IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 4, 2019, accepted September 18, 2019, date of publication September 23, 2019,
date of current version October 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942840

A Survey of Across Social Networks User Identification

LING XING[1], KAIKAI DENG[1], HONGHAI WU[1], PING XIE[1], H. VICKY ZHAO[2], AND FEIFEI GAO[2], (Senior Member, IEEE)
[1]School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China
[2]Institute for Artificial Intelligence, Tsinghua University, Beijing 100084, China

Corresponding author: Ling Xing (xingling_my@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771185, Grant 61772175, and Grant 61801171, in part by the Science and Technology Research Project of Henan Province under Grant 182102210044 and Grant 182102210285, in part by the Key Scientific Research Program of Henan Higher Education under Grant 18A510009, and in part by the Postdoctoral Science Foundation of China under Grant 2018M632772.

ABSTRACT With the popularization of the Internet and the arrival of the big data era, numerous different social networks (SNs) have emerged to satisfy users' social needs and offer them rich content and convenient services. Under these circumstances, identifying multiple social accounts belonging to the same user across different SNs is of great importance for many applications. Across social networks user identification (ASNUI) can help perfect user information, offer personalized service recommendation, and data mining, as well as provide support for scientific research. This paper first systematically introduces the application of ASNUI in the field of social computing, then states its applications and challenges, and reviews the adopted models, frameworks, and performance comparison state-of-the-art techniques used in ASNUI. Finally, we also identify a few future research directions in ASNUI, such as weight allocation of user attribute information, the fusion of multi-dimensional information, and large-scale user identification.

INDEX TERMS Big data, across social networks, user identification, entity user.

I. INTRODUCTION

We are currently experiencing an explosive growth in the amount of internet data. SNs provide a convenient platform for people to interact with information and are important providers of network big data. According to data released by Facebook in 2017, the platform's number of active users per month has exceeded 2 billion, making it the largest SN in the world. Moreover, according to WeChat[1] statistics from 2017, the number of monthly active users of this platform has reached 980 million. These data show that more and more netizens are using SNs for information interaction purposes in order to meet their different social needs.

Owing to the differences in the services provided by different SNs, people generally choose to selectively participate in various SNs. From the survey in [1], 42% of users had multiple SN accounts in 2013. Each SN reflects the real-life state of users from different perspective, and is an instance of the mapping of the real world onto the virtual network. Due to concerns regarding user privacy protection, user's information on SNs are isolated, and it is difficult for a SN service provider to obtain user's information in other SNs. Therefore, there is no direct link between an individual's various SN accounts, and a complete cross-SN map is difficult to obtain. As shown in FIG. 1, user information can be integrated and perfected to the maximum extent possible by identifying users' multiple social accounts across different SNs. Users' social data are mined to establish a relatively complete personal information for each user, after which a complete SN map of users can be constructed.

ASNUI aims to link accounts in different SNs that belong to the same person, and is thus important in many research fields. The main applications of this concept are as follows:

(1) User profile aggregation: The information that can be obtained about a user from a single SN is limited. However, if a user's accounts on multiple SNs can be identified, then the user history information and the latest dynamics can be more comprehensively mastered. For example, companies can utilize two types of social software, namely Microblog

The associate editor coordinating the review of this manuscript and approving it for publication was Wen Chen.

[1]https://weixin.qq.com/

and LinkedIn, to obtain information about an applicant's interests and work experience before an interview. Use of these network applications in conjunction with ASNUI technology makes it possible to perfect user information.

(2) Personalized service recommendation: Recommendation systems can be used to recommend friends and contents that might be of interest to specific users. Although it is easy to collect data via a single SN, this data tends to be relatively sparse, and the recommendation results are not ideal. On the other hand, if accounts from multiple SNs can be integrated to construct a complete user data, then recommendation performance will be significantly improved [2], [3]. Some experts and scholars were invited by Microsoft to analyze the behavior of Dianping,[2] Sina Microblog [3] and Douban[4] users and to draw social behavior maps to provide personalized service guidance for different types of users [4]. They designed LifeSpec, a data-driven framework for exploring and hierarchically categorizing users' lifestyles. Given user's behavior data as digital footprints, they formalized the lifestyle spectrum of a group and promoted a probabilistic model to learn the lifestyle spectrum. LifeSpec reveals lifestyle commonalities and variations of groups with different demographic attributes, such as place of residence, education and gender. Experimental results show that the proposed method is more stable and robust than other methods when the number of levels changes from 3 to 4. According to users' feedback, when the number of levels continues to increase, it is not so easy for them to choose the most relevant lifestyles. Research on user friend recommendation conducted by the Chinese Academy of Sciences shows that friend relationships and friend behaviors of associated users (identified users) have good consistency/similarity on Twitter and Flickr. Based on this feature, a across SNs friend recommendation system was designed [5].

(3) Data mining: More valuable information can be obtained via data mining through the association of multiple SNs. For example, some researchers used the migration learning method to verify the consistency of user behavior on different SNs. Moreover, common users as a bridge between SNs can be applied to user modeling and behavior prediction in order to help users understand the latest developments of influential people [6], [7].

(4) Support scientific research: The complex network of relationships between users is one of the most important features of SNs. The characteristics of complex networks have been studied intensively in a single SN. However, the question of whether new features will be generated when considering multiple SNs merits further exploration. Some researchers have utilized the meta-path method to achieve cross-SNs link prediction [8]; these results verify that the more common users are identified, the more effective the prediction performance of cross-SNs links will be. ASNUI is also useful in many research fields, such as cross-SNs user modeling, information dissemination [9], cyber security, etc.

While ASNUI provides huge benefits to people, it also carries the associated harm of potentially revealing personal information. An attacker can mine the user's identity information from various SNs without the user's permission by using various attributes of the network. The improper use of user identification techniques can result in a serious threat to all aspects of user privacy. For example, a malicious user could exploit location data to infer some sensitive information about other users [10]–[12]. More and more people are willing to submit their personal information to network applications only after SNs provide users with secure and guaranteed privacy protection [13].

ASNUI technology has made great progress in various applications. Many researchers have proposed a variety of algorithms to facilitate user identification. However, there are still challenges that need to be addressed.

(1) The problem of poor robustness exists in assigning user attribute weights via a subjective weighting method. There is a difference in the contribution of user-generated attribute information to ASNUI. Therefore, each attribute needs to be weighted. The subjective weighting method mainly assigns weights to attributes by human will, which is tightly coupled with the field of attributes and has poor robustness. The objective weighting method requires sufficient sample data to support it, and its universality is poor. In view of the above problems, using the concept of information entropy to weight each attribute will be a research direction in the future.

(2) User identification algorithms that work by exploiting user attribute information lack a time-variant analysis of user information. Some datasets contain attribute information that spans a long period of time. However, in real life, some attribute information changes over time, which is called temporal attribute information. The time period spanned by the temporal attribute information in different SNs is not the same. If the attribute values are blindly required to be the same when matching, it is likely to produce false negative. Therefore, it is necessary to perform time-variant analysis on the attribute information of the user.

(3) User identification algorithms that work by using network topology information lack an analysis of non-friend relationships. User identification methods based on network topology use the number of shared seed nodes to identify users. This type of method defines the relationship between nodes as mutual-following connections and single-following connections (non-friend relationship). When ASNUI occurs in some heterogeneous SNs, non-friend relationships are often ignored. We can equate these two relationships into user features for analysis, and weight them according to their contribution, which might achieve a good identification performance.

(4) User identification algorithms that work by utilizing user behavior information lack a dynamic evolution analysis of text information. The behavior information posted by the user in SNs can intuitively reflect the user's interest topics.

[2]https://www.dianping.com/
[3]https://weibo.com
[4]https://www.douban.com/

The traditional topic mining model is limited to static analysis and lacks dynamic analysis when analyzing text information, which leads to poor identification performance. Using the time window to divide the text information, not only the user's overall interest distribution, but also all the information in the corpus is divided into segments according to time, and the information in different time segments is analyzed locally. The user's matching can be judged by analyzing the regularity of its dynamic evolution, thereby improving the overall performance of ASNUI.

(5) Dataset challenge. To obtain an integral dataset for research purpose, we need face the following problems. i) User privacy, how to crawl and use user identity information without involving user privacy? ii) Ground truth, How to obtain matching user account pairs through SNs when some SNs deliberately prevent users from publishing content; iii) Limited access, when crawling user data using API provided by SNs, these SNs limit the rate and set permissions which make it difficult to obtain user data in large scale.

Moreover, many works only considered one of the above mentioned aspects, and utilized single-dimensional information for user identification without fusion of multi-dimensional information, the precision rate of these algorithms can be further improved.

The goal of this paper is to give a comprehensive review of ASNUI technology research and provide a guidance for future research directions. The contributions of our work are summarized as below:

1) We describe in detail the positive significance of ASNUI in various fields, and summarize the challenges and possible solutions for ASNUI;

2) The ASNUI problem has various models and identification frameworks. We provide three general and formal models for ASNUI according to different user data, and give a unified identification framework;

3) User identification mainly consists of two aspects: similarity calculation and account matching. We summarize the similarity calculation methods for different identification techniques and analyze the complexity between matching algorithms;

4) The general evaluation metrics of ASNUI are summarized. The research status of three identification technologies is analyzed in detail, and their identification performance is compared and analyzed;

5) ASNUI is still an active area, and there are many issues to be solved. We further discuss the future research directions of ASNUI.

The remainder of this paper is organized as follows. In Section II, we introduce the model and framework of ASNUI. In Section III, we summarize similarity calculation methods and matching algorithms. We review the state-of-the-art methods and compare the performance of different technologies for ASNUI in Section IV. Finally, we discuss the future directions in Section V and conclude this paper in Section VI.

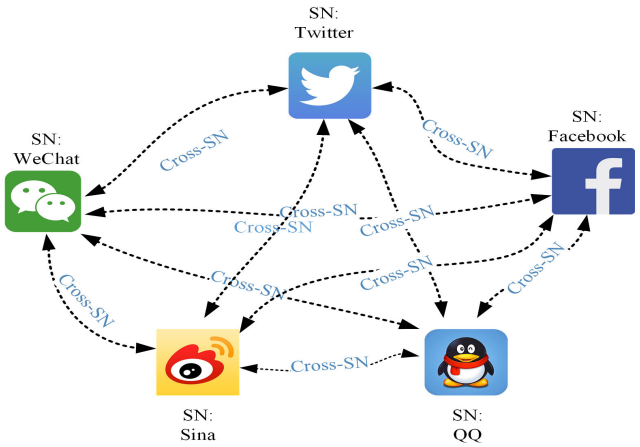

FIGURE 1. Cross-network research to merge various SNs.

II. MODEL AND BASIC FRAMEWORK OF ASNUI

The basic ASNUI model is as follow: a SN is represented by $G\{V, E\}$, where V is the set of users and E is the set of relationships between users. Given an account v_i^X in the source network $G^X\{V^X, E^X\}$, find the account v_j^Y in the target network $G^Y\{V^Y, E^Y\}$ such that they belong to the same user, that is, identify the pair (v_i^X, v_j^Y) that correspond to the same person.

A. USER ATTRIBUTE INFORMATION MODEL

Research based on user attribute information for the purpose of solving user identification problems primarily exploits personal information [34]–[56]. Multi-attribute information about users can be transformed into a multi-dimensional vector, which is used to characterize the user's identity on a specific SN.

Definition 1: If the user attribute information contains n user attribute items in a SN [17], then the multi-dimensional vector can be defined as $F_x = (a_1^x, a_2^x, \ldots, a_n^x)$, where a_n^x denotes the nth attribute item of the account x.

The term "identifiable user" refers to an account in the target network that can be matched with an account in the source network. Accounts from different SNs are selected to form feature vectors, and the similarity vector is constructed by calculating the similarity between various attributes of different accounts.

Definition 2: The similarity vector of accounts is defined as $V(F_A, F_B) = (v_1^{AB}, v_2^{AB}, \ldots, v_n^{AB})$, where v_n^{AB} denotes the similarity of the nth attribute between two accounts, which can be calculated by the *simfunc*() function, and $0 \leq |v_n^{AB}| \leq 1$.

The similarity of a user's multi-attributes can be calculated using the *simfunc*() function [40], [80], [81]. If the two attribute items being calculated have the same or similar values, then *simfunc*() returns '1'; otherwise, *simfunc*() returns '0'. Moreover, since there are differences in the user's multi-attribute information, the form of the *simfunc*() function should be different.

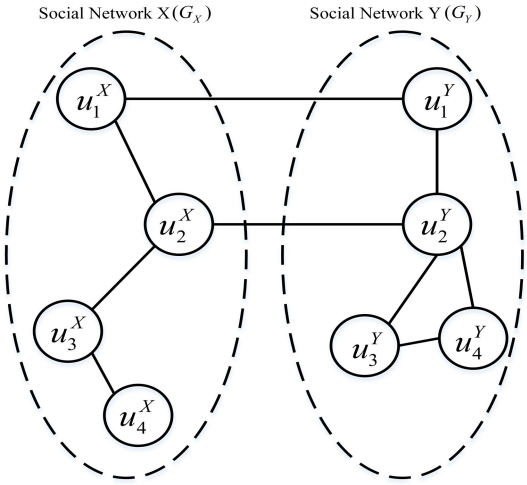

FIGURE 2. Cross-SNs node mapping.

The overall similarity vector of the accounts is constructed by exploiting multiple attributes of the user, after which each attribute item is assigned a reasonable weight, enabling the similarity of the two accounts to be measured. That is, the value of $similarity(F_A, F_B) = \sum_{i=1}^{n} (w_i^{AB} \times v_i^{AB})$ determines whether the two accounts from different SNs match or not [1], [32], [54].

B. NETWORK TOPOLOGY INFORMATION MODEL

Network topology information-based studies mainly utilize the user's circle of friends to identify different accounts belonging to the same user [57]–[71]. The user accounts are equivalent to network nodes to find the similarity relationship. A SN is represented as $G\{U, E\}$, where U denotes a set of user nodes on the SN, and E denotes a set of edges between SN nodes.

Definition 3 (Entity User) [67]: The real user behind the SN account.

Definition 4 (Seed Node) [68], [70]: The user node whose user information is identified in advance.

Definition 5 (The Mapping Function ϕ): ϕ denotes that the virtual account on the SN is mapped to the entity user p, that is, $\phi(u_i) = p$.

Input: Given n SNs $\{G_1, G_2, \ldots G_n\}$ and a few seed nodes whose identities have been identified a priori. The users are equivalent to the network nodes, and the model formulates the match degree using in- and out-degrees [69] in SNs.

Output: The identified user match pairs. For example, in FIG. 2, the account on SN X is represented as $u_i^X \in G_X$, while the account on SN Y is represented as $u_j^Y \in G_Y$. If $\phi(u_i^X) = \phi(u_j^Y) = p$, then the output matching result is (u_i^X, u_j^Y) [52], indicating that the entity user behind the two accounts is the same person.

C. USER BEHAVIOR INFORMATION MODEL

Since most traditional user identification algorithms ignore the effect of the user's published content, the precision of user identification is low. Some researchers have proposed user identification algorithms based on user behavior information, which mainly utilize the account's blog data to identify the user [74]–[90]. The set $A = \{V, C\}$ represents the blog data of the SN user, where V denotes a set of n users and C denotes n sets of user data.

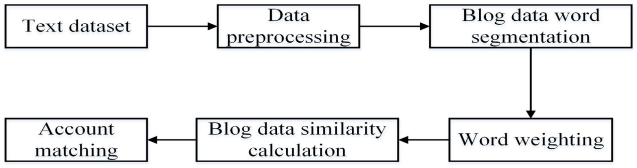

FIGURE 3. Flowchart of account matching based on user blog data.

As shown in FIG. 3, a blog post published by an account on a SN can be processed as a text dataset. Since the blog data published by the user on different SNs will have different formats, it is necessary to perform data pre-processing, and filter out information unrelated to user identification from the text via noise processing on the blog data. The relevant mining algorithm can then be employed for word segmentation [80], [81]; this can reduce the index amount and the degree of calculation required. The correlation weighting technique is used to quantify the importance of certain words. Finally, whether two accounts belong to the same user is determined by the similarity between their blog data.

D. BASIC FRAMEWORK OF ASNUI

As shown in FIG. 4, most of the existing ASNUI techniques have a unified framework [52], [83]. The main part of the user identification process can be divided into two steps:

Step 1: User account similarity calculation: FIG. 4 shows an example of two accounts on two SNs (S_1, S_2). Each network has one user. Given two accounts on two SNs, respectively, let TW_i be the user data on SN S_1, which can be denoted as a triple tuple $(V_{UPI}, V_{NTS}, V_{UGC})$, where V_{UPI} is the user attribute information, V_{NTS} is the network topology information, and V_{UGC} is the user behavior information. The similarity between two accounts is calculated using the methods that will be discussed in the following section, and is the input to *Step 2*.

Step 2: User account matching: The optimal matching scheme can be obtained by combining the correlation matching algorithm with the similarity between the accounts in *step 1*. The pruning filters [70] removes the wrongly matched pairs, and outputs the final matching pairs of user accounts on different SNs.

III. FUNDAMENTAL TECHNIQUES IN ASNUI

A. USER ACCOUNT SIMILARITY CALCULATION

1) USER ATTRIBUTE INFORMATION SIMILARITY CALCULATION

Since user data is stored as strings, the similarity value of a corresponding item of user data can be obtained by

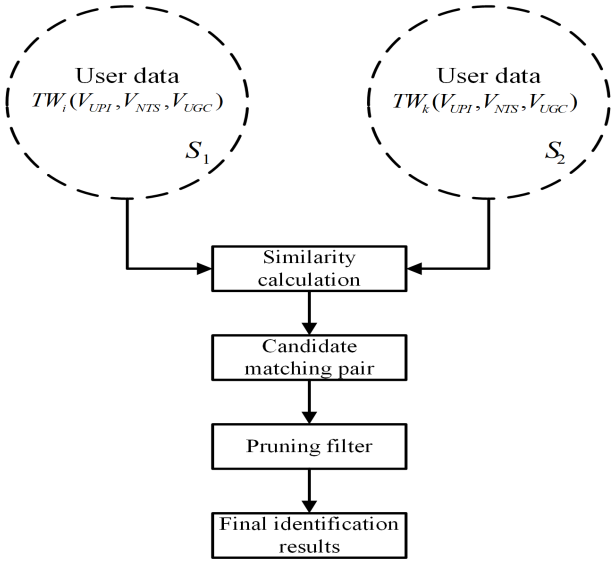

FIGURE 4. Basic framework of ASNUI.

calculating the similarity between string sequences. A common similarity calculation is as follows:

(1) Levenshtein Distance [14]: The number of character edit steps required to calculate the equality of two strings is used as an operational cost to measure the difference between strings. Given two strings n_i and n_j, their similarity is:

$$Simfunc(n_i, n_j) = 1 - \frac{d(n_i, n_j)}{\max(|n_i|, |n_j|)} \tag{1}$$

where $d(n_i, n_j)$ denotes the Levinshtein Distance between the strings n_i and n_j, and $\max(|n_i|, |n_j|)$ denotes the maximum value of the characters contained in the strings n_i and n_j.

(2) Dice Coefficient [15]: When calculating strings, they can be divided into two categories. When calculating the multi-valued strings n_i and n_j, the sum of the two times of the intersection information and divided by the sum of the elements of n_i and n_j yields the two strings of Dice coefficients, Then, their similarity is calculated using the Dice coefficient as follows:

$$Simfunc(n_i, n_j) = 2\frac{|n_i \cap n_j|}{|n_i| + |n_j|} \tag{2}$$

For example, in two multi-valued attribute strings ''vivid music movie'' and ''movie travel'', the intersection information is ''movie'', so the similarity is 2/5=0. 4.

For single-valued attribute strings, moreover, the Dice coefficient is calculated as above, except that the intersection information is different. For example, in the single-valued strings ''johe'' and ''joh'', the intersection information is ''jo, oh'', so the similarity is $4/5 = 0.8$.

(3) Jaro Distance [16]: This method is commonly used to measure the similarity of two strings, making it very suitable for the calculation of user name similarity. The formula for the Jaro Distance of the string n_i and n_j is as follows:

$$d(n_i, n_j) = \frac{1}{3}\left(\frac{m}{|n_i|} + \frac{m}{|n_j|} + \frac{m-t}{m}\right) \tag{3}$$

where m is the number of characters matched, and t is the number of transpositions.

The similarity of strings is calculated as follows:

$$Simfunc(n_i, n_j) = 1 - \frac{d(n_i, n_j)}{\max(|n_i|, |n_j|)} \tag{4}$$

(4) Matching Name (MN) distance: This method is a username matching algorithm [17] that first performs data preprocessing on the username, then combines the exact matching and the partial matching to obtain the final matching result.

(5) Term Frequency-Inverse Document Frequency (TF-IDF) [18]: This method is mainly utilized to measure the importance of a certain word in a certain document, and is often used to deal with multi-word attribute fields such as personal profiles. The specific steps involved are as follows:

Step 1: Calculate the term frequency (TF) of each word in the document:

$$TF = \frac{n}{N} \tag{5}$$

where n denotes the number of occurrences of a certain word, and N denotes the total number of words in the document.

Step 2: Calculate the inverse document frequency (IDF) of each word in the document:

$$IDF = \log\left(\frac{D}{P+1}\right) \tag{6}$$

where D denotes the total number of documents in the corpus, P denotes the number of documents containing a word in the document, and '1' is added to avoid cases in which the denominator is '0'.

Step 3: Calculate the TF-IDF of each word in the document:

$$TF - IDF = TF \times IDF = \frac{n}{N} \times \log\left(\frac{D}{P+1}\right) \tag{7}$$

Step 4: Select keywords in each document to construct a term frequency vector for calculating similarity.

Step 5: Calculate the similarity value by cosine similarity [19]:

$$\cos\theta = \frac{\sum_{i=1}^{k}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{k}(A_i)^2} \times \sqrt{\sum_{i=1}^{k}(B_i)^2}} \tag{8}$$

where A_i and B_i denote term frequency vectors.

2) NETWORK TOPOLOGY INFORMATION SIMILARITY CALCULATION

The user identification algorithm based on network topology information mainly relies on the similarity between different topology structure of the networks around two nodes to determine whether or not the accounts belong to the same entity user. That is, the more similar the network topology

TABLE 1. Similarity indicators based on common neighbors.

Name	Definition
Adamic-Adar indicator	$S(v_i^X, v_j^Y) = \sum_{z \in \Gamma(v_i^X) \cap \Gamma(v_j^Y)} \frac{1}{\log(\lvert\Gamma(z)\rvert)}$
Jaccard indicator	$S(v_i^X, v_j^Y) = \frac{\lvert\Gamma(v_i^X) \cap \Gamma(v_j^Y)\rvert}{\Gamma(v_i^X) \cup \Gamma(v_j^Y)}$
Resource Allocation (RA)	$S(v_i^X, v_j^Y) = \sum_{z \in \Gamma(v_i^X) \cap \Gamma(v_j^Y)} \frac{1}{\lvert\Gamma(z)\rvert}$
Salton indicator	$S(v_i^X, v_j^Y) = \frac{\lvert\Gamma(v_i^X) \cap \Gamma(v_j^Y)\rvert}{\sqrt{k(v_i^X) k(v_j^Y)}}$
LHN-I indicator	$S(v_i^X, v_j^Y) = \frac{\lvert\Gamma(v_i^X) \cap \Gamma(v_j^Y)\rvert}{k(v_i^X) k(v_j^Y)}$
Preferential Attachment (PA)	$S_{xy} = k_x \times k_y$

of network nodes, the greater the probability that the owners of the two accounts are the same in real life.

When calculating the similarity between node structures, it is necessary to represent the network topology of the node. In general, the set of a node's neighbor nodes can be used to represent the network topology. Given nodes v_i^X and v_j^Y from SNs X and Y, let $\Gamma(v_i^X)$ and $\Gamma(v_j^Y)$ represent the set of neighbor nodes of v_i^X and v_j^Y. There are many ways to calculate the similarity between network topologies, the most common of which is the Common Neighbor (CN) method [20]. The formula is as follows:

$$S(v_i^X, v_j^Y) = |\Gamma(v_i^X) \cap \Gamma(v_j^Y)| \tag{9}$$

There are also other similarity indicators, such as the Adamic-Adar indicator [21], the Jaccard indicator [22], Resource Allocation (RA) [23], the Salton indicator [24], the LHN-I indicator [25], Preferential Attachment (PA) [26], etc. The definitions of the above are in Table 1.

3) USER BEHAVIOR INFORMATION SIMILARITY CALCULATION

User identification based on user behavior information analyze the content published by users on SNs, then compare the similarity of behavior information between different social accounts to determine whether the user identities match or not. The most widely utilized method of calculating user behavior similarity is the Latent Dirichlet Allocation (LDA) model, a topic mining model [27], which is proposed based on Latent Semantic Analysis (LSI) [28] and Probabilistic Latent Semantic Analysis (PLSI) [29]. The basic idea is that each document can be considered equivalent to a mixed distribution of a series of topics, so that a three-layer Bayesian model of "document-topic-word" can be constructed [30]. Each document in the document set is categorized via probability distribution. According to the document generation rules and explicit data in the LDA model, the topic distribution is derived via expected value propagation [31]. The method of generating a document via the LDA model is represented in FIG. 5.

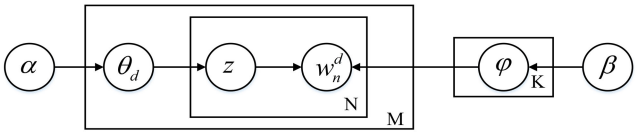

FIGURE 5. LDA model architecture.

Here, φ is the word distribution, θ_d is the topic probability distribution of the dth document, α is the parameter of the Dirichlet distribution of θ_d, β is the parameter of the Dirichlet distribution of φ, z denotes the topic of the word, and w_n^d denotes the nth word in the dth document. Moreover, M denotes the number of documents, N denotes the length of the documents, and K denotes the number of topics.

The document d can be represented as a joint distribution $P(z)$ of a series of topics, and the probability distribution of words on each topic is $P(w|z)$. The joint probability distribution of all words with their topics in the document is as follows:

$$P(w, z|\alpha, \beta) = \int P(w|z, \varphi) P(\varphi|\beta) d\varphi \cdot \int P(z|\theta) P(\theta|\alpha) d\theta \tag{10}$$

where d is posted by user.

The topic distribution of the microblog corpus can be obtained by using the maximum likelihood estimation method and combining account A with account B after modeling the microblog corpus using LDA. The probability distribution is used to measure the similarity of the topic distribution of the user accounts after the topic probability distribution has been obtained. KL divergence is often used to measure the similarity between the topic distributions, which is an asymmetry calculation method for calculating the degree of difference in probability distribution. For the vectors $a(i)$ and $b(i)$ of the two probability distributions, the KL divergence is calculated as follows:

$$D(A \| B) = \sum_i a(i) \ln \left[\frac{a(i)}{b(i)} \right] \tag{11}$$

Given two accounts v_x, v_y in the SN, then the vectors of their topic probability distribution are $a_x(\theta)$, $b_y(\theta)$, respectively. The similarity can be defined as:

$$S(A \| B) = \left\{ \sum_i a_{xi}(\theta) \cdot \ln \left[\frac{a_{xi}(\theta)}{b_{yi}(\theta)} \right] \right\}^{-1} \tag{12}$$

B. USER ACCOUNT MATCHING

Given the above calculated similarity values between two accounts on different SNs. The matching between the accounts can then be completed by using the relevant matching algorithm. When matching the accounts on the two SNs, a classic matching algorithm can be employed, such as a weighted bipartite graph matching algorithm [32], stable marriage matching [33], etc. Let n be the number of accounts to be matched, the time complexity of the bipartite graph maximum

weight matching algorithm is $O(n^3)$; By contrast, the time complexity of the stable marriage matching algorithm is $O(n^2)$, which is relatively low. However, the matching precision of this algorithm needs to be improved. Some researchers have improved the stable marriage matching algorithm. They proposed ranking-based cross-matching (RCM) algorithm. The purpose of the RCM algorithm is to accurately find more matching pairs, which decompose the seed user's identification into a step-by-step iterative process. In the iteration process of each step, the calculation process of the algorithm is divided into three stages: account selection, account matching and cross matching. The matched accounts are verified in each iteration, and the iteration stops when no user matched pairs are identified.

When the number of SNs being compared is greater than two, clustering methods are often used to find the initial matching accounts, and then thresholds are used to remove those pairs with small similarity values in each cluster. Since density-based spatial clustering of applications with noise (DBSCAN) does not require the number of clusters to be set in advance and can also identify noise points, DBSCAN is an ideal choice for clustering algorithms.

C. EVALUATION METRICS

In order to measure the effectiveness of different user identification algorithms and to compare their advantages and disadvantages, the most commonly used evaluation metrics are precision rate, recall rate and F-measure (F1). Since there are differences in the number of SNs targeted for user identification, the evaluation indicators can be divided into two categories: dual network evaluation metrics and multi-network evaluation metrics.

When the number of SNs is two, dual network evaluation metrics are used to evaluate the performance of the algorithm using the following formulae.

$$precision = \frac{tp}{tp + fp} \tag{13}$$

$$recall = \frac{tp}{tp + fn} \tag{14}$$

$$F1 = \frac{2 \times precision \times recall}{prcision + recall} \tag{15}$$

where tp denotes account pairs belonging to the same user and are correctly matched, fp denotes the number of pairs where the two corresponding accounts belong to different users but are identified as a matching pair, and fn denotes the number of users that are not matched but are the same users.

When the number of SNs is greater than two, multi-network evaluation metrics are adopted. The matching result $C = \{C_1, C_2, \ldots, C_n\}$ can be given by the identification algorithm. C_i is composed of several accounts in different SNs; the user identification algorithm then considers that the user accounts belonging to the same C_i match each other. The previously known matching result is denoted as $R = \{R_1, R_2, \ldots, R_m\}$, where R_i is also composed of several accounts in different SNs. The performance of the algorithm can be obtained by calculating the degree of matching between C and R. The multi-network evaluation metrics formulae are as follows:

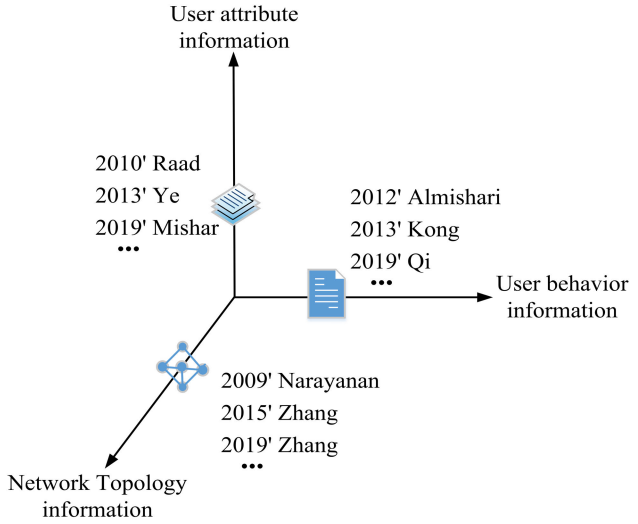

FIGURE 6. State of research into ASNUI.

$$precision = \frac{\sum_{i=1}^{N} tp_i}{\sum_{i=1}^{N} tp_i + \sum_{i=1}^{N} fp_i} \tag{16}$$

$$recall = \frac{\sum_{i=1}^{N} tp_i}{\sum_{i=1}^{N} tp_i + \sum_{i=1}^{N} fn_i} \tag{17}$$

$$F1 = \frac{2 \times precision \times recall}{prcision + recall} \tag{18}$$

where tp_i is the number of accounts contained in both C_i and R_i, fp_i is the number of accounts that are falsely put in C_i, and fn_i is the number of accounts that should be but have not been put in C_i.

IV. CLASSIFICATION AND PERFORMANCE EVALUATION OF ASNUI

User identification algorithms can be roughly classified into three categories according to the type of information used: user identification based on user attribute information, network topology information, and user-generated information. The current state of research on the three types of algorithms is summarized in FIG. 6.

A. USER ATTRIBUTE INFORMATION-BASED USER IDENTIFICATION

User attribute information refers to the data that a user needs to enter or select when registering a SN account, such as their username, gender, birthday, etc. Users can be easily identified using such personal information. This attribute information provides strong support for ASNUI. Moreover, the precision

of user identification will be improved if such attribute information can be obtained and utilized. There have been many studies on this topic, which can be divided into two categories depending on the way user attribute information is used for user identification: single attribute-based and multi-attribute-based methods.

1) SINGLE ATTRIBUTE-BASED USER IDENTIFICATION

User identification based on single attribute information mainly relies on user's single profile information, and the most commonly used attribute is username. Zafarani and Liu [34] first proposed this kind of method for user identification research, which is generally utilized to add or remove prefixes and suffixes of appellations and to map usernames from one community to another for user identification purposes. To verify the uniqueness of usernames, Perito *et al.* [35] have introduced a language-based model and a Markov chain technique by training the data of two SNs, after which they further estimated the rarity or commonality of the username via n-gram probability. Moreover, the edit distance is used to calculate the similarity between usernames, which verifies the intuitive observation. They also explains the high precision associated with the task of linking the public profile with the username. Then the username attribute is used to construct a training instance for user identification after the uniqueness of the username is analyzed. To connect user accounts across different SNs, the work in [36] combined user attributes with social activities and other relevant content to identify the entity users behind different accounts. For single attribute-based identification, most of the extracted features are not applicable when we identify only a priori username. Liu *et al*. [37] utilized the user naming patterns to extract the features required for user identification, and adopted support function to determine whether users have the same identity or not. However, the actual performance of support function on datasets has not achieved the desired results. In addition, in some SNs, such as QQ, Foursquare, user's username is a string of numbers automatically assigned by the SN. Under the above circumstances, existing username-based methods fail to achieve good identification performance.

Moreover, the username can consist of different characters. The researchers first modeled and analyzed the behavior patterns of users by defining some reasonable complex features when users select usernames, and then determined whether the two accounts were the same [38]. Wang *et al.* [39] conducted in-depth research on username attributes and extracted thousands of features such as alphanumeric combination features, date features, etc. As a username feature can be represented as a self-information vector, the similarity between vectors can thus be computed and analyzed via the relevant similarity calculation method to determine whether the entity users are the same. Li *et al.* [40] analyzed the differences in username naming on different SNs and constructed features that exploit information redundancies. The method of supervised machine learning was adopted to further confirm the identified matching pairs, which effectively increased the precision of the user identification method. However, even with such improved methods, user identification using a single attribute cannot achieve the required precision. Moreover, since a single piece of user attribute information has high potential for imitation and cannot correctly represent the actual situation of the user, there is a problem of poor robustness. However, the low computational complexity of single attribute information is also an advantage that cannot be ignored.

2) MULTI-ATTRIBUTE-BASED USER IDENTIFICATION

The username reflects a user's characteristics to a certain extent. However, many algorithms will exhibit a negative correlation trend in that the precision of identification decreases as the number of users increases when the identified dataset is too large. A solution to this critical problem requires the integration of multiple user information attributes. Vosecky *et al.* [17] first proposed to transform multiple attribute item information of users into n vectors, adopt different similarity calculation methods for each attribute item of users, and then select different matching weights for different attribute items. The disadvantage of this approach is that there is a tight coupling between the attribute and the domain, and weights assigned to attributes need to be recalculated whenever the application scenario changes. Motoyama and Varghese [41] have crawled and analyzed users' personal information on different SNs, represented it as a set of words, and then calculated the similarity between the words to obtain the similarity between different accounts. However, the multi-attribute user information is prone to forgery in SNs, which has a negative impact on the final identification results. Owing to the issue of user profile information having different formats across different SNs, the data format needs to be processed before the similarity between each attribute can be calculated. Raad *et al.* [42] designed a matching method based on the Friend-of-a-Friend (FOAF) vocabulary, transferred user profile data to the FOAF vocabulary, and implemented a decision algorithm to obtain the similarity between two social accounts. However, since e-mail address is used as a unique identifier in the user identification process, and this attribute cannot be easily obtained, the proposed algorithm suffers from the problem of poor universality.

Moreover, Iofciu *et al.* [43] jointly considered usernames and user tags and utilized a simple subjective weighting method to weight them. However, as it is necessary to redefine the user's attribute weights when the method is applied to a new SN, this increases the time complexity. Ye *et al.* [44] also proposed an objective weighting method based on subjective orientation to calculate the similarity among multiple user attributes; however, this method relies on a lot of sample data, and is not universal in practical applications. To solve the above problems, an objective weighting method based on information entropy was proposed, which use the entropy value of user attributes to assign weights to each attribute item. This method has high rationality in assigning weights

to multiple user attributes. However, this needs to calculate a weight allocation scheme for each source account in the SN, and thus its computational complexity increases drastically when there are more users in the SNs. Ma *et al.* [32] proposed a joint learning model that combines user profile information, online time distribution and interest to analyze the similarities between user accounts and adjust the weight of the information used by balancing factors. In addition, they also designed a KM algorithm-based user identification method and weighted bipartite graph maximum matching, and extended the application of KM algorithm. In order to reduce the amount of calculation, they set a threshold for user profile similarity to prune and filter some error matching pairs, which improves the performance of user identification.

After obtaining the similarity vector between different user accounts, a classifier is used to determine whether the user accounts have the same identity. Goga *et al.* [45] selected the multi-attribute user information on five SNs for user identification, and trained the data via supervised classification models such as decision tree [46], SVM [47], etc. To verify which classification model produced better results, experiments were conducted on a truth dataset showed that the performance of the Bayesian classification model is similar to that of other models in terms of true positive rates (TPR) and false positive rates (FPR). However, the calculation time is much shorter than other models. Accordingly, this paper selects the Bayesian model as the final classification result. Zamani *et al.* [48] took the user's unit, interests and other attribute information into full consideration, and integrated the similarity of multi-user attributes via the equal evaluation model and complex mixed training model. Experimental results proved that it was possible to correctly identify users due to the personalized characteristics of many users' attribute information. Esfandyari *et al.* [49] proposed an overlapping attribute items method to select user information that would make the trained model more applicable when compared to the traditional random selection method; however, this method increases the computational complexity. Considering complexity is an important indicator of user identification performance. Yin *et al.* [50] developed a Probabilistic Tensor Factorization (SPTF) model for processing user data, and designed a novel negative sampling method to optimize the model by utilizing both observed and unobserved examples with much lower computational complexity and higher modeling precision. This method is good reference to current studies, and can be applied to classifier training labeled and unlabeled instances.

By analyzing the above two methods of user identification. Peled *et al.* [51] found through analysis that users utilize the same or different profile information to register accounts on multiple SNs. They fused the available data to users into a single profile and exploited machine learning to implement user identification. In addition, they designed a classifier by using the acquired features and monitoring techniques, and tested it on a ground truth dataset. The AUC value reached 98.2%, which fully demonstrated the effectiveness of the proposed method. Since the increased difficulty in obtaining user profile information and the consideration of computational complexity, more and more studies have begun to use a small amount of profile information to identify users. Li *et al.* [52] proposed a user identification across SNs based on username and display name (UISN-UD) model using usernames or display names [53], which contain rich information redundancy, and the proposed method could conceivably reduce the use of attributes as well as the degree of computational complexity. The most prominent advantage of this approach is that it protects personal privacy and is highly accessible. The comprehensive evaluation index exceeds 90%. Although these solutions can achieve good performance, the biggest challenge is authenticity and integrity.

To optimize the performance of user identification process, Ma *et al.* [54] proposed a new joint method called MapMe that takes user profile information and network structure characteristics into full consideration to improve the universality of the method. This method uses Doc2vec to convert user profile information into vectors and utilizes the corresponding calculation methods to obtain similarity. Then, it analyzes the user's ego network features to obtain similarity. Finally, a smooth factor α is used to balance the features used to achieve better identification. He and Li [55] proposed a dynamic preferences-based identification method (DPUI). They analyzed the preferences of the user's naming display name and extracted five features to measure similarity. Furthermore, user interest is one of the factors they consider. The user's topic distribution is obtained through the LDA model, and long-term interests are selected as the features of the analysis. Finally, they combined the display name with long-term interests to identify the user and achieved good results in terms of precision rate and recall rate. Mishra [56] designed a system that employs pairwise comparison string matching method. The method analyzes multiple profile information of users on LinkedIn and Facebook, and uses corresponding similarity calculation methods to measure the similarity value of user information in different dimensions. Finally, matching pairs are generated by comparing with the set threshold. Since users usually do not provide their real and complete attribute information due to privacy concerns, methods based on user attributes only cannot obtain good matching results when the precision of the attribute information is not guaranteed.

B. NETWORK TOPOLOGY INFORMATION-BASED USER IDENTIFICATION

ASNUI based on network topology information refers to methods in which the friend relationships between users are treated as equivalent to the network topology so that similarity matching between nodes can be performed. Friendship relationships can be easily obtained through an open application programming interface (API) in SNs.

Narayanan and Shmatikov [57] proved for the first time that user identification could be accomplished by relying on network topology information. Starting from a small number of known seed nodes, iterative updating is then used to find

new matching nodes, after which ASNUI between two SNs can be realized. However, the precision rate and recall rate of user identification via this method needs to be improved. The work in [58] proposed to combine the user's profile information with the similarity of the graph to achieve mapping from an email network to a Facebook network. However, this mapping relationship has a one-to-one mapping conflict problem, which affects the precision of the final user identification results. To solve these problems, Kong *et al.* [59] transformed the problems identified in the literature [58] into prediction problems of directed links. However, the proposed method also has some disadvantages, such as mapping only for matching pairs, limited application scenarios (i.e., they assume heterogeneous information in the SNs is available, and only use social links, location and temporal distributions to infer the account similarity. User attribute information is excluded in their study), etc.

In addition, Korula and Lattanzi [60] abstracted the user identification problem into mathematical form, arguing that different SNs are generated by user graph structure through probability. Their work also considered that the selection process of graph edges in the network topology is approximate probability and has a cascading effect. Using this mathematical thinking, they utilized the iterative method to determine whether two accounts had the same entity user. Tan *et al.* [61] proposed the concept of hypergraph, and a novel subspace learning method, manifold alignment on hypergraph (MAH), is designed. They model high-order relations via hypergraph. For a target user in a SN, the proposed method ranks all users in the other SN by their possibilities of being the corresponding user. Moreover, methods based on username comparison can be incorporated into the proposed algorithm easily to further improve the mapping precision and map users in SNs to a low-dimensional space to reduce the complexity. Zhang *et al.* [62] proposed a method based on the energy model COSNET. This method utilizes energy level to distinguish between different user attributes and structural matching methods, and also uses a sub-gradient algorithm to train the energy model: the energy is lowest when the obtained matching result is optimal. In this way, these researchers transformed the user-matching problem into the parameter-seeking problem of an energy model, and also the dual problem decomposition thinking is used to improve the precision of user account matching.

With the development of machine learning, network-embedding techniques are gradually applied to ASNUI. Liu *et al.* [63] developed IONE algorithm, which embeds user accounts into low-dimensional spaces for user identification. A similar study designed a model based on embedding and matching [64], whose network embedding is purely unsupervised, and the observed anchor links are not used when encoding the network structure as embeddings. However, neither of the two methods can fully utilize network structure attributes. To address this problem, Zhang and Yu [65] formulated the similarity of friendship as a graph alignment problem and proposed a novel unsupervised multi-network alignment (UMA) framework. This paper also studies the inference of relationship type in cooperative networks, which will help to measure the similarity of friendship networks and achieve better identification. Compared with profile information and behavioral information, the friendship networks are more difficult to forge. Recently, Wang *et al.* [66] designed a semi-supervised network-embedding model, in which the nodes in SNs are embedded into low-dimensional spaces. On the one hand, the low-dimensional form of the node can be used to predict its context in the network, i.e., random walk sequences generated by network structure and adjacent nodes that share same attribute information. On the other hand, it is utilized to predict anchor links between SNs (semi-supervised). Nodes (users) with similar context and attribute information are very close in the embedding space.

Nitish and Silvi [60] designed an efficient SN propagation algorithm whose basic principle is to identify user identities based on neighbor nodes. For the first time, the author formulates the user identification problem in mathematical form. The proposed method adopts the idea of recursive algorithm and uses shared friends to analyze the friendship between two different users. Similar to literature [67], Zhou *et al.* [67] utilized the number of seed nodes shared by user nodes as a measure of similarity across different SNs, where the ones with the largest similarity were selected for matching. This method verifies that user identification can be better accomplished based on network topology information. The proposed friend relationship-based user identification (FRUI) algorithm can be applied to multiple online SNs with friend relationships. However, the user matching results obtained by these algorithms depend on the seed nodes. Accordingly, to address cases in which seed nodes are not available, Zhou *et al.* [68] designed an unsupervised scheme termed friend relationship-based user identification algorithm without prior knowledge (FRUI-P). The continuous bag-of-words (CBOW) model based on the negative sampling technique is adopted to learn the network vector by learning from the thinking of a random walk. The algorithm extracts the friend features of each account in the SN as a feature vector, then calculates the similarities between all candidate users across the two SNs via in- and out-degrees [69]. The main advantage of this method is that it does not need to know seed nodes and can provide reliable prior knowledge for user identification.

With the deepening of relevant research work. Qu *et al.* [70] used the user's profile similarity to obtain some valid priori matching pairs, and then proposed an identification algorithm based on friendship learning. The algorithm calculates the matching degree between two different users based on the prior knowledge, adopts the gradient descent algorithm to optimize the weight of the features used and thereby achieves accurate account matching. Zhang *et al.* [71] developed a graph neural network framework-based user identification framework. They encode the graph topology formed by the SN as a node feature. This feature learning process is called node embeddings. The purpose of embedding is to map the network structure

to the low-dimensional node space, in which both local and global graph connection patterns are preserved, so that the reconstruction SN based on the learning node features is close to the original SN. Moreover, they present a deep graph model to learn node embeddings of some large SNs and construct a non-linear mapping of nodes with the same identity across SNs. This semi-supervised learning method helps to effectively identify users in actual cases.

User identification based on network topology information mainly involves representing relationships between users as a network node graph, then realizing the task of graph matching between different SNs. Graph matching refers to that via a method to measure the similarity between graphs, which is then combined with the matching algorithm to select the best matching results from the massive amount of available data. To date, this method has been applied in many fields [72] and has helped complete some specialized tasks in the field of social security [73]. However, due to the heterogeneity of real SNs, which is often ignored in such methods, this type of approach has an impact on the precision of user identification.

C. USER BEHAVIOR INFORMATION-BASED USER IDENTIFICATION

User behavior information refers to all kinds of actions that users take on different SNs, such as commenting, forwarding and liking, which reflects their interests and hobbies. The user-generated behavioral data further expands the information dimension of ASNUI. If this data can be exploited appropriately, it can provide a new direction for user identification.

Almishari and Tsudik [74] took advantage of the different writing styles of users to connect them to different online SNs, which verifies the linkability between different SNs. However, this method has the problem of violating user's privacy and security. In order to improve the linkability between SNs, prior works in [99], they proposed to integrate user attribute information, user behavior information and other social behaviors to further improve the precision of SN links. Tu *et al.* [75] developed a profession-based identification model according to user's behaviors in SNs. Compared with user profile information, user behavior information is not easy to fake, and more credible. However, the proposed method requires natural language processing algorithms, so its implementation is more complicated than other methods. Nie *et al.* [76] subsequently proposed a dynamic core interest mapping algorithm (DCIM), which considers user topology and topic model based on user-generated content and ego-networks. This algorithm has mainly been used to analyze the dynamic rules of user interests. However, the methods proposed above are restricted to certain specific online SNs, making it impossible to generalize to all online SNs. Sha *et al.* [77] utilized status and comments posted by users to implement user identification across multiple SNs. These authors represented the user-generated text content as a vector via Doc2Vec, then performed user identification by calculating the similarity between these vectors. By starting with user-published content, together with the mining of other relevant behavior information, it is possible to achieve better user identification results.

Furthermore, Kong *et al.* [59] proposed a multi-network anchoring (MNA) algorithm by using multiple SNs to make directed anchor links in order to map multiple virtual accounts of users across different SNs. They obtained four kinds of information from the SN, namely geographical location, published content, check-in time, and published content topics and links. Then, a SVM classifier was trained by using the corresponding feature values of different information types, which was followed by the calculation of the similarity of different accounts in time, space, and other metrics across different SNs. In order to guarantee the constraint of one-to-one mapping, the algorithm gives priority to matching users with high matching scores. If the matching is successful, then other users will not be considered, which improves the precision of the matching result. However, this algorithm is not guaranteed to be completely error-free; as the number of users becomes larger, the similarity between users will also increase, which will have a certain impact on the final matching result. Roedler *et al.* [78] used the timestamp information generated by users on SNs in conjunction with the location information generated by mobile devices to construct a personalized social behavior pattern to solve the problem of user identification. The relevant research works are also reflected in the literature [79]–[81]. However, due to the privacy protection applied to user attribute information, the obtained user attribute data is often incomplete. To address this issue, Zhao *et al.* [82] proposed a semantic-based BM25 (Bag-of-words retrieval function) method, which is mainly used to calculate the semantic similarity between two different tags. The method selects the candidate matching pairs by calculating the similarity of the user display names, and then combines the extracted label features with the greedy algorithm to optimize the candidate matching pairs and output the matching results. Li *et al.* [83] designed a user generated content-based user identification model (U-UIM), in which several algorithms are developed to measure the similarity of UGC in space, time and content dimensions. Moreover, supervised machine learning algorithms were used to match users, which improved the comprehensive user identification performance. Recently, Chen and Tan [84] proposed the concept of fuzzy time window, mapping user access behavior to different time windows, extracting time, text and sequence features of user access behavior in web access logs, and fusing the extracted features to obtain semantically rich web usage contexts. Finally, multi-layered perception network-based user behavior identification model is constructed to determine the user identity.

Some recent works also employ user-generated trajectory information for user identification. Since the user trajectory information reflects the movement trajectory of the user in real life, it can be regarded as a form of behavior information reflecting the identity of the user. Algorithms for user identification that exploit user trajectory information have gradually attracted the attention of researchers. Cao *et al.* [85] proposed

a user identification algorithm to process multi-source location data by observing the co-occurrence frequency of positions across two user trajectories. However, this algorithm employs too many parameters, and parameter tuning becomes challenging, and its computational complexity increases drastically as well. Hao *et al.* [86] proposed to transform user trajectories into sequence of multiple grids, which are in turn transformed into vectors using a TD-IDF model. Then, the similarity of user trajectories is calculated via cosine similarity to determine whether the accounts match. Han *et al.* [87] proposed that each geographic coordinate point should be represented as a corresponding semantic position. In this way, the user's trajectory could be represented by the text composed of the semantic position. The LDA model was used to represent the user's topic distribution. Finally, the similarity of the user trajectories was calculated via KL divergence to determine whether the two users are the same.

Furthermore, the location that users often appear is easily identifiable. Riederer *et al.* [88] assumed that the number of times a user visited a certain geographic location within a certain time is subject to a Poisson distribution, and used a probability function to represent the probability that two accounts were the same entity user. Then, the objective function was optimized to obtain the best matching result. Han *et al.* [89] combined the spatial and temporal information of the user trajectory and converted the original multi-source spatio-temporal data into a tripartite graph. The best user matching scheme was obtained by finding the optimal partition of this graph. Qi *et al.* [90] proposed an identification solution based on the most frequently distributed TOP-N (the most frequently distributed N regions) regions of user trajectories. They first find TOP-N regions whose user's trajectories are most frequently distributed to reduce the computational complexity. Then, the similarity method based on probability deviation, angle cosine, as well as weighted Jaccard similarity are employed to calculate similarity of two trajectories and thereby achieve user identification.

Through the above-mentioned three ASNUI technologies are discussed and compared. We can clearly see that ASNUI has important applications in many fields. However, the issue of user privacy risk in the process of user identification is also a research direction that can not be ignored. Backstrom *et al.* [91] proposed a sub-graph search mode-based active and passive attack method in SNs to learn the relationship between friends. The method processes data required for user identification from the perspective of user privacy risk. However, this method is effective in identifying relationships between nodes in small-scale SNs, but not in large-scale SNs. Zhou *et al.* [92], [93] analyzed the neighborhood-based de-anonymity attack method, and proposed k-anonymity and l-diversity methods to protect user privacy. Hu *et al.* [94] proposed a collaborative management of user data method in SNs, which is a flexible mechanism. This mechanism provides conflict resolution in both privacy risk and data sharing.

Unfortunately, when the attackers gains lots of generated information from the users, these anonymization techniques are vulnerable, so that the user's personal information might be leaked. most existing approaches assume specific and restrict network structure as background knowledge and ignore semantic level prior belief of attackers, which are not apply to any privacy scenarios. To address these problems, Qian *et al.* [95] presented that knowledge graph is an effective model to achieve ASNUI and solve user privacy issues in SNs. Aggregating user information across multiple SNs will inevitably lead to serious privacy leakage problem. Du *et al.* [96] designed an attack-defense tree-based privacy risk analysis model to describe user privacy protection strategy. What they do is unlike any previous studies, which focuses on how to model and measure privacy risk in SNs. To our knowledge, most of the SNs are now focusing on the privacy of users, some sensitive data of users cannot be obtained without the authorization of users. When we elaborate on ASNUI technology, we also clearly stated that more and more technologies only use accessible data in the process of user identification, such as display name, username and other public data. However, the balance between ASNUI and user privacy protection is still a future research direction, and we have a description in Section V.

D. PERFORMANCE EVALUATION OF ASNUI

ASNUI should also take into account the availability of services and the computational overhead. Accordingly, in this paper, we evaluate the performance of ASNUI from the following three aspects:

(1) Degree of user identification: This is usually reflected by the evaluation metrics mentioned in subsection 3. 3. The higher the precision rate, recall rate and F1 value in the evaluation metrics, the higher the degree of user identification.

(2) Availability of services: This refers to the accurateness and timeliness of ASNUI, which in turn reflects the quality of service that users obtain after user identification. There is a trade-off between the availability of services and the degree of privacy protection, as improving the degree of privacy protection sometimes reduces the availability of services.

(3) Calculation overhead: The computational overhead of ASNUI includes pre-computation, storage required at runtime, and computational costs. The storage cost mainly arises in pre-computation, which is generally acceptable, and is ignored in ASNUI. Computation costs at runtime can be measured in terms of the time complexity of the algorithm.

ASNUI has a wide range of applications, and is an emerging research direction. This paper surveys recent progress in this area, summarizes existing research results, and introduces three types of user identification algorithms. It can thereby be seen that each type of identification algorithm has its own unique characteristics. The analysis and comparison results of various user identification algorithms for different application requirements are listed in Table 2. As shown in Table 2, the scope of their application and performance

TABLE 2. Performance evaluation of ASNUI.

Techniques	Degree of user identification	Calculation overhead	Missing data
User attribute information-based user identification	high	medium	high
Network topology information-based user identification	high	high	low
User behavior information-based user identification	high	low	low

TABLE 3. Comparative analysis of ASNUI.

Techniques	Advantages	Disadvantages	Representative methods
User attribute information-based user identification	High identification precision and simple implementation	Amount of missing data is very high; user information is easy to forge	FOAF framework [42] MADM [44] UISN-UD [52]
Network topology information-based user identification	Data integrity and easily obtained information	Network heterogeneity; high complexity	Re-identification algorithm [57] COSNET model [62] FRUI-P [68]
User behavior information-based user identification	Perfect data and low complexity	Some users lack effective behavioral information	Bayesian model [74] MNA [59] U-UIM [83]

can vary. The following techniques have shown good performance in terms of Degree of user identification. The performance of user behavior information-based user identification is the best in terms of computational overhead. Since user attribute data missing in the three data types mentioned is more serious, the missing data needs to be filled to better achieve the user identification. Moreover, we provide a further comparative analysis of ASNUI in Table 3.

V. FUTURE WORKS

In the era of big data, the channels of access to information and the types of user information are also becoming more diverse. This section points out future research directions in three aspects: weight allocation of user attribute information, fusion of multi-dimensional information, and large-scale user identification.

A. WEIGHT ALLOCATION OF USER ATTRIBUTE INFORMATION

Since different attribute items have different influences on the degree of user identification, it is necessary to analyze how to optimally assign weights to different attribute items. The traditional weight allocation algorithm [44] has some limitations in terms of robustness and universality. Some similar works also have identified each source account via variant entropy values. However, under these circumstances, each instance of user identification requires redefinition of the attribute weight allocation scheme. When the number of accounts to be matched in the SN is large, the computational complexity required to determine the weight of the user attributes will also increase.

$$W(x) = -p(y_s|s) \sum_{x \in X} p(x) \log(p(x)) \tag{19}$$

where $W(x)$ denotes the attribute weight value, $p(x)$ is the possible value probability for the attribute, $X = \{x_1, x_2, \ldots, x_n\}$ denotes a set of n attributes of user, and $p(y_s|s)$ is the posterior probability of the attribute.

In information theory, the entropy value reflects the degree of information uncertainty. The smaller the entropy value is, the more orderly the information is, and the more valuable this attribute is; on the contrary, the more disordered the information is, the lower the value of this attribute is. Therefore, information entropy can be used to evaluate the effectiveness of the attributes employed. The entropy value is obtained by calculating the probability of the user attribute. In order to make the probability description of attributes more precise, and more effective weights are assigned to each attribute. On the basis of information entropy, the posterior probability of user attributes is further calculated, which aids in improving the precision of user identification. By combining the posterior probability and information entropy, the attribute information weight of the user can be expressed as follows:

The weight assignment can be effectively performed for a given attribute by combining the posterior probability of the user attributes with the information entropy. When the identified user changes, the posterior probability of the corresponding attribute item is constant. In this way, the amount of calculation in the identification process can be greatly reduced.

B. FUSION OF MULTI-DIMENSIONAL INFORMATION

User identification based on the fusion of multi-dimensional information refers to comprehensively utilizing two or three types of user data for user identification purposes. For some special applications, it may be necessary to identify the user with a high level of precision. In these cases, the use of single-dimension information for user identification has certain limitations. To address this issue, recent works in [97], this work have considered non-friend relationships when exploiting network topology for user identification by proposing an intimacy function, which is used to determine the importance of both friend and non-friend relationships in order to identify users. Moreover, some matching algorithms have been adopted to unify the user attribute information and the link relationships, including friend and non-friend relationships, in a further effort to solve the problem of user identification.

Recent studies in [98] and [99] have also considered the information interaction among network topology, user attribute information, and user behavior information to achieve accounts matching across different SNs. We believe that ASNUI that integrate multi-dimensional user

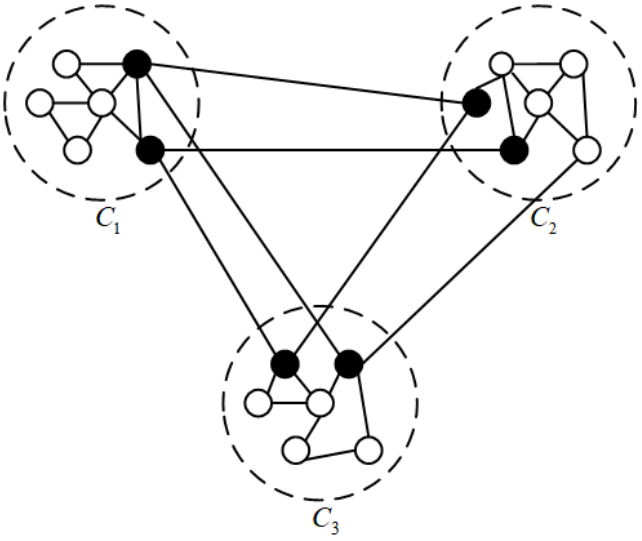

FIGURE 7. Community detection-based user node structure.

information are an important part of future research in user identification. Furthermore, in order to prevent malicious users from attacking normal users, the tradeoff between user privacy and user identification is also one of the future research directions that should be pursued.

C. LARGE-SCALE USER IDENTIFICATION

When the network size increases, the precision will deteriorate. Community detection in complex networks can effectively divide large-scale users into different communities for user identification. The application of community detection technology to ASNUI is a potential research direction for the future.

Definition 6 (Community): A set of all users on a SN, each community consisting of users with the same characteristics. These communities have close internal links, and there are sparse links between communities [34].

Given a SN $G = (U, E, A)$, U represents a set of users, E represents a set of relationships between users, and A represents a set of user attributes. $C_i \subseteq U$ represents a community, and $C\{C_1, C_2, \ldots, C_n\}$ represents the community division of n communities belonging to a single SN.

As shown in FIG. 7, in this approach, users on a SN are represented as nodes in the network topology. These nodes are then divided into communities. The seed users of each community are represented as black nodes, and the purpose of using community discovery is to maximize the identification of unmatched white nodes. By doing so, the negative correlation between the precision rate and the number of users exhibited by existing algorithms can be better balanced. Therefore, the application of community discovery technology to ASNUI is a promising research direction.

VI. CONCLUSION

User data will be collected from different perspectives and channels in the era of big data. ASNUI has revolutionized the ways in which we live and do business, as well as our scientific research. To date, researchers have expended significant effort on the topic of ASNUI. This paper reviews the results of research into ASNUI, surveys the state of the art of user identification, and summarizes the model and basic framework of ASNUI. We categorize the current research works according to the different types of user information employed, then analyze and compare these three types of algorithms. Moreover, we further describe potential directions for future research in the field. In conclusion, ASNUI is an emerging research field brought about by the rise of big data, and many key issues in this field that require in-depth and meticulous research.

REFERENCES

[1] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explor. Newslett.*, vol. 18, no. 2, pp. 5–17, 2017.
[2] X. Liu, T. Xia, Y. Yu, C. Guo, and Y. Sun, "Cross social media recommendation," in *Proc. Int. AAAI Conf. Web Social Media*, 2016, pp. 1–10.
[3] R. Zafarani, L. Tang, and H. Liu, "User identification across social media," *ACM Trans. Knowl. Discovery Data*, vol. 10, no. 2, 2015, Art. no. 16.
[4] N. Yuan, F. Zhang, D. Lian, S. Yu, and X. Xie, "We know how you live: Exploring the spectrum of urban lifestyles," in *Proc. 1st ACM Conf. Online Social Netw.*, 2013, pp. 3–14.
[5] M. Yan, J. Sang, T. Mei, and C. S. Xu, "Friend transfer: Cold-start friend recommendation with cross-platform transfer learning of social knowledge," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2013, pp. 1–6.
[6] E. Zhong, W. Fan, and Q. Yang, "User behavior learning and transfer in composite social networks," *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 1, 2014, Art. no. 6.
[7] W. Shen, J. Wang, P. Luo, and M. Wang, "Linking named entities in Tweets with knowledge base via user interest modeling," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 68–76.
[8] J. Zhang, P. Yu, and Z. Zhou, "Meta-path based multi-network collective link prediction," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1286–1295.
[9] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks," in *Proc. 4th Int. Conf. Weblogs Social Media*, 2010, pp. 90–97.
[10] S. B. Wicker, "The loss of location privacy in the cellular age," *Commun. ACM*, vol. 55, no. 8, pp. 60–68, 2012.
[11] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy*, May 2008, pp. 111–125.
[12] L. Xing, Q. Ma, J. P. Gao, and S. Chen, "An optimized algorithm for protecting privacy based on coordinates mean value for cognitive radio networks," *IEEE Access*, vol. 6, no. 1, pp. 21971–21979, 2018.
[13] L. Xing, Q. Ma, H. Wu, and P. Xie, "General multimedia trust authentication framework for 5G networks," *Wireless Commun. Mobile Comput.*, vol. 2018, Jun. 2018, Art. no. 8974802.
[14] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1091–1095, 2007.
[15] G. Kondrak, D. Marcu, and K. Knight, "Cognates can improve statistical translation models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2003, pp. 46–48.
[16] D. Caliano, E. Fersini, P. Manchanda, M. Palmonari, and E. Messina, "UniMiB: Entity linking in tweets using Jaro-winkler distance," in *Proc. 6th Int. Workshop Making Sense Microposts*, 2016, pp. 70–72.
[17] J. Vosecky, D. Hong, and V. Y. Shen, "User identification across multiple social networks," in *Proc. 1st Int. Conf. Netw. Digit. Technol.*, Jul. 2009, pp. 360–365.
[18] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Document.*, vol. 60, no. 5, pp. 493–502, 2004.
[19] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 709–720.
[20] J. Rexford and C. Dovrolis, "Future Internet architecture: Clean-slate versus evolutionary research," *Commun. ACM*, vol. 53, no. 9, pp. 36–40, 2010.

[21] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Soc. Netw.*, vol. 25, no. 3, pp. 211–230, 2003.
[22] A. Greenhalgh, F. Huici, M. Hoerdt, P. Papadimitriou, M. Handley, and L. Mathy, "Flow processing and the rise of commodity network hardware," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 2, pp. 20–26, 2009.
[23] E. Maccherani, M. Femminella, J. W. Lee, R. Francescangeli, J. Janak, G. Reali, and H. Schulzrinne, "Extending the NetServ autonomic management capabilities using OpenFlow," in *Proc. IEEE Netw. Oper. Manage. Symp.*, Apr. 2012, pp. 582–585.
[24] H. Kim and N. Feamster, "Improving network management with software defined networking," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 114–119, Feb. 2013.
[25] Z. A. Qazi, C.-C. Tu, L. Chiang, R. Miao, V. Sekar, and M. L. Yu, "SIMPLE-fying middlebox policy enforcement using SDN," in *Proc. ACM SIGCOMM Conf. SIGCOMM*, 2013, pp. 27–38.
[26] S. Milojević, "Modes of collaboration in modern science: Beyond power laws and preferential attachment," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 7, pp. 1410–1423, 2010.
[27] L. He, Y. Jia, W. Han, and Z. Ding, "Mining user interest in microblogs with a user-topic model," *China Commun.*, vol. 11, no. 8, pp. 131–144, Aug. 2014.
[28] G. Recchia and M. N. Jones, "More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis," *Behav. Res. Methods*, vol. 41, no. 3, pp. 647–656, 2009.
[29] M. Masseroli, D. Chicco, and P. Pinoli, "Probabilistic latent semantic analysis for prediction of gene ontology annotations," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–8.
[30] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 542–550.
[31] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proc. 18th Conf. Uncertainty Artif. Intell.*, 2002, pp. 352–359.
[32] J. Ma, Y. Qiao, G. Hu, T. Li, Y. Huang, Y. Wang, and C. hang, "Social account linking via weighted bipartite graph matching," *Int. J. Commun. Syst.*, vol. 31, no. 7, p. e3471, 2018.
[33] S. Modi, N. M. Shagari, and B. Wadata, "Implementation of stable marriage algorithm in student project allocation," *Asian J. Res. Comput. Sci.*, vol. 1, no. 4, pp. 1–9, 2018.
[34] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *Proc. Int. AAAI Conf. Weblogs Social Media*, vol. 9, 2009, pp. 354–357.
[35] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *Proc. Int. Symp. Privacy Enhancing Technol. Symp.*, 2011, pp. 1–17.
[36] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name?: An unsupervised approach to link users across communities," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 495–504.
[37] D. Liu, Q. Y. Wu, W. H. Han, and B. Zhou, "User identification across multiple Websites based on username features," *Chin. J. Comput.*, vol. 38, no. 10, pp. 2028–2040, 2015.
[38] R. Zafarani and H. Liu, "Connecting users across social media sites: A behavioral-modeling approach," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2013, pp. 41–49.
[39] Y. Wang, T. Liu, Q. Tan, J. Shi, and L. Guo, "Identifying users across different sites using usernames," *Procedia Comput. Sci.*, vol. 80, pp. 376–385, Jan. 2016.
[40] Y. Li, Y. Peng, W. Ji, Z. Zhang, and Q. Xu, "User identification based on display names across online social networks," *IEEE Access*, vol. 5, pp. 17342–37353, 2017.
[41] M. Motoyama and G. Varghese, "I seek you: Searching and matching individuals in social networks," in *Proc. 11th Int. Workshop Web Inf. Data Manage.*, 2009, pp. 67–75.
[42] E. Raad, R. Chbeir, and A. Dipanda, "User profile matching in social networks," in *Proc. 13th Int. Conf. Netw.-Based Inf. Syst.*, Sep. 2010, pp. 297–304.
[43] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 522–525.
[44] N. Ye, L. Zhao, L. Dong, G. Bian, E. Liu, and G. J. Clapworthy, "User identification based on multiple attribute decision making in social networks," *China Commun.*, vol. 10, no. 12, pp. 37–49, 2013.
[45] O. Goga, D. Perito, H. Lei, R. Teixeira, and R. Sommer, "Large-scale correlation of accounts across social networks," Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep. TR-13-002, 2013.
[46] H. X. Li, H. J. Zhu, S. G. Du, X. H. Liang, and X. M. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 646–660, Jul./Aug. 2018.
[47] R. H. Veni, A. H. Reddy, and C. Kesavulu, "Identifying malicious Web links and their attack types in social networks," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 3, no. 4, pp. 1060–1066, 2018.
[48] K. Zamani, G. Paliouras, and D. Vogiatzis, "Similarity-based user identification across social networks," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, 2015, pp. 171–185.
[49] A. Esfandyari, M. Zignani, S. Gaito, and G. P. Rossi, "User identification across online social networks in practice: Pitfalls and solutions," *J. Inf. Sci.*, vol. 44, no. 3, pp. 377–391, 2016.
[50] H. Yin, H. Chen, X. Sun, H. Wang, Y. Wang, and Q. V. Nguyen, "SPTF: A scalable probabilistic tensor factorization model for semantic-aware behavior prediction," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2017, pp. 585–594.
[51] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Matching entities across online social networks," *Neurocomputing*, vol. 210, pp. 91–106, Oct. 2016.
[52] Y. Li, Y. Peng, Z. Zhang, H. Yin, and Q. Xu, "Matching user accounts across social networks based on username and display name," *World Wide Web*, vol. 22, no. 3, pp. 1075–1097, 2018.
[53] Y. Li, Y. Peng, Z. Zhang, Q. Xu, and H. Yin, "Understanding the user display names across social networks," in *Proc. Int. World Wide Web Conf. Committee (IW3C2)*, 2017, pp. 1319–1326.
[54] J. Ma, Y. Qiao, G. Hu, Y. Huang, M. Wang, A. Sangaiah, C. Zhang, and Y. Wang, "Balancing user profile and social network structure for anchor link inferring across multiple online social networks," *IEEE Access*, vol. 5, pp. 12031–12040, 2017.
[55] Z. He and W. Li, "Research on user identification across multiple social networks based on preference," in *Proc. 5th IEEE Int. Conf. Cloud Comput. Intell. Syst.*, Nov. 2018, pp. 122–128.
[56] R. Mishra, "Entity resolution in online multiple social networks (@Facebook and LinkedIn)," in *Emerging Technologies in Data Mining and Information Security*, vol. 813. Singapore: Springer, 2019, pp. 221–237.
[57] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proc. 30th IEEE Symp. Secur. Privacy*, May 2009, pp. 173–187.
[58] Y. Cui, J. Pei, G. Tang, W.-S. Luk, D. Jiang, and M. Hua, "Finding email correspondents in online social networks," *World Wide Web*, vol. 16, no. 2, pp. 195–218, 2013.
[59] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 179–188.
[60] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *VLDB Endowment*, vol. 7, no. 5, pp. 377–388, 2014.
[61] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen, "Mapping users across networks by manifold alignment on hypergraph," in *Proc. 28th AAAI Conf. Artif. Intell.*, vol. 14, 2014, pp. 159–165.
[62] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. Yu, "COSNET: Connecting heterogeneous social networks with local and global consistency," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1485–1494.
[63] L. Liu, W. K. Cheung, X. Li, and L. J. Liao, "Aligning users across social networks using network embedding," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1774–1780.
[64] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1823–1829.
[65] J. Zhang and P. S. Yu, "Multiple anonymized social networks alignment," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 599–608.
[66] S. Wang, X. Li, Y. Ye, S. Feng, R. Y. K. Lau, X. Huang, and X. Du, "Anchor link prediction across attributed networks via network embedding," *Entropy*, vol. 21, no. 3, p. 254, 2019.
[67] X. Zhou, X. Liang, H. Zhang, and Y. Ma, "Cross-platform identification of anonymous identical users in multiple social media networks," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 2, pp. 411–424, Feb. 2016.
[68] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1178–1191, Jun. 2018.

[69] J. Lee, R. Hussain, V. Rivera, and D. Isroilov, "Second-level degree-based entity resolution in online social networks," *Social Netw. Anal. Mining*, vol. 8, p. 19, Dec. 2018.
[70] Y. Qu, S. Yu, W. Zhou, and J. Niu, "FBI: Friendship learning-based user identification in multiple social networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2019, pp. 1–6.
[71] W. Zhang, K. Shu, H. Liu, and Y. Wang, "Graph neural networks for user identity linkage," 2019, *arXiv:1903.02174*. [Online]. Available: https://arxiv.org/abs/1903.02174
[72] N. Wang, Q. Sun, Y. Zhou, and S. Shen, "A study on influential user identification in online social networks," *Chin. J. Electron.*, vol. 25, no. 3, pp. 467–473, 2016.
[73] Z. Yu, M. Li, X. Yang, and X. Li, "Palantir: Reseizing network proximity in large-scale distributed computing frameworks using SDN," in *Proc. IEEE 7th Int. Conf. Cloud Comput.*, Jun./Jul. 2014, pp. 440–447.
[74] M. Almishari and G. Tsudik, "Exploring linkability of user reviews," in *Computer Security—ESORICS*. Berlin, Germany: Springer, 2012, pp. 307–324.
[75] C. Tu, Z. Liu, and M. Sun, "PRISM: Profession identification in social media with personal information and community structure," in *Proc. 4th Chin. Nat. Conf. Social Media Process.*, 2015, pp. 15–27.
[76] Y. Nie, Y. Jia, S. Li, X. Zhu, A. Li, and B. Zhou, "Identifying users across social networks based on dynamic core interests," *Neurocomputing*, vol. 210, pp. 107–115, Oct. 2016.
[77] Y. Sha, Q. Liang, and K. Zheng, "Matching user accounts across social networks based on users message," *Procedia Comput. Sci.*, vol. 80, pp. 2423–2427, Jan. 2016.
[78] R. Roedler, D. Kergl, and G. D. Rodosek, "Profile matching across online social networks based on geo-tags," in *Advances in Nature and Biologically Inspired Computing*. Cham, Switzerland: Springer, 2016, pp. 417–428.
[79] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 447–458.
[80] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A user identification algorithm based on user behavior analysis in social networks," *IEEE Access*, vol. 9, pp. 47114–47123, 2019.
[81] K. Deng, L. Xing, M. Zhang, H. Wu, and P. Xie, "A multiuser identification algorithm based on Internet of Things," *Wireless Commun. Mobile Comput.*, vol. 2019, May 2019, Art. no. 6974809.
[82] D. Zhao, N. Zheng, M. Xu, X. Yang, and J. Xu, "An improved user identification method across social networks via tagging behaviors," in *Proc. IEEE 30th Int. Conf. Tools Artif. Intell.*, Nov. 2018, pp. 616–622.
[83] Y. Li, Z. Zhang, Y. Peng, H. Yin, and Q. Xu, "Matching user accounts based on user generated content across social networks," *Future Gener. Comput. Syst.*, vol. 83, pp. 104–115, Jun. 2018.
[84] L. Chen and F. Tan, "Identity recognition scheme based on user access behavior," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf.*, May 2019, pp. 125–129.
[85] W. Cao, Z. Wu, D. Wang, J. Li, and H. Hu, "Automatic user identification method across heterogeneous mobility data sources," in *Proc. IEEE 32nd Int. Conf. Data Eng.*, May 2016, pp. 978–989.
[86] T. Hao, J. Zhou, Y. Cheng, L. Huang, and H. Wu, "User identification in cyber-physical space: A case study on mobile query logs and trajectories," in *Proc. ACM SigSpatial*, 2016, Art. no. 71.
[87] X. Han, L. Wang, S. Xu, G. Liu, and D. Zhao, "Linking social network accounts by modeling user spatiotemporal habits," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, Jul. 2017, pp. 19–24.
[88] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 707–719.
[89] X. Han, L. Wang, L. Xu, and S. Zhang, "Social media account linkage using user-generated geo-location data," in *Proc. IEEE Conf. Intell. Secur. Inform.*, Sep. 2016, pp. 157–162.
[90] M. Qi, Z. Wang, Z. He, and Z. Shao, "User identification across asynchronous mobility trajectories," *Sensors*, vol. 19, no. 9, 2019, Art. no. 2102.
[91] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou R3579x?: Anonymized social networks, hidden patterns, and structural steganography," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 181–190.
[92] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 506–515.
[93] B. Zhou and J. Pei, "The K-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks," *Knowl. Inf. Syst.*, vol. 28, pp. 47–77, Jul. 2011.
[94] V. C. Hu, D. R. Kuhn, and D. F. Ferraiolo, "Attribute-based access control," *Computer*, vol. 48, no. 2, pp. 85–88, Feb. 2015.
[95] J. Qian, X.-Y. Li, C. Zhang, and L. Chen, "De-anonymizing social networks and inferring private attributes using knowledge graphs," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.
[96] S. Du, X. Li, J. Zhong, L. Zhou, M. Xue, H. Zhu, and L. Sun, "Modeling privacy leakage risks in large-scale social networks," *IEEE Access*, vol. 6, pp. 17653–17665, 2018.
[97] S. Bartunov, A. Korshunov, S. T. Park, W. Ryu, and H. Lee, "Joint link-attribute user identity resolution in online social networks," in *Proc. 6th SNAKDD Workshop*, 2012, pp. 1–9.
[98] P. Jain, P. Kumaraguru, and A. Joshi, "@i seek 'fb.me': Identifying users across multiple online social networks," in *Proc. 22nd Int. Conf. World Wide Web Companion*, 2013, pp. 1259–1268.
[99] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 51–62.

LING XING received the B.S. degree in electronic engineering from the Southwest University of Science and Technology, China, in 2002, the M.S. degree in electronic engineering from the University of Science and Technology of China, in 2005, and the Ph.D. degree in communication and information system from the Beijing Institute of Technology, in 2008.

In 2007, she was a Visiting Scholar with the Illinois Institute of Technology, IL, Chicago, USA. She is currently a Professor with the School of Information Engineering, Henan University of Science and Technology, China. Her research interests include multimedia semantic mining, private preserving, and social computing.

KAIKAI DENG received the B.S. degree in electronic information engineering from the Anyang Institute of Technology, China, in 2017. He is currently pursuing the M.S. degree in information and communication engineering with the Henan University of Science and Technology, China. His research interests include data mining, social media trust analysis, and social computing.

HONGHAI WU received the B.S. degree from Zhengzhou University, China, in 2001, and the M.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, China, in 2007 and 2015, respectively.

He was with China United Telecommunications Company Ltd., from 2007 to 2011. He is currently an Associate Professor with the School of Information Engineering, Henan University of Science and Technology, Luoyang, China. His research interests include delay/disrupted tolerant networks, opportunistic networks, and video delivery.

PING XIE was born in Hunan, China, in January 1984. She received the B.Sc. degree in communication engineering and the M.S. degree in communication and information systems from the Kunming University of Science and Technology, Kunming, China, in 2007 and 2011, respectively, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2014. She is currently a Lecturer with the Henan University of Science and Technology. Her research interests include cognitive radio networks, physical layer security, and resource allocation for fading channels.

H. VICKY ZHAO received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1997 and 1999, respectively, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2004, all in electrical engineering.

She was a Research Associate with the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, from 2005 to 2006. From 2006 to 2012, she was an Assistant Professor with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, TB, Canada, and then an Associate Professor from 2012 to 2016. Since 2016, she has been an Associate Professor with the Department of Automation, Tsinghua University. She has coauthored the *Multimedia Fingerprinting Forensics for Traitor Tracing* (Hindawi, 2005) and *Behavior Dynamics in Media-Sharing Social Networks* (Cambridge University Press, in 2011). Her research interests include media-sharing social networks, information security and forensics, and digital communications and signal processing. From 2010 to 2012, she was a member of the IEEE Signal Processing Society Information Forensics and Security Technical Committee, and the Multimedia Signal Processing Technical Committee, from 2011 to 2013. She received the IEEE Signal Processing Society 2008 Young Author Best Paper Award. She is the General Co-Chair of the 2014 IEEE International Workshop on Information Forensics and Security, the Technical Program Co-Chair of the 2012 IEEE International Workshop on Multimedia Signal Processing, the Publication Co-Chair of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, and the Finance Co-Chair of the 2015 IEEE Global Conference on Signal and Information Processing and the 2013 IEEE International Conference on Multimedia and Expo. From 2013 to 2015, She was an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the *Elsevier Journal of Visual Communication and Image Representation*, from 2009 to 2015, and a Guest Editor of a special issue on Signal and Information Processing for Social Learning and Networking of the IEEE *Signal Processing Magazine*.

FEIFEI GAO (M'09–SM'14) received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China, in 2002, the M.Sc. degree from McMaster University, Hamilton, ON, Canada, in 2004, and the Ph.D. degree from the National University of Singapore, Singapore, in 2007.

He was a Research Fellow with the Institute for Infocomm Research (I2R), A*STAR, Singapore, in 2008, and an Assistant Professor with the School of Engineering and Science, Jacobs University, Bremen, Germany, from 2009 to 2010. In 2011, he joined the Department of Automation, Tsinghua University, Beijing, China, where he is currently an Associate Professor. He has authored/coauthored more than 100 refereed IEEE journal articles and more than 100 IEEE conference proceeding articles, which have been cited over 4000 times from Google Scholar. His research interests include communication theory, signal processing for communications, array signal processing, and convex optimizations, with particular interests in MIMO techniques, multicarrier communications, cooperative communication, and cognitive radio networks.

Dr. Gao has also served as the Symposium Co-Chair for the 2015 IEEE Conference on Communications, the 2014 IEEE Global Communications Conference, and the 2014 IEEE Vehicular Technology Conference Fall, and a Technical Committee Member for many other IEEE conferences. He has served as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE COMMUNICATIONS LETTERS, the IEEE SIGNAL PROCESSING LETTERS, the IEEE WIRELESS COMMUNICATIONS LETTERS, *International Journal on Antennas and Propagation*, and *China Communications*.