CROSS-PLATFORM RECOGNISATION OF UNKNOWN IDENTICAL USERS IN MULTIPLE SOCIAL MEDIA NETWORKS

N. Naga Priyanka, N. Geetha and A. Viji Amutha Mary
Department of Computer Science and Engineering School of Computing Sathyabama University, Chennai, India
Email: vijiamumar@gmail.com

ABSTRACT

From very recent past years have witnessed the requirement and evolution of a vibrant research Crew on a large variation of online Social Media Network (SMN) platforms. Recognizing anonymous, same yet identical users among multiple SMNs is still a major problem. Clearly, saying that cross-platform exploration may help solve too many problems in social computing in both theory and applications. Up to now public profiles can be duplicated and easily impersonated by users with different purposes, most current user identification resolutions, which mainly focus on text mining of users 'public profiles, fragile. Some studies have attempted to match users based on the location and timing of user content as well as writing style. However, the locations are sparse in the majority of SMNs, and writing style is difficult to discern from the short sentences of leading SMNs such as, Sina Micro blog and Twitter. Moreover, up to now online SMNs are quite symmetric, existing user identification schemes based on network structure are not effective. The real-world friend cycle is highly individual and virtually no two users share a congruent friend cycle. So that, it is more accurate to use a friendship structure to analyze cross-platform SMNs. Up to now anonymous users were influenced to set up similar friendship structures in the different SMNs, here they proposed the Friend Relationship-Based User Identification (FRUI) algorithm. FRUI Algorithm calculates a match degree for all candidate User Matched Pairs (UMPs) only, UMP with top ranks are considered as identical users. We also developed two propositions to improve the efficiency of the algorithm. The Results of these extensive experiments demonstrate that FRUI performs much better than current network structure-based algorithms.

Keywords: friend relationship algorithm, user identification, cross-platform, social media network, anonymous identical users.

1. INTRODUCTION

In the past decade, there are many types of social networking websites have started and contributed immensely to too many volumes of real-world data on people's behaviors. Twitter was the 1, the largest micro-blog service, has more than 650 million of users and produces more than 375 million tweets per day [1]. Sina Micro-blog 2, the primary Twitter-style Chinese micro-blog site, has more than 555 millions of accounts and generates over 100 million tweets per single day [2]. Due to that diversity of various online social media networking sites (SMNs), people interested to use different SMNs for different purposes. For instance, RenRen 3, a Face book-style but antonymous SMN, is used in China for blogs, while Sina Micro-blog is used to share statuses. In other means, for every existent SMN fulfils some user requirements. In other terms of SMN management, matching these anonymous users across different SMN platforms that can be provide integrated components on the each user and inform corresponding regulations, such as targeting services of provisions. In the theory, these cross-platform explorations will allow a bird's-eye view of all SMN user behaviours. However, those nearly all recent SMN-based studies that focuses on the single SMN platform, yielding incomplete data.

User identification is also known for user recognition; user identity resolution is a, user matching, and the anchor linking program. Although there is no solution that can recognise all the identical anonymous SMN users, where some SMN elements may be used to recognise a portion of users across multiple SMNs. Many studies up to now have addressed the user reorganisation problem by examining public user profiles attributes, including screen name, birth-day, place, gender, profile photo, etc. [3], [4], [6], [7], [9], [10], [11], [13], [14], [15], [16], [17]. Up to now these attributes do not require exclusivity and are easily copied by users to different purposes (including malicious users), these schemes were regularly fragile. Some researchers have to leverage public user activities are to recognize users using post time, location & writing style [18], [19], [20], and [21]. Up to now location data is difficult to obtain and writing style is difficult to extract from short these sentences, these techniques were plagued by few limitations. Even though connections can be collected and are difficult to impersonate in nearly all SMNs, our literature review revealed only a few studies that explored in employing many user friends to recognise their users [22], [23], [24].

For the terms of data security and privacy, Narayanan [22] de-anonym zed a social network graph by correlating it with known identities. NS was the first effort to recognize users purely by using connections, and successfully matched 30% of the accounts with a 12% error rate. [23] Proposed a new Joint Link-attribute Algorithm (JLA) for match two social networks and obtained a new portion of identical users. The all new SMN connections fall into two categories: single-following connections as well as mutual-following connections. Single-following connections are also called following relation-ships or following links. If user A follows user B, then user A and user B have a following relationship (single-way fans in which one knows the other, but not vice versa). Following relationships are common in micro-blogging SMNs, such as Twitter and

Sina Micro-blog. Likewise, mutual-following connections are called friend relationships. In micro blogging SMNs, a friend relationship refers to the mutual following relationships between two users. In most other SMNs, such as Face book, RenRen and We chat, a friend relationship forms only if a friend request is sent by one user and confirmed by the other user. Friend relation-ships are difficult to fake by malicious users, and so that reflect real-world relationships much better. Because to their reliability and consistency, friend relationships are more robust in user identification tasks.

In that present study focused on friend relationships on SMNs and developed a new algorithm based on the networks. That algorithm can only be recognising a part of the identical users in real-world SMN. Anyways it can be applied jointly with other element based user identification algorithm to get more accurate identification result. That study makes the following contributions

a) To developing a uniform solution framework to the network structure based user identification. First of all, a set of seed mapped users that are provided manually or other-wise being identified. Iteration is used for re-recognise as many users as possible, using that seed or priori mapped users to along with network structures. In the other current literature, all network structure-based solutions perform in that manner.

b) While Proposing a novel Friend Relationship-based User Identification (FRUI) algorithms. In that analysis of cross-platform SMNs, we deeply mined friend relationships and network structures for that. In the real world, people interested to have mostly the same friends in different SMN, or the friend cycle is highly individuals. That more matches in the two un-mapped users' known friends, the higher the probability that they belong to the same individual in real world. Based on that fact, we proposed the FRUI algorithm. Up to now FRUI employs a unified friend relationship; it is apt to recognise users from a different network structure. Unlike existing algorithms [22], [23], [24], FRUI chooses candidate matching pairs from currently known identical users rather than not matched ones. That operation always reduces computational complexity, up to now only a very small portion of unmapped users are involved in each iteration. Moreover, up to now only mapped users are exploited, our solution is scalable and can be easily extended to online user identification applications. In contrast with current algorithms [22], [23], [24], that FRUI requires no more control parameters.

c) For providing concrete demonstrations of FRUI performance with three synthetic networks and two major online SMNs in China: Micro-blog and RenRen. The synthetic networks include Erdős - Rényi (ER) [25] random net-works, Watts - Strogatz (WS) [26] small-world networks and Barabási - Albert preferential attachment model (BA) [27] Network. The Findings show that FRUI is superior to NS in these networks. Moreover, FRUI is effective for the de-anonymization task, up to now the user identification task is similar to the de-anonymization problem.

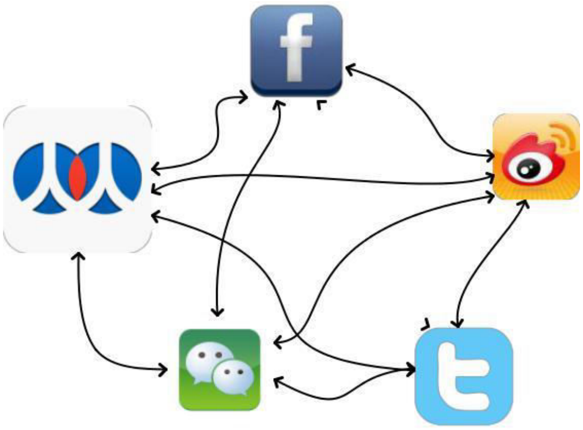


Figure-1. Social networks.

Table-1. Comparisons of FRUI to JLA and NS.

Differences	JLA	NS	FRUI
Type	Network Undirected	Directed	undirected
Additional control parameters	-	Eccentricity Threshold	-
Matching Method	Unmapped neighbours of nodes from single graph	Unmapped users from different networks	Mapped users
Attributes Used	Identified users and their degrees	Identified users and in/out - degree of the unmapped users	Identified users
Match Degree	Dice coefficient	Shared known outgoing /incoming neighbors and in/out degree	Shared identified friends

2. RELEVANT WORK ON CROSS-PLATFORM USER IDENTIFICATION

The Profiles, and contents and network structures are three cardinal components in an SMN. Accordingly, current studies on cross-platforms users identifications can be divided into three major categories: profile-based, content-based and network structure-based approaches.

2.1 Profile-based user identification

Several studies addressing anonymous user identification have focused on public profile attributes, including screen name, gender, birthday, city and profile image.

A screen name is the publically required profile feature in almost all SMNs. It has been widely explored as a way to recognize users across different SMNs. Perito *et al.* [3] calculated the similarity of the screen names and identified users by using binary classifiers. Similarly, Liu *et al.* [4] matched users in an overall unsupervised approach using screen names. Zafarani and Liu [5] proposed a method for the map identities across different SMN platforms, empirically validating several hypotheses. On that top of that work, they [6] further developed a user mapping method by modelling user behaviour on screen names. Among public profile attributes, the profile image is another feature that has received considerable study. Acquisitive al. [7] addressed the user identification task with a face recognition algorithm. Although both screen name and profile image that can recognise users, they cannot be applied to large SMNs. That is because of some users may have the same screen name and profile images. For example, many users have the screen name "John Smith" on Face book.

Obviously, leveraging a combination of profile features can result is much better than user identification. Iofciu *et al.* [8] pro-posed an approach by measuring the distance between user profiles. Motoyama and Varghese [9] gathered attributes (education, occupation, etc.) as sets of words and matched users by calculating the similarity of users. Goga [10] linked accounts belonging to the same person identity, based solely on the profile information. Curtis [11] pro-posed a weighted ontology-based user profile resolution technique. Abel *et al.* [12] aggregated user profiles and matched users across systems. Similar studies across multiple platforms are also found in [13], [14], and [15].

2.2 Content-based user identification

The Content-Based User Identification solutions attempt to recognize users based on the times and locations that users post content, as well as the other writing style of the content. Zhen *et al.* [18] proposed a framework for authorship identification using the writing style of online messages and those classification techniques. They calculated the combined similarities of user's social, spatial, temporal and that text information in different SMNs, and examined a stable matching problem between two sets of user account. Jog *et al.* [21] exploited the geo-location attached to users' posts, that timestamp of posts, and users' writing style to address user identification tasks. That Geo-location appears to have forceful components for user recognition. However, that information is often sparse in SMNs, up to now only a small portion of the users are willing to post their locations. Although writing style solutions to per-form well in scenarios involving long content, these techniques are not applicable to the SMNs such as Twitter and Sina Micro-blog, in which short sentences were most likely to be posted.

2.3 Network structure-based user identification

Network structure-based studies [22], [23], [24] on user identification across the multiple SMNs are used for recognize identical users solely by user network structures and seed, or priori, identified users.

The task on the user identification is closely related to the de-anonymization problem [28] for the privacy-preserving social network analysis, in which re-identifies individuals in online published SMN datasets. In that context, SMN data are anonymised before release. Zhou and Pei analyzed the neighbourhood attacks of de-anonymization and proposed privacy preservation approaches using k-anonymity and l-diversity [29], [30]. Other de-anonymization attacks have also been analyzed [31], [32], and [33]. Up to now cross-platform user identification is similar to the de-anonymization tasks, that it can be applied to address the de-anonymization problem. As demonstrated in the experiments of Section 5, FRUI performs much better than NS, that de-anonymization algorithm.

So that the joint use of profile information, user behaviors hid-den content and network structures may lead to better results. The Jain and Kumara guru [16], [17] developed Finding Memo, a method that matches Face book and Twitter ac-counts. However, that text-based network search method has low accuracy and the high complexity in terms of user identification, up to now only the texts of the same nicknames were recognized when searching for the friend sets of friends [12, 14]. Bruno *et al.* [23] integrated all profiles with a network structure using a Conditional Random Fields model and obtained better user identification results.

Network structure based user identification is hard nut to crack, and that can be used for to recognise only a portion of identical users. NS was the first network structure-based user recognition algorithm across SMNs, can carry out user recognition tasks by using only the network structure, and identified 31% identical users in the ground-truth dataset [22].

The FRUI differs from the two existing algorithms, JLA and NS, in the following aspects (see Table-1):

a) That NS is suitable for directed networks, while JLA and FRUI focus on undirected networks. JLA is restricted in un-directed networks by Conditional Random Fields, while that FRUI relies on friend relationships, as that is of more reliable and consistent with real-life friendship.

b) That NS requires an additional control parameter (eccentricity threshold) to recognise user match and if the eccentricity is above a pre-determined threshold, NS accepts a candidate User Matched Pair (UMP). So, clearly the thresh-old is a free parameter and should be provided in advance. In the contrast, there were no extra free parameters are required by JLA and FRUI.

c) That JLA compares unmapped neighbors of nodes from one of the two SMNs, while that NS matches unidentified users from different networks by comparing the mapped neighbors of each node. FRUI aims to recognise the most matched pairs among mapped users, but does not iterate unmapped users. So that, it markedly reduces computational complexity.

d) That NS employs unmapped users' in- and out-degrees, as well as the identified users, to calculate scores in directed networks. JLA, in contrast, employs identified users and their degrees. Any mapped user has different degrees in different SMNs. So that, component son how these degrees are obtained should be discussed in advance. Comparatively, only identified users are required in our FRUI.

e) NS computes the match degree by shared known outgoing/incoming neighbors and out-/in-degrees, with the assumption that users in different SMNs have similar outgoing/incoming neighbors. JLA uses a dice coefficient to calculate the match degree, which may mismatch users with high probability when two users share only a few known friends. In contrast, FRUI takes into account the number of shared known friends in match degree calculation. Moreover, the calculation method to match degree in FRUI has been shown to be simpler and more effective than those of JLA and NS (see Section 4.2).

3. PROBLEM DEFINITIONS IN CROSS-PLATFORM USER IDENTIFICATION

That section defines related terminologies, presents a uniform solution framework for user identification solely by using network structure, and defines the problem of friend relationship-based user mapping.

3.1 Terminology

Social media refers to virtual communities and networks in which people create, share, and/or exchange information and ideas [34]. In social media, people are allowed to (1) construct public or semi-public profiles within a bounded system, and (2) articulate with a set of other users with whom they share connections [35]. From that description, it is evident that an SMN is composed of three crucial elements: users with public or semi-public profiles, interaction information among users (or content), and connections (or network). Below are formal definitions of these terms.

Define 1 (SMN). An SMN is defined as SMN = {U, C, I}, where U, C and I denote the users, connections and interactions among users, respectively. By Delving into the main components of an SMN, one can easily find that both C and I are generated by U. For that extent, C and I can be treated as the attributes of U, and U is the core item in the SMN. So that, user identification is of paramount importance in cross-SMN studies.

In that study, SMNA is used to represent SMN A. With-out a specification; SMNA denotes the pure network structure of SMN A.

Define 2 (User Entity). A User Entity (UE) is a user in combination with his or her profile, connections and interaction content. So that An SMN is a set of UEs which has the same number as the accounts in the SMN. Similarly, UEA is used to indicate the UE list of SMNA, and UEAi is taken as the token of the i-th element in UEA. Define 3 (User Matched Pair). The Given that SMNA and SMNB, if UEAi and UEBj belong to the same individual in real-life, which is denoted as Ψ, then we hold that UEAi and UEBj match on Ψ, and they compose a User Matched Pair UMPΨ. UMPΨ can also be expressed as UMPA~B(i, j) or UMP(UEAi, UEBj), equivalently.

Define 4 (Fully Overlapped SMNs). If a UMP set covers all users in both SMNA and SMNB, and the two SMNs have the same network structure, then all UEs and their relations overlap in the two SMNs and we can state that the SMNA and SMNB are fully matched.

Define 5 (Priori UMP). Priori UMPs are UMPs given in advance, before user identification resolution work is executed. These Priori UMPs are often used as the condition to recognise more UMPs.

Define 6 (Valid Priori UMP). Valid Priori UMPs are Priori UMPs that are useful to user identification. In FRUI, only the Priori UMPs connected to unmapped users in both SMNs are Valid Priori UMPs.

Define 7 (Adjacent Users). These Unmapped users with connections to Priori UMPs are defined as Adjacent Users. Then Only Adjacent Users are involved in an iteration process in FRUI.

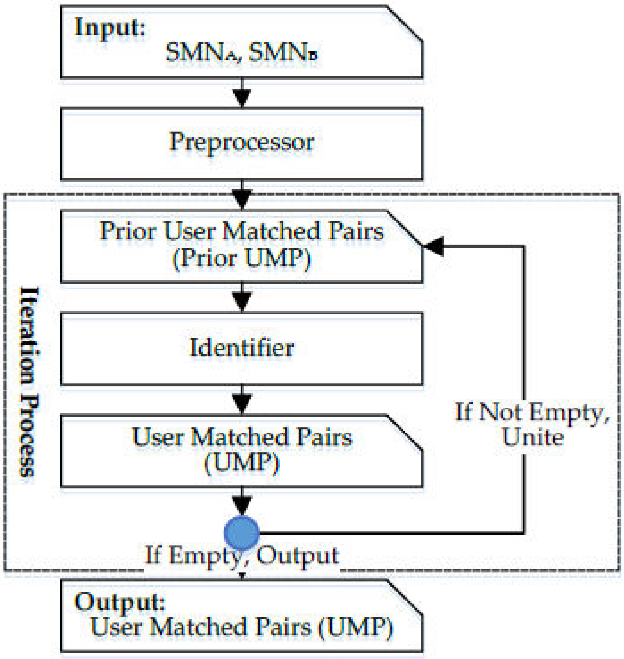


Figure-2. Uniform solution framework. The network structure-based user identification first obtains Priori UMPs through a Pre-processor, and then identifies more UMPs through the Identifier.

3.2 Uniform solution framework

According to the definitions above, recognising users across two SMNs yields a UMP set using Priori UMPs. Thus, network structure-based user identification algorithms are divided into two main steps: the Priori UMP set recognition and the iteration identification. The framework has two modules: the Pre-processor and Identifier. These Pre-processor is set to reveal Priori UMPs through a limited number of profiles. So that, the Identifier, the core component of our resolution, these recognizes UMPs through users' networks in an iterative manner. In the segment of uniform solution framework, the input is two SMNs in which user identification is performed, and the output is the UMP. After a set of Priori UMPs is to be identified, these Identifier is implemented to recognize a set of new UMPs using network structure in the iteration process element. The identified UMP set, if it is not empty, is in union with the Priori UMP set and yields to the new Priori UMPs for the next iteration. The iteration process ends when no UMP can be identified by the Identifier.

3.3 Problem definition

In the outside world, we can infer that each person has his own friend cycle, that which is highly individual. So that, if we know all of a person's friends, we probably know who he is. By using SMNA in Figure-3(a) as an example, if one has only user 1 as a friend, it is obvious that he must be user 3. If that is someone asks who has the friend set of users 1 and 2, it is obviously user 4. Users interested to have similar friends across different SMN. Wu's survey revealed that a user in QQ, majorly the most famous Instant Messengers in China, shared telephone numbers of about 60% of his QQ friends [36]. Now We investigated 129 individuals with both RenRen as well as Sina Micro-blog accounts and found that an average of 67.5% of their friends in Sina Micro-blog concurred in RenRen. These Numbers of their Sina Micro-blog friends varied from 4 to 317. Consequently, we can hypothesize that: (1) If some Valid Priori UMPs are given, then a set of candidate UMPs can be deduced, and (2) the more known friends are shared in that candidate UMP, the higher the probability that they belong to the same individual. Using the fully overlapped SMNA and SMNB, as an illustration, figure shows SMNA and SMNB with the Priori UMPs, UMPA~B(1, 1) and UMPA~B(2, 2). So that Intuitively, (UEA1, UEB1) and (UEA2, UEB2) are placed together. And now Then we find that UEA4 and UEB4 share the same friend set, which is the largest set based on the current UMP set. So all We can then conclude that UEA4 and UEB4 stand a good chance of forming a new UMP. After UMPA~B(4, 4)was identified, now it is added to Priori UMPs. By repeatedly using the above method, UMPA~B(5, 5), UMPA~B(6, 6), UMPA~B(7, 7), and UMPA~B(1, 1) were identified consecutively.

The main question in the above scenario is that the overlap of the users' friends. And to address that issue, and we discuss the overlap of SMNs, now including node and edge overlap, below.

a) These Node overlaps. Many studies have verified that numerous users are overlapped in different SMNs. And now nearly all cross-platform user identification studies mention node overlap, because it is the fundamental assumption to solve that issue. Now in the Early in 2007, 64% of Face book users had MySpace accounts [37]

b) The Edge overlap. Until very recently, no statistical studies quantified the relationship overlap in two SMNs. Anyways However, some studies noted that these relationships overlap to a certain extent. And NS [22], which identifies users purely through networks in ground-truth datasets, proved that users have similar relationships in Twitter and Flickr. Paridhi [16], [17] it also found that users interested to connect with a segment of the same people across SMNs, and he introduced network structure to improve the accuracy of user's identification between Twitter and Face book. All these network structure involved user identification solutions conceal the fact that edges are partially overlapped in different SMNs. The reasons for edge overlap may be that: (1) So many people were interested to set up relations with their real and real-life friends (e.g: classmates, workmates, and family members) in different SMNs. (2) now the People interested very interested to connect to those with their similar interests. (3) In directed SMNs, users were allowed to be followed by their "fans."

4. PREPROCESSOR

Always a pre-processor was designed to acquire as many Priori UMPs as possible. Now currently, there is no common approach available to obtain UMPs between two SMNs as well. Now the Specified methods must be formulated according to given SMNs. So Although no unified process is suitable for the Pre-processor, some algorithms can be adopted according to the application, e.g., email address, screen name, URL, etc. so that an email address appears to be a unique feature for each account, and can be used to collect Priori UMPs. Balduzzi. [38] It explored email addresses to find identical users among different SMNs with the "Friend Finder" mechanism. Now however, up to now email addresses are private, nearly all SMNs have disabled the "Friend Finder."

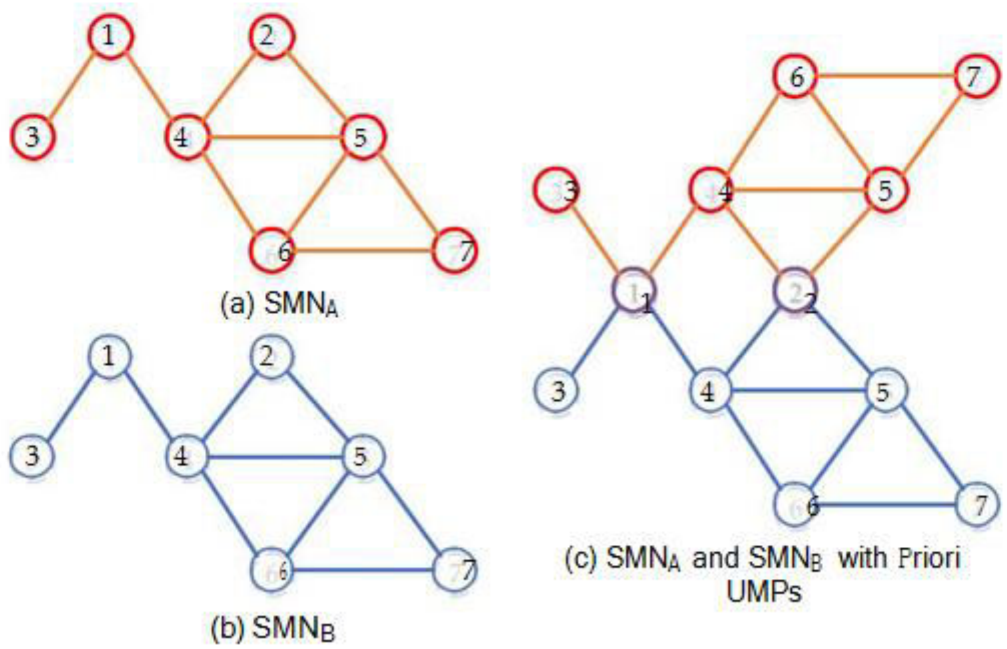


Figure-3. Examples of two basic SMNs with FRUI. (a) and (b) are two simple SMNs, SMNA and SMNB. (c) Shows that SMNA and SMNB have two Priori UMPs 1 and 2, denoted as UMPA~B (1, 1) and UMPA~B(2, 2). Up to now user 4 in both SMNs shares the same and the largest friend set, UMPA~B(4, 4) is identified and added to the Priori UMP set. So that Iteratively, UMPA~B(5, 5), UMPA~B(6, 6), UMPA~B(7, 7), and UMPA~B(1, 1) are identified consecutively.

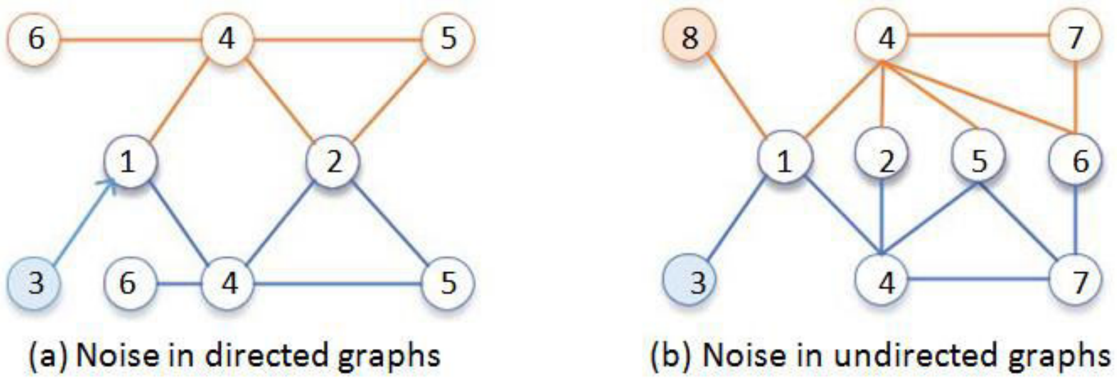


Figure-4. Noise in user identification. These nodes and links at the top, bottom and middle denote SMNA, SMNB and identified users, respectively. (a) It shows the noise introduced in directed graphs that to a link with no arrow indicates a mutual-following relationship as well. Now a following relationship from UEB3 to identified to the user 1 would prevent the identification of UMPA~B(4, 4), by using NS when user 4 has a large in- or out-degree in it.

That is the same as the friend relationship between users 3 and 1 in (b). In (b), the fact that M83 is larger than M44 using JLA results in the identification of an incorrect UMPA~B(8, 3). Now as stated in [17], one individual interested to use the same nickname in the different SMN platforms. So that and thus, when a nick-name can be taken as that UE, the nickname can be obtained and is unique. Now so that in these cases, the pair of UEs having the same nickname from two SMNs can be treated as that UMPs. So that, in these scenarios where the people are allowed to have the same (or) similar usernames such as the RenRen, so that method fails to recognising the user. Now the one solution is to verify candidate UMPs through the other accessible factors, such as description, location, and birthday. And now with the advances in SMN services, more than SMNs allow users are to bind their accounts with other major SMNs. So that now In that case, priori knowledge can be obtained with bound information. For example, PaPa and ChangBa, the two major mobile applications (apps) in China, encourage users to link their Sina Micro-blog accounts for commercial interests, bridging their websites with the largest micro-blog services in China.

5. IDENTIFIER

In that section, we systematically discuss our solution to the user identification problem by leveraging users' friends, and develop two propositions to improve the efficiency of our algorithm.

5.1 Methodology

This identifier finds UMPs using connections among users on Priori UMPs. As noted above, a match degree for each candidate UMP should be calculated should be in advance. NS formulates the match degree of using in- and out-degrees in directed networks.

$$M_{ij} = s(\mathrm{UE}_{Ai}, \mathrm{UE}_{Bj}) = \frac{c_{\text{in}}}{\sqrt{d_{\text{in-}Bj}}} + \frac{c_{\text{out}}}{\sqrt{d_{\text{out-}Bj}}}$$

Where this cin and cout denote the numbers of shared of incoming and outgoing neighbours of UEAi and UEBj respectively, and din-Bj and dout-Bj stand for the in- and out-degrees of UEBj. These NS operates under the assumption that the same user in different SMNs has the same amount of in and out degrees. Now in NS, Mij depends heavily on din-Bj and dout-Bj. In the single-following connections, users can follow any other users freely, which would introduce noise for the user identification task simply. We take Figure-4(a) as an example, when UEB3 follows UEB1, M43 = 1 in NS. Once UEB4 has a large in- or out-degree, NS has difficulty recognising UMPA~B(4, 4). Nevertheless, our datasets and [39] indicate that real-world SMNs are symmetric, with many nodes sharing a portion of neighbours in SMNs. That would prevent identification of many identical users. Figure-4(b) displays undirected graphs. So, Although UEB3 and UEA4 share only one known friend, that may hinder identification of UMPA~B(4, 4). In other words, even though as many as 10 out of 100 friends are observed between UEA100 and UEB100, any UE with a much lower degree that happens to share one identified user with UEA100 may hinder identification of UMPA~B(100, 100). Now Consequently, NS may miss numerous matched users, especially in the sparse SMNs. As we discussed above, the friend relationship needs the confirmation by the two users, and which is much more reliable and consistent in SMNs. Thus, it can reduce the noise introduced by a discretionary single-following relationship.

5.2 Algorithm

Now clearly, the number of the shared known friends is the key value to be calculated in FRUI. To lower the complexity, we present the following two propositions. Proposition 1. Given two SMNs, SMNA and SMNB, with s pairs of Priori UMPs, the m × n matrix R = QAPB contains the numbers of shared known friends, where m and n are the numbers of Adjacent Users in SMNA and SMNB, rij stands for the number of shared friends of UIAi and UIBj in the s pairs of UMP, QA and PB denote the connections between Adjacent Users and identified users in SMNA and the connections between identified users and Adjacent Us-ers in SMNB, respectively.

Proof. QA represents the connections between Adjacent Users and identified users, and can be written as QA= [∟1T, ∟2T, …, ∟mT]T where (•)T is the transposition of (•), ∟i ⅃ { 0, 1 }1×s denotes the connections of the i-th Adjacent User to identified users. Similarly, PB can be written as PB = [β1, β2, …, βn], where βj ⅃ { 0, 1 } s×1 denotes the connections of the j-th Adjacent User to identified users. As a result, ma-trix R can be converted to

$$R = Q_A P_B = \begin{bmatrix} □□_1 □ \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_n].$$

Considering □$_i$□$_j$ with □$_i$ = [a_{i1} a_{i2} … a_{is}] and β_j = [b_{j1} b_{j2}… b_{js}]^T, where a_{ik} and b_{jk} denote whether there is a connection between the *i*-th Adjacent User and the *k*-th identified user in SMNA, and the *j*-th Adjacent User and the *k*-th identified user in SMNB, respectively. If both the *i*-th Adjacent User in SMNA and *j*-th Adjacent User in SMNB are connected with the *k*-th identified user, $a_{ik}b_{jk}$ is assigned 1 and 0 otherwise,

$$a_{ik} b_{ik} = \begin{cases} 1, \text{ if } \mathrm{UE}_{Ai} \text{ and } \mathrm{UE}_{bj} \text{ share the } k\text{-th identified user;} \\ 0, \text{ otherwise.} \end{cases}$$

Then, $\alpha_i\beta_j$ can be calculated as

$$\alpha_i \beta_j = \sum_{k=1}^{s} a_{ik} b_{jk} \quad .$$

Obviously, the i-th Adjacent User in SMNA and the j-th Adjacent User in SMNB share ∟i∟j identified users. Up to now one and only Adjacent Users are involved, that greatly reduces computational complexity. Proposition 1 reduces complexity to O(∑sdAidBi) ≤ O(sdAdB) in calculating the numbers of shared known friends when the s pairs are calculated separately, where dA and dB denote the maximal degrees of users in SMNA and SMNB, respectively as well. Consider that k UMPs are identified in the t-th iteration. By taking these k Priori UMPs as the input for Proposition 1, we have another matrix ∟R= ∟QA⅃PB, where ∟R, ∟QA, and ⅃PB have a similar meaning as R, QA, and PB in Proposition

1. The combination of R(t) and ⌞R returns R(t+1), where R(t) denotes matrix R for the t-th iteration. That leads to Proposition 2.

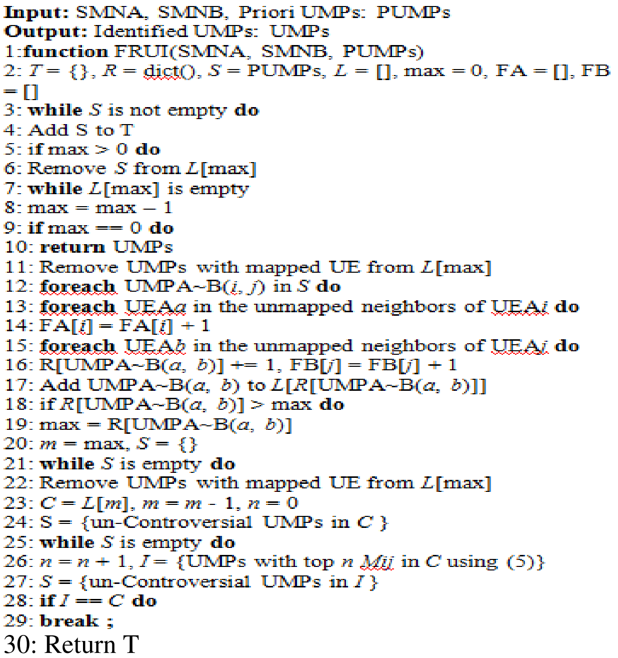
Proposition 2. In the t-th iteration process, if k UMPs are generated, then R(t+1) = combine(R(t), ⌞R)

Now Where the function combine removes the items with, any UE included in the k UMPs, and returns the union of the remaining ones in R(t) and ⌞R. on the other hand the union operation adds the value of the items in both R(t) and ⌞R, and joins the left items.

Now the Proof Up to now those items with any UE in the k UMPs that will not be used in subsequent identifications, they are re-moved. And for those candidate UMPs that occur in both R(t) and ⌞R, their shared known friends are the sum of those in R(t) and ⌟R. The shared known friends of the candidate UMPs only existing in R(t) are unchanged. Here the number of the shared known friends of the new candidate UMPs generated by the k UMPs is the one in ⌟R.

It is worth-noting that only users in both SMNs with connections to newly identified users are involved in each iteration process, based on Proposition 2. Furthermore, when t = 0, R(t) is a 0 matrix, the k identified UMPs turn to Priori UMPs, then, Proposition 2 is degenerated into Proposition 1.

FRUI Algorithm:



```
Input: SMNA, SMNB, Priori UMPs: PUMPs
Output: Identified UMPs: UMPs
1:function FRUI(SMNA, SMNB, PUMPs)
2: T = {}, R = dict(), S = PUMPs, L = [], max = 0, FA = [], FB
= []
3: while S is not empty do
4: Add S to T
5: if max > 0 do
6: Remove S from L[max]
7: while L[max] is empty
8: max = max – 1
9: if max == 0 do
10: return UMPs
11: Remove UMPs with mapped UE from L[max]
12: foreach UMPA~B(i, j) in S do
13: foreach UEAa in the unmapped neighbors of UEAi do
14: FA[i] = FA[i] + 1
15: foreach UEAb in the unmapped neighbors of UEAj do
16: R[UMPA~B(a, b)] += 1, FB[j] = FB[j] + 1
17: Add UMPA~B(a, b) to L[R[UMPA~B(a, b)]]
18: if R[UMPA~B(a, b)] > max do
19: max = R[UMPA~B(a, b)]
20: m = max, S = {}
21: while S is empty do
22: Remove UMPs with mapped UE from L[max]
23: C = L[m], m = m - 1, n = 0
24: S = {un-Controversial UMPs in C }
25: while S is empty do
26: n = n + 1, I = {UMPs with top n Mij in C using (5)}
27: S = {un-Controversial UMPs in I }
28: if I == C do
29: break ;
```

30: Return T

6. EXPERIMENTAL STUDIES

To evaluate the identification resolution, we verify FRUI in both synthetic and ground-truth networks. Now all the experiments were conducted in the computer with 8G memory and 2.8GHz. Now we used NS as a baseline because it is closest to FRUI as a state-of-art, as a network structure-based user identification algorithm, while JLA performs better when profile attributes are added and without a loss of generality, the eccentricity threshold of NS is set to 0.5 in the experiments.

6.1 Synthetic network experiments

For validate the performance of FRUI, we conducted experiments in (ER) [25] random networks, Watts Strogatz (WS) [26] small-world networks and the Albert preferential attachment model (BA) [27] networks. The degree distribution of the three synthetic networks with 10,000 nodes. For the degree distribution, and the ER, WS and BA networks followed a normal distribution, now a bell distribution and a power-law distribution, respectively. Now the degree distribution of this WS network was similar to that of the ER network, because both ER and WS networks are generated from the regular random network by rewiring each edge with a probability of it. If all the edges are rewired so that the probability of rewiring equals 1, now the network turns out to be an ER network; otherwise, it is a WS network. In the experiments, the probation rewiring in WS network was 0.5.

We generated 10 pairs of networks in experiments to illustrate the performance of FRUI in synthetic networks. In the ER and WS network experiments, five networks with 5000 nodes and another with five with 10,000 nodes were created, with p equalling 0.05, 0.1, 0.2, 0.3 and 0.4, respectively. Now similarly, in the BA network experiment, five networks with 10,000 nodes and another five with 20,000 nodes were produced. And the number of edges to attach a new node to the existing nodes, it denoted as m, increased from 20 to 100 by Table 2 displays results of empirical testing in ER networks. The FRUI identified almost all identical nodes in the 10 pairs of networks, with only 2% UMPs. The table 3 illustrates the performance of FRUI in WS networks. Analogous to the experiments in the ER network, FRUI recognized no less than 75.9% of all UMPs, with 2% UMPs, as well as and no less than 96.1% of all the UMPs with 5% UMPs. Table 4 shows that FRUI also has good performance in BA networks. No less than 89.4% of all UMPs can be identified by 5% UMPs. In all experiments on ER, WS and BA networks, FRUI revealed nearly all UMPs, with 5% UMPs. That indicates that FRUI can address the user identification task with a small portion of UMPs.

Now we also conducted experiments to compare that efficiency of FRUI and NS. Figure compares FRUI and NS in the three synthetic networks. Figure (a) and (b) displays results of experiments conducted with p = 0.05 in networks generated by ER and WS models, with 1000 nodes and 5000 nodes. These Comparisons of FRUI and NS in ER and WS networks with 10,000 nodes were not illustrated, up to now both FRUI and NS identified almost all the UMPs, with 2% Priori UMPs in these networks. Figure-(c) displays empirical testing results in the BA network with m = 20. In sum, these findings demonstrate that FRUI is more efficient than NS in recognising nodes in the now three synthetic networks. Figure- (a) and (b) indicates that when p = 0.05 and 0.4, the performance of both FRUI and NS decreased in the net-works with 1000 nodes. Now that is because extra Priori UMPs were the necessary to distinguish nodes in the same network in the ER network when p = 0.4, while more Priori UMPs were needed to ensure enough shared known nodes among identical nodes in the WS network when p = 0.05. The density of the BA network counts for my/ (v2) = 2m/ (v - 1), where v represents the number of nodes in the network. Finally, obviously the BA network is much sparser, and in that situation more Priori UMPs are required to ensure that the correct UMPs share enough known friends in the sparser networks. In other words, these smaller the m, the sparser the network, and the more Priori UMPs are required.

6.2 Social media network experiments

So overall that section, we use ground truth datasets to evaluate the user identification resolution. So, In order to verify FRUI in different types of SMNs, we collected data from two heterogeneous SMNs: Sina Micro-blog and RenRen. These Sina Micro-blog dataset was captured from that Sina Micro-blog search page, while the RenRen dataset was directly obtained from its Open API. As we shown in Table-5, the Sina Micro-blog dataset consisted of 1.17 million users and 1.9 million friend relationships, and we each user had an average of 3.2 friends. The RenRen dataset was comprised of 5.5 million nodes and 14.6 million edges, and each user had an average of 5.3 friends. So that, the RenRen dataset was much denser than Sina Micro-blog's. This Figure illustrates the degree distributions of the two graphs. Clearly, they are scale-free networks [27].

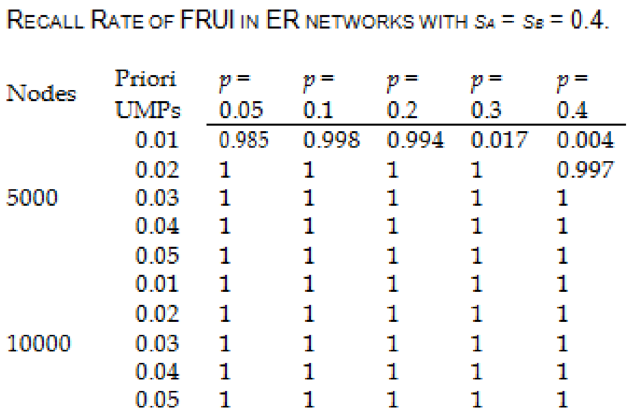
In this point of view of experiments, we randomly chose a number of shared nodes as well as Priori UMPs. Then we executed user identification in both NS and FRUI. Now we increased the percent-age of Priori UMPs in all UMPs from 0.01 to 0.1 by 0.01. Up to now the average degrees of both Sina Micro-blog and RenRen are fairly low, and only the nodes with no less than θ neighbors were selected as overlap node as well. Now to check the performance of FRUI, they increased θ from 20 to 100 by 20. Figure (a) compares the recall rates of FRUI and NS with θ= 80. The FRUI identified around 50% UMPs with 5% Priori UMPs, while NS returned no more than 40% UMPs with 10% Priori UMPs in the Sina Micro-blog dataset. The FRUI also identified many more UMPs in the RenRen dataset. So it is apparent that FRUI, performed much better. Figure (b) also shows that FRUI performs better in precision than NS in both datasets. Figure (c) shows that FRUI is less costly than NS in terms of running time, and which stands for the elapsed time during the identification process. These Results show that FRUI is more efficient in practice, which is consistent with the theoretical analysis. The Figure- (d) compares the recall rates of NS and FRUI with 8% Priori UMPs in both datasets generated from Sina Micro-blog and RenRen and the Results show that FRUI returned many more UMPs than NS in all scenarios in both datasets. So that In the Sina Micro-blog dataset, the recall rate of FRUI is around 0.5, which is much larger than that of NS, so it it's all about 0.3. The same trend occurred in datasets produced by the RenRen. So that indicates that FRUI has more capacity to find UMPs. It indicates that the distributions of the number of the number of shared neighbours between any two connected users in the Sina Micro-blog and RenRen follow an exponential distribution. Most users have much lower degrees in real-world datasets, so that most connected users share a few common users. Those low degree users introduce noise for the NS.

Figures (a) and (c) further reveals that both FRUI and NS yielded better results in the RenRen sub graphs than in the Sina Micro-blog sub graphs. That is because both Sina Micro-blog and RenRen are sparse SMNs. Analogous to discussions the synthetic networks in Section 5.1 and consistent with the analysis in Section 4.2, the denser network structure can benefit both FRUI and NS in those sparser SMNs.

To study the effects of noise, we conducted experiments shown in Figure. It is evident that as the noise increases, the evaluation indices decrease. Nonetheless, FRUI still identified a large volume of identical users in the noisy environment. We also evaluated how well FRUI identified users across Sina Micro-blog and RenRen by conducting three groups of experiments. So in each experiment, we selected a pair of sub graphs by starting with the identical users and extracting two-layer friends using a breadth-first search. Presently, then we manually labelled 150 users as Priori UMPs and we performed FRUI and NS. Up to now the exact number of identical users is unknown, only the precision was compared. We randomly chose 300 identified UMPs to check the precision. If the table 6 illustrates the empirical results.

Both FRUI and NS were identified a portion of the total nodes, so up to now most users have only one neighbour. However, FRUI returned many more UMPs and we obtained much higher precisions in all three experiments as well. Now these findings reveal that FRUI is much more proficient for recognizing identical users across Sina Micro-blog and RenRen.

Table-2.



RECALL RATE OF FRUI IN ER NETWORKS WITH $S_A = S_B = 0.4$.

Nodes	Priori UMPs	$p =$ 0.05	$p =$ 0.1	$p =$ 0.2	$p =$ 0.3	$p =$ 0.4
5000	0.01	0.985	0.998	0.994	0.017	0.004
	0.02	1	1	1	1	0.997
	0.03	1	1	1	1	1
	0.04	1	1	1	1	1
	0.05	1	1	1	1	1
10000	0.01	1	1	1	1	1
	0.02	1	1	1	1	1
	0.03	1	1	1	1	1
	0.04	1	1	1	1	1
	0.05	1	1	1	1	1

Table-3.

RECALL RATE OF FRUI IN WS NETWORKS WITH $S_A = S_B = 0.4$.

Nodes	Priori UMPs	$p =$ 0.05	$p =$ 0.1	$p =$ 0.2	$p =$ 0.3	$p =$ 0.4
	0.01	0.040	0.023	0.017	0.009	0.008
	0.02	0.759	0.928	0.993	0.991	0.993
5000	0.03	0.780	0.984	0.998	0.998	1
	0.04	0.964	0.994	0.999	1	1
	0.05	0.961	1	1	1	1
	0.01	0.027	0.052	0.997	0.997	0.991
	0.02	0.899	0.997	1	1	1
10000	0.03	0.993	1	1	1	1
	0.04	0.998	1	1	1	1
	0.05	0.999	1	1	1	1

Table-4.

RECALL RATE OF FRUI IN BA NETWORKS WITH $S_A = S_B = 0.4$.

Nodes	Priori UMPs	$m =$ 20	$m =$ 40	$m =$ 60	$m =$ 80	$m =$ 100
	0.01	0.012	0.008	0.008	0.013	0.012
	0.02	0.637	0.980	0.983	0.977	0.962
10000	0.03	0.813	0.990	0.994	0.991	0.992
	0.04	0.882	0.996	0.998	0.998	0.996
	0.05	0.894	0.998	0.999	0.998	0.998
	0.01	0.834	0.975	0.053	0.936	0.888
	0.02	0.889	0.994	0.998	0.997	0.996
20000	0.03	0.909	0.997	0.999	1	1
	0.04	0.916	0.998	1	1	1
	0.05	0.919	0.999	1	1	1

Table-5.

NETWORKS OF THE GROUND TRUTH DATASET

Network	Nodes	Edges	Average Degree
Sina Microblog	1.17M	1.9M	3.2
RenRen	5.5M	14.6M	5.3

Table-6.

COMPARISON OF FRUI AND NS IN IDENTIFYING USERS ACROSS SINA MICROBLOG AND RENREN

# pair of Subgraphs			Identified UMPs		Precision	
(Nodes)			FRUI	NS	FRUI	NS
1	Sina RenRen	7926 26422	1962	691	0.453	0.203
2	Sina RenRen	7131 24052	1645	598	0.427	0.173
3	Sina RenRen	7733 24893	1734	713	0.430	0.217

CONCLUSIONS

The study addressed that problem of user identification across all SMN platforms and offered a proper innovative solution. As we need a key aspect of SMN, a network structure is of paramount importance and helps resolve de-anonymization user identification tasks. So that, we proposed a uniform network structure that is based on based user identification solution. As of now we developed a novel friend relationship-based algorithm called FRUI. To improve the efficiency of FRUI, we described two major propositions and addressed the complexity. Finally, we verified our algorithm in both synthetic networks and ground-truth networks. The results of our empirical experiments reveal that network structure can accomplish important user identification work. Our FRUI algorithm is simple, yet be efficient, and performed much better than NS as well, the present existing state-of-art network structure-based user identification solution it will be. In the scenarios when raw text data is sparse to, the incomplete, or hard to obtain due to privacy settings, the FRUI is extremely suitable for cross-platform tasks also. Moreover, our resolution can be easily applied to any SMNs with friendship networks, those including Twitter, Face book and Foursquare.

So that, only a portion of identical users with different nicknames can be recognized with that method. That study built the foundation for further studies on that issue. Ultimately, it is our hope that a final approach can be developed to recognise all identical users with different nicknames. Other user identification methods can be applied simultaneously to examine multiple SMN platforms. These methods are complementary and not mutually exclusive, up to now the final decision may rely on human user's involvement. So that, we suggest using these methods synergistically and considering strengths and weaknesses for the best results.

REFERENCES

[1] M. Hay, G. Makalu, D. Jensen and D. Owsley. 2008. Resisting structural identification in anonym zed social networks. Proc. of the 34th International Conference on Very Large Databases (VLDB'08). pp. 102-114.

[2] Xinhua net. 2014. Sina Micro-blog achieves over 500 Million Users. http://news.xinhuanet.com/tech/2012-02/29/c_122769084.htm.

[3] D. Period, C. Castelluccia, M.A. Kadar and P. Manilas. 2011. How unique and traceable are usernames? Privacy Enhancing Technologies (PETS'11). pp. 1-17.

[4] J. Liu, F. Zhang, X. Song, Y.I. Song, C.Y. Lin and H.W. Hon. 2013. What's in a name? An unsupervised approach to link users across communities. Proc. of the 6th ACM international conference on Web search and data mining (WDM'13). pp. 495-504.

[5] R. Maharani and H. Liu. 2009. Connecting corresponding identities across communities. Proc. of the 3rd International ICWSM Con-ferrous. pp. 354-357.

[6] R. Maharani and H. Liu. 2013. Connecting users across social media sites: a behavioral-modeling approach. Proc. of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13), pp.41-49.

[7] A. Acquits, R. Gross and F. Stutz man. 2011. Privacy in the age of augmented reality. Proc. National Academy of Sciences.

[8] T. Iofciu, P. Fankhauser, F. Abel and K. Bischoff. 2011. Recognising users across social tagging systems. Proc. of the 5th International AAAI Conference on Weblogs and Social Media. pp. 522-525.

[9] M. Motoyama and G. Varghese. 2009. I seek you: searching and matching individuals in social networks. Proc. of the 11th inter-national workshop on Web Information and Data Management (WIDM'09). pp. 67-75.

[10] O. Gaga, D. Period, H. Lei, R. Teixeira and R. Summer. 2013. Large-scale Correlation of Accounts across Social Networks. Technical report.

[11] K. Curtis, S. Sherri, I. Rivera and S. Handschuh. 2013. An ontology-based technique for online profile resolution. Social Informatics, Berlin: Springer. pp. 284-298.

[12] F. Abel, E. Herder, G.J. Hoban, N. Hence and D. Krause. 2013. Cross-system user modelling and personalization on the social web. User Modelling and User-Adapted Interaction. 23: 169-209.

[13] O. De Val, A. Anderson, M. Corny and G. Mohan. 2001. Mining e-mail content for author identification forensics. ACM Sigmoid Record. 30(4): 55-64.

[14] E. Read, R. Cheri and A. Diana. 2010. User profile matching in social networks. Proc. Of the 13th International Conference on Network-Based Information Systems (NBiS'10). pp. 297-304.

[15] J. Vosecky, D. Hong and V.Y. Sheen. 2009. User identification across multiple social networks. Proc. Of the 1st International Confer-once on Networked Digital Technologies. pp. 360-365.

[16] P. Jain, P. Kumara guru and A. Joshi. 2013. @ I seek 'fib. Me': recognise-in users across multiple online social networks. Proc. of the 22nd International Conference on World Wide Web Companion. pp. 1259-1268.

[17] P. Jain and P. Kumara guru. 2012. Finding Memo: searching and re-solving identities of users across online social networks. Arrive preprint arXiv: 1212.6147.

[18] R. Zhen, J. Li, H. Chen and Z. Huang. 2006. A framework for au-thorship identification of online messages: writing-style fee- tares and classification techniques. J. of the American Society for Information Science and Technology. 57(3): 378-393.

[19] M. Alistair and G. Studio. 2012. Exploring likability of user re-views. Computer Security–ESORICS 2012 (ESORICS'12). pp. 307-324.

[20] X. Kong, J. Zhang and P.S. Yu. 2013. Inferring anchor links across multiple heterogeneous social networks. Proc. of the 22nd ACM International Conf. on Information and Knowledge Management (CIKM'13). pp. 179-188.

[21] 2007. Art thou r3579x? Anonym zed social networks, hidden patterns, and structural steganography. Proc. of the 16th international con-ferrous on World Wide Web (WWW'07). pp. 181-190.

[22] B. Zhou and J. Pei. 2008. Preserving privacy in social networks against neighbourhood attacks. Proc. Of the 24th IEEE International Conference on Data Engineering (ICDE'08). pp. 506-515.