Article

Exploiting User Friendship Networks for User Identification across Social Networks

Yating Qu ¹ , Ling Xing ^{2,*} , Huahong Ma ² , Honghai Wu ² , Kun Zhang ¹ and Kaikai Deng ³

¹ School of Automotive and Rail Transportation, Luoyang Polytechnic, Luoyang 471099, China; 18438550596@163.com (Y.Q.); zhangkun9595@163.com (K.Z.)

² School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China; mhh@haust.edu.cn (H.M.); honghai2018@haust.edu.cn (H.W.)

³ School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China; dkk@bupt.edu.cn

* Correspondence: xingling_my@haust.edu.cn

Abstract: Identifying offline entities corresponding to multiple virtual accounts of users across social networks is crucial for the development of related fields, such as user recommendation system, network security, and user behavior pattern analysis. The data generated by users on multiple social networks has similarities. Thus, the concept of symmetry can be used to analyze user-generated information for user identification. In this paper, we propose a friendship networks-based user identification across social networks algorithm (FNUI), which performs the similarity of multi-hop neighbor nodes of a user to characterize the information redundancy in the friend networks fully. Subsequently, a gradient descent algorithm is used to optimize the contribution of the user's multi-hop nodes in the user identification process. Ultimately, user identification is achieved in conjunction with the *Gale-Shapley* matching algorithm. Experimental results show that compared with baselines, such as friend relationship-based user identification (FRUI) and friendship learning-based user identification (FBI): (1) The contribution of single-hop neighbor nodes in the user identification process is higher than other multi-hop neighbor nodes; (2) The redundancy of information contained in multi-hop neighbor nodes has a more significant impact on user identification; (3) The precision rate, recall rate, comprehensive evaluation index (F1), and area under curve (AUC) of user identification have been improved.

Keywords: across social networks; user identification; entity; friendship networks; multi-hop neighbor nodes

Citation: Qu, Y.; Xing, L.; Ma, H.; Wu, H.; Zhang, K.; Deng, K. Exploiting User Friendship Networks for User Identification across Social Networks. *Symmetry* **2022**, *14*, 110. <https://doi.org/10.3390/sym14010110>

Academic Editors: Vyacheslav Yukalov and Yin Zhang

Received: 6 December 2021

Accepted: 5 January 2022

Published: 8 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid development of social networks has been extensively affected by the Web 3.0 technology. More and more users have multiple social virtual accounts on social networks and conduct frequent information interactions [1]. According to the latest statistical report in 2021 [2], there are approximately 2.853 billion monthly active users on Facebook, and 1.242 billion active users on WeChat. Due to the differences in the application scenarios and services provided by major social networks, people have gradually begun to use different social networks to meet their social needs. For example, people keep in touch with their friends through instant messaging QQ and WeChat; follow existing hot issues through Weibo and Twitter; establish contacts in the workplace through LinkedIn; solve their doubts in life and share their insights on things through Zhihu. However, user information on social networks is not interoperable.

With the proliferation of social network users, the data types generated have gradually diversified. However, the different social networks are not closely connected. In this case, it is easy to cause some malicious users to perform illegal operations through registration, which seriously endangers network security. Therefore, it is necessary to identify the

entities behind multiple virtual accounts of users across social networks. If this technology can be implemented, it will have a very important impact in many different areas, such as the following.

- Aggregate and analyze information from multiple user accounts, allowing for more accurate recommendation services and reduced commercial overhead [3,4];
- The ability to better analyze and predict user behavior patterns are difficult to achieve with individual social platforms [5];
- Improve user account information to provide researchers with more realistic user data [6];
- Malicious users can be detected in a more timely manner, providing some assistance to the field of network security [7].

Existing research work has focused on three types of data generated by users for identification. First, researchers rely on user profile information [8,9], which contains mainly the user's username, interests, date of birth, etc. As the major social networks are now slowly becoming aware of the privacy issues of social users and the improvement of users' own privacy and security, the cost of obtaining such information in large quantities is high, and there is a certain degree of falsification of such information so that the performance of user identification is somewhat affected. There are also some researchers who use user behavior information for user identification [10–12], that is, the sum of user behavior information generated on social networks. This type of information can be personalized to reflect the user's behavior habits so that the performance of user identification can be improved. However, some users produce less content on social networks, which results in the sparseness of user data. Therefore, the universality of algorithms based on this type of user information needs to be improved.

In addition, there are also a large number of researchers using the user's friendship networks for user identification [13,14]. The user's friendship networks is easier to obtain, and the probability of forged information is low. However, the existing research work only focuses on the analysis of the single-hop node of the user node and ignores the role of the multi-hop nodes in user identification. Although the contribution of single-hop nodes in the process of user identification is relatively large, the contribution of multi-hop nodes cannot be ignored.

As existing work does not provide a deep analysis of the redundant information contained in the multi-hop nodes of a user's friendship network, this paper addresses this problem by proposing a user identification across social networks algorithm based on friendship networks. On the basis of single-hop nodes, multi-hop nodes in user friendship networks are analyzed, and a gradient descent algorithm is used to assign the degree of contribution of multi-hop nodes in the user identification process. This is ultimately combined with relevant account matching algorithms to better identify users.

Our major contributions are summarized as follows:

- We design a user identification across social networks framework based on friend relationships;
- We take the user's multi-hop nodes into consideration to quantify the contribution of multi-hop nodes in the user identification process;
- We use the gradient descent algorithm to weight user's multi-hop nodes;
- Experimental results prove that our proposed algorithm has an average 28.2% improvement in identification performance compared to two baselines.

2. Related Works

In current research work, user identification has been addressed primarily based on the similarity of user-accessible attributes, such as user profile information, user behavioral information, and friendship networks.

2.1. User Identification Based on Profile Information

The username is the profile information commonly used to identify the user. Zafarani et al. [15] first proposed the problem of user identification across social networks and designed an identity matching method based on usernames. Similarly, Perito et al. [16] measured the similarity of usernames between social networks and applied their calculation results to user identification. Liu et al. [1] focused on determining whether across social networks user accounts with the same username belong to the same person. Agarwal et al. [17] proposed a user identification method based on the uniqueness of usernames. Since the username is a unique attribute of users on social networks, these current research works have achieved good performance. However, on some social networks (such as QQ and Foursquare), usernames are numeric strings that are not suitable for user identification. Li et al. [8] proposed a user identification scheme based on user display names to solve this problem. Since the user display name is not unique, the display name may be shared by multiple users, which reduces the accuracy of user identification.

Combining the username or display name with other profile information can effectively improve the accuracy of identification. Motoyama et al. [18] combined some profile information to effectively match user accounts on Facebook and MySpace. Narayanan et al. [19] mapped user accounts on Flickr and Twitter based on usernames, display names, and user locations. Raad et al. [20] measured the similarity between different accounts based on user profile attributes. Bartunov et al. [21] proposed a method based on conditional random fields. Xing et al. [9] used a number of user profile information to identify users and involved more user information. When the user information is not complete, the detection precision is easily affected.

User profile information provides rich information redundancy for user identification. Therefore, most identification algorithms based on user profile information have good performance. However, some user information is usually lost or forged for specific purposes, which affects the identification rate of these methods.

2.2. User Identification Based on Behavioral Information

Generally, the user behavior information of the same user contains a lot of information redundancy, including time redundancy, position redundancy, and writing style redundancy. According to the above ideas, Narayanan et al. [22] measured the similarity of writing styles in user behavior information. Since then, user behavior information has attracted great attention from researchers. Zafarani et al. [23] used the naming pattern, language function, and writing style of user behavior information to identify the user. Goga et al. [24] measured similarities between user behavior in time, space, and content perception. Li et al. [25] also proposed similar work. Liu et al. [26] extracted points of interest and writing style from user behavior information to measure the similarity of two user accounts. Deng et al. [11] proposed a recognition method based on frequent pattern mining, which analyzes user behavior information personalized. Chen et al. [27] realized user identification across social networks by using convolutional neural network to process user-generated content.

User behavior information is easy to access and not easy to forge, so it is ideal data for user identification. However, spoken words and abbreviations are often used in user behavior information. Therefore, an excellent recognition method based on user behavior information must rely on excellent natural language processing algorithms to a large extent, which is a huge challenge.

2.3. User Identification Based on Friendship Network

Due to the symmetry of a user's friendship network on different social networks, use identification algorithms based on user friend networks mainly rely on the friend relationship between users to determine the account identity. Vosecky et al. [28] proposed an identification method based on user profile information and friend networks in the early work of user identification based on friendship networks. Since then, the information

redundancy of the friend networks has been applied to user identification. Zhang et al. [29] regarded the user identification problem based on the friend network as a graph alignment problem and proposed a corresponding solution. Zhou et al. [13,30] proposed a semi-supervised scheme and an unsupervised scheme, named FRUI and FRUI-P, to study user identification based on friend networks. Li et al. [31] proposed a minor change to make FRUI a more general user identification algorithm. Meanwhile, Li et al. [32] proposed a friend relationship-based identification network that recognizes the user's identity by calculating the similarity of the user's friend network.

These existing friend network-based methods can usually achieve good performance. However, in practice, the friend network presents heterogeneity, different semantics, and even serious deviations in quantity [25]. Due to user privacy settings, friend networks are also difficult to access, making it difficult to measure the similarity of friend networks accurately. In addition, the current research work rarely focuses on how much information redundancy in multi-hop neighbor nodes can help user identification.

3. Problem Definition

Across social networks, friend networks of the same user have rich information redundancy, which is essential for matching user accounts across social networks. When considering the problem of user identification based on the friend network, it is first necessary to measure the redundancy of this information and then answer the following questions based on the measurement results.

Problem 1. Given two user accounts v_A^i and v_B^j and their friend networks are denoted as $S(v_A^i)$ and $S(v_B^j)$, respectively, are these two user accounts the same user?

The above Problem 1 can be further formulated as Problem 2.

Problem 2. We suppose two social networks A and B are given, $V = \{(v_A^i, v_B^j, y_{ij}), v_A^i \in A, v_B^j \in B, y_{ij} \in \{0, 1\}\}$, $\forall (v_A^i, v_B^j, y_{ij}) \in V$. We know the value of the friend network l_{ij} of user accounts v_A^i and v_B^j . When $y_{ij} = 1$, user accounts v_A^i and v_B^j belong to the same user. Otherwise, the user accounts v_A^i and v_B^j are the accounts of different users. Given two user accounts v_A^m , v_B^n and their friend networks, where $(v_A^m, v_B^n, y_{mn}) \notin V$, the task of user identification based on friend networks is to determine $y_{mn} = 0$ or $y_{mn} = 1$.

The key goal of Problem 2 is to learn the user identification classification function based on the labeled data, that is, the value of y_{mn} . Therefore, we must accurately analyze and measure the information redundancy of the user's friend networks. In the remainder of this paper, we refer to the solution of Problem 2 as FNUI.

For various reasons, some users will have multiple virtual accounts on the same social network, but we usually think that these accounts are independent and belong to different people. In other words, we only identify one account of the user.

4. Similarity Calculation of Friend Network

Generally, a friend owned by a user on one social network may also maintain a friend relationship with him on another social network. Therefore, the friend network across social networks has certain similarities. In this section, our feature extraction is similar to the previous work performed by Li et al. [32] because we all analyze the user multi-hop nodes to improve the performance of user identification. However, our difference is that we mainly focus on innovation at the weight level and user identification level rather than features. We will analyze the similarity of friend networks and start with the following three aspects: (1) user friend relationship, (2) user friend circle, and (3) clustering coefficient of an ego-centric network.

4.1. Similarity of User's Friendship Networks

Assuming that v_A^i is the i th user of social network A , $N^n(v_A^i)$ represents the set generated by the user's n th hop neighbor node. In the absence of ambiguity, we also use $N(v_A^i)$ to represent the set of neighbor nodes in the general sense of user v_A^i . The friend relationship refers to the relationship between the user account to be matched and its n -hop neighbor nodes. Generally, the simplest way to characterize the similarity of the friend relationship between two users is to express it in terms of the number of common neighbors. The common neighbor index of the number of mutual friends of two user accounts has been proposed, such as the hub promoted index (HPI).

$$S(\alpha, \beta) = \frac{1}{\min\{k_\alpha, k_\beta\}} \sum_{u \in U_{\alpha\beta}} 1, \quad (1)$$

where α and β represent two sets, k_α and k_β represent the size of α and β , and $U_{\alpha\beta}$ represents the intersection of α and β .

Following the idea of HPI, we define the similarity of the friend relationship between two users, as shown in Definition 1.

Definition 1. (Similar friend relationship). Given two user accounts v_A^i and v_B^j , the similarity of the friend relationship between these two user accounts is defined as the ratio of the number of their mutual friends to the minimum set of their friends. The calculation formula is:

$$Sim_{fr} = \frac{|N(v_A^i) \cap N(v_B^j)|}{\min\{|N(v_A^i)|, |N(v_B^j)|\}}, \quad (2)$$

among them, $N(v_A^i)$ and $N(v_B^j)$ denote the set of neighbor nodes of users i and j on social networks A and B , respectively.

Based on Formula (2), the similarity between $N^p(v_A^i)$ and $N^q(v_B^j)$ can be further defined, as shown in the following formula:

$$Sim_{fr}^{pq} = \frac{|N^p(v_A^i) \cap N^q(v_B^j)|}{\min\{|N^p(v_A^i)|, |N^q(v_B^j)|\}}. \quad (3)$$

With the above analysis, given two different user accounts and their corresponding friend networks, we can extract 14 features from the friend networks of these two user accounts to calculate the corresponding similarity. The 14 friendships with the most significant features we extracted are 4. We cross-calculate the multi-hop nodes of the friend relationship of two different social network accounts to determine the similarity between the users' friendships. Table 1 lists the combinations of multi-hop friendships among different users.

Table 1. The 14 characteristics of friend relationship.

Feature Numbers	Feature Name	Feature Number	Feature Name
1	Sim_{fr}^{11}	8	Sim_{fr}^{32}
2	Sim_{fr}^{12}	9	Sim_{fr}^{33}
3	Sim_{fr}^{13}	10	Sim_{fr}^{34}
4	Sim_{fr}^{21}	11	Sim_{fr}^{41}
5	Sim_{fr}^{22}	12	Sim_{fr}^{42}
6	Sim_{fr}^{23}	13	Sim_{fr}^{43}
7	Sim_{fr}^{31}	14	Sim_{fr}^{44}

4.2. Similarity of User's Friend Circle

Here, we need, first, to clarify two terms: friend relationship and friend circle. User v_A^i friend relationship refers to his n -hop neighbor nodes, and his friend circle is a group of users interacting with him, including him. Theoretically speaking, friend circle is part of the friendship relationship. The users in the friend circles are more closely connected. However, from the perspective of friend relationship, we only pay attention to the relationship between his neighbors and not the relationship between his neighbors. Here we can divide the user's friend circles into the maximum and minimum friend circles.

Definition 2. (Maximum or minimum circle). Given a user account v_A^i , his maximum (or minimum) circle refers to the largest (or 3) in his friend circles.

Assume that user v_A^i maximum friend circles is $C^{\max}(v_A^i)$. $\forall v_A^r \in C^{\max}(v_A^i)$ and $\bigcup_{v_A^r} C^{\max}(v_A^r)$ are user v_A^i 2-hop maximum friend circles; similarly, we can obtain the 3-hop friend circles based on the 2-hop friend circles, and then we can obtain the n -hop circle of user v_A^i maximum friend circles in turn.

User v_A^i has many minimum circles. All the minimum circles are integrated into his 1-hop minimum circles. To be consistent with the description of the maximum circles, we extend the minimum circles in Definition 2 and redefine the user's minimum circles as his 1-hop minimum friend circles. User v_A^i minimum circles is denoted by $C^{\min}(v_A^i)$. $\forall v_A^r \in C^{\min}(v_A^i)$ and $\bigcup_{v_A^r} C^{\min}(v_A^r)$ are user v_A^i 2-hop minimum friend circles, similarly, user v_A^i n -hop minimum friend circles $C^{n-\min}(v_A^i)$ can be obtained.

Definition 3. (Similarity calculation of friend circle). Suppose two user accounts v_A^i and v_B^j are given, and their circle of friends is $C(v_A^i)$ and $C(v_B^j)$, respectively. The similarity of the friend circles is the minimum value of the size of the intersection of two users' friend circles over the size of each of these two users' friend circles, which is calculated as:

$$Sim_{fc} = \frac{|C(v_A^i) \cap C(v_B^j)|}{\min\{|C(v_A^i)|, |C(v_B^j)|\}}. \quad (4)$$

In the same way, we can further define the similarity calculation formula between different user multi-hop nodes $C^p(v_A^i)$ and $C^q(v_B^j)$:

$$Sim_{fc}^{pq-\max} = \frac{|C^{p-\max}(v_A^i) \cap C^{q-\max}(v_B^j)|}{\min\{|C^{p-\max}(v_A^i)|, |C^{q-\max}(v_B^j)|\}}. \quad (5)$$

$$Sim_{fc}^{pq-\min} = \frac{|C^{p-\min}(v_A^i) \cap C^{q-\min}(v_B^j)|}{\min\{|C^{p-\min}(v_A^i)|, |C^{q-\min}(v_B^j)|\}}. \quad (6)$$

According to the calculation formula of the maximum (minimum) friend circles, we can extract 24 features of two users for similarity calculation. These 24 features are shown in Table 2.

Table 2. The 24 characteristics of friend circle.

Feature Numbers	Friend Circles	
	Maximum	Minimum
1	Sim_{fc}^{11-max}	Sim_{fc}^{11-min}
2	Sim_{fc}^{12-max}	Sim_{fc}^{12-min}
3	Sim_{fc}^{13-max}	Sim_{fc}^{13-min}
4	Sim_{fc}^{21-max}	Sim_{fc}^{21-min}
5	Sim_{fc}^{22-max}	Sim_{fc}^{22-min}
6	Sim_{fc}^{23-max}	Sim_{fc}^{23-min}
7	Sim_{fc}^{31-max}	Sim_{fc}^{31-min}
8	Sim_{fc}^{32-max}	Sim_{fc}^{32-min}
9	Sim_{fc}^{33-max}	Sim_{fc}^{33-min}
10	Sim_{fc}^{34-max}	Sim_{fc}^{34-min}
11	Sim_{fc}^{43-max}	Sim_{fc}^{43-min}
12	Sim_{fc}^{44-max}	Sim_{fc}^{44-min}

4.3. Similarity of Clustering Coefficient

On social networks, the clustering coefficient refers to the probability that two friends of a user are also friends. The clustering coefficient of user v_A^i is defined as:

$$CC(v_A^i) = \frac{2e_i}{n(n-1)}, \quad (7)$$

among them, e_i denotes the number of all neighboring user nodes of user v_A^i that are connected to each other, and n represents the number of all neighboring user nodes of user v_A^i .

Given a user account v_A^i and its corresponding friend network, we can define the average clustering coefficient of user v_A^i and his friend network, the calculation formula is:

$$ACC(v_A^i) = \frac{1}{|N(v_A^i)| + 1} \sum_{v_A^r \in N(v_A^i) \cup \{v_A^i\}} CC(v_A^r). \quad (8)$$

After the above calculations, we defined two different methods to characterize the similarity of user clustering coefficients: the difference between the average clustering coefficients of different users and the ratio of the average clustering coefficients of two users. Suppose v_A^i and v_B^j are a pair of users. The following equations can define the difference and ratio of their average clustering coefficients,

$$Sim_{acc}^1 = |AC(v_A^i) - AC(v_B^j)|, \quad (9)$$

$$Sim_{acc}^2 = \frac{\min\{AC(v_A^i), AC(v_B^j)\}}{\max\{AC(v_A^i), AC(v_B^j)\}}. \quad (10)$$

After analyzing the characteristics of the multi-hop nodes in the above-mentioned user friend network, the results we obtained are similar to those of [32]. Therefore, we briefly extract these 40 features in the paper without introducing them in detail, mainly because [32] has visualized this result. Therefore, we have carried out new designs in the weight distribution and user identification algorithms to obtain better identification performance.

5. Friendship Networks-Based User Identification Algorithm

Based on the above analysis of the user's friend network characteristics, the gradient descent algorithm is used to characterize the degree of contribution of different features in user identification. The relevant account matching algorithm is used for user identification.

5.1. Overview

User identification across social networks is designed to identify the physical users behind multiple virtual social network accounts and is essential in many research areas. This paper focuses on analyzing the role of user friendship networks in user identification and further improving the precision of user identification by combining the gradient descent algorithm and the *Gale–Shapley* matching algorithm. As shown in Figure 1, we crawl the user's friendship networks from mainstream social networks. Analyze the characteristics of the friendship networks from three different perspectives. We extract 40 features to calculate the similarity between different user accounts mentioned above. We also use a gradient descent algorithm to weigh different features and form a similarity based on weights. Finally, the *Gale–Shapley* matching algorithm identifies user accounts on different social networks.

Figure 1. Schematic diagram of user identification across social networks algorithm.

5.2. Gradient Descent Algorithm

Since the single-hop node of the user's friend network is not enough to fully characterize the role of the user's friend network in the user's identity recognition process. Therefore, this paper measures and analyzes the multi-hop nodes in the friend network to illustrate the contribution of multi-hop nodes in the process of user identification. Since our feature extraction work is similar to [32], our main innovation is on the weight distribution of user friendship network features. Our goal is to find a better and lighter solution to reflect the weight relationship between users' different features. In contrast, in [32] uses multiple classifiers to determine the weight, and this design has a higher computational complexity. We can see the above analysis of the redundant information contained in the multi-hop node in the friend network, we can see that the multi-hop node's impact on the user identification process cannot be ignored. However, nodes with different hop counts have different contributions to user identification. Therefore, the gradient descent algorithm is used to deal with the weight problem of nodes with different hop counts.

Through the above analysis of the friend network, a total of 40 features were extracted. The weights obtained by these features are optimized further to improve the performance

of user identification across social networks. The calculation formula for the network similarity of user friends based on weight is as follows:

$$\begin{aligned} Sim_w &= w_1 Sim_1 + w_2 Sim_2 + \dots + w_n Sim_n, \\ 0 &\leq w_1, w_2, \dots, w_n \leq 1, \end{aligned} \quad (11)$$

where w_n denotes the weight of the n th feature extracted by the friend network.

To better identify the user's identity, we need to optimize the weight of each extracted feature and use the gradient descent algorithm to optimize the weight. First, the prediction function needs to be constructed, and its related expressions are as follows:

$$\phi = h_w(x) = w_1 x_1 + w_2 x_2 + \dots + w_n x_n, \quad (12)$$

among them, ϕ denotes the prediction function, and x_n represents the n th extracted feature.

Then we establish the corresponding loss function, the calculation formula is as follows:

$$J(w_1, w_2, \dots, w_n) = \frac{1}{2n} (h_w(x_i) - y_i)^2, \quad (13)$$

among them, $J(w_1, w_2, \dots, w_n)$ denotes the loss function, and y_i represents the true value of the i th extracted feature.

We need to minimize the loss function, mainly through the following two stages to optimize the relevant parameters.

Phase 1: We use partial differential operations to process and obtain general expression formulas:

$$\frac{\partial J(w_1, w_2, \dots, w_n)}{\partial w_n} = \frac{1}{n} \sum_{i=1}^n (h_w(x_i) - y_i) x_i. \quad (14)$$

Phase 2: We need to continuously iteratively update w_n , and obtain its general formula:

$$w_n = w_n - \alpha \frac{1}{n} \sum_{i=1}^n (h_w(x_i) - y_i) x_i, \quad (15)$$

where α denotes the learning rate.

After optimizing the weights of the extracted features through the above two phases, each feature in the friend network can be assigned the optimal weight to optimize Sim_w further.

5.3. User Account Matching

Through the above-mentioned feature extraction, analysis, measurement, and weighting of the user's friend network, a weight-based similarity value can be obtained. In [32] use nine classifiers to identify user accounts and summarize the identification results to give a final result. However, this kind of user identification algorithm involves more complicated calculations and does not limit the one-to-many and many-to-many problems in the user account matching process. In contrast, we design a novel user account matching algorithm and restrict the problems encountered in user account matching. The essence of user identification is a classification problem. Therefore, a user identification function l_{ij} is needed to characterize the question of whether the account pair matches.

As shown in Figure 2, given two social networks A and B , and their corresponding three user accounts v_A^i , v_A^j , and v_B^k , where $v_A^i, v_A^j \in S_A$, $v_B^k \in S_B$. The weight-based similarity corresponding to user account v_A^i , v_A^j , and v_B^k are Sim_A^{wi} , Sim_A^{wj} , and Sim_B^{wk} . Taking user accounts v_A^i and v_B^k as an example, this pair of user accounts is mapped to node r_{ik} . Therefore, the identification problem of user accounts v_A^i and v_B^k is transformed into a classification problem.

Figure 2. User account matching framework.

If y_{ik} is the classification result of r_{ik} , then $y_{ik} = 1$, user accounts v_A^i and v_B^k are regarded as the same user, otherwise, the two user accounts are not the same natural person. Therefore, we use supervised machine learning algorithms to solve such problems. For the seed node, $\forall i, k (v_1^i, v_2^k) \in E, y_{ik} = 1$. We can obtain the feature vector X_{ik} from r_{ik} , which contains the user's friend network feature information. We can construct training data (X_{ik}, y_{ik}) and train a supervised classifier.

It can be seen from Figure 1 that if the recognition result is not restricted during the recognition process, $y_{ik} = 1$ and $y_{jk} = 1$ will appear. To avoid this one-to-many or many-to-many problem in the identification result, this paper uses Formulas (16) and (17) to realize the one-to-one constraint of user account identification.

$$z(y_{ij}, y_{kl}) = \begin{cases} 1 & j = k \text{ or } i = l \\ 0 & \text{others} \end{cases}, \quad (16)$$

$$\forall v_A^j \in S_A, \sum_{v_B \in S_B} y_{jk} = 1. \quad (17)$$

If $z(y_{ij}, y_{kl}) = 1$, it means that only one of the two pairs of user accounts identified has a value of 1, that is, the result of user identification must satisfy Formula (17). On the issue of user account matching, this paper is also affected by the idea of stable marriage matching, combined with the *Gale–Shapley* algorithm to further improve the performance of user identification, the detailed process is as Algorithm 1.

Algorithm 1 User identification algorithm

```

1: Input: The feature vector  $s(X_{jk})$  of the user account in  $S_A$  and  $S_B$ ,  $R = \phi$ .
2: Output: User account identification result set  $R = \{(v_A^j, v_B^k) | y_{jk} = 1 \wedge \text{formula}(17)\}$ .
3: for each user accounts  $v_A^j$  and  $v_B^k$  belong to  $S_A$  and  $S_B$  respectively
4: two probability sets  $\{rank^A\}$  and  $\{rank^B\}$  are respectively generated by the classifier
5: while  $\exists v_A^j \in S_A$  or  $\{rank^A\} \neq \phi$ 
6: from  $\{rank^B\}$  to select  $v_B^k$ ,  $v_B^k = \arg \max(probability)$ 
7: if  $v_B^k$  does not match
8: add  $(v_A^j, v_B^k)$  to  $R$ 
9: else
10: compare the priority of user accounts  $v_A^j$  and  $v_A^i$  in  $\{rank^B\}$  (assuming that user
    accounts  $v_B^k$  and  $v_A^i$  have been paired in advance)
11: if  $v_A^j > v_A^i$ 
12: remove  $(v_A^i, v_B^k)$  from  $R$ 
13: add matching pair  $(v_A^j, v_B^k)$  to  $R$ 
14: else
15: ignore  $(v_A^i, v_B^k)$ 
16: return  $R$ 

```

6. Analysis of Experimental Results**6.1. Dataset Acquisition**

Some social networks (such as Foursquare) provide the function of cross-network linking, which can help the work of this paper to obtain information and data of the same user across networks for experimentation. Let us take Foursquare as an example. In Foursquare, users are allowed to display links to their Facebook and Twitter accounts on their pages. Before displaying the link, the user must authorize Foursquare to access the user's Facebook and Twitter accounts. Only after Foursquare verifies ownership, users can provide account information [25,32]. These account links are voluntarily disclosed by users and authorized by Foursquare, so the links are extremely reliable.

As shown in Figure 3, we use the Python to write the corresponding code to crawl the friendship networks of different users. Based on the across social networks link function, we can obtain real data of the same user from Facebook, Twitter, and Foursquare. For each user, we obtain its friend network on these three social networks. The reason why we choose Foursquare to obtain real data is its popularity and unique digital user ID. If we know the ID of a user, we can access the user's profile page through the URL <https://foursquare.com/user/ID>, accessed on 5 December 2021. Moreover, Foursquare users' IDs are automatically assigned by the site in ascending order. According to the above ideas, we constructed three crawlers for these three social networks, as shown in Figure 3. To effectively obtain user data, we design a distributed framework of crawlers, in which each crawler is responsible for crawling a portion of users. We obtained data on 29,583 Foursquare users, 19,654 Facebook users, and 23,352 Twitter users. For convenience, the user data sets obtained from Foursquare, Facebook, and Twitter will be referred to as FS, FB, and TW in this paper.

After obtaining FB, FS and TW, we construct three cross-network ground truth datasets, FB-FS, FB-TW, and FS-TW, respectively. To make the measurement results more reliable and convincing, we compare the friendship network of the same user with the friendship network of different users. Therefore, there are two types of instances in each constructed dataset: positive and negative instances.

Figure 3. User data acquisition framework.

6.2. Evaluation Metrics

In the era of big data, researchers will design different user identification algorithms based on actual application scenarios. To verify the effectiveness of these algorithms, some standard evaluation indicators are needed to measure them. The most commonly used evaluation indexes are precision, recall, F1, and AUC. The calculation formula corresponding to the above evaluation index is as follows.

$$\text{Precision} = \frac{tp}{(tp + fp)} \quad (18)$$

$$\text{Recall} = \frac{tp}{(tp + fn)} \quad (19)$$

$$F1 = 2 \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})} \quad (20)$$

AUC: It refers to the area formed under the receiver operating characteristic (ROC) curve, which can be obtained by calculating the area. The X and Y axes of the ROC curve are represented by false positive rate (FPR) and true positive rate (TPR), respectively. The calculation formulas for these two indicators are as follows:

$$\text{TPR} = \frac{tp}{tp + fn}, \quad (21)$$

$$\text{FPR} = \frac{fp}{fp + tn}, \quad (22)$$

where tp denotes the number of matched users who are the same user, and tn denotes the number of unmatched users, and the user account is not the same user simultaneously; fp denotes the number of matched, but the user account does not belong to the same user; fn denotes the number of unmatched users, but the user account is the number of the same user.

6.3. Impact of Multi-Hop Nodes in the Friends Network on User Identification

To clearly clarify the effectiveness of the algorithm proposed in this paper, the acquired user data were constituted into three experimental datasets, namely FB-FS, FB-TW, and FS-TW, to carry out the validation of the algorithm. In addition, in order to compare the performance of the proposed algorithm FNUI with the existing research work, we also compare and analyze FNUI with the following two representative algorithms.

- FRUI [13]: A friend relationship-based user identification algorithm (FRUI), which calculates the friend matching degree of all candidate user matching pairs (UMP), and selects the UMP with the highest ranking as the same user;

- FBI [33]: A friendship learning-based user identification algorithm (FBI), which aims to provide a potential attack mechanism for subsequent privacy protection research. A new identification method based on friend relationship matching is designed, and a weighting mechanism is adopted to empower user profile information and friend networks. In addition, machine learning methods are also used to optimize parameters.

As shown in Figure 3, it can be seen that the algorithm in this paper is much better than the two compared algorithms in terms of four evaluation metrics, namely precision, recall, F1, and AUC, and presents good identification performance on different experimental datasets. As shown in Table 3, we highlight the performance of our algorithm on the four metrics. It can be seen that our algorithm has better universality. Specifically, the precision of FNUI is higher than FRUI and FBI on different experimental datasets (FB-FS, FB-TW, FS-TW), 35.5%, 53%, 41.3% and 15.6%, 10%, 14%, respectively; In terms of recall rate, FNUI is 60%, 67.25%, 54.9% and 13.4%, 13.1%, 11.6% higher than FRUI and FBI, respectively; In F1, FNUI is 51.8%, 62.5%, 49.7% and 14.5%, 11.6%, 12.8% higher than FRUI and FBI, respectively; In terms of AUC, FNUI is 36.9%, 56.7%, 39.9% and 16.2%, 15.3%, 15.7% higher than FRUI and FBI, respectively.

For a baseline, FRUI, we can see from Figure 4 and Table 3 that the performance of this algorithm is not very good. We can see from the above experimental results that the identification performance of the FRUI algorithm is much lower than that of the other two algorithms. The main reason is that the identification performance of FRUI largely depends on the proportion of mutual friends in the three experimental datasets. Since the matching degree of FRUI is mainly calculated by common neighbour nodes, i.e., single-hop nodes. If the number of single-hop nodes in the acquired user datasets is small, this will reduce the overall identification performance of the algorithm, which also illustrates the poor generality of the FRUI algorithm. Therefore, it is necessary for us to conduct multi-hop node analysis on different social network users.

Figure 4. Comparison of recognition performance of various algorithms. (a) FB-FS; (b) FB-TW; (c) FS-TW.

Table 3. A detailed comparison of the algorithm proposed with two baselines.

Datasets	Algorithms	Precision	Recall	F1	AUC
FB-FS	FNUI	0.941	0.915	0.928	0.960
	FRUI	0.586	0.315	0.410	0.591
	FBI	0.785	0.781	0.783	0.798
FB-TW	FNUI	0.912	0.885	0.898	0.954
	FRUI	0.382	0.213	0.273	0.387
	FBI	0.812	0.754	0.782	0.801
FS-TW	FNUI	0.936	0.901	0.918	0.962
	FRUI	0.523	0.352	0.421	0.563
	FBI	0.796	0.785	0.790	0.805

The FBI algorithm, on the other hand, combines user profile information and friend networks, the possible forgery of the user profile information, the lack of detail in the division of friend networks and the failure to take into account the contribution of multi-hop nodes also lead to the poor performance of the algorithm in terms of identification. Although the FBI algorithm combines user profile information with the friend network, it does not take into account the forgery of user profile information, the coarse-grained division of the friend network, and the lack of analysis of the contribution of multi-hop nodes, which also leads to the poor identification performance of the algorithm. In addition, the FBI will also have the problem of user privacy leakage in identifying users. Our method dramatically protects the privacy of users, which is another advantage of our method compared to the FBI.

The FNUI proposed in this paper takes into account the role of multi-hop nodes in user identification, which subdivides the friend networks, and uses different similarity calculation methods to measure the similarity of different network features, followed by appropriately weights each network feature. In addition, the one-to-many and many-to-many problems arising in the matching process are constrained in the user identity matching stage. The account matching is implemented in conjunction with the *Gale–Shapley* algorithm. Experimental results also fully demonstrate the reasonableness and effectiveness of the algorithm proposed in this paper.

7. Discussion and Future Directions

The existing user identification across social networks algorithms still has some problems to solve in the application process. This paper will conduct more in-depth exploration in future research to find a model suitable for application in actual scenarios. Therefore, the future research work will be combined with actual social network scenarios to explore, mainly reflected in the following three aspects:

- The information generated by users in social networks is diversified, and it has not been deeply explored and analyzed [7]. For special scenes in actual scenes, the accuracy of user recognition needs to be strictly controlled. At this time, it is necessary to mine and analyze multiple information of users. Therefore, in future scenarios, it is necessary to consider developing the attribute information of the user's trust;
- When the existing identification models face large-scale users, their recognition performance will gradually decline, which is also reflected in the experimental part of this paper. The community discovery technology in complex networks has inspired our research, and we can solve this large-scale problem by doing some distributed work [27]. Use community discovery to divide large-scale datasets and focus on recognizing communities so that large-scale user identification can be effectively performed. The existing identification model can further effectively alleviate the problem that the

identification performance of the existing identification decreases with the increase in the dataset;

- Existing models involve a large amount of user information in the process of identifying users. Using a small amount of useful information to achieve better multi-user identification is an urgent challenge. There will be some malicious attackers in social networks [34,35], and the use of a small amount of user information will further reduce such problems. Therefore, future work should proceed from the perspective of game theory, which guarantees the user's information security and meets the needs of identification performance.

8. Conclusions

User identification has attracted widespread attention in the academic world, and it can promote the development of many fields. The existing user identification methods are based on the similarity of user profile information, user behavior information, and friend networks. Due to the high cost and difficulty of forging a friend network, using a user friend network to match user accounts has become a research hotspot. Different from the existing work, this paper analyzes the similarity of multi-hop nodes in the friend networks and studies their contribution to user identification. In addition, the contribution of different neighbor nodes is reasonably weighted for them, and the one-to-many and many-to-many account matching problems are solved in the account matching process. Finally, the *Gale–Shapley* algorithm is combined to realize user identification. Experimental results show that FNUI is superior to the comparison algorithm in terms of precision, recall, F1, and AUC.

In the future, we will focus on protecting user privacy issues based on user identification. Since user identification and user privacy are in a game relationship, the trade-off between privacy and identification performance in future work will be an urgent problem to be solved.

Author Contributions: Conceptualization, Y.Q. and L.X.; methodology, Y.Q. and L.X.; validation, Y.Q.; data curation, L.X.; writing—original draft preparation, Y.Q., K.Z. and K.D.; writing—review and editing, L.X., H.M., H.W. and K.D.; supervision, L.X., H.M. and H.W.; project administration, L.X., H.M., H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the National Natural Science Foundation of China [62171180, 62072158], and in part by the Key Science and Research Program at the University of Henan Province [21A510001], Program for Innovative Research Team in University of Henan Province(21IRTSTHN015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the research team members for their contributions to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, J.; Zhang, F.; Song, X.; Song, Y.L.; Lin, C.Y.; Hon, H.W. What's in a name? An unsupervised approach to link users across communities. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013; pp. 495–504.
2. Most Popular Social Networks Worldwide as of July 2021, Ranked by Number of Active Users [EB/OL]. Available online: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> (accessed on 5 December 2021).
3. Zheng, J.; Li, D.; Arun Kumar, S. Group user profile modeling based on neural word embeddings in social networks. *Symmetry* **2018**, *10*, 435. [CrossRef]

4. Xing, L.; Deng, K.; Wu, H.; Xie, P. Review of User Identification across Social Networks: The Complex Network Approach. *J. Univ. Electron. Sci. Technol. China* **2020**, *49*, 905–917.
5. Li, C.Y.; Lin, S.D. Matching users and items across domains to improve the recommendation quality. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; pp. 801–810.
6. Nie, Y.; Jia, Y.; Li, S.; Zhu, X.; Li, A.; Zhou, B. Identifying users across social networks based on dynamic core interests. *Neurocomputing* **2016**, *210*, 107–115. [\[CrossRef\]](#)
7. Xing, L.; Deng, K.; Wu, H.; Xie, P.; Zhao, H.V.; Gao, F. A survey of across social networks user identification. *IEEE Access* **2019**, *7*, 137472–137488. [\[CrossRef\]](#)
8. Li, Y.; Peng, Y.; Ji, W.; Zhang, Z.; Xu, Q. User identification based on display names across online social networks. *IEEE Access* **2017**, *5*, 17342–17353. [\[CrossRef\]](#)
9. Xing, L.; Deng, K.; Wu, H.; Xie, P.; Zhang, M.; Wu, Q. Exploiting Two-Level Information Entropy across Social Networks for User Identification. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1082391. [\[CrossRef\]](#)
10. Shu, K.; Wang, S.; Tang, J.; Zafarani, R.; Liu, H. User identity linkage across online social networks: A review. *ACM Sigkdd Explor. Newsl.* **2017**, *18*, 5–17. [\[CrossRef\]](#)
11. Deng, K.; Xing, L.; Zheng, L.; Wu, H.; Xie, P.; Gao, F. A user identification algorithm based on user behavior analysis in social networks. *IEEE Access* **2019**, *7*, 47114–47123. [\[CrossRef\]](#)
12. Xing, L.; Deng, K.; Wu, H.; Xie, P.; Gao, J. Behavioral habits-based user identification across social networks. *Symmetry* **2019**, *11*, 1134. [\[CrossRef\]](#)
13. Zhou, X.; Liang, X.; Zhang, H.; Ma, Y. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE Trans. Knowl. Data Eng.* **2015**, *28*, 411–424. [\[CrossRef\]](#)
14. Mishra, R. Entity resolution in online multiple social networks (@ Facebook and LinkedIn). In *Emerging Technologies in Data Mining and Information Security*; Springer: Singapore, 2019; pp. 221–237.
15. Zafarani, R.; Liu, H. Connecting corresponding identities across communities. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media, San Jose, CA, USA, 17–20 May 2009.
16. Perito, D.; Castelluccia, C.; Kaafar, M.A.; Manils, P. How unique and traceable are usernames? In *International Symposium on Privacy Enhancing Technologies Symposium*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 1–17.
17. Agarwal, A.; Toshniwal, D. Smpft: Social media based profile fusion technique for data enrichment. *Comput. Netw.* **2019**, *158*, 123–131. [\[CrossRef\]](#)
18. Motoyama, M.; Varghese, G. I seek you: Searching and matching individuals in social networks. In Proceedings of the Eleventh International Workshop on Web Information and Data Management, Marina Del Rey, CA, USA, 5–9 February 2009; pp. 67–75.
19. Narayanan, A.; Shmatikov, V. De-anonymizing social networks. In Proceedings of the 2009 30th IEEE Symposium on Security and Privacy, Berkeley, CA, USA, 17–20 May 2009; pp. 173–187.
20. Raad, E.; Chbeir, R.; Dipanda, A. User profile matching in social networks. In Proceedings of the 2010 13th International Conference on Network-Based Information Systems, Takayama, Japan, 14–16 September 2010; pp. 297–304.
21. Bartunov, S.; Korshunov, A.; Park, S.T.; Ryu, W.; Lee, H. Joint link-attribute user identity resolution in online social networks. In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis, Istanbul, Turkey, 26–29 August 2012.
22. Narayanan, A.; Paskov, H.; Gong, N.Z.; Bethencourt, J.; Stefanov, E.; Shin, E.C.R.; Song, D. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*; IEEE: New York, NY, USA, 2012; pp. 300–314.
23. Zafarani, R.; Liu, H. Connecting users across social media sites: A behavioral-modeling approach. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; pp. 41–49.
24. Goga, O.; Lei, H.; Parthasarathi, S.H.K.; Friedland, G.; Sommer, R.; Teixeira, R. Exploiting innocuous activity for correlating users across sites. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 447–458.
25. Li, Y.; Zhang, Z.; Peng, Y.; Yin, Y.; Xu, Q. Matching user accounts based on user generated content across social networks. *Future Gener. Comput. Syst.* **2018**, *83*, 104–115. [\[CrossRef\]](#)
26. Liu, S.; Wang, S.; Zhu, F.; Zhang, J.; Krishnan, R. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, UT, USA, 22–27 June 2014; pp. 51–62.
27. Chen, H.; Yin, H.; Sun, X.; Chen, T.; Gabrys, B.; Musial, K.C. Multi-level graph convolutional networks for cross-platform anchor link prediction. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 6–10 July 2020; pp. 1503–1511.
28. Vosecky, J.; Hong, D.; Shen, V.Y. User identification across multiple social networks. In Proceedings of the 2009 First International Conference on Networked Digital Technologies, Ostrava, Czech Republic, 28–31 July 2009; pp. 360–365.
29. Zhang, J.; Philip, S.Y. Multiple anonymized social networks alignment. In Proceedings of the 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14–17 November 2015; pp. 599–608.

30. Zhou, X.; Liang, X.; Du, X.; Zhao, J. Structure based user identification across social networks. *IEEE Trans. Knowl. Data Eng.* **2017**, *30*, 1178–1191. [[CrossRef](#)]
31. Li, Y.; Su, Z. A Comment on ‘Cross-Platform Identification of Anonymous Identical Users in Multiple Social Media Networks’. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1409–1410. [[CrossRef](#)]
32. Li, Y.; Su, Z.; Yang, J.; Gao, C. Exploiting similarities of user friendship networks across social networks for user identification. *Inf. Sci.* **2020**, *506*, 78–98. [[CrossRef](#)]
33. Qu, Y.; Yu, S.; Zhou, W.; Niu, J. FBI: Friendship learning-based user identification in multiple social networks. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–6.
34. Reza, K.J.; Islam, M.Z.; Estivill-Castro, V. Privacy protection of online social network users, against attribute inference attacks, through the use of a set of exhaustive rules. *Neural Comput. Appl.* **2021**, *33*, 12397–12427. [[CrossRef](#)]
35. Xing, L.; Jia, X.; Gao, J.; Wu, H. A Location Privacy Protection Algorithm Based on Double K-anonymity in the Social Internet of Vehicles. *IEEE Commun. Lett.* **2021**, *25*, 3199–3203. [[CrossRef](#)]