities, different online social network platforms attract users for different purposes such as information seeking/sharing and social connection maintenance. For example, users may use Twitter to publish opinions on political events while adopting Instagram to share their leisure activities [44]. To better take advantage of services provided by each social network, users tend to join multiple online social networks. It has become increasingly popular for users to have accounts (also called user identities) on multiple social networks. As reported by a social media study, by the end of 2013, 42% of online adults are using multiple social media sites at the same time [1]. For example, 93% of Instagram users are involved in Facebook concurrently and 53% Twitter users are using Instagram as well [2].

Implications of Linking User Identities. The increasing popularity of users with accounts on multiple social media sites brings new opportunities and challenges to various mining and learning tasks. First, users with accounts on multiple social media sites give potentials to fully understanding users' interests and provide better recommendations or services [13; 33]. As user uses different online social networks for different purposes, analyzing a user identity on a single social media may not give a comprehensive understanding of his/her personalities and interests. However, if we can link a person's user identities on multiple datasets, collect and analyze his/her data on these social media sites together, we may have a more comprehensive view about the user and provide better services. Second, users with accounts on multiple social media sites allow us to integrate patterns among online social network sites and solve some problems unsolvable by data from only one site such as cold-start and data sparsity problems in many predictive tasks-[18; 62]. For example, a newly founded social media service may not have enough historical data for recommendations to users. If we can identify these users on other well established social media sites, then we can transfer knowledge from the mature social media to the new social media and thus mitigate the data sparsity or cold start problems. Finally, users with accounts on multiple social media sites can also help analyze user migration patterns and guide web developments [33]. Users migrating from one social network to another often reflects the user experience of web development. The linkage of user identities across different social media sites provides a great chance to study use migration behaviors. In addition, linking user identities allows for:

I. Enhancing Friend Recommendation. Online user engagement can increase with better friend recommendations. However, most friend recommendation algorithms recommend (1) non-connected users that (2) share mutual friends, as potential friends. Consider two users u_1 and u_2 that are not connected and are both friends of u_3 on site S_1. Thus, u_1 seems a good candidate for recommendation to u_2 on S_1. u_1 and u_2 are also members of social network S_2 and are also not connected on S_2. Assume that u_1 and u_2 share no mutual friends on S_2. With the information that we have

[1] http://www.pewinternet.org/2015/01/09/social-media-update-2014/

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

from S_1, the recommendation algorithm could recommend u_1 to u_2 on S_2, even though they share no mutual friends on S_2. This type of recommendation is only possible when there is cross-site complementary information.

II. Information diffusion. Information diffusion has been traditionally studied within a single social network. In reality, information and rumors can travel within and across different social networks. Thus, it is interesting to investigate whether information diffuses more within one network or across networks. Moreover, what type of information propagates more within a network and what type propagates more across networks?

III. Analyzing Network Dynamics. Dynamics of single-site social networks are well-studied in the literature. These networks are known to have a power-law degree distribution, a small average path length, and being highly clusterable [63]. However, users belong to multiple sites and these network properties need to be generalized to multiple networks. In particular, it is interesting to determine how close the dynamics of single networks are to that of multi-networks. Recent studies have looked at the types of sites that users join [67] and how degree distributions (i.e., the number of friends) and friends that users have vary across sites [68].

Identity Linking Challenges. Although users with accounts on multiple social media sites bring many opportunities, taking advantage of these opportunities is not a trivial task. It's obvious that all the aforementioned opportunities require us to link users' accounts on multiple online social networks. However, the task of linking users accounts on multiple social media sites, also called user identity linkage, is a challenging task because: (1) user identity information can be rather diverse across different online social network sites for the same person in the real world [49] and (2) online social network data is big, noisy, incomplete and highly unstructured [58]. In particular, user identity data in online social networks has the following unique properties:

Profile Inconsistency: Different online social network sites have different structures and schemes to present user profiles. The user profile attributes can reveal user's basic information such as screen name, real name, age, biography, gender, location, education background, contact information, etc. Online social networks may allow users to selectively show profile attributes publicly and keep some sensitive information (e.g. age or contact information) private. In addition, the same attribute can be filled up with different information, e.g. location, depending on the site and user's purpose. Even in a single platform, a user profile may be deliberately counterfeited similarly to impersonate other users [24], which increases the uncertainty and ambiguity of profile features.

Content Heterogeneity: User's generated content can reflect his/her behavior properties as *when, where* and *what* he/she is posting. The content may involve in various medium types such as text, image, video, check-in, etc. The heterogeneous content information makes it extremely difficult to leverage them simultaneously to accurately link user identities [40]. In addition, online social network platforms may deliberately prohibit users to exchange information with others, resulting in the "Data Isolated Islands" phenomenon.

Network Diversity: Online social network structures for a specific user can be rather diverse on different social media platforms. Each social network structure is constructed for the user's specific objective and only reflects a subset of his/her real world social circle. For example, a PhD student looking for jobs has connections with hiring managers on LinkedIn that does not necessarily mean they are friends on Twitter. In practice, we cannot get the complete network structures for all users as well, due to the large scale and privacy issues maintained by online social network companies. This may prevent us from using graph structure patterns to match user entities as traditional entity resolution tasks [15].

As user identity linkage across online social networks is a very important and challenging problem, it has become a trending research area and attracted more and more research attention. The goal of this article is to provide a comprehensive review of recent studies of user identity linkage methods across online social networks and give a guidance on future research directions. The contributions of this paper are summarized as below:

- The user identity linkage problem has various definitions and formulations. [3] We provide a general and formal definition of the user identity linkage that covers most existing definitions;
- Existing approaches share similar characteristics in the problem-solving process that allows us to present a unified framework for user identity linkage task. The unified framework consists of two phases – feature extraction and model construction. We summarize different aspects of existing feature extraction and model construction techniques;
- Empirical evaluation can quantitatively assess and guide different algorithms. We discuss different datasets and evaluation metrics proposed by existing approaches;
- Linking user identities across online social networks is still an active area and there are many research opportunities. We compare the related research areas and discuss some open issues and possible future research directions.

The rest of this article is organized as follows. In Section 2, we introduce notations and formally define user identity linkage. In Section 3, we present the general framework of user identity linkage approaches. Specifically, we review the details of the feature extraction process in Section 3.1 and illustrate various types of model construction mechanisms in Section 3.2. In Section 3.3, we review the state-of-the-art methods for user identity linkage task. We discuss the datasets and evaluation metrics used by existing methods in Section 4. We briefly introduce the areas related to user identity linkage problem in Section 5. Finally, we discuss the open issues and future directions in Section 6 and conclude this article in Section 7.

2. PROBLEM DEFINITION

In this section, we introduce some basic notations and definitions which are used for user identity linkage task. Without loss of generality, we focus on a single real-world natural

[3] This problem is also known as Social Identity Linkage [40], User Identity Linkage [47], User Identity Resolution [5], Social Network Reconciliation [32], User Account Linkage Inference [54], Profile Linkage [70], Anchor Link Prediction [30] and Detecting me edges [11].

person on two online social network sites. Note that the settings of two online social networks and a single natural person can be easily extended to multiple sites and persons. The basic notations are defined as below,

- Let *User Identity* u refer to the unique social account representation on a social media site for a real natural person $\mathcal{P}$. It can consists of three components: *Profile, Content* and *Network*. Profile $\vec{p_u}$ includes a set of user description features such as username, location, age, among other attributes. Content $\vec{c_u}$ consists of a set of attributes that represent the activities that the user is involved in and includes time, location, text, image, etc. Network $\vec{n_u}$ consists of a set of attributes that describe the user's social connections with other users such as the friends in the ego-network.

- An *Online Social Network* $\mathcal{G}$ is represented as a graph $\mathcal{G}(\mathcal{U}, \mathcal{E})$ where $\mathcal{U} = \{u_1, u_2, ..., u_N\}$ is the set of user identities and $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{U}$ is the set of links in the network.

Definition 1 (User Identity Linkage[4]) *Given two online social networks $\mathcal{G}^s$ (source site) and $\mathcal{G}^t$ (target site), the task of user identity linkage is to predict whether a pair of user identities u^s and u^t chosen from $\mathcal{U}^s$ and $\mathcal{U}^t$ respectively belong to a same real natural person, i.e., $\mathcal{F} : \mathcal{U}^s \times \mathcal{U}^t \to \{0, 1\}$ such that,*

$$\mathcal{F}(u^s, u^t) = \begin{cases} 1, & \text{if } u^s \text{ and } u^t \text{ belong to same person,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where $\mathcal{F}$ is the prediction function we want to learn.

3. A GENERAL FRAMEWORK FOR USER IDENTITY LINKAGE

Many user identity linkage methods are proposed and most of the existing methods can be generalized into a unified framework as shown in Figure 1. This framework is composed of two major phases: i) Feature extraction and ii) Model construction. In the feature extraction phase, for a pair of users, features are extracted from users' profile, content and network structures. The extracted features are then used as inputs for the model construction phase, where a supervised, semi-supervised or unsupervised model is trained according to the availability of labeled pairs. Finally the trained models are used to predict whether two user identities match or not. Next, we will give details of feature extraction and model construction phases.

3.1 Feature Extraction

As previously mentioned, an user identity is composed of profile, content and network components. Next, we will introduce the detail of how to extract features from these components and how they can be represented and leveraged for the model construction phase.

[4]Note that even though it is possible that one person can have more than one user account in each online social network site, most previous work, if not all, assume that one person can only have one user account in one site.

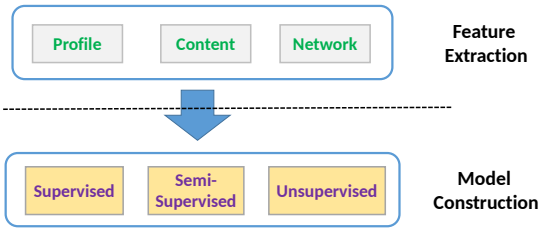

Figure 1: A General Framework for User Identity Linkage.

3.1.1 Profile Features

Profile features $\vec{p_u}$ for a user u are the set of profile fields that describe the user's basic information. Profile features $\vec{p_u}$ can usually be represented by an m dimensional vector i.e., $\vec{p_u} = (p_u^1, p_u^2, ..., p_u^m)$ where each dimension is a profile field. Different online social network sites have different structures and schemes to present user profile features. A list of representative public profile fields are as follows[5],

- **Username**: This refers to the unique identifier that can represent a user in the online social network. A real natural person can choose different usernames on different online social network sites.
- **Screen name**: It is usually formed from the first name and the last name that a user has entered in his profile.
- **Location**: The locations provides information on where the the user lives. Location may come in various forms: detailed addresses, lat-long coordinates, or city names.
- **Biography**: A free-form short text description written by a user as an introduction in online social networks. It often includes the user's occupation, organization, interest, among other attributes.
- **Education**: This refers to the education background for a user and often contains education history such as the names of universities, high schools, middle schools, etc.
- **Avatar**: The thumbnail or image provided by the user to visually present herself.

Other profile fields include gender, age, occupation, email, URL, etc. For two user identities u^s and u^t from source and target online social networks $\mathcal{G}^s$ and $\mathcal{G}^t$, denote their profile features as $\vec{p_{u^s}}$ and $\vec{p_{u^t}}$. Profile features can be utilized in different ways to decide whether u^s and u^t belong to the same person. Existing approaches that use profile features can be categorized into *Distance-based* and *Frequency-based* methods.

Distance-based: The similarity between profile fields of two user identities can be measured by comparing the "distance" between them. For text fields (e.g. username), string similarity schemes such as Jaro-Winkler distance [16], Jaccard similarity, and Levenshtein (Edit) distance are applied [54; 5; 56; 8; 27; 40]. For visual fields (such as the avatar), mean square error, peak signal-to-noise ratio, and Levenshtein distance are utilized [42] to calculate the similarities. After we

[5]Public profile fields are those can be accessed through API without authentication, while private profile fields need authentication.

calculate the distances of each field, we can further compute a weighted similarity score $sim(\vec{p_{u^s}}, \vec{p_{u^t}})$ of u^s and u^t to represent profile features [60].

Frequency-based: Instead of looking at profile attribute values directly and comparing their distances, one can investigate their frequency patterns. For example, profile text fields can be put into a bag-of-word model [47; 46] or the TF-IDF model [76; 52]. Other approaches estimate the uniqueness of profile fields (such as username) via a probabilistic model such as the Markov-chain model [52; 65].

3.1.2 Content Features

Content features $\vec{c_u}$ for user u reveal her activities such as posting, commenting, or replying in online social networks. Each content feature can consists of three types of information: temporal, spatial, and post. In other words, content features are defined as the multi-set of (temporal, spatial, post) bins: $\vec{c_u} = \{(t_1, s_1, p_1), (t_2, s_2, p_2), ...\}$.

- **Temporal**: The temporal information provides the timestamps of user's activities. Temporal information is usually automatically recorded by online social networking sites.
- **Spatial**: Spatial information often comes from the geo-tags attached to user posts, which can be transformed to accurate latitude and longitude values. Without geo-tags, spatial information can also be extracted from the posted texts or images.
- **Post**: Posts come in two forms, namely texts and images. On Twitter, people are more likely to post short texts. While images are preferred by Instagram users.

Content features are often jointly represented to capture special characteristics of user identities. Specifically, existing approaches use content features from the following aspects,

Interest-based: The temporal and post information can collectively reflect the topical interests of user identities. Thus, a long-term topic modeling [50; 40] can be performed to extract the user's core interests.

Style-based: The goal is to use posts to extract the writing style of users. The writing style includes personalized words and emoticons which can help distinguish user identities. Usually an n-gram language model [23] or term-frequency analysis [40; 27; 65; 30] is performed to extract words that distinguish one's identity from all the posts.

Trajectory-based: Trajectory can be extracted from a set of timestamped location data and modeled to capture the unique footprints of users' activities [53].

3.1.3 Network Features

Network features $\vec{n}_u$ for user u refer to the social network interactions with other users in the same online social network. Based on the completeness and connectivity of network topology structures, we can categorize networks into two types: *local network* and *global network*.

- **Local network**: These network features can be built from the *ego-networks* of user identities. The ego-network for each user identity is obtained through the one-hop neighborhoods (e.g. following/followee/friend relationships). In the real world situation, social network API often provides access permission of user's direct friendship if we know the user information. This holds for Facebook API with the appropriate permission set.
- **Global network**: This kind of network often indicates arbitrary merging graph such as a large sample or even complete social networks [5]. In global network, all user identities need to be connected. both immediate (i.e. one-hop) neighborhoods and non-immediate neighborhoods are considered in global networks.

With respect to the two aforementioned different network types, various network features can be constructed. For two user identities u^s and u^t from source and target online social networks $\mathcal{G}^s$ and $\mathcal{G}^t$, their immediate neighbor nodes are denoted as $\Gamma(u^s)$ and $\Gamma(u^t)$.

Neighborhood-based: Based on initial pairs of user identities that match $\mathcal{M}$, neighborhood-based features aim to capture the *match degree* of $\Gamma(u^s)$ and $\Gamma(u^t)$. For example, match degree can be computed using the number of shared identified friends [78; 69; 32], known in/out neighbors and in/out degree [48], and Dice coefficient [5]. Other metrics such as common neighbors, Jaccard's coefficient and Adamic/Adar score are extended to measure the neighborhood similarities as well [76; 30].

Embedding-based: Network embedding techniques can be utilized to learn latent network features that can preserve the original network structure, such as first-order proximity and second-order proximity [57]. In first-order proximity, a pair of nodes u_1 and u_2 in graph $\mathcal{G}$ can be represented as two vectors $\vec{z}_1$ and $\vec{z}_2$ and the probability that an edge is observed is computed by the sigmoid function [43]. In second-order proximity, each node plays two roles, namely the node itself and the "context" of other nodes (such as follower-followee relationship) [39]. Other approaches regard source network $\mathcal{G}^s$ and target network $\mathcal{G}^t$ as an entire network and map it to a hypergraph to learn latent network features [56]. Note that since local network only contains the ego-network structures of user identities, only neighborhood-based network features can be applied. For global network, both neighborhood-based and embedding-based network features can be extracted.

3.1.4 Discussion

We have demonstrated that profile, content and network features can be extracted and represented in different ways. In practical scenarios, these features have their specific characteristics: i) Profile features are relatively easy to obtain since they are usually publicly available; however, different online social network sites may allow users to fill profile fields selectively, which leads to many missing and inconsistent values. Moreover, profile features may be easily impersonated deliberately by other users [24]; ii) Content features can be very sparse for those users who are not active in posting their activities; Thus a continuous process is needed to obtain easy-to-use content features; iii) Network features can also be very noisy because not all edges represent true "friend" relations [40]. In addition, some network features can only be utilized when fully-aligned networks (e.g. global networks) are obtained, which is not practical in real-world scenarios.

3.2 Model Construction

In the previous section, we detail different aspects of feature extraction phase. Here we review the model construction

phase. Following traditional ways of classifying data mining and machine learning models, we summarize existing models into three groups: supervised, semi-supervised and unsupervised models.

3.2.1 Supervised Model

For a typical binary classification problem, there are two types of instances: positive instances (matching user identity pairs) and negative instances (non-matching user identity pairs). Suppose $\mathcal{Q} = \{(u^s, u^t), u^s \in \mathcal{U}^s, u^t \in \mathcal{U}^t\}$ denotes all the possible user identity linking pairs and $\mathcal{M} \subset \mathcal{Q}$ represents the positive instances, where u^s and u^t belong to the same natural person. The set of negative instances $\mathcal{N}$ satisfies $\mathcal{N} = \mathcal{Q} - \mathcal{M}$. The positive and negative instances $(\mathcal{M}, \mathcal{N})$ can be divided into the training set $(\mathcal{M}', \mathcal{N}')$ and the test set $(\mathcal{M}'', \mathcal{N}'')$. The goal of a supervised model is to learn a function $\mathcal{F} : \mathcal{U}^s \times \mathcal{U}^t \to \{0, 1\}$ on training set and then evaluation can be performed on test set. Existing supervised approaches fall into the following categories:

Aggregating methods: Aggregating methods combine the similarity scores of different features into a hybrid weighted form.

$$\mathcal{F}(u^s, u^t) = \alpha S_p(\vec{p_{u^s}}, \vec{p_{u^t}}) + \beta S_c(\vec{c_{u^s}}, \vec{c_{u^t}}) + \gamma S_n(\vec{n_{u^s}}, \vec{n_{u^t}}) \tag{2}$$

where α, β, γ are *weight* parameters and S_p, S_c, S_n are *similarity* functions of profile, content and network features [60; 27; 50]. Note that α, β, γ could be 0 if the algorithm excludes the corresponding features.

Probabilistic methods: A probabilistic classifier aims to predict a probability distribution of class labels. The basic assumption is giving a pair of user identities u^s and u^t, the probability of linkage is conditionally dependent on the feature vector $\vec{x}$ extracted from u^s and u^t, training set $(\mathcal{M}', \mathcal{N}')$, and the specific probability model M.

$$\mathcal{F}(\vec{x}) = \arg\max \Pr(y = 1 | \vec{x}, (\mathcal{M}', \mathcal{N}'), M) \tag{3}$$

where $y = 1$ indicates that u^s and u^t are matched. This is also known as *Maximum a posteriori* (MAP) estimate and can be solved based on Bayes theorem [70; 52; 71].

Boosting methods: Boosting methods build a conjunction of several types of weak hypotheses to learn a strong hypothesis [46].

$$\mathcal{F}(\vec{x}) = \text{sign}(\sum_{i=1}^{T} \alpha_i h_i(\vec{x})) \tag{4}$$

where $\vec{x}$ is the feature vector extracted for each pair of user identities (u^s, u^t), h_i is a weak classifier, α_i is the weight assigned to h_i and T is the count of weak classifiers.

Projection methods: Projection methods aim to learn a projection function Φ to map between original feature space (e.g. profile, content and network features) and a latent feature space of user identities in each online social network [43; 47]. Two projection functions Φ_s and Φ_t for source and target online social networks can be learned through the training process. Giving a user identity u^s from source social network $\mathcal{G}^s$, the target matching user identity $\hat{u^t}$ is chosen by following formula,

$$\hat{u^t} = \arg\min_{u^t \in \mathcal{U}^t} \mathbb{D}(\Phi_s(u^s), \Phi_t(u^t)) \tag{5}$$

where $\mathbb{D}$ is function to measure the distance for u^s and u^t in the latent space. Note that projection functions can be achieved simultaneously for the scenario of multiple social networks [47].

Some other supervised approaches try to find a best classifier from multiple traditional classifiers. Usually several popular classifiers are trained and validated such as Naïve Bayes, Decision tree, Logistic regression, KNN and SVM, then the best classifier is selected finally [65; 23; 51; 42; 25; 77].

3.2.2 Semi-Supervised Model

For semi-supervised approaches, both labeled and unlabeled user identities are taken into account and unknown user identity pairs are predicted during the learning process. A set of *seed* matching user identity pairs $\mathcal{M}'$ is obtained beforehand. We also denote $\mathcal{M}'_s = \{u^s | (u^s, u^t) \in \mathcal{M}'\}$ and $\mathcal{M}'_t = \{u^t | (u^s, u^t) \in \mathcal{M}'\}$ as the corresponding user identities in $\mathcal{G}^s$ and $\mathcal{G}^t$. Usually, unknown user identity matching pairs $\mathcal{M}''$ are discovered by utilizing the *topological structure* of the network and the feedback from seed user identity matching pairs $\mathcal{M}'$. Note that for simplicity, we treat transductive learning as a special type of semi-supervised learning [79]

Propagation methods: Propagation methods discover unknown user identity pairs in an iterative way from seed matching user identity pairs. Let $\Gamma(u^s)$ and $\Gamma(u^t)$ denote the neighborhood of user identity u^s and u^t, respectively, the key idea is to define a function Ψ to compute the *match degree* of user identities using *known neighborhood* (i.e. $\Gamma(u^s) \cap \mathcal{M}'_s$ and $\Gamma(u^t) \cap \mathcal{M}'_t$) information and other features (such as profile or content features) extracted from u^s and u^t. In each iteration, the user identity pairs with the highest match degree score (or with score exceeding a threshold) are selected. The propagation process terminates until no more user identity pairs can be found. Usually, two kinds of propagation order are used: i) *Exhaust comparison*, where each candidate matching pair is selected from the remaining unmatched user identities $\mathcal{S}$ and $\mathcal{T}$ [69; 32; 48; 40]; ii) *Local expansion*, where candidate matching pairs are locally expanded from neighbors of existing matched user identities [78; 8; 76; 77].

Embedding methods: Semi-supervised embedding methods usually learn the latent features of user identities collectively in source and target networks $\mathcal{G}^s$ and $\mathcal{G}^t$. To map user identities from the original feature space to a *common embedding space*, the labeled user identities $\mathcal{M}'_s$ and $\mathcal{M}'_t$ of seed matching pairs $\mathcal{M}'$ are constrained to have the same latent representations. In [56], a hypergraph is built and seed matching pairs $\mathcal{M}'$ are projected to a node to ensure the aforementioned constraint. In [39], the following/followee relations are approximated in the latent space with an explicit constraint to ensure that latent feature vectors for $(u^s, u^t) \in \mathcal{M}'$ are equal (i.e. $\vec{z_{u^s}} = \vec{z_{u^t}}$).

3.2.3 Unsupervised Model

Due to the high cost to obtain labeled matching user identity pairs, unsupervised models are performed with only unlabeled data. Only few unsupervised approaches exist since a small set of seed matching user identity pairs can be acquired from user self-posting websites such as Google+ or About.me[6], or through human annotating by comparing profile, content and network features. Existing unsupervised approaches fall into following categories:

[6] https://about.me/

Aligning methods: Aligning methods usually consist of three steps: i) Compute an *affinity score* for every candidate pair of user identities $(u^s, u^t) \in \mathcal{U}^s \times \mathcal{U}^t$ based on profile, content or network features; ii) Build a bipartite graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}', w)$, where $\mathcal{V}' = \mathcal{U}^s \cup \mathcal{U}^t$, $\mathcal{E}' = \{(u^s, u^t) | u^s \in \mathcal{U}^s, u^t \in \mathcal{U}^t\}$ and the weight $w(u^s, u^t)$ of edge (u^s, u^t) is given by the aforementioned affinity score; iii) Based on the resulting graph $\mathcal{G}'$, an optimized problem is formalized to achieve one-to-one matching for all user identity pairs. For example, an affinity score is computed by modeling content (i.e. time-location trajectories) similarities and a maximum weight matching scheme is performed to link user identities in [53]. In [34], affinity score is computed from profile and network features (i.e. the username overlap of common friends) and then overlaps are maximized to match all user identities.

Progressive methods: Progressive methods try to classify user identity pairs in a progressive way: i) A feature with strong discriminability is used to find *partial* ground truth and ii) all extracted features are considered to perform classification on remaining unlabeled user identities. In [38], n-gram probability is used to automatically acquire training data and then a SVM classifier is utilized to classify remaining unmatched user identity pairs.

3.2.4 *Discussion*

Here, we discuss some key aspects of aforementioned models for the user identity linkage problem. Since social network data is huge and people always involve in multiple online social networks, we need to consider the *scalability* and *multiplicity* of those methods. i) Scalability: For source and target online social networks $\mathcal{G}^s$ and $\mathcal{G}^s$, the complexity for exhausting comparison is $|\mathcal{U}^s| \times |\mathcal{U}^t|$. Similar to traditional entity resolution problems, some *blocking functions* can be pref-defined to reduce the computation cost [22]. One way is to construct blocking keys from profile, content or network features, so that user identities not matching on the key are not compared [28; 11]. The other way is neighborhood based, where only the neighborhood of known matching user identities are compared [78; 8; 76; 77]. ii) Multiplicity: most existing approaches consider pair-wise user identity linkage problem and the case for multiple platforms are can be extended by integrating pair-wise linkage results in a transitive manner. However, it may suffer the problem of inconsistent results when the order of transitivity changes. A possible solution is to learn the projection function Φ from original feature spaces to latent feature spaces for each online social network individually [43; 47] or simultaneously [56; 39].

3.3 A Summary of User Identity Linkage Algorithms

In the previous subsection, we give an overview about the two important components of algorithms for user identity linkage – feature extraction, and model construction. In this subsection, we further give a summary about representative user identity linkage methods in Table 1. A brief introduction about these algorithms is given below:

3.3.1 *Supervised*

• MOBIUS [65]: This paper explores the minimum part of profile features (i.e. username) insightfully by modeling user behaviors from human limitation, exogenous factors and endogenous factors, and the Naïve Bayes classifier is used.

• ULink [47]: This paper uses basic profile features and a projection algorithm is proposed. An online version has been developed to handle dynamic scenario of social network datasets.

• Perito'11 [52]: This paper is the first to only use specific profile features, i.e., username, to perform user identity linkage task. The algorithm is a Markov-Chain based probabilistic model.

• LU-Link [23]: This paper utilizes style-based content features and then a logistic regression classifier is applied to predict matching user identity pairs.

• OPL [70]: This approach focuses on linking user identities in cost-sensitive setting. Extensive profile features are extracted and then a probabilistic classifier is applied.

• DCIM [50]: This approach jointly models neighborhood-based network features and interested-based content features by discovering *core interests* of users, which can capture users' characteristics more accurately. An aggregating method is applied for this method.

• Peled'13 [51]: This approach extracts distance-based profile features and neighborhood-based network features. Many popular classifiers are performed in the experiments such as Adaboost, Random Forest, etc.

• Malhotra'12 [42]: This paper utilizes distance-based profile features (also referred to "User footprint") and then the Naïve Bayes classifier is applied. Comprehensive comparison is performed to analyze feature discriminative capacities and username and display name are found to be the most discriminative features.

• Vosecky'09 [60]: This approach uses distance-based profile features and a supervised aggregating method to link user identities. Weights can be learned adaptively for different profile attributes fields.

• Lofciu'11 [27]: This paper focuses on online social tagging systems. It presents a model that use frequency-based profile (i.e. username) features and style-based content (i.e. tags) features, and then a supervised aggregating algorithm is implemented.

• Motoyama'09 [46]: This is a systematic approach for searching and matching user identities on multiple online social networks. It uses frequency-based profile features and then a boosting algorithm is leveraged to classify user identity matching.

• Goga'13 [25]: This method leverages public distance-based profile features then applies several popular classifiers to decide whether user identity pairs are matched.

• PALE [43]: This supervised framework employs embedding-based network features to map social network structures into low dimension space. Based on latent features of user identities, a projection method is applied.

• MNA [30]: This approach extracts style-based content features and neighborhood-based network features. A supervised aggregating algorithm is built on seed matching user identity pairs and weighted maximum matching scheme is utilized to rank all potential user identities.

3.3.2 *Semi-supervised*

• IONE [39]: This approach extracts embedding-based network features to learn the follower-ship/followee-ship of each user simultaneously. Seed user identity pairs are constrained to transfer the context of social relation network structure. The embedding algorithm is developed to match unknown

Table 1: Comparison of User Identity Linkage Approaches: Algorithms.

Feature \ Model	Supervised	Semi-Supervised	Unsupervised
Profile	[65] [47] [70] [52] [42] [60] [46] [25]		
Content	[23]		[53]
Network	[43]	[78][48] [32] [39]	
Profile, Content	[27]		
Profile, Network	[51]	[76] [54] [5] [8] [56] [77] [11]	[34]
Content, Network	[50] [30]		
Profile, Content, Network		[40]	[38]

user identity pairs.

• COSNET [76]: This paper presents an energy-based model to link user identities by considering both local and global consistency. The local consistency refers to situation of only two social networks, while global consistency is the mapping consistency on multiple networks. COSNET first extracts distance-based profile features and neighborhood-based network features and then use an aggregating algorithm to obtain local consistency.

• HYDRA [40]: This paper proposes a semi-supervised multi-objective framework jointly modeling heterogeneous behaviors and structure consistency. Heterogeneous behaviors including distance-based profile features, style-based content features, trajectory-based content features and neighborhood-based network features are modeled dynamically in a propagation algorithm.

• FRUI [78]: This paper presents a Friend Relationship-Based User identification framework, which first extracts neighborhood-based network features and then develops a semi-supervised propagation algorithm.

• JLA [5]: This approach is based on Conditional Random Fields which jointly models distance-based profile features and neighborhood-based network features. JLA uses propagation method to identify new matching pairs.

• Shen'14 [54]: This approach considers distance-based profile features and neighborhood-based network features, and a propagation method is used to iteratively identify unknown user identity pairs.

• Zhang'16 [77]: This approach leverages distance-based public profile features and neighborhood-based network features to link user identities by a local expansion propagation algorithm.

• User-Matching [32]: This paper proposes an efficient propagation algorithm for online social networks, which uses neighborhood-based network features.

• Bennacer'14 [8]: This paper presents a rule-based propagation algorithm by using network and profile features. It consists two major steps: 1) Selecting candidate user identity pairs based solely on neighborhood-based network features; 2) Determining the exact match by comparing the distance-based profile features.

• MAH [56]: This paper incorporates hypergraph to model the embedding-based network features and proposed a embedding method mapping user identities to lower dimension spaces. Distance-based profile features (e.g. username) are also leveraged to MAH model to achieve better performance.

• NS [48]: This is the first approach to use a graph theoretic model based on neighborhood-based network structure to perform user identity linkage task. NS implements a propagation algorithm.

• DetectMe [11]: This approach combines distance-based profile features and neighborhood-based network features to classify user identity linkages. It first select a candidate list with a threshold to filter similarity score of profile features. Then another threshold is set to filter network structure similarity score.

3.3.3 *Unsupervised*

• Alias-Disamb [38]: This paper focuses on task to decide whether cross-platform user identities with same username belongs to same natural person. First, a feature with strong discriminability is used to find partial ground truth and then all extracted distance-based profile features are considered into a SVM classifier.

• POIS [53]: This approach utilizes trajectory-based content features to link user identities. The model is an aligning algorithm, where affinity score is computed based on time-stamped location data, and then the maximum weighted matching scheme is utilized to find the most likely matching user identities.

• Labitzke'11 [34]: This paper tries to compare neighborhood-based network features (i.e. public available mutual friends) for user identity linkage. The model is a typical aligning algorithm. It first builds a "comparison set" of users' friends and then defines metrics to maximize the overlap and minimize the distance of friends list find the most likely corresponding user identities.

4. EVALUATION

In this section, we discuss how to assess the performance of algorithms for user identity linkage across online social networks. We focus on the datasets and evaluation metrics for this task. It's worth mentioning that there is no "best" user identity linkage method due to the variety of data sources and application domains.

4.1 Datasets

Real data: Since most online social network sites provide Application Program Interface (API) to grant access to their data, there are lots of datasets available to do *single* social network research. However, there are no agreed benchmark datasets for the user identity linkage task across online social networks. On one hand, it's very difficult to obtain the ground truth of known user identity linkage pairs. On the other hand, most of existing work attempt to tackle this problem from different feature spaces and it is hard to obtain a comprehensive dataset with all kinds of features. We list some publicly available datasets as below, and a more

Table 2: Comparison of User Identity Linkage Approaches: Datasets.

Feature \ Availability	Public	Request	Not public
Profile		[24]	[65] [47] [70] [52] [42] [60] [46] [25]
Content		[23]	[53] [23]
Network			[78][48] [32] [39]
Profile, Content			[27]
Profile, Network	[11] [76]		[51] [54] [5] [8] [56] [77] [34]
Content, Network			[50] [30]
Profile, Content, Network			[40] [38]

comprehensive data set comparison for existing methods is described in Table 2.

- *Goga1315* [7]: This data resource consists of two datasets collected based on google+ sitemap data. The first dataset is used in [23] and contains the data of three social networks, i.e., Twitter, Flickr and Yelp. Both profile and content features are provided in this dataset. The second dataset is used in [24] and public profile features are provided from Twitter, Facebook, LinkedIn and Flickr.
- *Buccafurri12* [8]: This dataset was originally collected in 2012 by the authors in [11] and it was further enriched in 2014 [8]. It contains the data of four online social networks, i.e., LiveJournal, Flickr, Twitter and Youtube. The dataset is composed of 93,169 user identities, 145,580 friendship links and 462 matching user identity links. Profile features such as username and network features such as friendship are explicitly provided in this dataset.
- *Zhang15* [9]: This dataset was collected during 2015 with two kinds of online social networks, namely social network sites and academia network sites [76]. Social network sites include Twitter, LiveJournal, Flickr, Last.fm, MySpace and academia networking sites are ArnetMiner, VideoLecture and LinkedIn. For each online social network, the username and friendship network are provided for user identities. The ground truth of matching user identities are obtained from [52] and the crowdsourcing service on ArnetMiner.

Synthetic data: Some approaches only utilize network features to link user identities. The performance can be evaluated on simulated synthetic networks. According to whether a real social network exists or not, synthetic data sets have two types of sampling strategies:

- *Full synthetic*: Social networks are sampled only based on existing graph generating algorithms and no real social network is provided. In [78; 32], different types of synthetic networks are built such as Erdős-Réyi random graph [20], Watts-Strogatz small-world graph [61], Affiliation Network [35], Barabási Albert preferential attachment model [4] and RMAT [14].
- *Partial synthetic*: A real social network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is given, and two sub-networks can be constructed based on *node sampling* or *edge sampling* strategies. Node sampling aims to extract two subset of nodes $\mathcal{V}^s$ and $\mathcal{V}^t$ from $\mathcal{V}$ such that $\mathcal{V}^s \bigcup \mathcal{V}^t = \mathcal{V}$ and $\mathcal{V}^s \bigcap \mathcal{V}^t \neq \phi$ [48]. Three common ways for edge sampling [69] are i) Randomly adding edges with a predefined probability; ii) Randomly removing edges with a predefined probability; 3) randomly rewiring edges with a pre-defined probability.

In addition, we summarize the datasets used by representative methods in Table 2 from the feature and availability perspectives. For availability, i) Public means that the dataset can be directly downloaded from the website; ii) Request means that the dataset is provided upon request from the authors; iii) Not public means no explicit sources are provided to acquire dataset.

4.2 Evaluation Metrics

To evaluate the performance of algorithms for the user identity linkage problem, different metrics are proposed. Here, we review some widely used metrics such as prediction metrics and ranking metrics.

Prediction metrics: Most previous approaches consider user identity linkage problem as a binary classification task that, given two user identities u^s and u^t from source and target online social networks $\mathcal{G}^s$ and $\mathcal{G}^t$, determine whether u^s and u^t are matching or not:

- True Positive (**TP**): when predicted matched user identities belong to same natural person;
- True Negative (**TN**): when predicted unmatched user identities belong to different natural persons;
- False Negative (**FN**): when predicted unmatched user identities belong to same natural person;
- False Positive (**FP**): when predicted matched user identities belong to different natural persons.

Based on aforementioned possible classification results, we can define following metrics,

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{6}$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{7}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \tag{9}$$

[7] http://www.mpi-sws.org/~ogoga/data.html
[8] http://www.ursino.unirc.it/pkdd-12.html
[9] http://aminer.org/cosnet

The metrics above are based on pair-wise matching results. In [70], a set-wise metric called *Identity-based Accuracy* (I-Acc) is defined as follows:

$$I\text{-}Acc = \frac{\#\text{ correctly identified user identities}}{\#\text{ground truth user identities}} \tag{10}$$

Note that for $Precision, Recall, F1$ and $I-Acc$, the higher the value, the better the performance.

Ranking metrics: Some approaches may provide a top-k ranking list of potential matching user identities rather than only one. Giving a user identity u^s from source social network $\mathcal{G}^s$, all K candidate identities in target social network $\mathcal{G}^t$ are ranked based on the matching degree with u^s, i.e. $\mathcal{R}_{u^s} =< u_1^t, u_2^t, ..., u_K^t >$. The goal is to rank true matching user identities as top as possible.

The *Receiver Operating Characteristics* (ROC) curve can be drawn by plotting *False Positive Rate* (FPR) and *True Positive Rate* (TPR) as x and y axes respectively. It can compare the performance of different classifiers by changing class distributions via a threshold. TPR and FPR are defined as follows,

$$TPR = \frac{|TP|}{|TP| + |FN|} \tag{11}$$

$$FPR = \frac{|FP|}{|FP| + |TN|} \tag{12}$$

Based on ROC, we can compute the Area Under ROC curve (AUC) value which can measure the overall performance of how well the classifier can rank the positive linking user identity higher than any negative user identity. Based on [26], AUC is defined as below,

$$AUC = \frac{\sum(n_0 + n_1 + 1 - r_i) - n_0(n_0+1)/2}{n_0 n_1} \tag{13}$$

where r_i is the rank of i_{th} positive matching user identities and n_0 (n_1) is the number of positive (negative) user identities.

By assuming only one user identity in $\mathcal{R}_{u^s}$ that can match u^s (i.e. $n_0 = 1$), there is only one positive matching denoted as r_1 and AUC can be computed by,

$$AUC = \frac{n_1 + 1 - r_1}{n_1} \tag{14}$$

A similar measure called $Hit-Precision$ [47] is defined as follows,

$$Hit\text{-}Pricision = \frac{n_1 + 2 - r_1}{n_1 + 1} \tag{15}$$

Other metrics are proposed such as *Mean Reciprocal Rank* (MRR) [27] and *Mean Average Precision* (MAP) [43] which are calculated by the average performance of Reciprocal Rank (RR) and Average Precision (AP) over all user identities that need to be classified. Under the assumption that $n_0 = 1$, Reciprocal Rank and Average Precision are,

$$AP = RR = \frac{1}{r_1} \tag{16}$$

In [27], $Success@k$ measures whether the positive matching user identity will occur in top-k ($k <= K$) list or not. Note that for $AUC, Hit-Precision, MAP, MRR$ and $Success@k$, a higher value indicates better performance.

5. RELATED AREAS

In this section, we discuss areas related to the problem of user identity linkage across online social networks. We introduce these areas by briefly explaining the task goals, highlighting some popular methods and pointing out the differences from the user identity linkage problem.

5.1 Record linkage

Record linkage (or entity resolution) refers to the process of finding related entries in one or more related relations in a database and creating links among them [10]. This problem has been extensively studied in the database area and applied to data warehousing and business intelligence. Based on this survey [31], existing methods exploit features in three ways, namely numerical, rule-based and workflow-based. Numerical approaches combine the similarity score of each feature into a weighted sum to decide linkage [21]; Rule-based approaches derive match decision through a logical combination of testing separate rules of each feature with a threshold; Workflow-based methods apply a sequence of feature comparison in an iterative way. Both supervised such as TAILOR [19] and MARLIN [9], and unsupervised approaches such as MOMA [59] and SERF [7] are studied in the literature. Note that user identity linkage differs from the record linkage problem due to the specialty of online social network scenarios.

5.2 Network alignment

The network alignment task is to find a common subgraph across multiple input networks and can be categorized into local network alignment and global network alignment problems [55]. Local network alignment tries to multiple unrelated regions of isomorphism, while global network alignment maintains a consistent overall alignment for all nodes among networks. This problem has been widely applied in many application areas such as database matching [45], bioinformatics [29], computer vision [17], etc. Representative algorithms include IsoRank [55], NetAlign [6], etc. Recent approaches study network alignment problems under online social network scenarios with [75] or without attribute information [74]. The problem settings of network alignment and user identity linkage are very similar when network features are considered. However, they are different because: i) User identity linkage has its specialty which can be performed without network features ii) Network alignment aims to find a partial or overall alignment of subgraphs while user identity linkage focuses on node alignment.

5.3 De-anonymizing social networks

Social network anonymization refers to the process to replace each user identity's unique identifier (e.g. username) with a random string, but the network structure remains revealed [3]. From the attacker's perspective, both active and passive attacks can be performed on a single online social network with limited information to de-anonymize user identities [3]. By pointing out several drawbacks of active attacks, [48] proposes a large-scale passive social network de-anonymization method. Specifically, it utilizes known anchor links from the source social network as auxiliary information to de-anonymize user identities in the target social network. In this sense, social network de-anonymization problem are actually user identity linkage problem when only network structure information is leveraged.

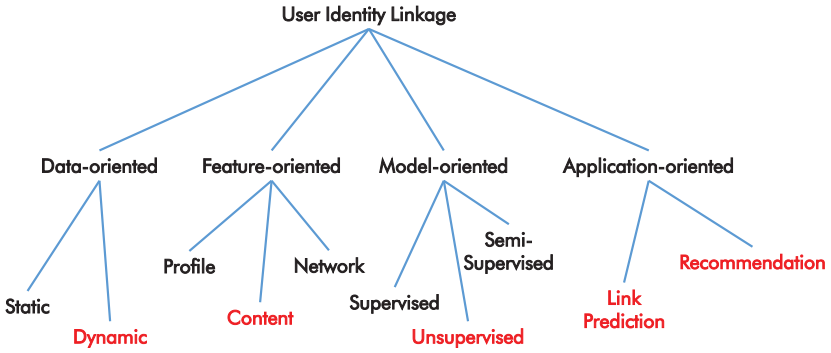

Figure 2: An overview of research aspects for User Identity Linkage. Those areas highlighted in red have not been extensively studied.

5.4 Link prediction

In the context of online social networks, traditional link prediction aims to predict missing links or future links between two user identities in a single social network [2]. Both supervised [1] and unsupervised [37] approaches are proposed to solve the link prediction task. User identity linkage can also be treated as link prediction problem. The difference is that for user identity linkage, we predict "link" between user identities of the same natural person on multiple social network media sites, while for link prediction, we usually predict links between two different users/objects on single homogeneous or heterogeneous network.

6. OPEN ISSUES AND FUTURE RESEARCH DIRECTIONS

In this section, we propose some remaining open issues for the user identity linkage problem and future research directions along this line. We will introduce the dataset challenge and evaluation challenge for the user identity linkage problem. Then, we will discuss some future research directions which have potential to attract more and more attention. As shown in Figure 2, we categorize the research aspects into four parts: data, feature, model, and application. We have already given detailed descriptions on feature and model in the previous sections, thus we focus on data-oriented and application-oriented research directions.

Data challenge: The user identity linkage problem is becoming popular in recent years and more and more methods are developed. However, as mentioned in Section 4.1, there is no agreed benchmark dataset to evaluate and compare existing methods. Existing publicly available datasets may contain *partial* features (such as network structures and user names), but dataset equipped with *comprehensive* profile, content and network features is limited. To obtain a comprehensive dataset for research purpose, we face following challenges: i) *User privacy*, how to access and use user identity features without invasion of user privacy? ii) *Ground truth*, how to obtain matching user identity pairs across online social networks when some social network sites may intentionally prevent user sharing contents; iii) *Limited access*, some online social network sites provide API to access their data for proper use, but they often set rate limits and restricted permission which make it hard to acquire data in large scale.

Evaluation challenge: In practical situation, we cannot get the entire social network datasets to perform the user identity task. Note that the matching and non-matching user identity pairs are very *imbalanced*, which may affect performance evaluation significantly [24]. Goga *et al.* also recommend precision and recall to be more reliable evaluation metrics [24]. In addition, different problem settings (such as exact matching and top-k matching) may cause the demand for choosing different evaluation metrics.

Dynamic user identity linkage: Social networks are dynamically changing over time. Profile, content and network features for user identities keep changing or being accumulated as time goes by. Thus, a more practical solution is to build online user identity linkage methods by extracting features dynamically. In [47], an online learning algorithm is developed to take advantage of incremental data to efficiently improve linkage performance. Nie *et al.* [50] shows that by modeling the "core interest" of user identities with accumulated content features, user identity linkage performance can be significantly improved. Along this line, we can consider to use other types of feature (e.g. network structure) properties dynamically. Moreover, existing approaches for dynamic (attribute) network analysis can also be extended to the multiple network scenario to perform dynamic user identity linkage task.

Jointly user identity linkage and recommendation: Cross-domain recommendations have attracted much attention from researchers recently. It aims to jointly leverage knowledge from source and target domains to build recommendation systems. Existing approaches focus on exploiting cross domain knowledge in following ways: linking, aggregating, sharing and transferring [12]. *Domain* can range in different levels such as attribute, type, item and system level. Note that social networks can be treated as a system level domain. For aggregating and sharing knowledge approaches, user and/or item *overlap* is needed among different social networks. Cross-domain recommendations can benefit from linking user identities from following aspects. First, when user overlap information is obtained via the user identity linkage task, user profiles can be enriched and social relations can be transferred to boost recommendation performance such as video [18], friend [66; 62] and product recommendations [41]. Second, user identity matching and recommendation task can be modeled jointly into a matrix factorization model [36]. For example, Li *et al.* provided a new viewpoint via collaborative filtering for user identity

linkage as well as item recommendations.

Jointly user identity linkage and link prediction: Link prediction problem has been proven to be an important research topic for decades. The major task of link prediction is to predict missing or future formed links in different social networks (homogeneous or heterogeneous social networks). In recent years, link prediction on *aligned networks*, where social networks share common users, are becoming popular [30]. User identity links (also refer to anchor links) can play very important roles in link prediction across aligned networks. Zhang *et al.* formalized the problem of *collective link prediction*, which jointly predicts anchor and social links together across heterogeneous online social networks [73; 64]. They demonstrated that link prediction can be performed simultaneously in multiple networks through meta-path based methods. Meanwhile, in [72] multiple anchor link (such as user and location anchor links) prediction has been explored with an unsupervised model recently.

7. CONCLUSION

Nowadays, people tend to join multiple online social networks for different purposes. Linking user identities across online social networks is of great value in many application areas such as recommendations, link prediction, etc. In this paper, we introduce a unified framework for the user identity linkage problem, which consists of two phases: i) Feature extraction and ii) Model construction. In detail, features can be obtained from profile, content and network information; while models can be conducted in supervised, semi-supervised and unsupervised ways. We further highlight some state-of-the-art approaches and introduce representative datasets and metrics . Moreover, we discuss data and evaluation challenges for this task. For future directions, many tasks in single social network can be properly adjusted and applied in cross network scenarios, and more practical problem settings for user identity linkage can be further explored.

8. REFERENCES

[1] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.

[2] Mohammad Al Hasan and Mohammed J Zaki. A survey of link prediction in social networks. In *Social network data analytics*. 2011.

[3] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW*, 2007.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 1999.

[5] Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, and Hyungdong Lee. Joint link-attribute user identity resolution in online social networks. In *ACM (SNA-KDD)*, 2012.

[6] Mohsen Bayati, Margot Gerritsen, David F Gleich, Amin Saberi, and Ying Wang. Algorithms for large, sparse network alignment problems. In *ICDM*, 2009.

[7] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: a generic approach to entity resolution. *VLDB*, 2009.

[8] Nacéra Bennacer, Coriane Nana Jipmo, Antonio Penta, and Gianluca Quercini. Matching user profiles across social networks. In *International Conference on Advanced Information Systems Engineering*. Springer, 2014.

[9] Mikhail Bilenko and Raymond J Mooney. Adaptive duplicate detection using learnable string similarity measures. In *KDD*, 2003.

[10] David Guy Brizan and Abdullah Uz Tansel. A. survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, 2015.

[11] Francesco Buccafurri, Gianluca Lax, Antonino Nocera, and Domenico Ursino. Discovering links among social networks. In *ECML/PKDD*, 2012.

[12] Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berkovsky, and Paolo Cremonesi. Cross-domain recommender systems. In *Recommender Systems Handbook*. 2015.

[13] Francesca Carmagnola and Federica Cena. User identification for cross-system personalisation. *Information Sciences*, 2009.

[14] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. R-mat: A recursive model for graph mining. In *SDM*, 2004.

[15] Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[16] William Cohen, Pradeep Ravikumar, and Stephen Fienberg. A comparison of string metrics for matching names and records. 2003.

[17] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 2004.

[18] Zhengyu Deng, Jitao Sang, and Changsheng Xu. Personalized video recommendation based on cross-platform user modeling. In *ICME*, 2013.

[19] Mohamed G Elfeky, Vassilios S Verykios, and Ahmed K Elmagarmid. Tailor: A record linkage toolbox. In *ICDE*, 2002.

[20] P ERDdS and A R&WI. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[21] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 1969.

[22] Lise Getoor and Ashwin Machanavajjhala. Entity resolution: theory, practice & open challenges. *VLDB*, 2012.

[23] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting innocuous activity for correlating users across sites. In *WWW*, 2013.

[24] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P Gummadi. On the reliability of profile matching across large online social networks. In *KDD*, 2015.

[25] Oana Goga, Daniele Perito, Howard Lei, Renata Teixeira, and Robin Sommer. Large-scale correlation of accounts across social networks. 2013.

[26] David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 2001.

[27] Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. Identifying users across social tagging systems. In *ICWSM*, 2011.

[28] Paridhi Jain and Ponnurangam Kumaraguru. Finding nemo: searching and resolving identities of users across online social networks. *arXiv preprint arXiv:1212.6147*, 2012.

[29] Gunnar W Klau. A new graph-based method for pairwise global network alignment. *BMC bioinformatics*, 2009.

[30] Xiangnan Kong, Jiawei Zhang, and Philip S Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, 2013.

[31] Hanna Köpcke and Erhard Rahm. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 2010.

[32] Nitish Korula and Silvio Lattanzi. An efficient reconciliation algorithm for social networks. *VLDB*, 2014.

[33] Shamanth Kumar, Reza Zafarani, and Huan Liu. Understanding user migration patterns in social media. In *AAAI*, 2011.

[34] Sebastian Labitzke, Irina Taranu, and Hannes Hartenstein. What your friends tell others about you: Low cost linkability of social network profiles. 2011.

[35] Silvio Lattanzi and D Sivakumar. Affiliation networks. In *STOC*, 2009.

[36] Chung-Yi Li and Shou-De Lin. Matching users and items across domains to improve the recommendation quality. In *KDD*, 2014.

[37] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 2007.

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

[39] Li Liu, Cheung K. William, Xin Li, and Lejian Liao. Aligning users across social networks using network embedding. In *IJCAI*, 2016.

[40] Siyuan Liu, Shuhui Wang, Feida Zhu, Jinbo Zhang, and Ramayya Krishnan. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *SIGMOD*, 2014.

[41] Chun-Ta Lu, Sihong Xie, Weixiang Shao, Lifang He, and Philip S Yu. Item recommendation for emerging online businesses. 2016.

[42] Anshu Malhotra, Luam Totti, Wagner Meira Jr, Ponnurangam Kumaraguru, and Virgilio Almeida. Studying user footprints in different online social networks. In *ASONAM*, 2012.

[43] Tong Man, Huawei Shen, Shenghua Liu, Xiaolong Jin, and Xueqi Cheng. Predict anchor links across social networks via an embedding approach. In *IJCAI*, 2016.

[44] Lydia Manikonda, Venkata Vamsikrishna Meduri, and Subbarao Kambhampati. Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media. *arXiv preprint arXiv:1603.02718*, 2016.

[45] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *ICDE*, 2002.

[46] Marti Motoyama and George Varghese. I seek you: searching and matching individuals in social networks. In *Proceedings of the eleventh international workshop on Web information and data management*, 2009.

[47] Xin Mu, Feida Zhu, Zhi-Hua Zhou, Ee-Peng Lim, Jing Xiao, and Jianzong Wang. User identity linkage by latent user space modeling. In *KDD*, 2016.

[48] Arvind Narayanan and Vitaly Shmatikov. Deanonymizing social networks. In *ISSP*, 2009.

[49] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of personally identifiable information. *Communications of the ACM*, 2010.

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

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

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

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

[54] Yilin Shen and Hongxia Jin. Controllable information sharing for user accounts linkage across multiple online social networks. In *CIKM*, 2014.

[55] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 2008.

[56] Shulong Tan, Ziyu Guan, Deng Cai, Xuzhen Qin, Jiajun Bu, and Chun Chen. Mapping users across networks by manifold alignment on hypergraph. In *AAAI*, 2014.

[57] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *WWW*, 2015.

[58] Jiliang Tang, Yi Chang, and Huan Liu. Mining social

media with social theories: a survey. *ACM SIGKDD Explorations Newsletter*, 2014.

[59] Andreas Thor and Erhard Rahm. Moma-a mapping-based object matching system. In *CIDR*, 2007.

[60] Jan Vosecky, Dan Hong, and Vincent Y Shen. User identification across multiple social networks. In *2009 First International Conference on Networked Digital Technologies*. IEEE, 2009.

[61] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 1998.

[62] Ming Yan, Jitao Sang, Tao Mei, and Changsheng Xu. Friend transfer: cold-start friend recommendation with cross-platform transfer learning of social knowledge. In *ICME*, 2013.

[63] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[64] Reza Zafarani and Huan Liu. Connecting corresponding identities across communities. *ICWSM*, 2009.

[65] Reza Zafarani and Huan Liu. Connecting users across social media sites: a behavioral-modeling approach. In *KDD*, 2013.

[66] Reza Zafarani and Huan Liu. Finding friends on a new site using minimum information. In *SDM*, 2014.

[67] Reza Zafarani and Huan Liu. Users joining multiple sites: Distributions and patterns. In *ICWSM*. Citeseer, 2014.

[68] Reza Zafarani and Huan Liu. Users joining multiple sites: Friendship and popularity variations across sites. *Information Fusion*, 28:83–89, 2016.

[69] Reza Zafarani, Lei Tang, and Huan Liu. User identification across social media. *TKDD*, 2015.

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

[71] Haochen Zhang, Minyen Kan, Yiqun Liu, and Shaoping Ma. Online social network profile linkage based on cost-sensitive feature acquisition. In *Chinese National Conference on Social Media Processing*, 2014.

[72] Jiawei Zhang and Philip S Yu. Pct: partial co-alignment of social networks. In *WWW*, 2016.

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

[74] Jiawei Zhang and Philip Yu S. Multiple anonymized social networks alignment. In *ICDM*, 2015.

[75] Si Zhang and Hanghang Tong. Final: Fast attributed network alignment. In *KDD*. ACM, 2016.

[76] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S Yu. Cosnet: connecting heterogeneous social networks with local and global consistency. In *KDD*, 2015.

[77] Yuxiang Zhang, Lulu Wang, Xiaoli Li, and Chunjing Xiao. Social identity link across incomplete social information sources using anchor link expansion. In *PAKDD*, 2016.

[78] Xiaoping Zhou, Xun Liang, Haiyan Zhang, and Yuefeng Ma. Cross-platform identification of anonymous identical users in multiple social media networks. *TKDE*, 2016.

[79] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 2009.