A Systematic Review on User Identity Linkage across Online Social Networks

Submitted By
Rishabh Kaushal
PhD15008

For Comprehensive Examination
February, 2020

Prof. Ponnurangam Kumaraguru
Advisor

Contents

PhD Course Work Completed

First Year

CSE 648, Privacy and Security in Online Social Media, grade B

CSE 508, Information Retrieval, grade A-

Second Year

CSE 545, Foundations of Computer Security, grade B

CSE 651, Topics in Adaptive Cyber Security, grade A-

A Systematic Review on User Identity Linkage across Online Social Networks

Rishabh Kaushal

April 6, 2020

Abstract

Online Social Networks (OSNs) present a wide variety of information to their users in terms of different types of content (text, video, pictures) and different kinds of network (friends). To avail this diverse information, users register and maintain their accounts (hereafter referred to as user identities) across multiple OSNs. This situation leads to the problem of User Identity Linkage (UIL) across multiple OSNs. More formally, given a user identity on one OSN (referred to as source network), the goal is to find user identity on another OSN (referred to as target network). In this report, we present a systematic review of issues related to the UIL problem from different viewpoints. Collecting ground truth user identities across multiple OSNs that belong to the same person is the first step in the study of the UIL problem. We refer to the collected identities belonging to the same person as *linked user identities*. We perform a detailed study and comparative evaluation of prior data collection methods that collect such identities by leveraging user behaviors. Once we collect linked user identities, typically the next step is to formulate the UIL problem as a machine learning driven classification task. Prior works compute hand-crafted features derived from user behaviors based on settings on user profile, content posted by users and friend network maintained by users. Subsequently, they build supervised, semi-supervised and unsupervised machine learning models on these features. Recent trend to solve the UIL problem is to leverage graph representation techniques and construct embedding vectors corresponding to user nodes in the social network graph, thereby automating the feature computation task. In this work, we perform a detailed study of both conventional machine learning based approaches and more recent graph representation based approaches. Linking users across OSNs have implications on many other problems in social networks. For instance, given that the UIL problem helps in building a comprehensive user behavior expressed across OSNs, therefore, it naturally helps in recommendation systems, targeted advertisements, link prediction across OSNs, and many other applications. From a user's privacy perspective, the linkage of user identities would impact particularly those users who segregate their personal and professional activities across different OSNs. Therefore, in the last part of this report, we present a discussion on implications, and applications that benefit from the solution to the UIL problem.

1 Introduction

Online Social Networks (OSNs) have become a popular medium for socializing in the online world. Users post, share, and view content on OSNs. Novel ways are being devised by OSNs to attract users to use their platforms. With over 2.2 billion monthly active users, Facebook [53] is one of the most popular platforms. Twitter, with 330 million registered users [28], is a fast-paced, concise, and easy way to connect with your audience. LinkedIn focuses more specifically on business and professional communities [50], with 660 million user base. YouTube is the leading video-sharing platform with 2 billion monthly active users [25] viewing and/or sharing video content. In terms of content, some OSN platforms offer video (like YouTube and Vimeo), some offer image (like Instagram and Flickr) and others offer a combination of text with image & video (Facebook and Twitter). Users view and engage with the content of their friends. In terms of friend connections, some OSN platforms offer professional network (like LinkedIn) while others offer friends in general (like Facebook). Owing to privacy concerns, some social networks (like Whisper and Reddit) allow users, by design, to post messages anonymously. Some ephemeral social networks (Snapchat) keep user content temporarily for some time and then remove it.

Given that many OSNs are offering different services, it is natural for users to create accounts (referred to as *user identities*) on more than one OSN platform. As per Pew Research Center [58], more than half of online users (56%) use more than one OSM platform, a trend that has been consistent in the past few years. Furthermore, among these users who use more than one OSM platform [10], the average number of social media accounts which each such user maintains have increased from 4.3 to 7.6 from the year 2013 to 2017. As users join multiple OSN platforms as depicted in Fig 1, the problem of *User Identity Linkage* (UIL) becomes of significant interest.

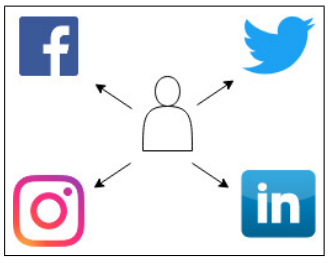

Figure 1: A typical scenario where the same user has accounts (referred to as user identities) across many social networks.

We define UIL as a problem of finding user identity on target OSN when that user's identity is known on source OSN, as depicted in Fig 2. We refer to the user identities on different OSN platforms belonging to the same person as *linked user identities*.

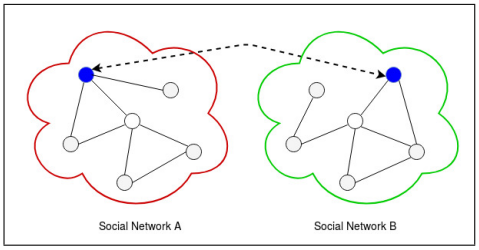

Figure 2: Two social networks A and B are given along with users (represented by circles) in each of them. The goal is to link (represented by dotted lines) user identities belonging to the same person across the two social networks.

Next, we discuss the motivation, challenges, and methodology adopted to survey the literature related to the UIL problem.

1.1 Motivation for User Identity Linkage

Linking user identities on many OSNs is significant for several reasons. *Firstly*, it provides a more comprehensive description of a user by aggregating user information on different OSNs, in terms of profile attributes, content posted or engaged, and network (friends) maintained. This comprehensive view of user facilitates a better understanding of users' interests, thereby enabling better recommendations. *Secondly*, it helps in predicting user behaviors, network dynamics and information diffusion on a relatively newer OSN platform based on user behaviors in well established existing OSNs, an issue commonly referred to as cold-start problem. *Thirdly*, user migration from one OSN platform to another OSN can be studied, and reasons for migration studied.

1.2 Challenges in Linking User Identities

However, there are several challenges in user identity linkage. *First*, given the diversity of content offered by different OSNs, the content generated (*what*) and the content generation patterns in terms of time (*when*) and location (*where*) vary a lot from one OSN to another. *Second*, since the network maintained by users often also varies from one OSN to another, so the friend network of users is quite diverse on multiple OSNs. *Third*, the amount of profile information made available by user varies from person to person and

from platform to platform. Some users are open to disclosing most of their profile attributes whiles others would be skeptical. On the other hand, some OSNs do not allow too many profile attributes to be configured than other OSN platforms.

1.3 Methodology of Survey

This problem is known in the literature by multiple names such as Social Identity Linkage [39], User Identity Resolution [3], Social Network Reconciliation [31], User Account Linkage Inference [59], Profile Linkage [78], Anchor Link prediction [30], and Detecting *me* edges [5]. We collect all prior works by searching these names as the search keys, which are indexed on Google Scholar for the past ten years. After examining those prior works, we present a systematic review of the problem user identity linkage from different perspectives as below.

- Problem formulation: We find that there are subtle variations in which the UIL problem has been formulated. Predominant formulation of the UIL problem in prior works [51, 76, 19, 37, 7, 59] is to decide whether the two given user identities on two different OSNs belong to the same person or not. However, other variations exist [24, 57] where the goal is to find top-k most likely matching identities in the target network corresponding to the given identity in the source network. Alternatively, given a large collection of user identities on two social networks with few known linked identities (also referred to as anchor links) and the goal is to find more linked identities.

- Data collection methods across OSNs: In the UIL problem, the primary challenge is to collect ground truth user identities across multiple OSNs belonging to the same individual, referred to as *linked user identities*. We study the various user behaviors namely social aggregation [51, 37, 19, 80, 78], self-disclosure [76, 34, 7, 59, 84, 51, 87, 30, 82, 57], friend-finder [18], and snowball sampling [3, 40] that are leveraged in past research in order to obtain ground truth linked user identity pairs. We also highlight some of the most commonly studied social networks in the context of the UIL problem, namely Twitter, Facebook, Instagram, and FourSquare.

- Proposed approach adopted: Conventionally, prior works address the UIL problem by looking at it as a machine learning problem and then developing supervised, semi-supervised,

and unsupervised machine learning models. We find that past works propose novel ways to hand-craft features derived from profile [51, 76, 34, 19, 37], network (friends) [87, 86, 40] and content [18, 7, 1] posted by users across OSNs. However, with the recent advancements in graph representation learning, we also find works [63, 52, 68, 20, 38, 42, 62] that automatically learn features as embedding vectors without the need to hand-craft the features. We perform a detailed study of both conventional approaches, and recent graph representation approaches to solve the UIL problem.

- Future Directions: In the last part, we discuss various classical problems in the area of social networks that would benefit from the solution of UIL problem. Problems of recommendation [49, 48, 45], link prediction [79, 82, 54], and many more can be more effectively solved once a comprehensive user behavior is obtained through the user's linked identities. We also discuss privacy implications [13, 67, 15] owing to the linkage of user identities and biases in identity linkage datasets.

2 Problem Formulations and Evaluation

In this section, we present two key formulations of the UIL problem and their evaluation procedures.

2.1 Identity Linkage

The most commonly explored problem formulation in prior works [51, 76, 19, 37, 7, 59] is to learn an identity linkage function that *predicts* or *classifies* whether two given user identities belong to the same individual or not. In this formulation, we model the function as a conventional machine learning-based binary classifier, which takes features related to user identities as input. We derive these features from user profile attributes, user content posting (and engagement with content), and network (friends) maintained by the user. More formally, we define the problem as follows.

Definition 2.1 *Given two user identities I_a and I_b on OSNs a and b, respectively, the goal is to learn a function F, which predicts whether I_a and I_b belong to the same individual or not.*

$$F(I_a, I_b) = \begin{cases} 1, & \text{if } I_a \textit{ and } I_b \textit{ belong to the same user.} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

We learn the function in two ways. The *first* approach is to create handcrafted features derived from the user's profile, content, and network. These features are then fed as input to the machine learning algorithms, as shall be explained later in Section 4. The *second* approach is to learn user identity representation, as discussed in Section 5, in the form of an embedding vector and then apply machine learning algorithms on the learned embeddings. Given that we cast the problem as a binary classification problem, the standard evaluation metrics namely Precision (P), Recall (R), F1-score, True Positive Rate (TPRs), and False Positive Rate (FPRs) are employed. In the context of user identity linkage, we follow the evaluation approach as below.

1. We consider all possible identity pairs $< I_a, I_b >$ comprising of identities belonging to two social networks a and b as part of the input dataset D.

2. Each identity pair $< I_a, I_b >$ has a *label* associated with it, whose value is binary, either 1 or 0, indicating whether two identities I_a and I_b on OSNs a and b, belong to the same or different individuals, respectively.

3. We split the dataset D into training and test datasets. We use the label as supervisory information for learning of the function F. Evaluation is done based on standard metrics, as discussed in Table 1.

Evaluation Metric	**Interpretation in context of UIL problem**
True Positive (TP)	User identities I_a and I_b belong to the same person and the learned function F also predicts the same person.
True Negative (TN)	User identities I_a and I_b do not belong to the same person and the learned function F also says they do not belong to the same person.
False positive (FP)	User identities I_a and I_b do not belong to the same person but the learned function F says they belong to the same person.
False negative (FN)	User identities I_a and I_b belong to the same person but the learned function F says they do not belong to the same person.

Table 1: Explanation of evaluation metrics in the context of the UIL problem.

4. Consequently, we redefine the standard classification metrics as below.

- Precision (P): It is defined as the proportion of times the learned function F correctly predicts the input user identity pairs I_a and I_b to belong to the same individual.
- Recall (R): It is defined as the proportion of user identity pairs I_a and I_b to belong to the same individual that the learned function F can retrieve out of total identity pairs belonging to the same person.

2.2 Linked Identity Extractor

The other way of problem formulation is to learn a *ranking function* which given a single user identity on one social network (source), *orders* the identities on another social network (target) such that correct linked identity appears among the *top-k* identities extracted from the target network. In this formulation, prior works [24, 57] model the ranking function as a conventional ranked retrieval problem from the field of information retrieval (or extraction). Like, the binary classifier function, we compute this ranked retrieval function using the features derived from profile, network and content of user identities, details are presented in Section 4. More formally, we define the problem as follows.

Definition 2.2 *Given a user identity I_a on source OSN_a , the goal is to learn a function F_{rank} that finds top-k user identities $< I_b^1, I_b^2, ..., I_b^k >$, one out of which is likely to belong to the same individual whose identity I_a on OSN_a is already known.*

Alternatively, in recent times, we learn embedding vectors that represents user identity and we compare these embeddings to obtain a rank score which is used to rank identities, details are presented in Section 5.

Evaluation Metric	**Interpretation in the context of UIL problem**
Success/Hit at Rank k (S@k)	The proportion of times that correct linked identity I_b is present among the top-k identities that we retrieve.
Mean Reciprocal Rank (MRR)	The average rank at which the linked identity I_b occurs in the top-k identities that we retrieve.

Table 2: Explanation of evaluation metric in the context of user identity linkage

Given that we cast the UIL problem as a *ranked retrieval problem*, we adopt the following evaluation approach.

1. We consider all possible identity pairs $< I_a, I_b >$ comprising of identities belonging to the two social networks a and b to be part of input dataset D.

2. For each user identity I_a in linked identity pair $< I_a, I_b >$, using different ranking functions, we find an ordered list of identities $< I_b^1, I_b^2, ..., I_b^k >$.

3. Subsequently, we perform evaluation on the basis of metrics discussed in Table 2.

Having discussed the two key formulations for the UIL problem, we discuss methods for collecting linked user identities in Section 3.

3 Data Collection Approaches

Users create their identities (accounts) on multiple Online Social Networks (OSNs) to access a variety of content on offer and connect to their friends. In order to solve the UIL problem, an essential first step is to collect ground truth user identities that belong to the same person across different OSNs, referred to as *linked user identities*.

3.1 Methods for Linked User Identities Collection

In this section, we organize and present methods to collect linked user identities. In Fig 3, we depict a generic framework for data collection, data integration, and data extraction & indexing. The first

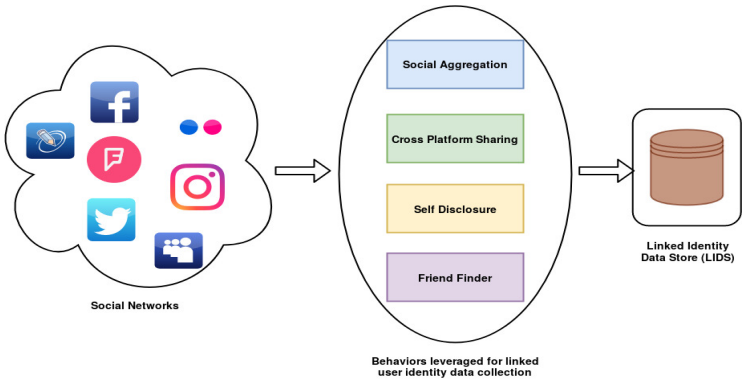

Figure 3: Generic framework for collecting linked user identities across social networks.

step is data collection, in which we identify a source of data (social networks) followed by a selection of data collection methods. We follow it by data integration in which we store user identities collected from all methods at a single data store point, which we refer to as Linked Identity Data Store (LIDS). Next, we present a detailed methodology adopted to perform data collection which lever-

ages different behaviors, namely Social Aggregation (SA), Cross-Platform Sharing (CPS), Self-Disclosure (SD) and Friend Finder Feature (FFF).

3.1.1 Social Aggregation (SA)

There are several websites on which users create an account and provide details of their multiple OSN accounts. We refer this behavioral phenomena as *social aggregation* and such websites as social aggregators. Many prior works [51, 37, 19, 80, 78] exploit this behavior to obtain linked user identities across social networks. One such social aggregator is *about.me*[1] which provides a platform for users to mention numerous user identities, external websites, and well-known social networking websites such as Facebook, Flickr, Google+, Pinterest, LinkedIn, Twitter, Tumblr, and YouTube. Users put their one-page descriptions giving details of their social media profiles along with their background image and abbreviated biography. The website *about.me* provides an option to search user profiles using the interest-based keywords (referred to as *discovery feature*, as depicted in Fig 4). Given an interest-topic as input, it would return all the user profiles having that interest. In addition to the above, datasets available in the public domain can be found, like the ScholarBank at NUS, which we can use.[2] Lastly, we can also use the online web search method using the *site* as *about.me* and interests as *intext* to obtain more user profiles.

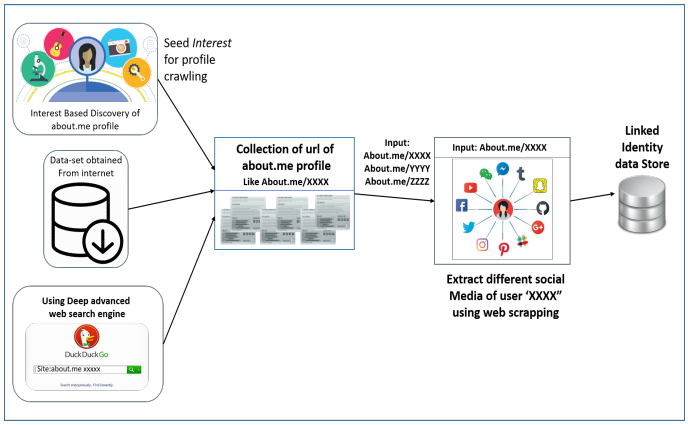

Figure 4: Typical pipeline for Social Aggregator (SA) method using aboutme website.

While we have discussed *about.me*, it may be noted that there are

[1]About.me: https://about.me/
[2]http://scholarbank.nus.edu.sg/bitstream/10635/137403/2/about_me.sql

other similar sites, like Google profile, where users list their social media accounts. Perito et al. [51] performed large scale crawling on public Google profiles and eBay accounts to obtain 3.5 million and 6.5 million usernames. Liu et al. [37] crawled 75,472 public profiles on a social media aggregator site called about.me where users mention details of their identities on at least two social media sites. They found that a total of 15 different social media sites are mentioned by these users, with each user mentioning on average 3.92 social media sites on their About.me profile page. Besides, they also conducted a survey comprising of 153 participants and found that around 82% of them participated in 1-4 online social media sites. One of the contributions of their work was to find the rareness or commonness of usernames, for which they collected usernames by searching through 69 million question-answer threads in Yahoo! answers. From these, they sampled 299,716 usernames mentioned by 673,037 unique users. Goga et al. [19] crawled 3 million Google+ accounts to find ground truth and leveraged the fact that users on Google+ can mention their social media accounts on other websites. Besides above, they also obtained ground truth of 19,000 user pairs on Flickr and Twitter using friend finder feature based on emails. Zhang et al. [80] obtained ground truth by leveraging the fact that users on Question-Answer social networking sites mention details of their other accounts on their home pages on these sites. Around 10,000 users from three sites, namely Stack Overflow, Super User, and Programmer Q& A are obtained, out of which around 20-30% users match pair-wise. Zhang et al. [78] sampled 152,294 Twitter profiles from the tweets posted by users and parse 154,379 profiles from LinkedIn. For ground truth, they looked at Google+ profiles of users and found 9,750 user identities that belong to both Twitter and LinkedIn.

3.1.2 Cross-Platform Sharing (CPS)

Many OSNs provide an option to share content across other (target) OSNs, which we refer to as *cross-platform sharing* (CPS). As depicted in Figure 5, a user makes an update on the source OSN (Instagram) and then shares the same update on the target OSN (Twitter).

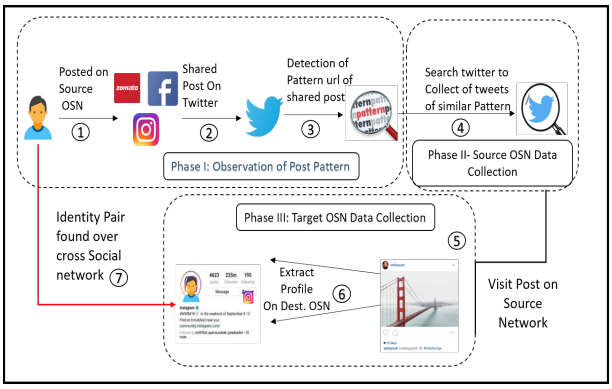

Figure 5: Pipeline for Cross Platform Sharing (CPS) method.

Such shared content on the target network appears with a specific pattern, which in this case, is $\backslash instagram.com\backslash \boldsymbol{p}\backslash$. Using the API provided by the target OSN (Twitter), we search for posts that contain such patterns. We also specifically check for the *source* field present in the Tweet JSON object and make sure that it has the name of the source OSN (Instagram). This check ensures that we filter out the scenario in which a user might copy-paste the link to an update (post) on source OSN while making a post on target OSN. Once we obtain the collected posts from the target network, we identify the URL and expand the URL to reach the desired content on the source social network. On reaching the source social network, we either use source social network API or scrap the post page to obtain the tagged user (mentioned user) in the post on the source social network. In this way, we obtain the linked identity pair between source and target social network. Jain et al. [26, 27] and Correa et al. [11] have used this approach of cross-posting, referred as *self-mention*, to collect identities belonging to the same person.

3.1.3 Self-Disclosure (SD)

Whenever a user signs up on OSN, there is an option to provide a user description. At times, users provide details of their identities on other OSNs, which we refer to as *self-disclosure*, a method leverage by numerous prior works [76, 34, 7, 59, 84, 51, 87, 30, 82, 57]. In Fig 6, we specifically focus on the user's *bio* field in the Twitter network. We first use *Twiangulate* web tool to collect all those twitter profiles which have at least one social network mentioned in their bio-field.[3] Then, we observe various patterns in the bio-field on Twitter because a user can specify other OSN details in multiple ways. For

[3]Twiangulate: http://twiangulate.com/search/

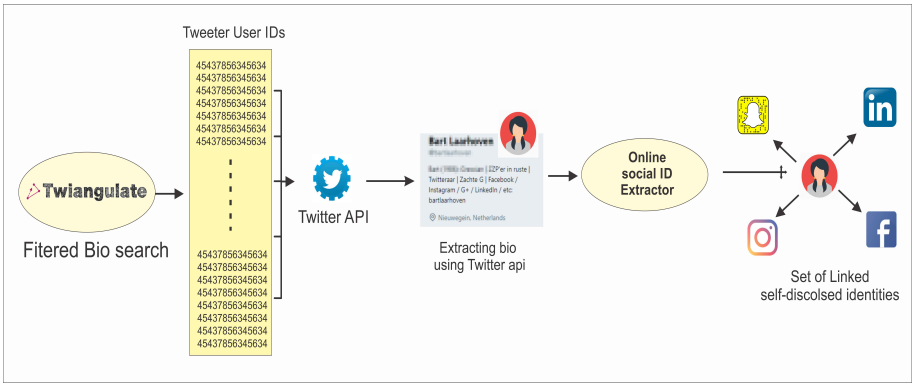

Figure 6: Typical pipeline for Self-Disclosure (SD) method on Twitter.

instance, a user can mention *TV Host and Media Trainer - Instagram: @NeshanTVxyz Snapchat: @Neshaxyz* while another user can use acronyms like *TV Host and Media Trainer - IG: @NeshanTVxyz SP: @Neshaxyz FB: nashbin123*. To address these variations, we tokenize all text and check for the occurrence of URL, which could lead to other OSNs. Besides Twitter, there are other sources, namely blogs, forums, and social networking sites from where Zafarani et al. [76] collected 100,179 username and prior usernames. Their work is unique in the sense that they obtained these usernames pairs from 32 different sites, the maximum coverage any work has done so far. Li et al. [34] leveraged the unique numeric user ID of users on location based social network, Foursquare. On their profile page on Foursquare, some users mention URLs of their Twitter and Facebook profiles. Out of the 1.3 million identities crawled on Foursquare, they could get only 597,822 profiles that were public and available. Among these, 288,480 profiles mentioning Facebook identity, 102,315 profiles mentioned Twitter identity and 67,826 profiles mentioned both Facebook and Twitter identity. Chen et al. [7] leveraged trajectory and check-in data in three real-world datasets. The first dataset comprises walk trajectories of users capturing their outdoor movements like cycling, shopping, driving and site-seeing. This data comprises of 182 user pairs containing 14,337 walk trajectories with 2,190,957 locations and 5,475 car trajectories with 925,380 locations. The second dataset comprises of 89 user pairs from Twitter-Foursquare containing 3,924 check-ins on Foursquare and 35,384 check-ins on Twitter. The third dataset consist of 908 pair of users from Instagram and Twitter, comprising of 267,029 check-ins in Instagram and 357,949 check-ins on Twitter. Shen et al. [59] focused on three social networks, namely Google+, Twitter,

and Foursquare. They collected data using the APIs of these networks and also use crawling to collect more details of users like their neighborhood information. They used common screen names across Twitter and Google+ to find linked Twitter - Google+ user pairs. Some users mention details of their Twitter and Google+ account on their Foursquare profiles, which they used to construct linked Twitter - Foursquare and Google+ - Foursquare user identity pairs. Zhang et al. [84] considered five social networks, namely Twitter, LiveJournal, Flickr, Last.fm, and MySpace. They obtained ground truth linked identity dataset from the prior work of Perito et al. [51]. In addition to social networks, they also used datasets comprising of academic data, namely Arnet-Miner, LinkedIn, and VideoLectures. Arnet-Miner is a platform where users mention details of their other networks (like LinkedIn), which helped in ground truth data for these academic social networking platforms. Zhou et al. [87] evaluated their FRUI (Friendship Relationship Based User Identification) algorithm on both synthetic and real-world datasets. For synthetic datasets, they used random networks [16], small-world networks [70] and preferential attachment model based networks [2], with each network comprising of 10,000 nodes. For real networks, they captured data from the Sina Microblog search page and use OpenAPI to collect RenRen dataset. Kong et al. [30] used the self mention information of Twitter identities on the Foursquare profile of users to link their identities on Foursquare with Twitter. In total, they obtained 500 ground truth matching users on both Foursquare and Twitter. Zhang et al. [82] crawled two social networks Foursquare and Twitter, around November 2012. They crawled 5,392 users from Foursquare to obtain 48,756 tips and 38,921 locations. From Twitter, they crawled 5,223 users and retrieve 9,490,707 tweets. Sajadmanesh et al. [57] used 3456 Foursquare users and 5223 Twitter users as the two social networks. Ground truth comprises 3282 out of which 1900 users join the target network after joining the source network.

3.1.4 Friend Finder Feature (FFF)

Whenever a user joins a new OSN, they sign up using their unique identifier, say email or phone number. This information is used by OSN to find our friends in our email contacts or phone contacts. Using this information, OSN offers a *friend finder* option to help connect to those friends who already have an account in OSN. Figure 7 depicts the entire sequence of steps that we followed in this method. In the first step, we use a deep web search engine like Duck-

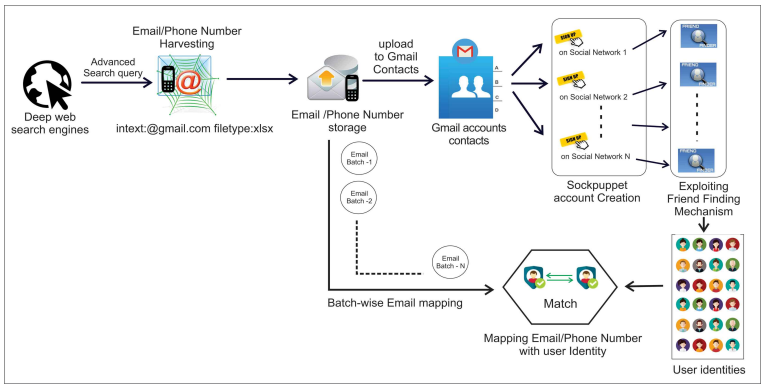

Figure 7: Typical pipeline for Friend-Finder Feature method.

duckgo [4] or any other method for retrieving emails present over the web. Next, we create an email account and add the extracted emails in the contact list of this email account. Then we sign-up in a social network using this email account and leverage the friend finder feature to find whether anyone from contact list is also already a user of the social network. We use string matching on display name of users to find identity belonging to the same user. Goga et al. [18] leveraged the mechanism of *friend-finder* in social networking sites. An extensive collection of 10 million emails were used to link accounts belonging to these emails on three social media sites namely Twitter, Flickr and Yelp. Number of linked users in Twitter-Flickr, Twitter-Yelp and Flickr-Yelp obtained were 13,629 , 1,889 and 1,199 , respectively. Subsequently, they reorganized this data across five localities in US namely Los Angeles, New York, Chicago, San Francisco and San Diego). To get metadata associated with tweets and photos, they used Twitter API and Flickr API, respectively. In the case of Yelp, profile pages were crawled and parsed to extract relevant information.

3.1.5 Snowball Sampling (SS)

In the context of a collection of linked identities, snowball sampling would refer to the process where we increase the linked identities collection by searching in the neighborhood of known linked identities (referred to as seed pairs). Bartunov et al. [3] started with a seed of 16 users on Twitter and Facebook, and use a snowball sampling to collect 398 and 977 users on these two social networks, respectively. For Twitter, they use mutual following as an equivalent of friend-

[4]Deep Web: www.duckduckgo.com

ship relation in Facebook. Liu et al. [40] accessed user behavior data on Douban using it's API which is Chinese social networking site allowing users to create content related to books, movies, music, and local events in cities. A random set of 20 users were selected and their network is crawled using breadth first search approach to increase the number of users to 50,000.

3.1.6 Miscellaneous

Besides the above methods, few prior works have adopted data collection methods that do not fall under any of the methods mentioned above, therefore, we discuss them in this miscellaneous category. Almishari et al. [1] extracted two small subsets from the set of tweets collected by a prior study done by Yang et al. [73] across six month period in 2009. The first subset comprises of 8,262 users who have tweeted more than 2,000 tweets and the second subset contains tweets (around 300 - 400 per user) from 10,000 randomly selected users. They divided each user's tweets into two sets namely Identified Record (IR) and Anonymous Record (AR). Further, they used stylometric features to *link* user's tweets across IR and AR. Zhou et al. [87] evaluated their Friendship Relationship Based User Identification (FRUI) algorithm on both synthetic and real-world datasets. For synthetic datasets, they used random networks [16], small-world networks [70] and preferential attachment model based networks [2], with each network comprising of 10,000 nodes. For real networks, they captured data from the Sina Microblog search page and used OpenAPI to collect the RenRen dataset. Zhang et al. [86] used the Facebook dataset provided by Viswanath et al. [66] comprising of 63,731 nodes and 817,090 edges and synthetically generated two sub-graphs. Nie et al. [47] identified the core interests of users based on tweets from 1,000 random Twitter users over 12 months period. Further, for evaluating linking of profiles across social networks, they targeted 1,213 user pairs from TWitter and BlogCatalog, a social site that allows users to join communities, thereby indicating user interests. Zhang et al. [85] collected details of 20,448 and 40,618 users on two popular Chinese social networks namely Sina Weibo (similar to Twitter) and Renren (similar to Facebook), respectively. For ground truth, they manually linked user identities from these two social networks.

To summarize, Table 3 provides the distribution of prior works among the various data collection methods discussed in this section. Most of the works have used social aggregation or self-disclosure as their data collection methods. Prior work rarely use the friend finder

Name of Method	Prior Works
Social Aggregator (SA)	[51], [37], [19], [80], [78]
Cross-Platform Sharing (CPS)	[26], [27], [11]
Self-Disclosure (SD)	[76], [34], [7], [59], [84], [51], [87], [30], [82]
Friend Finder Feature (FFF)	[18]
Snowball Sampling (SS)	[3], [40]
Miscellaneous	[1], [73], [87], [86], [66], [47], [85]

Table 3: Distribution of prior works among the data collection approaches for collecting linked identities.

method because of the dependence on the availability of emails. Cross-platform sharing is a promising method but sparingly used because it requires API support for post filtering in target OSN, which is not common.

3.2 Social Network Diversity

Prior works cover several social networks. In this section, we present the distribution of social networks covered by researchers to solve the problem of user identity linkage in the past. Table 4 provides the list of social networks, it may be noted that each prior work appears two or more times because each work collects user identities from two or more OSN platforms. From Table 4, we observe that most of the prior works use Twitter as the social media platform because data on Twitter is public by default and it provides excellent support for Application Programming Interface (API), which is a collection of pre-defined functions used to obtain Twitter data through computer programs. After Twitter, we find that many prior works collect data from location-based social network Foursquare and image-based social network Flickr. Following them, we observe that social networks, namely Google+, Facebook, MySpace, and LiveJournal, are the platforms for data collection. While Facebook is the most widely used social network, the reason for the low adoption of Facebook in the research community is because the Facebook graph API is restrictive owing to the nature of private content, which is mostly present on Facebook. Prior works sparingly use remaining social networks.

We provide below a few indicative prior works along with the details of social networks being used by them. Perito et al. [51] conducted studies on using only usernames. They investigated large lists of usernames comprising of 3.5 million usernames obtained from public Google profiles, 6.5 million from eBay accounts. They used the information expressed on Google profiles to derive linked user

Social Network	Prior Works
Twitter	[3] (2012), [19] (2013), [30] (2013), [18] (2013),[1] (2014), [59] (2014), [4] (2014), [82] (2014), [78] (2014), [84] (2015), [81] (2016), [57] (2016), [34] (2017), [7] (2017),
Foursquare	[30] (2013), [59] (2014), [82] (2014), [81] (2016), [57] (2016), [7] (2017), [34] (2017)
Flickr	[24] (2011), [18] (2013), [19] (2013), [4] (2014), [84] (2015)
Google+	[51] (2011), [19] (2013), [59] (2014)
Facebook	[3] (2012), [19] (2013), [34] (2017)
MySpace	[19] (2013), [84] (2015)
LiveJournal	[4] (2014), [84] (2015)
About.me	[37] (2013)
Blogs	[76] (2013)
Delicious	[24] (2011)
Douban	[40] (2017)
Instagram	[7] (2017)
Last.fm	[84] (2015)
LinkedIn	[78] (2014)
Stack Overflow	[80] (2015)
StumbleUpon	[24] (2011)
Super User	[80] (2015)
YouTube	[4] (2014)
Yelp	[18] (2013)

Table 4: Distribution of social networks from where user identities are collected by prior works.

identities. Zafarani et al. [76] did not restrict themselves to only social networking sites. They obtained username pairs from various other sources like web blogs and forums. In total, they collected usernames from 32 online sites. Li et al. [34] leveraged the incremental numeric user IDs on Foursquare to collect ground truth. From the Foursquare profile pages of users, they gathered self-disclosed identities of users on two other social media sites, namely Facebook and Twitter. Liu et al. [37] crawled 75,472 public profiles on About.me and collected a total of 15 different social media sites mentioned by these users. Goga et al. [18] considered data from three social media sites, namely Yelp, Twitter, and Flickr offering various content sharing services to users in terms of service reviews, microblogs, and photo sharing, respectively. Chen et al. [7] obtained datasets on Instagram-Twitter and Foursquare-Twitter from prior work of Riederer et al. [56] and pruned the data to only those data instances which contain sufficient trajectories. Besides these, they also evaluated their approach to walk and car trajectories data from

Beijing's GeoLife project.[5] Almishari et al. [1] looked at the problem of linking content posted by users within single social network, namely Twitter. They divided the tweets posted by the user into two parts and recast the linkability problem as detecting the same user's posts across these two parts. Shen et al. [59] focused on three social networks namely Google+, Twitter, and Foursquare. Zhang et al. [84] worked on data from five social networks, namely Twitter, LiveJournal, Flickr, Last.fm, and MySpace. Additionally, they also use datasets comprising of academic content, namely Arnet-Miner, LinkedIn, and VideoLectures. Iofciu et al. [24] linked users across three social networks, namely Flickr, Delicious, and StumbleUpon. While Flickr is an image sharing platform, the remaining two help users organize their publicly available web documents. Kong et al. [30] collected user data from Foursquare, and Twitter. They employ breadth-first search strategy using the 7,504 tips (location updates) information as a seed to obtain 500 users on Foursquare. Further, corresponding to these users, another 500 users on Twitter are collected with 741,529 tweets. Bartunov et al. [3] collected 398 and 977 user identities on Twitter and Facebook, respectively, starting with 16 seed pairs of nodes. Goga et al. [19] studied five popular social networks namely Facebook, Twitter, Flickr, Google+, and MySpace. Bennacer et al. [4] worked on four social networks YouTube, Flickr, Twitter, and LiveJournal. They extended the dataset provided by Buccafurri et. al. [5] by filling the missing attribute information and adding new friend connections using the APIs of these networks. Zhang et al. [82] evaluated their Multi-Network Link Identifier (MLI) framework on Foursquare and Twitter social networks, comprising of around 5,000 users from each of the network. Zhang et al. [80] focused on linking users across Question-Answer based social networks namely Stack Overflow, Super User and Programmers Q&A. Zhang et al. [81] used Foursquare and Twitter as the two social networks with both users and locations co-aligned as the ground truth. Sajadmanesh et al. [57] also used Foursquare and Twitter as the two social networks. Zhang et al. [78] performed profile linkage using cost-sensitive features on Twitter and LinkedIn social networks. Liu et al. [40] used Douban, which is a Chinese social network that provides facility to user to create content related to films, music, books, and events in various cities.

[5]https://www.microsoft.com/en-us/research/people/yuzheng

4 Machine Learning Approach

In this section, we discuss the machine learning approach to solve the UIL problem. As per this approach, we leverage profile, content, and network information of the users to create features. We next describe these features.

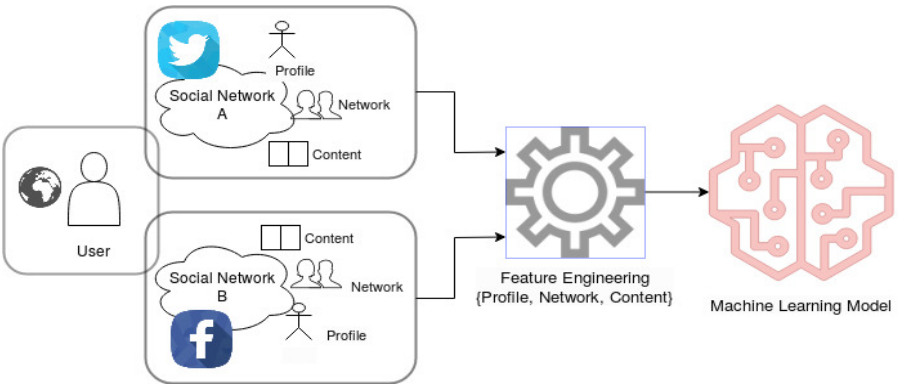

Figure 8: Depiction of machine learning approach for solving user identity linkage problem.

4.1 Profile Features

Profile features comprise of user's basic information like username, display name, location, and profile picture. OSNs have different options and interfaces with varying degrees of details to represent user profile features. Given that access to user's content and network (friends) has been dwindling due to privacy considerations, there are works in the past that have restricted themselves to the use of only profile features.

One of the earliest works was by Perito et al. [51] who proposed to connect user identities only based on usernames. They applied the concept of *information surprisal*, which quantifies the amount of information that the outcome of an experiment conveys. For random variable X and x as one of the outcome, the information surprisal is defined as $I(x) = -logP(x)$, which suggests that low probability gives a higher surprisal. They found that usernames alone express much information quantified by information surprisal. Besides, they argued that the probability of two usernames belonging to the same person depends on the shared information conveyed by these usernames and likelihood of user changing username from one form to another. They proposed three approaches to compute this likelihood. The first approach modelled it as a Markov-Chain

process, in which the goal is to predict the next character of the username. The second approach used TF-IDF, where they considered characters as terms and all possible substrings of given usernames as documents. In the third approach, they used string-only similarity metric, namely levenshtein distance to measure the similarity between two strings. Zafarani et al. [76] proposed a framework called MOBIUS (modeling behavior for identifying users across sites) for connecting user identities across social media sites. The framework comprised of three steps. In the first step, users were identified by their unique behaviors, thereby resulting in redundancies across social media sites. In the second step, they generated features that were based on these redundancies. Finally, in the third step, the features were fed into machine learning classifiers. MOBIUS used the most basic information, that is, username as the user attribute to measure user behaviors. It created an extensive set of features based on patterns due to human limitations, exogenous factors and endogenous factors. While creating usernames, humans are constrained by knowledge limitation, memory & time limitations. The exogenous factors affecting users' decision to create usernames are typing and language patterns adopted by humans. They extracted a total of 414 features by leveraging these factors, out of which top-10 features are finally considered after performing feature importance. Li et al. [34] investigated the redundant information associated with usernames of users across social networks. They captured the redundant information in terms of length of username, similar characters in the username, and similarity in the distribution of letters in username. As per their findings, around 45% of users keep the same usernames across social networks. Goga et al. [19] found a correlation between readily available attributes, namely username, profile pic, location, and real name. They obtained classification features from comparisons of these attributes on five different social networks. If two accounts belonging to the same user do not exhibit a high correlation for a particular pair of the social network, then the chain of correlation is explored to link user accounts using correlation of attributes with third social network. Liu et al. [37] looked at the problem from the perspective of *alias-disambiguation* which tells whether two same usernames belong to the same person or not. They solved the problem by proposing a methodology for automatic labeling of usernames. They hypothesized that usernames which are rare would belong to the same individual whereas username which is common would belong to different person. They computed the rareness or common-ness of usernames using the n-gram username probability. To this end, they segmented the given

username into words and then find the probability of the words in the given corpora. Logistic regression function was applied to the n-gram username probability to find whether two given usernames belong to the same person or not. Furthermore, they claimed that this model outperforms the model which was using features derived from user meta-data like avatar, location and user's post based features. Iofciu et al. [24] leveraged the user assigned tags to the user profile on different social networks. They used TF-IDF based vectorization to consider each user's profile as a vector of tags associated with the profile. Cosine distance was the metric used to compare two vectors representing two user profiles.

4.2 Content Features

In this section, we discuss prior works that derive features from the content posted by users on various OSNs. Goga et al. [18] studied the content posted by users across different social networks and proposed a solution using which adversaries can match accounts belonging to the same individual. They investigated three characteristic features associated with posted content, which include the timestamp of post, the writing style of the user, and the geo-location with the post. For locations, they used the zip code of users. Histogram representing the frequency of visits of users to a particular location is used as a *location profile* of the user. TF-IDF weights on zip codes for a user are used to construct location features for the user. For the timestamp of the post, authors exploited the automated cross-posting behavior of users across social networks. Posts made within a short time period, obtained from ground truth, were considered coming from the same users. Lastly, they considered the content of the post made by users across social media sites. Language models were constructed based on the histograms of unigrams occurring in the user posts. Features derived from posts, timestamps and locations are passed as input to binary logistic regression classifier. They found that location and timestamp play a more critical role in identifying users than the content of posts. Chen et al. [7] proposed a novel STUL (spatio-temporal user linkage) model, which extracted the spatial and temporal features of users to link user identities across social networks. They considered both time and space as continuous variables. They used an extension of density-based clustering to obtain spatial features of users, which are captured as *stay regions* as places where user has stayed. To extract temporal features of users, they used Gaussian Mixture Model (GMM), which contains global and local time distri-

butions. Features from space and time were assigned weights based on the TF-IDF approach. Two types of user data were monitored namely trajectory of the user and the check-in data from the user. Almishari et al. [1] showed that users maintaining multiple accounts on Twitter can be linked to the same person in the presence of large number of Twitter users provided they were actively posting tweets. Two categories of text features were extracted, namely unigrams comprising of all english letters and bigrams consisting of all possible two-letters found in tweets. These features were used in Naive Bayes classifier to decide the user who has posted the tweet.

4.3 Profile and Network Features

Shen et al. [59] focused on raising awareness of the risks associated with linking user identities across social networks. In particular, they proposed a User Account Linkage Inference (UALI) framework, which helped in making users aware of the risks due to the linkage of user identities. Subsequently, they introduced a mechanism to enable users control the risks associated with identity leakage through their proposed framework, referred to as the Information Control Mechanism (ICM). The UALI framework used basic features obtained from profile (name, gender, location) and neighborhood (friends, followers, and followees). Zhang et al. [84] proposed a novel energy-based model, referred to as COnnecting heterogeneouss Social NETwork (COSNET) which incorporates local user matching based on the profile information of the user and network matching based on neighborhood information of the user. Besides, since the work focuses on more than two social networks, they considered global consistency which states that if I_a, I_b and I_b, I_c were linked user identity pair on social networks a, b and b, c, respectively, then by transitivity, I_a, I_c is also linked pair across networks a, c. They obtained an objective function by combining local, network, and global consistency. Zhang et al. [78] proposed an approach to profile linkage that leverages cost-sensitive features, namely profile avatar and geocode using Google Maps API, besides the common friend information. Their approach used local features, namely username, language, profile description, and network popularity. Bartunov et al. [3] introduced a probabilistic approach based on conditional random fields, referred to as Joint Link-Attribute (JLA), to find user identities of single-user across social networks. They used *scheme mapping* [33] to align two key user attributes, namely screen name and URLs provided by the user in their profiles of social networks. For comparing common network structures,

they used the *dice coefficient*, which is the normalized form of common nodes directly connected to the given node pair. Zhang et al. [85] proposed a local expansion strategy based on the breadth first search to find user identities belonging to the same user. They used profile and network based features, namely username, home town, and friend network to expand the initial small seed linked users, referred to as known anchor links. Bennacer et al. [4] leveraged publicly available profile information along with topology of users' friend network to link user accounts across the social networks. The first step involved the selection of candidate pairs of users who are likely to belong to the same individual based on network topology. In the second step, they used public attributes to create matching rules to compare two user accounts. Zhang et al. [81] linked not just common users across social networks, but also common locations being referred across social networks. They proposed unsupervised concurrent alignment (UNICOAT), which leverages attribute and link information to recast the alignment problem as a joint optimization problem. Their work relied on the observation that users have common neighbors and profile attribute information across social networks, the quality of this common-ness was captured in the cost function.

4.4 Content and Network Features

Nie et al. [47] proposed a Dynamic Core Interest Mapping (DCIM) algorithm that builds upon human behavioral limitations in social networks. As a consequence of human limitations, the core interests of users are limited. Moreover, the DCIM algorithm computed core interests of users and then used it to map user identities across social networks. Content posted by users, along with the structural connections shared by users with their friends, were jointly used in the algorithm. Zhang et al. [80] focused on multiple anonymized social network alignment problem in which an unsupervised approach which relies on transitive relation among user accounts across social networks. They referred their proposed approach as Unsupervised Multi-network Alignment (UMA) to align multiple networks in which users are anonymized to protect their identity. UMA leverageds the fact that social networks have few common users across them, referred to as *anchor nodes*. Question-Answer types of social networks were considered, and an edge between two users is considered if they both post on the same question. This edge information was used to cast a pairwise network alignment problem as optimization problem. Kong et al. [30] proposed a Multi-Network Anchoring

(MNA) framework, which captured heterogeneous features of users across social networks. They derived the first set of features from the social connections of users across social networks. In particular, the notion of a common network (friend circle) was captured in three different metrics, namely common neighbors, Jaccard coefficient, and Adamic/Adar measure. They considered the content posted by users as weighted TF-IDF vectors. Additionally, they also considered the location and time of the user posts as features derived from content. Zhang et al. [82] proposed a Multi-Network Link Identifier (MLI) framework, based on the creation of intra-network and inter-network social meta paths. The social network was modeled as a graph comprising of nodes of different kinds - users, posts, words appearing in posts, the time stamp of posts, and locations from where posts are made. Homogeneous meta paths captured the relationship between the same type of node, in this case, user-user relationships based on follow-followee relationships. Heterogeneous meta paths captured the relationship between dissimilar types of nodes, in this case, user-content relationships based on location, timestamp, and words appearing in a post. Mutual information based on information theory is used as the ranking metric to identify important meta paths. They used the features from these meta paths to build link prediction models. Sajadmanesh et al. [57] also used meta-path based approach, in particular, they proposed two types of meta-paths namely Connector and Recursive Meta-Paths (CRMP). Like Zhang et al. [82], they also created paths comprising of user nodes, user posts, words in the post, time and location of the posts. They constructed six different types of meta-paths based on user social connections (follower-followee relationship). Other types of meta-paths were based on the temporal, spatial, and textual similarity of posts made by users. Path count, in other words, a number of meta-paths for each node in the target network, was used as the feature for the SVM classifier with a linear kernel.

4.5 Network Features

One of the fundamental principles of social networking is the concept of *homophily*, which implies similar users connect with each other. User's network information is an essential feature for linking user identities. Zhou et al. [87] proposed FRUI (Friendship Relationship Based User Identification) algorithm, which used the fact that identical users set up common friendship structures in different social networks. Given two user identities I_a and I_b from two social networks a and b as input, the algorithm aimed to find the

match degree $M_{i,j}$ which was defined in terms of common neighborhood. Zhang et al. [86] observed that users have different tie strength across social networks with their friends, which they refer to as heterogeneous relationships. The degree of interaction among two users decided the tie strength. They proposed network reconciliation algorithm (*NR-GL*) that leverages this heterogeneous relationship among users, into a unified framework, *UniRank*, comprising of local and global features. Proposed algorithm starts by exploring seed user pairs (similar user identities across social networks) and then for each such pair, uses a breadth first strategy with local matching to find more such seed pairs. UMA leverages the fact that social networks have few common users across them are called as *partially aligned networks*, and such users are referred to as *anchor nodes*. Liu et al. [40] approached the problem of linking users across different social networks by proposing a model that measures the distance of users across social networks, referred to as the Adaptive User Distance Measurement (AUDM) model. Model casts the problem as a convex optimization problem, converts each social network into a common embedding space, leverages metric learning, and boosting to find the distance between users.

5 Representation Learning Approach

In the representation learning approach, features are learned implicitly rather than explicitly from profile, content, and network. The implicit learning of features is made possible by implementing methods for learning network embeddings. These network embeddings are inherently low dimension representation of network nodes.

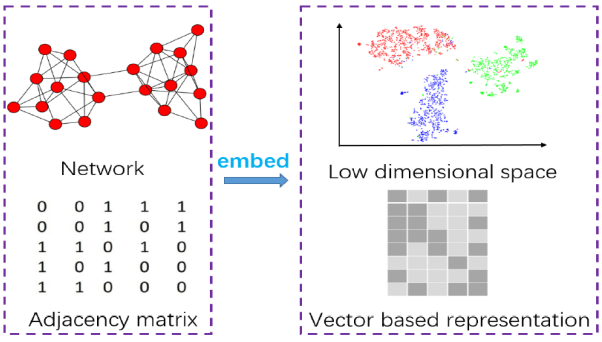

Figure 9: Depiction of representation learning in which low dimension node embeddings are learned [12].

These low dimension representations are the features learned,

which is an alternative to the approach where hand-crafted features are computed explicitly. Recently, there are a few works that have emerged which address the problem of user identity linkage using the network embedding approach. We categorize these works into two main categories, namely, problem-independent and problem-dependent approaches.

5.1 Problem independent approaches

These approaches aim to learn generic low-dimensional representations without focusing on any specific problem. In other words, their goal is construct representations without optimizing them for the specific problem of user identity linkage. Rather, the objective is to learn effective node representations in low dimensions, using mostly the structural information present in a graph. The reason we study these works is that many of the approaches (as we shall discuss in the next sub-section) that focus on identity linkage problem draw inspirations from the optimization frameworks proposed in these works. Given that these works do not directly focus on the user identity linkage problem, we discuss only a few well-known works in this category. Depending upon the kind of information used for learning node embeddings, we divide the works in this section in two parts, one which uses only structural information and second, which uses both structure and content (semantic) information present with nodes in the network.

5.1.1 Network based

Tang et al. [63] proposed a framework, referred to as LINE for network embedding in large graphs. Their approach can be applied to different types of graphs, namely directed, undirected, and weighted. They preserved first order node proximity, which means nodes that are directly connected with each other have their embeddings closer than other nodes. Besides, they also preserved second-order node proximity, to capture the notion that related nodes can also be present at two-hop distance. In order to make stochastic gradient descent based optimization computationally feasible, they proposed negative edge sampling technique to learn the embeddings at a faster rate, thereby ensuring the LINE works well on large scale graphs. Perozzi et al. [52] proposed the DeepWalk framework to learn node representations in a given network. The key difference from LINE was the adoption of an alternative approach to learn node embeddings. They performed random walks over the graph in

a truncated manner and leveraged the notion of the skip-gram model in language modeling to learn latent representations of nodes in a graph. The nodes which appeared in the truncated random walk are considered to be closer (or similar) to the starting node from where the walk started. Wang et al. [68] proposed SDNE (Structural Deep Network Embedding) method, which departs from the earlier methods based primarily on shallow methods. Given that network structures are complex and non-linear, SDNE learnt node embedding using a semi-supervised deep learning approach. As a result, non-linear relationships in graph structures were captured in the SDNE approach. In order to take care of sparsity and preserve network structure, the SDNE framework leveraged first and second-order node proximity as proposed by prior works like LINE. Grover et al. [20] extended the notion of a random walk proposed in the DeepWalk framework ([52]) by introducing biased-ness in the random walks. They proposed node2vec framework for learning node features in a given network. The notion of biased-ness captures the diversity in the network neighborhood. More specifically, the biased walk controlled the graph exploration strategies, whether to walk in a depth-first manner or a breadth-first manner. They introduced a new parameter, referred to as search bias which is used to control the exploration of a random walk. Chen et al. [6] proposed PME (Projected Metric Embedding) model. As per this model, they learnt the node embeddings and their relationship embeddings in separate embedding space. They projected node embeddings onto the relations embedding space and then measured the relationship proximities. For optimization, an adaptive sampling approach that is loss-aware was employed. Matsuno et al. [43] solved the user identity linkage by recasting a network into multiple layers. More specifically, they modelled social networks as multiplex networks representing multiple layers, each of which depicts a specific type of relationship. They proposed the MELL framework, which was an embedding method for multiplex networks. MELL converted each node in each layer into low dimensional vectors and then leveraged edge probabilities to learn node embeddings in the multi-layer scenario.

5.1.2 Network & Content based

Methods discussed till now leverage only the structural information in a network to learn node embeddings. However, there are works which, in addition to the network information, also utilize the semantic relationships between nodes to create node representations. Xu et al. [72] proposed two embeddings for each node that capture

the structural proximity of nodes as well as the semantic similarity, which they express in terms of common interests. More specifically, they considered two types of links, namely structural-close links and content-close links, to capture structural closeness and common interests. Liang et al. [35] proposed Dynamic User and Word Embedding model (DUWE) that monitors over some time, the relationship between user and words. Both user and word embeddings were learned in the same embedding space, thereby effectively capturing their similarities. The learned embeddings helped in the retrieval of top-k most relevant users with given interests. Like Xu et al. [72], this work also captured both network and content proximities in the given network. Liu et al. [36] presented a Self-Translation Network Embedding (STNE) framework that is a sequence-to-sequence framework taking into consideration both content and network features of the node. They performed random walks to generate sequences. The goal of the STNE framework was to translate content sequence to node sequence.

5.2 Problem dependent approaches

In this section, we discuss prior works that learn low-dimensional embedding focusing on the specific problem, which in our case is to detect cross-network linkages representing user identities across social networks. Like the categorizations in the previous section, we divide prior works in this section as well based on the type of information used to learn node representations.

5.2.1 Network based

Liu et al. [38] proposed an Input-Output Node Embedding (IONE) framework to align user identities across social networks belonging to the same person by learning node representations that preserve follower-followee relationships. IONE framework brought the embedding vectors of nodes closer in embedding space who have similar followers and followees. To capture follower-followee relationship, they defined input and output context for each node. Input context defined the contribution of a given node to each of the neighbors of the node. Output context defined the contribution of neighbors of a given node to the node. For learning node representations, they used negative sampling with stochastic gradient descent. Man et al. [42] introduced a framework referred to as PALE (Predicting Anchor Links via Embedding), which predicted anchor links via embeddings. They used few known linked identities referred to as

anchor links as supervisory information. First, it converted a social network into a low dimensional node representation. They followed it up by learning a matching function that was supervised by known anchor links. Sun et al. [62] addressed the issue of lack of labeled data and the unavailability of seed anchor node pairs. They proposed a bootstrapping approach that labels node pairs that were likely to belong to the same user in an iterative manner. A network of users was represented as a knowledge graph, and the process of assigning labels was referred to as entity alignment. Chu et al. [9] proposed CrossMNA, which refers to the cross network embedding method. They addressed the issue of linking users across multiple social networks rather than two social networks only. CrossMNA used only the structural information of nodes to create node embeddings. They used two types of information, namely intra-vector, which reflected structural information inside a given network and inter-vector, which captured the common-ness among the potential node pairs belonging to the same user. Yasar et al. [75] proposed a global structure assisted network aligner (GSA-NA) method. Rather than using local information, they leveraged global structure present in graphs to align nodes belonging to the same user. From the given set of anchor nodes, they identified a small subset of anchors referred to as vantage points, which act as reference points for large graphs. Instead of working on the entire graph, computations were performed on these vantage points, thereby reducing the computational costs considerably. Yang et al. [74] proposed Graph-Aware Embedding Method (GAEM), which modelled the relationships between two or more social networks into a single unifying framework. They used only the network's structure information to learn node embedding for the user identity linkage problem. For second-order structural similarities, they made use of the K-nearest neighbor algorithm to identify nodes at second order proximities. Cheng et al. [8] proposed USAIP (User Alignment via Structural Interaction and Propagation) model which captured the information interactions among users in a structural manner. USAIP can learn from the new structural information formed by newly added nodes in the network along with existing structural information.

5.2.2 Network and Attribute based

Heimann et al. [21] proposed the REGAL framework, which stands for representation learning-based graph alignment and was based on the cross-network matrix factorization method (xNetMF). To speed up the computations, they employed approximations of dense and

large matrices, which were of low-rank, as proposed by Drineas et al. [14]. Each node is represented as a vector that is formed from structural information and attribute information available in the node. A combined node similarity function that captured attribute-based distance and structure-based distance was employed. Su et al. [61] proposed MASTER framework to overcome the three shortcomings of robustness, comprehensiveness, and multiplicity in the prior works. The MASTER framework worked across multiple social networks and combines information from node structure and node attribute information. They proposed constrained dual embedding (CDE) model that simultaneously aligned more than two social networks and learn node embeddings at the same time. Zhang et al. [83] aimed to address the problem of diversities in the node neighborhood and error propagation by proposing MEgo2Vec node embeddings. It was based on graph-based neural networks to represent the immediate neighborhood of nodes across two social networks. Attribute information associated with each node was considered as a list of words. They converted each word into embedding vector and subsequently create character embeddings using CNN. A combined objective function that concatenates the difference between structure embeddings and attribute embeddings was employed.

5.2.3 Network and Content

Wang et al. [69] proposed LHNE mode referred to as linked heterogeneous network embedding model. It created a unified framework to leverage structure and content posted by users for learning node representations. From the content posted by users, they extracted the topics representing user interests using Latent Dirichlet Allocation (LDA). For the structure, friend based node proximities were preserved across the social networks. They learnt a joint optimization function combining interests and friends' information. Sajadmanest et al. [57] proposed CRMP (Connector and Recursive Meta-Path) framework, which was a meta-path based approach. In addition to the actual friendship network, they created a content based network taking into account location, keywords, and time of the post. They projected friends information and post information on a heterogeneous graph and meta-paths capture walk on user nodes and content nodes in such a graph. Nechaev et al. [46] proposed a graph embedding framework to link users in the knowledge base (DBpedia) with Twitter users. They constructed co-occurrence matrices using the words present in content posted by users. For constructing graphs, they considered retweet and mention behavior on Twitter.

Xie et al. [71] used the concept of factoid embedding, which was an unsupervised approach to perform user identity linkage. A factoid is a triple containing two users and the relationship between them. For instance, a user following another user. Their approach learnt factoid embedding by taking into consideration that each user had diverse attributes, content updates, and neighborhood.

6 Future Directions

In this section, we discuss various directions for future work in the context of users joining multiple social networks. Prior works address most of the problems in social networks in the context of a single social network by monitoring user behavior in one single social network. However, with the availability of linked user identities, more comprehensive information about user's behaviors over several social networks can be obtained [41]. This information would help solve many problems in social networks; we discuss some of them in this section.

6.1 Recommendations

Making recommendations for different aspects by using user's behavioral preferences on more than one social network is an important application. Ozsoy et al. [49] collected data from different online platforms, namely Twitter, BlogCatalog, Facebook, Flickr, LastFm, and YouTube to help in recommendations. They compared recommendation systems built from only one social network with those built using many social networks and found that recommendations done using the later approach are more robust and comprehensive. Ostuni et al. [48] and Musto et al. [45] performed recommendations by leveraging Linked Open Data (LOD) platforms like DBpedia. However, most of these prior work made use of data-level linkages across the social network, which does not involve privacy issues typically associated with users. It would be interesting to explore in the direction of user identity linkage to improve user recommendations.

6.2 Link Prediction

In the context of two or more social networks, the problem of link prediction helps in finding out whether a user would join a new social network or not. Zhang et al. [79] presented a survey of prior works that focus on link prediction across social networks. More specifically, they focused on user-user links and user-location links

across social networks as well for the prediction tasks. Zhang et al. [82] also proposed meta-path based approach for collective link prediction across multiple social networks. Qi et al. [54] proposed to solve link prediction in the presence of sparse connectivity of users in a given network. In such a scenario, they made use of the inter-connections in other social networks of the users to help in link prediction. While there are prior works which predict links across social networks, we need to extend the idea beyond links. More specifically, predicting the social behavior of users by leveraging their behaviors in multiple social networks.

6.3 Social Capital of User

In the context of online social networks, the social capital [29, 60] of users refer to their popularity and acceptance in the social network world which prior works have measures in different ways in terms of likes, shares, engagements, and followers that users receive. Quantifying social capital is helpful for many applications like influence prediction and propagation in the political domain [32] and human resource management [23]. Zafarani et al. [77] studied variations in popularity and friendship for the same users across different social networks. They use this information to predict if a given user is going to be popular on a target social network or not. Most of the prior work has quantified social capital by using only a single social network. There is a need to measure a user's social capital using that engagement received by the user across multiple social networks.

6.4 Social network forensics

Malicious users perform online crimes, and very often they leave behind digital footprints across social networks [41]. Michel et al. [44] proposed an ontology based methodology for the detection of salient traits of users across social networks, which can help in cyber forensics. In a typical scenario, a user who indulges in online crime on a particular social network would not leave any identification trace in the network where the crime was committed. However, if we can link that user's identity to another social network where his behaviors are more apparent, then it would help in tracing the culprits. Given the widespread prevalence of cybercrimes, obtained linked user identities of suspects across multiple social networks would help in better understanding of their behavior and would facilitate investigators in decision making.

6.5 User Privacy

In this section, we discuss privacy implications on users owing to the linkage of their identities across social networks. As we know, some OSNs provide access to the professional network (like LinkedIn) while others provide access to a more personal network (like Facebook). Managing one's identity on multiple such OSN platforms are tricky. A user is likely to post about personal life events on a network like Facebook, but would probably refrain from doing the same on a professional network like LinkedIn. In other words, a user tries to maintain different contexts on different OSN platforms. With online social networks, there is a collapse of user context [13, 67], which has privacy implications. Elias et al. [15] performed a detailed study on the implications of OSNs on the personal and professional life of users, particularly learners in educational settings. On the other hand, using a personal network in the professional domain comes with its share of challenges. Ranieri et al. [55] studied the use of Facebook by teachers for professional purposes. Fox et al. [17] investigated the challenges faced by professionals, particularly teachers, in managing their personal and professional identities in social media. Besides, there are other factors as well that complicate and affect users' participation in these networks. For instance, an incoming friend request on a professional network tends to be accepted even if a requester is not personally known (referred to as 'others') whereas, on a personal network, a user would not like to accept such a request. Most instances discussed above are commonplace for a majority of social media users today. However, when a user's identity is linked across such social networks, then it gives rise to a variety of privacy implications which are seldom addressed or acknowledged. It would be worthwhile to explore the impact of user identity linkage on users who are conscious about their privacy.

6.6 Dataset Biases

A number of data collection approaches, which we discuss in Section 3, have been used in the past to collect user identities belonging to the same user across social networks. Each of those approaches relies on specific characteristic behaviors of users who maintain identities across multiple social networks. Consequently, behavioral biases exhibited by users often get infested in these linked identity datasets. Dataset biases, in general, are being extensively studied. For instance, in the domain of computer vision, there are several prior works [64, 65, 22] that investigate the biases in image datasets. How-

ever, the study of behavioral biases that manifest in the linked user identity datasets has not been explored. Such a study will ensure that the learned models are free from biases and are more robust to different kinds of the dataset being used for their training.

To sum up, there are many applications that stand to benefit from linked user identities because it will provide a more comprehensive information about the users under study.

Contributions and Publications

The main contributions along with the publications are given below.

Methods for User Profiling Across Social Networks: Users have their accounts across multiple Online Social Networks (OSNs). To obtain a comprehensive view of user activities, an important first step is to link user accounts (identities) belonging to the same individual across OSNs. To this end, we provide a detailed methodology of five methods useful for user profiling, which we refer to as Advanced Search Operator (ASO), Social Aggregator (SA), Cross-Platform Sharing (CPS), Self-Disclosure (SD) and Friend Finding Feature (FFF). Taken all these methods together, we collect linked identities of 208,120 individuals distributed across 43 different OSNs. We compare these methods quantitatively based on social network coverage and the number of linked identities obtained per-individual. We also perform a qualitative assessment of linked user data, thus obtained by these methods, on the criteria of completeness, validity, consistency, accuracy, and timeliness.

[1] R. Kaushal, V. Ghose, and P. Kumaraguru. Methods for user profiling across social networks. In Proceedings of the 12th IEEE International Conference On Social Computing (SocialCom). IEEE, 2019.

Investigation of Biases in Identity Linkage DataSets: Previous works on linking user identities across OSNs typically perform two steps. First, they collect ground truth datasets of user identities across social networks belonging to the same individuals and then build a machine learning model whose features are derived from user identities. User behaviors on different social networks drive the construction of these datasets, and as a consequence, behavioral biases get manifested in them. We perform a detailed investigation into these dataset biases, a work which has mostly remained underexplored in the identity linkage research. More specifically, we characterize, detect, and quantify behavioral biases in these datasets. We find that biases manifest in the form of lexical differences in user-generated content, particularly in usernames and display names configured by users.

[2] R. Kaushal, S. Gupta, and P. Kumaraguru. Investigation of biases in identity linkage datasets. In Proceedings of the 35th ACM/ SIGAPP Symposium on Applied Computing. ACM, 2020.

NeXLink: Node Embedding Framework for Cross-Network Linkages Across Social Networks: Users create accounts on multiple social networks. A pair of user identities across two different social networks belonging to the same individual is referred to as Cross-Network Linkages (CNLs). In this work, we model the social network as a graph to explore the question, whether we can obtain effective social network graph representation such that node embeddings of users belonging to CNLs are closer in embedding space than other nodes, using only the network information. To this end, we propose a modular and flexible node embedding framework referred to as NeXLink, which comprises of three steps. First, we obtain local node embeddings by preserving the local structure of nodes within the same social network. Second, we learn the global node embeddings by preserving the global structure, which is present in the form of common friendship exhibited by nodes involved in CNLs across social networks. Third, we combine the local and global node embeddings, which preserve local and global structures to facilitate the detection of CNLs across social networks. We evaluate our proposed framework on an augmented (synthetically generated) dataset of 63,713 nodes & 817,090 edges and real-world dataset of 3,338 Twitter-Foursquare node pairs. Our approach achieves an average Hit@1 rate of 98% for detecting CNLs across social networks and significantly outperforms previous state-of-the-art methods.

[3] R. Kaushal, S. Singh, and P. Kumaraguru. Nexlink: Node embedding frame- work for cross-network linkages across social networks. In Proceedings of the International Conference On Network Science (NetSciX), pages 61-75. Springer, 2020.

Nudging Nemo: Helping Users Control Linkability across Social Networks: Numerous techniques to link user identities across different OSNs have been proposed. However, this linking poses a threat to the users' privacy; users may or may not want their identities to be linkable across networks. In this work, we propose *Nudging Nemo*, a framework which assists users to control the linkability of their identities across multiple platforms. Nudging Nemo has two components; a linkability calculator which uses state-of-the-art identity resolution techniques to compute a normalized linkability measure for each pair of social network platforms used by a user, and a soft paternalistic nudge, which alerts the user if any of their activity violates their preferred linkability. We evaluate the effectiveness of the nudge by conducting a controlled user

study on privacy-conscious users who maintain their accounts on Facebook, Twitter, and Instagram. Outcomes of user study confirm that the proposed framework helped most of the participants to make informed decisions, thereby preventing inadvertent exposure of their personal information across social network services.

[4] R. Kaushal, S. Chandok, P. Jain, P. Dewan, N. Gupta, and P. Kumaraguru. Nudging nemo: Helping users control linkability across social networks. In International Conference on Social Informatics, pages 477–490. Springer, 2017.

Detecting of Misbehaviors in Clone Identities in Online Social Networks: The account registration steps in Online Social Networks (OSNs) are simple to facilitate users to join the OSN sites. Alongside, Personally Identifiable Information (PII) of users is readily available on- line. Therefore, it becomes trivial for a malicious user (attacker) to create a spoofed identity of a real user (victim), which we refer to as clone identity. While a victim can be an ordinary or a famous person, we focus our attention on clone identities of famous persons (celebrity clones). These clone identities ride on the credibility and popularity of celebrities to gain engagement and impact. In this work, we leverage the identity linkage approaches to detect clone identities and then analyze celebrity clone identities to extract an exhaustive set of 40 features based on posting behavior, friend network and profile attributes. Accordingly, we characterize their behavior as benign and malicious. On detailed inspection, we find benign behaviors are either to promote the celebrity which they have cloned or seek attention, thereby helping in the popularity of celebrity. However, on the contrary, we also find malicious behaviors (misbehaviors) wherein clone celebrities indulge in spreading indecent content, issuing advisories and opinions on contentious topics. We evaluate our approach on a real social network (Twitter) by constructing a machine learning based model to automatically classify behaviors of clone identities, and achieve accuracies of 86%, 95%, 74%, 92% & 63% for five clone behaviors corresponding to promotion, indecency, attention-seeking, advisory, and opinionated.

[5] R. Kaushal, C. Sharma, and P. Kumaraguru. Detection of misbehaviors in clone identities on online social networks. In Proceedings of the 7th In- ternational Conference On Mining Intelligence and Knowledge Engineering, Springer, 2019.

References

[1] Mishari Almishari, Dali Kaafar, Ekin Oguz, and Gene Tsudik. Stylometric linkability of tweets. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 205–208. ACM, 2014.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[3] Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, and Hyungdong Lee. Joint link-attribute user identity resolution in online social networks. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis*. ACM, 2012.

[4] Nacéra Bennacer, Coriane Nana Jipmo, Antonio Penta, and Gianluca Quercini. Matching user profiles across social networks. In *International Conference on Advanced Information Systems Engineering*, pages 424–438. Springer, 2014.

[5] Francesco Buccafurri, Gianluca Lax, Antonino Nocera, and Domenico Ursino. Discovering links among social networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 467–482. Springer, 2012.

[6] Hongxu Chen, Hongzhi Yin, Weiqing Wang, Hao Wang, Quoc Viet Hung Nguyen, and Xue Li. Pme: projected metric embedding on heterogeneous networks for link prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1177–1186, 2018.

[7] Wei Chen, Hongzhi Yin, Weiqing Wang, Lei Zhao, Wen Hua, and Xiaofang Zhou. Exploiting spatio-temporal user behaviors for user linkage. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 517–526. ACM, 2017.

[8] Anfeng Cheng, Chun-Yi Liu, Chuan Zhou, Jianlong Tan, and Li Guo. User alignment via structural interaction and propagation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[9] Xiaokai Chu, Xinxin Fan, Di Yao, Zhihua Zhu, Jianhui Huang, and Jingping Bi. Cross-network embedding for multi-network alignment. In *The World Wide Web Conference*, pages 273–284, 2019.

[10] J. Clement. Number of social media accounts (2019), January 2019. [Online; posted January-2019].

[11] Denzil Correa, Ashish Sureka, and Raghav Sethi. Whacky!-what anyone could know about you from twitter. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, pages 43–50. IEEE, 2012.

[12] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[13] Jenny L Davis and Nathan Jurgenson. Context collapse: Theorizing context collusions and collisions. *Information, communication & society*, 17(4):476–485, 2014.

[14] Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(Dec):2153–2175, 2005.

[15] Scott Elias. *Implications of online social network sites on the personal and professional learning of educational leaders*. PhD thesis, Colorado State University, 2012.

[16] P Erd6s and A Renyi. On random graphs. *Publ. Math. Debrecen*, 6:290–297, 1959.

[17] Alison Fox and Terese Bird. The challenge to professionals of using social media: Teachers in england negotiating personal-professional identities. *Education and Information Technologies*, 22(2):647–675, 2017.

[18] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*, pages 447–458. ACM, 2013.

[19] Oana Goga, Daniele Perito, Howard Lei, Renata Teixeira, and Robin Sommer. Large-scale correlation of accounts across social networks. *University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002*, 2013.

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

[21] Mark Heimann, Haoming Shen, Tara Safavi, and Danai Koutra. Regal: Representation learning-based graph alignment. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 117–126. ACM, 2018.

[22] Luis Herranz, Shuqiang Jiang, and Xiangyang Li. Scene recognition with cnns: objects, scales and dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–579, 2016.

[23] John R Hollenbeck and Bradley B Jamieson. Human capital, social capital, and social network analysis: Implications for strategic human resource management. *Academy of Management Perspectives*, 29(3):370–385, 2015.

[24] Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. Identifying users across social tagging systems. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[25] Mansoor Iqbal. Youtube revenue and usage statistics (2019), August 2019. [Online; posted 8-August-2019].

[26] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. @ i seek'fb. me' identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1259–1268, 2013.

[27] Paridhi Jain, Tiago Rodrigues, Gabriel Magno, Ponnurangam Kumaraguru, and Virgílio Almeida. Cross-pollination of information in online social media: A case study on popular social networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 477–482. IEEE, 2011.

[28] J.Clement. Number of monthly active users in twitter, August 2019. [Online; posted 14-August-2019].

[29] Hao Jiang and John M Carroll. Social capital, social network and identity bonds: a reconceptualization. In *Proceedings of the fourth international conference on Communities and technologies*, pages 51–60, 2009.

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

Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pages 179–188. ACM, 2013.

[31] Nitish Korula and Silvio Lattanzi. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment*, 7(5):377–388, 2014.

[32] Ronald La Due Lake and Robert Huckfeldt. Social capital, social networks, and political participation. *Political Psychology*, 19(3):567–584, 1998.

[33] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM, 2002.

[34] Yongjun Li, You Peng, Zhen Zhang, Quanqing Xu, and Hongzhi Yin. Understanding the user display names across social networks. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1319–1326. International World Wide Web Conferences Steering Committee, 2017.

[35] Shangsong Liang, Xiangliang Zhang, Zhaochun Ren, and Evangelos Kanoulas. Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1764–1773. ACM, 2018.

[36] Jie Liu, Zhicheng He, Lai Wei, and Yalou Huang. Content to node: Self-translation network embedding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1794–1802. ACM, 2018.

[37] 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 *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 495–504. ACM, 2013.

[38] Li Liu, William K Cheung, Xin Li, and Lejian Liao. Aligning users across social networks using network embedding. In *IJCAI*, pages 1774–1780, 2016.

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

2014 ACM SIGMOD international conference on Management of data, pages 51–62. ACM, 2014.

[40] Yufei Liu, Dechang Pi, and Lin Cui. Learning user distance from multiple social networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3280–3287. IEEE, 2017.

[41] Anshu Malhotra, Luam Totti, Wagner Meira Jr, Ponnurangam Kumaraguru, and Virgilio Almeida. Studying user footprints in different online social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 1065–1070. IEEE Computer Society, 2012.

[42] Tong Man, Huawei Shen, Shenghua Liu, Xiaolong Jin, and Xueqi Cheng. Predict anchor links across social networks via an embedding approach. In *IJCAI*, volume 16, pages 1823–1829, 2016.

[43] Ryuta Matsuno and Tsuyoshi Murata. Mell: effective embedding method for multiplex networks. In *Companion Proceedings of the The Web Conference 2018*, pages 1261–1268, 2018.

[44] Mary C Michel, Marco Carvalho, Heather Crawford, and Albert C Esterline. Cyber identity: Salient trait ontology and computational framework to aid in solving cybercrime. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 1242–1249. IEEE, 2018.

[45] Cataldo Musto, Pierpaolo Basile, Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Linked open data-enabled strategies for top-n recommendations. In *CBRecSys@ RecSys*, pages 49–56. Citeseer, 2014.

[46] Yaroslav Nechaev, Francesco Corcoglioniti, and Claudio Giuliano. Sociallink: exploiting graph embeddings to link dbpedia entities to twitter profiles. *Progress in Artificial Intelligence*, 7(4):251–272, 2018.

[47] Yuanping Nie, Yan Jia, Shudong Li, Xiang Zhu, Aiping Li, and Bin Zhou. Identifying users across social networks based on dynamic core interests. *Neurocomputing*, 210:107–115, 2016.

[48] Vito Claudio Ostuni, Tommaso Di Noia, Eugenio Di Sciascio, and Roberto Mirizzi. Top-n recommendations from implicit feedback leveraging linked open data. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 85–92, 2013.

[49] Makbule Gulcin Ozsoy, Faruk Polat, and Reda Alhajj. Making recommendations by integrating information from multiple social networks. *Applied Intelligence*, 45(4):1047–1065, 2016.

[50] Rachel Palmateer. Key linkedin statistics to know from 2019, October 2019. [Online; posted 19-October-2019].

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

[52] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[53] Facebook Inc. Press. Company info: Facebook, February 2019. [Online; posted 11-February-2019].

[54] Guo-Jun Qi, Charu C Aggarwal, and Thomas Huang. Link prediction across networks by biased cross-network sampling. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 793–804. IEEE, 2013.

[55] Maria Ranieri, Stefania Manca, and Antonio Fini. Why (and how) do teachers engage in social networks? an exploratory study of professional use of f acebook and its implications for lifelong learning. *British journal of educational technology*, 43(5):754–769, 2012.

[56] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 707–719. International World Wide Web Conferences Steering Committee, 2016.

[57] Sina Sajadmanesh, Hamid R Rabiee, and Ali Khodadadi. Predicting anchor links between heterogeneous social networks. In *Proceedings of the 2016 IEEE/ACM International Conference*

on Advances in Social Networks Analysis and Mining, pages 158–163. IEEE Press, 2016.

[58] Andrew Perrin Shannon Greenwood and Maeve Duggan. Social media updates (2016), November 2016. [Online; posted November-2016].

[59] Yilin Shen and Hongxia Jin. Controllable information sharing for user accounts linkage across multiple online social networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 381–390. ACM, 2014.

[60] Charles Steinfield, Nicole B Ellison, and Cliff Lampe. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6):434–445, 2008.

[61] Sen Su, Li Sun, Zhongbao Zhang, Gen Li, and Jielun Qu. Master: across multiple social networks, integrate attribute and structure embedding for reconciliation. In *IJCAI*, pages 3863–3869, 2018.

[62] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping entity alignment with knowledge graph embedding. In *IJCAI*, pages 4396–4402, 2018.

[63] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[64] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[65] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[66] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[67] Jessica Vitak. The impact of context collapse and privacy on social network site disclosures. *Journal of broadcasting & electronic media*, 56(4):451–470, 2012.

[68] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2016.

[69] Yaqing Wang, Chunyan Feng, Ling Chen, Hongzhi Yin, Caili Guo, and Yunfei Chu. User identity linkage across social networks via linked heterogeneous network embedding. *World Wide Web*, pages 1–22, 2018.

[70] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[71] Wei Xie, Xin Mu, Roy Ka-Wei Lee, Feida Zhu, and Ee-Peng Lim. Unsupervised user identity linkage via factoid embedding. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1338–1343. IEEE, 2018.

[72] Linchuan Xu, Xiaokai Wei, Jiannong Cao, and Philip S Yu. On exploring semantic meanings of links for embedding social networks. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 479–488. International World Wide Web Conferences Steering Committee, 2018.

[73] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[74] Yang Yang, De-Chuan Zhan, Yi-Feng Wu, and Yuan Jiang. Multi-network user identification via graph-aware embedding. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 209–221. Springer, 2018.

[75] Abdurrahman Yasar and Ümit V Çatalyürek. An iterative global structure-assisted labeled network aligner. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2614–2623, 2018.

[76] Reza Zafarani and Huan Liu. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 41–49. ACM, 2013.

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

[78] 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*, pages 117–128. Springer, 2014.

[79] Jiawei Zhang and S Yu Philip. Link prediction across heterogeneous social networks: A survey. *Social networks*, 2014.

[80] Jiawei Zhang and S Yu Philip. Multiple anonymized social networks alignment. In *2015 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2015.

[81] Jiawei Zhang and Philip S Yu. Pct: partial co-alignment of social networks. In *Proceedings of the 25th International Conference on World Wide Web*, pages 749–759. International World Wide Web Conferences Steering Committee, 2016.

[82] Jiawei Zhang, Philip S Yu, and Zhi-Hua Zhou. Meta-path based multi-network collective link prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1286–1295, 2014.

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

[84] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S Yu. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1485–1494. ACM, 2015.

[85] Yuxiang Zhang, Jiamei Fu, Chengyi Yang, and Chunjing Xiao. A local expansion propagation algorithm for social link identification. *Knowledge and Information Systems*, 60(1):545–568, 2019.

[86] Zhongbao Zhang, Qihang Gu, Tong Yue, and Sen Su. Identifying the same person across two similar social networks in a unified way: Globally and locally. *Information Sciences*, 394:53–67, 2017.

[87] Xiaoping Zhou, Xun Liang, Haiyan Zhang, and Yuefeng Ma. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE transactions on knowledge and data engineering*, 28(2):411–424, 2016.