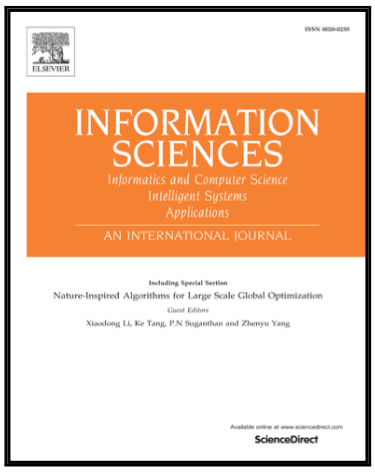


A Deep Dive into User Display Names across Social Networks

Yongjun LI , You Peng , Zhen Zhang , Mingjie Wu , Quanqing Xu , Hongzhi Yin

PII: S0020-0255(18)30146-4
DOI: 10.1016/j.ins.2018.02.072
Reference: INS 13488

To appear in: *Information Sciences*

Received date: 6 May 2017
Revised date: 12 December 2017
Accepted date: 26 February 2018

Please cite this article as: Yongjun LI , You Peng , Zhen Zhang , Mingjie Wu , Quanqing Xu , Hongzhi Yin , A Deep Dive into User Display Names across Social Networks, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.02.072

Highlights

- A display name acquisition framework across social networks is presented.
- The display names of a user contain the abundant information redundancies.
- The information redundancies of display names are time-independent.
- The display names are of great benefit to user identification across social sites.

A Deep Dive into User Display Names across Social Networks

Yongjun LI[*1] You Peng[1] Zhen Zhang[1] Mingjie Wu[1]
Quanqing Xu[2] Hongzhi Yin[3]

[1]*School of Computer, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China*
[2]*Data Storage Institute, A*STAR, Singapore138632, Singapore*
[3]*School of ITEE, The University of Queensland, St Lucia Campus, Brisbane, QLD 4072, Australia*

Abstract:

The display names from an individual across Online Social Networks (OSNs) always contain abundant information redundancies because most users tend to use one main name or similar names across OSNs to make them easier to remember or to build their online reputation. These information redundancies are of great benefit to information fusion across OSNs. In this paper, we aim to measure these information redundancies between different display names of the same individual. Based on the cross-site linking function of Foursquare, we first develop a distributed crawler to extract the display names that individuals used in Facebook, Twitter and Foursquare, respectively. We construct three display name datasets across three OSNs, and measure the information redundancies in three ways: length similarity, character similarity and letter distribution similarity. We also analyze the evolution of redundant information over time. Finally, we apply the measurement results to the user identification across OSNs. We find that 1) more than 45% of users tend to use the same display name across OSNs; 2) the display names of the same individual for different OSNs show high similarity; 3) the information redundancies of display names are time-independent; 4) the AUC values of user identification results only based on display names are more than 0.9 on three datasets.

Keywords: online social network; information redundancies; display name; measurement and analysis

1. Introduction

Nowadays, online social networks (OSN), such as Facebook, Foursquare and Twitter, have been very popular communication tools in our daily life. We almost daily share our ideas, photos, reviews, and get the latest news on these sites. According to the statics report [4], until April 2017, there are 1,968 million active users on Facebook, 319 million active users on Twitter, 600 million active accounts on Instagram. There are also more than 50 million users on Foursquare [2]. However, no one social network is universal. The functionality of different popular social networks varies

* corresponding author, lyj@nwpu.edu.cn (Yongjun LI)

differently, so an individual often joins various social networks for different purposes. Liu et al. [23] found that an individual joined 3.99 social networks on average.

The personal information on a single site is often incomplete. If we integrate these sites, a better profile of a user can be built. The information redundancies are of great benefit to information fusion across OSNs. In this paper, we mainly measure the information redundancies between the display names, which come from different social networks but belong to the same individual. To make our display names easier to remember[34] or to build our online reputation, we often has consistent behavior when selecting our display names on different social networks, which brings redundant information between our different display names, such as our commonly used strings, career or educational experiences etc. For instance, a user, whose display name is 'San Francisco Dad' on Twitter, has display name 'Bay Area Dad' on Foursquare, which both reflect his role in family as well as his location.

The existing works about redundant information mainly focus on username, which is different from display name. Vosecky et al. [32] analyzed two users' similarity based on their profile information, including the similarity of two usernames based on the vector-based name matching algorithm. Perito et al. [30] estimated uniqueness of usernames by the entropy. Iofciu et al. [15] summarized the methods used for comparing two usernames, such as edit distance, Jaccard similarity, and the longest common subsequence. Liu et al. [22] analyzed usernames characteristic including length, special character, numeric character, character input mode, character combine, English character similarity etc. Zafarani et al. [34] proposed presented a MOBIUS method to analyze the usernames that belong to the same individual.

However, the usernames are not always alphanumeric string in social networks, such as Foursquare and QQ[1], the username is a numeric string and assigned by the site. In this situation, it has little information redundancies between the usernames. On the other hand, the user's display name, which is set by the user, is often alphanumeric string and also is obtained easily. The display names an individual selects for different OSN sites often also have redundant information, therefore we focus on the measurement and analysis on the display names across social networks. We make the following four main contributions.

Display Name Acquisition Framework on Cross-OSNs. Based on our previous work in display name [19, 20], we first adopt three real social network datasets for our measurement and analysis. Based on the cross-site linking function of Foursquare, we developed a distributed crawler to extract the display names individuals selected for Facebook, Twitter and Foursquare, respectively. In addition, we sampled a fraction of Foursquare users that are registered at different time instead of at random. It is helpful for evolution analysis. This is the foundation of measurement and analysis on display name.

Display Name Overview on Single OSN. We evaluate the size of dataset we obtained from each social network, and give an overview on three datasets, including

[1] QQ is a very popular instant messenger in China.

ratio of duplicate display names, length distribution, letter distribution, specific character distribution, numeric character distribution and percentage of the same display name, etc. Our observation indicates that 1) the duplication of display name is rare in social network, with the highest probability of 0.0045 that name appearing more than once. 2) The display name on different networks is completely accordant with the naming rules in length, specific and numeric character distribution. 3) The letter distribution of an individual display name is similar with his real name. 4) More than 45% of users tend to use the same display name across different social networks.

Display Name Attribute Analysis. We measure the display names' redundant information in three ways: the length similarity, the character similarity and the letter distribution similarity. We find that 1) there is no obvious difference between the display name lengths of the same individual. 2) The character similarity between a user's display names is very high. For example, for more than 76% of users, excluding those users who select the completely same display name, the length of longest common subsequence between his display names is more than half of the shortest length of his display names. 3) The letter distribution of display names is very similar. It should be mentioned that our measurements only consider the positive instances that two names are different.

Display Name Evolution over Time. We divide our real data into nine datasets based on the chronological order of registration, and demonstrate whether our measured display name attributes are relevant with the user registration time. Except in a period of time when Foursquare changed its privacy policy, the display name attributes are time-independent. These findings provide insights into individual identification across social networks.

The structure of this paper is as follows. In Section 2, we present the related works. We describe the data acquisition process and then give an overview on our obtained datasets in Section 3. We detail the measurement on the display name in Section 4 and analyze data consistency as time evolution in Section 5. The cross-name discovery is presented in Section 6 and we apply the measurement results into user identification in Section 7. Finally we conclude this paper in Section 8.

2. Related works

Over the past few years, researchers have studied many of the properties of various online social networks. Li et al. [21] measured the similarity of User Generated Content across Facebook, Twitter and Foursquare. Motoyama et al. [26] proposed a method for matching individuals based on user's profiles on Facebook and Myspace. Wang et al. [33] analyzed user activities across Facebook, Twitter, and Foursquare. Chen et al. [9] presented a holistic measurement on Foursquare based on its cross-site linking function. Ottoni et al. [29] studied the user behavior on Twitter and Pinterest, and found that the global patterns of use across the two sites differ significantly. These existing works give us a good view on cross-sites analysis.

In this paper, we mainly measure display names across social networks. There are several similarity algorithms related to our works, such as Jaro distance [31][12] [16][10], Jaro-Winkler [24][8][18] [10] and TF-IDF algorithm [24][18][10], which are

always employed to compute the similarity of two usernames. Buccafurri et al. [8] also used Levenshtein, QGrams, Monge-Elkan and Soundex algorithm to compute the similarity of two usernames. Zafarani et al. [34] utilized Longest Common Substring, edit distance, Dynamic Time Warping distance, Jensen-Shannon divergence and *n*-gram algorithm etc. to compare usernames similarity. Liu et al. [22] proposed a similarity algorithm based on the Longest Common Substring. Jain et al. [16] adopted Cosine similarity to measure the similarity of two tweets. Hussain et al. [14] also introduced Cosine similarity into medications to identify and correct the misspelled drugs' names. A display name could be considered as a short string. We make some improvements based on the above basic algorithm for calculating the similarity of two display names.

3. Data Collection and Overview

3.1 Collection method

To obtain the users' display names, we first need to know the social network sites that users have joined. There are several ways to obtain personal information across social networks, such as questionnaire survey, leaked data and web crawler. Liu et al. [23] conducted a survey and asked users to provide their information across social networks. Liu et al. [22] used the information disclosed in 2011. Zafarani et al. [34] collected data by the account URL that users revealed on Google+, blog, forum, etc. However, the data obtained by questionnaire and disclosure is limited and costly. Currently, some social network sites support the cross-site linking function, which allows a user to link his accounts to other social network site, such as Foursquare, Google+, Pinterest. We choose Foursquare to obtain the user information because of its great popularity and unique numerical user ID. This ID is assigned in an ascending order. If we know the ID of a user, we can access his profile page with URL *https://foursquare.com/user/ID*.

Fig.1 shows the public profile pages of two users on Foursquare. We can see their display names and cross-site links. One links his Twitter account, and the other links both Facebook account and Twitter account. These account links are user-authorized and have extremely high reliability. Based on this cross-site linking function, we could obtain an individual's display names on Foursquare, Facebook and Twitter, respectively.

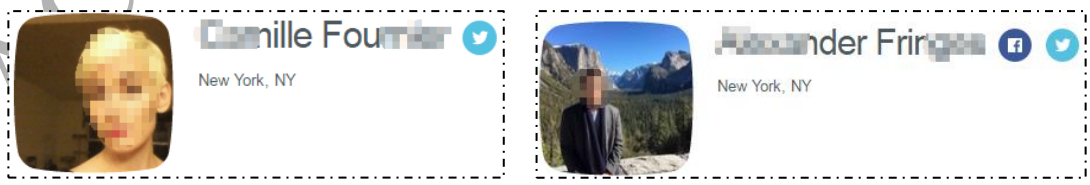


Fig.1 Two Foursquare user's public profile pages

Fig.2 illustrates our basic framework for data collection. We obtain the display names in three steps: 1) access a user's Foursquare profile page with the given ID via *https://Foursquare.com/user/ID*; 2) parse the obtained profile page to get the user's Foursquare display name, as well as Twitter link and Facebook link if this user has

revealed them publicly; 3) extract his corresponding display names on Facebook and Twitter by API, respectively. Finally we get three display names that this user selects for Foursquare, Facebook and Twitter, respectively.

To access the Foursquare users' profile pages, we first need a number of Foursquare user IDs. Unlike Wang et al. [33] sampling a fraction of Foursquare IDs at random, we get a fraction of IDs in segmentation. To measure the display name evolution over time, we obtain the real data in two stages. In the first stage, we access the profile pages corresponding to the first 100,000 IDs. In the second stage, we extend the scale of user's IDs to1.3 million. To solve the limitation of request number from the same IP, we develop a distributed crawler, in which each sub-crawler is responsible for crawling a part of IDs. In total, 1.3 million Foursquare IDs are crawled during April and May in 2016 [19, 20].

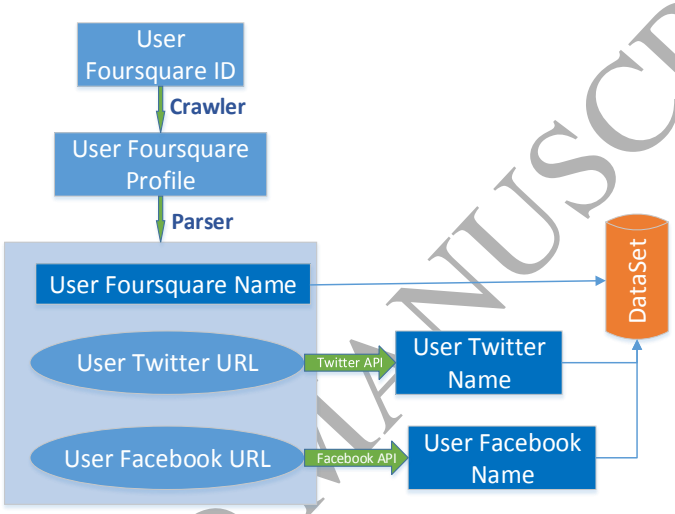


Fig.2 Data collection framework

The sizes of the real datasets we obtained are shown in Table 1. Overall, we successfully obtained 597,822 display names on Foursquare among 1.3 million IDs. The actual obtained ratio is only about 46%. This is mainly due to users' deactivated pages or privacy protection pages.

Table 1. Display name collection statistics

	Planed	Obtained
Foursquare	1,300,000	597,822
Facebook	327,609	288,480
Twitter	113,951	102,315
Facebook-Twitter	-	67,826

As shown in Table 1, we actually obtain 102,315 display names on Twitter and 288,480 display names on Facebook, respectively. The number of users, who have

revealed both Facebook and Twitter URLs, is 67,826. Specifically, we find that 54.80% of users have disclosed their Facebook URLs, and only 19.06% of users exposed their Twitter URLs. The former disclosed ratio is nearly three times of the latter [20]. We take a further analysis and find it is mainly caused by their popularity. The former number of active users is 5.47 times as the latter.

3.2 Data overview

The dataset consisting of the display names obtained from Facebook is denoted by FB. Based on the same method, we could get the dataset TW and dataset FS. We first give an overview of the display names on three datasets.

Duplication of Display Name Unlike the username, the display name is not necessarily unique in social network. It is possible that one display name belongs to multiple different users. We count the appeared times of every display name on three datasets, respectively, and show the CCDF of appeared times in Fig.3.

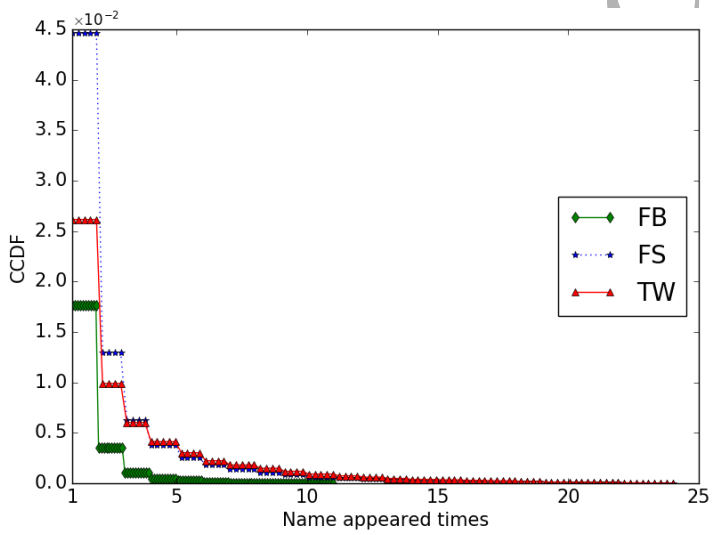


Fig.3. CCDF of display name appeared times

From Fig.3, we can see that all three probabilities of appeared times more than 1 are less than 0.045. In other words, the duplication of name is rare in social network. The probability of appeared times more than 1 in FS dataset reaches the biggest value with 0.045. In TW dataset and FB dataset, the probabilities are about 0.025 and 0.016, respectively. From Table 1, we can see the size of the FS dataset is much larger than the sizes of the other two datasets, so the probability of duplication of name in FS dataset is relatively higher. We also further analyze the users' naming habit, and find the individuals prefer to select their real names or similar names as their display names in Facebook, but they are relatively free to choose a display name for Twitter. Some commonly used display names appear frequently in TW dataset, so the probability of duplication of name in TW dataset is slightly higher than in FB dataset.

Length Distribution We all know that different social networks have different rules on the length of display names. Is there significant difference in the length of the display names? By computing the length of display names on each social network, we show their length distribution in Fig.4.

The length distribution of display name is quite similar. On three social networks, the lengths of display names are concentrated from 10 to 16. On Facebook, Foursquare and Twitter, the percentages of length distribution in this interval are 51.3%, 70.9%, 60.78%, respectively. From Fig.4, we can see 1) the distribution on Twitter is more uniform, its maximum length is just 20. 2) On Facebook, the display names less than 6 in length are rare. Its distribution shape just likes a peak, the left side is steep, and the right is gentle. 3) On Foursquare, the distribution exhibits an obvious peak.

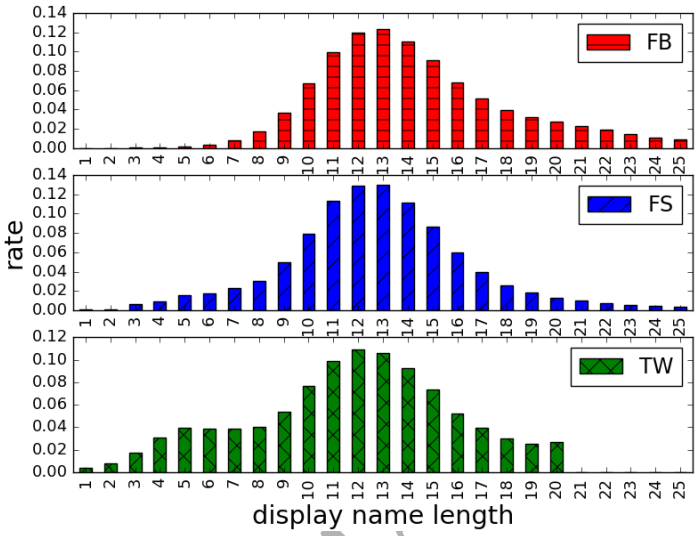


Fig.4. Length distribution of display name on three datasets

We display the detailed length information in Table 2. The average length of display names on Facebook is the largest, and the maximum length of display names on Facebook is much larger than on Twitter. This is just because Twitter has limitation on the maximum length (20) of display name, while Facebook and Foursquare do not have.

Table 2. Display Name Length Statistics

	Avg. Length	Min. Length	Max. Length
FB	14.54	2	70
FS	12.93	1	111
TW	11.68	1	20

Letter Distribution To compare the letter distribution of display names on different sites, we calculate the frequency of each letter in display names, and compare the obtained display names with the commonly used names in life. These real names are collected from the data hall [1] and named as common dataset.

Fig.5 presents the percentages of 26 letters on FB, FS, TW, and common datasets, respectively. Letters ‘e’ and ‘l’ appear more frequently in common dataset, and the

percentages of other letters on four datasets are very similar. The higher percentages are followed by ‘a’,’e’,’n’,’i’,’r’,’o’,’l’,’s’,’t’, and ’m’, accounting for about 70.4%, 71.21%, 65.63%, 75.67% of all characters on FB, FS, TW, common datasets respectively. We observe that the letter distribution of display names in three social networks is similar to in real life, and the letter distribution on FB, FS, and TW are almost completely consistent [20].

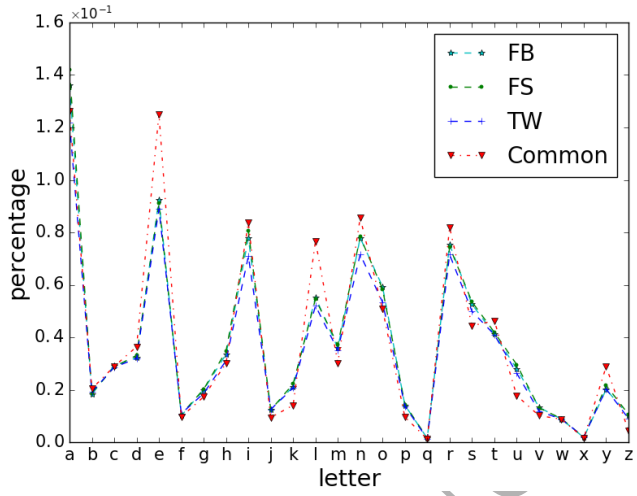


Fig.5. Letter Distribution Comparison with Real Names

Special Character Distribution Different from names in real life, the display names on social networks often contain special characters. We compute the percentage of each special character. As shown in Fig.6, we find that the special characters used in Twitter display names are more massive and diverse than in other two networks. In the Twitter, some special characters, including ‘.’, ‘-‘, ‘_’, ‘’’, ‘!’, ‘(‘, ‘)’, ‘#’, ‘*’, ‘,’, ’:’ and ‘@’ etc., appear frequently. Their frequency accounts for 0.94% of all characters. In the Foursquare, the proportion of special characters only accounts for 0.41%. These characters mainly include ’.’, ‘-‘, ‘?’, ‘’’, ‘_’, ‘&’, ‘@’, ‘)’ and ‘,’. In the Facebook, there are only three characters emerged, ’-‘, ‘.’, and ’”, accounting for 0.29%. This is consistent with the naming rule on the corresponding site. Facebook only allows users to use the three special characters '-', '.', and "' in their display names, while Twitter and Foursquare have no restrictions on the use of special characters.

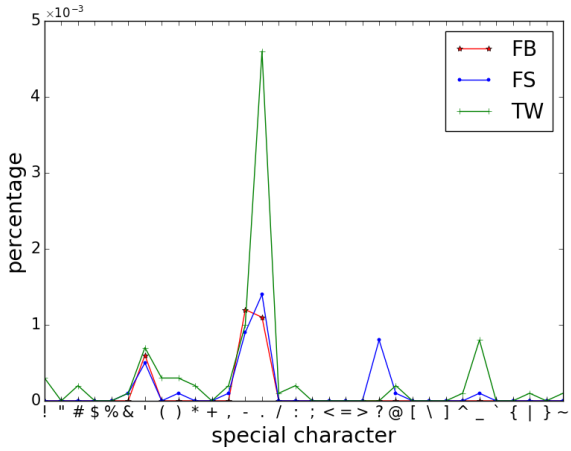


Fig.6. Special Character Distribution Comparison among Different Datasets

The Twitter and the Foursquare allow users to use any character. Therefore, except the character mentioned above, the display name also contain many unreadable characters. For example the display names “Donny (σ`▽´)-σ” and “☆☆ Dav Yaginuma ☆☆” are legal in the Twitter and Foursquare, but illegal in the Facebook. Therefore, it would be better to neglect these unreadable characters when comparing display names on Twitter and Facebook, or on Foursquare and Facebook.

Numeric Character Distribution In addition, we also count the numeric character distribution in all display names, the results are shown in Fig.7. We find that 1) there are no numeric character in Facebook display names, because the site prohibits the numbers from appearing in display name. 2) The Twitter display names have higher rate than the Foursquare display names. 3) The most frequency numeric is “0”, “1” and “2” both in the Twitter and Foursquare display names.

Although some online social networks allow users to use numbers and special characters in display name, the users still rarely use them in their display names with the using rate less than 1%.

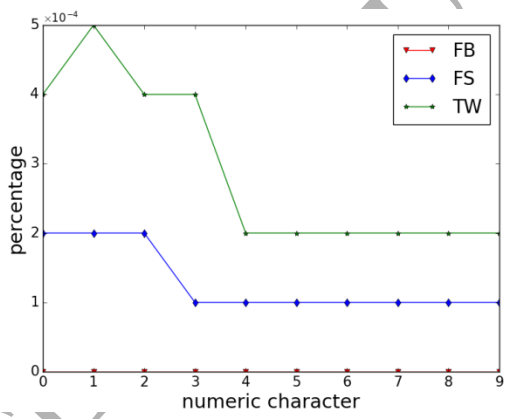


Fig.7. Numeric Character Distribution Comparison among different datasets

Ratio of Same Display Name We combine two display names that the same individual uses in two different sites as a pair and construct three datasets. These datasets are denoted by FB-TW, FS-TW, and FB-FS, respectively.

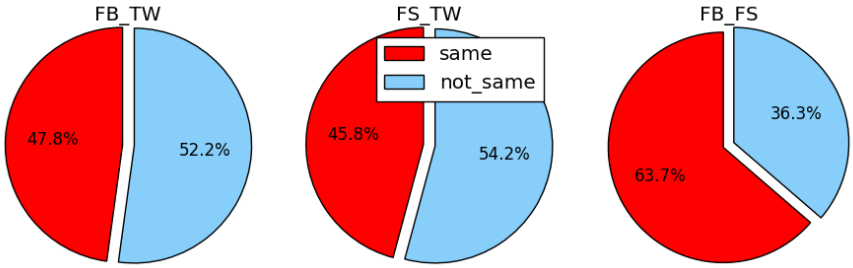


Fig.8. Ratio of Same Name on Different Datasets

Some individuals usually use the same name in multiple social networks to avoid memory trouble. It is also a good way for us to maintain our personal image on the

Web. We calculate the percentages of the same display names in three datasets, respectively. We ignore the letter case when we count the same display names. The results are illustrated in Fig.8. The percentages on FB-TW, FS-TW, and FB-FS are 47.84%, 45.84%, 63.68%, respectively. Liu et al [23] have found that 59% of individuals prefer to use the same username. For display name, we reach the similar conclusion that more than 45% individuals prefer to use the same display name across the social networks [20].

From the overview of the obtained data, we can conclude that the display names are closely related to the social network's naming rule. The character distribution of the display names is similar with the real names and more than 45% of users usually use the same display name in several social network sites [20].

4. Data analysis

The display name is a literal sign or mark which is used for identifying an individual on an OSN site and is not necessarily unique. In order to make their own display names easy to be identified, most users choose some unique letters as part of their display names, such as their nicknames, hobbies, favorite words or numbers etc. An application [6] that aims to help a user generate his Twitter display name, just imitates his psychological characteristics when he is selecting his display name. Thus, the display names an individual selects for different social network sites might contain some redundant information. In this section, we further measure and analyze the redundant information in three ways, including length similarity, character similarity and letter distribution similarity.

We conduct the display name analysis on datasets FB-TW, FB-FS and FS-TW, respectively. To make our analysis more reliable and convincing, we construct three negative datasets named negFB-TW, negFB-FS and negFS-TW. For negFB-TW, we take display names of one FB and one TW account of different users to build 80% negative instances. For the rest 20%, we take display name pairs which share either surname or given name. We employ the similar method to generate the negative instances in FB-FS and FS-TW.

4.1 Length Similarity

Based on our previous works [19, 20], we conduct a detailed measurement and analysis on the length difference and length ratio of the display names.

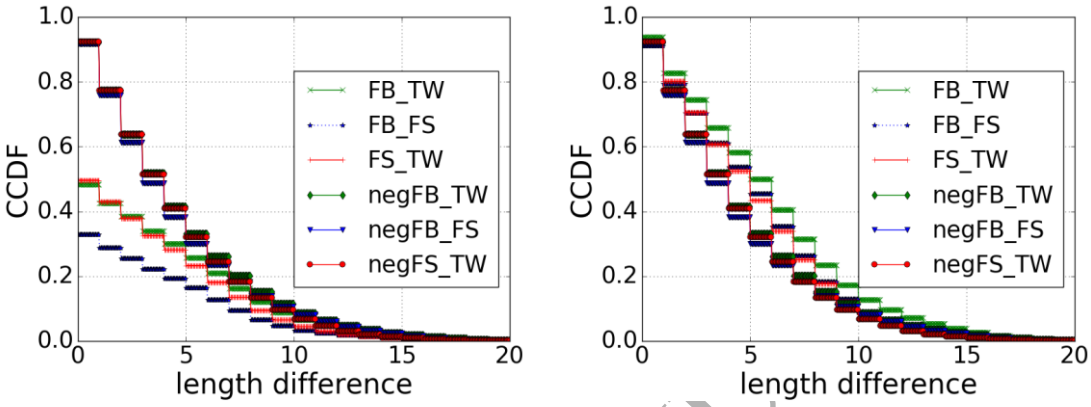
Length difference We assume that $name_1$ and $name_2$ are two display names of an individual. The length difference of $name_1$ and $name_2$ is expressed by Eq. (1). The results are shown in Fig.9 (a).

$$\Delta Len_{name} = abs(len(name_1) - len(name_2)) \quad (1)$$

From Fig.9(a), we can see that most display name pairs has length difference less than 20. More than 90% of the negative instances have length difference larger than 0, while less than 50% of the positive instance with value larger than 0. This is mainly due to the fact that more than 45% individuals use the same display names on different social networks [20]. For further observation, we remove these positive instances which two display names are completely same, and repeat the above

measurement. The results are shown in Fig.9(b).

Fig.9(b) show that the CCDF curves based on positive and negative datasets are very close, and the curves of positive datasets are just slightly higher than negative datasets. That is to say, regardless of the same display name pairs, the same user's display name length difference is slightly larger than two random users' in general. Besides, the curves on FB-TW, FB-FS, FS-TW are almost completely coincide. This means the length difference have no significant correlation with the social networks.



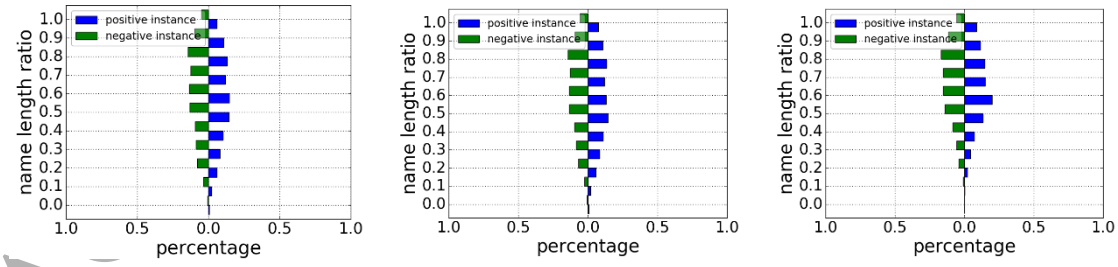
(a) with same name-pairs (b) without same name-pairs

Fig.9. The length difference distribution of the display names

Length ratio Length ratio is the ratio of the length of short display name to the length of long display name and is expressed by Eq.(2). The length ratio ranges from 0 to 1.

$$Ratio_{len} = \frac{\min(len(name_1), len(name_2))}{\max(len(name_1), len(name_2))} \quad (2)$$

The smaller is the length ratio, the larger is the length difference. The value of 1 indicates these two display names have the same length.



(a) FB-TW (b) FS-TW (c) FB-FS

Fig.10. Distribution of Name Length Ratio on three Datasets

We remove the same display name pairs from our datasets, and calculate the length ratio on FB-TW, FB-FS, FS-TW and the corresponding negative datasets. We divide the value space of length ratio into 11 slots, [0-0.1), [0.1-0.2), ... , [0.9-1.0),1.0. The results of percentage on each slot are illustrated in Fig.10. We can see the length ratio distribution is similar between the positive instances and the negative instances, centering from 0.5 to 0.8. On FB-FS, 83.65% of the instances' length ratio is over 0.5.

On negFB-FS, there also have 79.09% cases which length ratios are over 0.5. The similar patterns can be found on the other four datasets. We can easily reach the conclusion that length ratio of the positive instance has no significant difference with the length ratio of the negative instance.

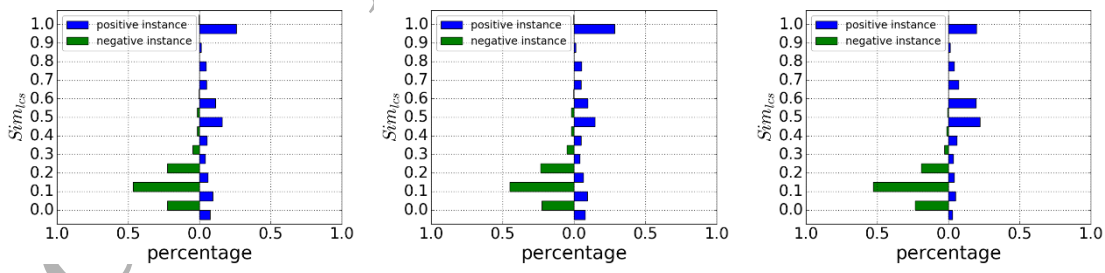
4.2 Character Similarity

Two display names from different social networks are two special strings. Each string is composed of 1-3 words. Thus, we can combine the characteristics of string and name to measure the character similarity. In this subsection, we present five attributes based on the longest common substrings, the longest common subsequences, and edit distance.

Length of LCS/Short Length The longest common substring problem [25] is to find the longest string that is a substring of two strings. It is a good metric to measure the similarity of two different strings. We define this metric as the ratio of the length of the longest common string to the minimum length between two strings. Its value ranges from 0 to 1. The greater the value is, the more similar two display names are. Assume two display names are $name_1$ and $name_2$, respectively. This metric is expressed by Eq. (3).

$$Sim_{lcs} = \frac{len\big(lcs(name_1, name_2)\big)}{\min(len(name_1), len(name_2))} \tag{3}$$

For example, $name_1$ is "JingLee", $name_2$ is "j1nglee". We first convert two names to lower case, respectively. The longest common substring of two names is "nglee", and its length is 5. The minimum length of two names is 7. The metric Sim_{lcs} is 0.7143 (= 5/7). Zafarani [34] and Iofciu [15] also use the longest common substring when calculating the similarity of usernames. However they used the average length of usernames to standardize the LCS.



(a) FB-TW (b) FS-TW (c) FB-FS

Fig.11. Distribution of Sim_{lcs} on three Datasets

Similar to the length ratio, we divide the value space of Sim_{lcs} into 11 slots. Based on Eq.(3), we calculate Sim_{lcs} of every pair of display names on three positive datasets and the corresponding negative datasets, respectively. The distributions of Sim_{lcs} are illustrated in Fig.11. The left side is the negative instance and right side is the positive instance. Generally, the Sim_{lcs} values of the negative instances are concentrated at range [0, 0.2], and its proportion is larger than 91%. However, the values of the positive instances are distributed in each slot, and most positive instances are located

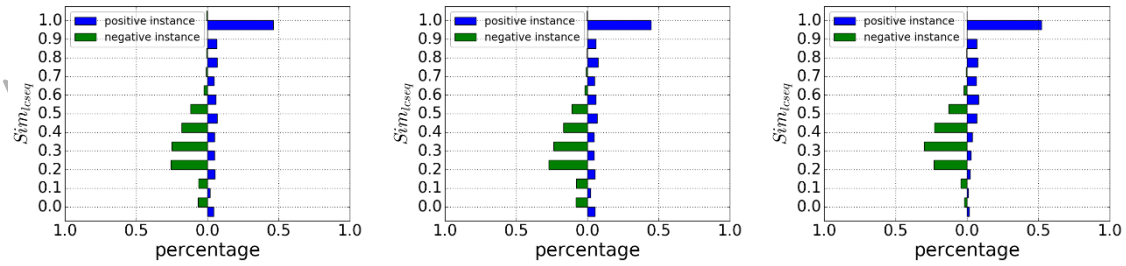
in [0.5, 1], with the proportions more than 64%, 64%, 74% on three datasets, respectively. By contrast, there are only less than 2% negative instances whose Sim_{lcs} values are larger than 0.5. In other words, the Sim_{lcs} value of the positive instances in most cases is bigger than the value of the negative instance. In FB-FS and FS-TW, there are over 26% instances whose Sim_{lcs} values are 1.0 and over 20% instances on the FB-FS. However, on the negative datasets, there is no instance whose Sim_{lcs} value is equal to 1.0. Thus, it is clear that the users have their own fixed naming habit, rather than completely random selecting.

Length of LCSequence/Short Length The longest common subsequence problem [13] is to find the longest subsequence common to two sequences, also can be used for measuring the similarity between two strings. Unlike the longest common substring, the longest common subsequence is not required to occupy consecutive positions within the original sequences. Take <"Jeffrey Donenfeld", "Jeffzilla Don"> for instance. Its longest common subsequence is 'Jeff Don', while its longest common substring is 'Jeff'. Similar to the metric Sim_{lcs}, we measure and analyze the ratio of the longest common subsequence length to the minimum name length, and this metric is expressed by Eq.(4).

$$Sim_{lcseq} = \frac{len\big(lcseq(name_1, name_2)\big)}{\min(len(name_1), len(name_2))} \tag{4}$$

The measurement results are shown in Fig.12. We easily see that the values of 97% of negative instances are under 0.5. However, it is a very common case that the value of positive instance is greater than 0.5 on three datasets, accounting for more than 77% on FB-TW, 76% on FS-TW and 87% on FB-FS, respectively.

We make a further analysis on the positive instances. The percentages of positive instances with Sim_{lcseq} value 1.0 are 40% on FB-TW, 45% on FS-TW, 52% on FB-FS, respectively, but the corresponding percentages on metric Sim_{lcs} are 26%, 28% and 20%, respectively. The gaps of two metrics are all greater than 14% on three datasets. This is mainly because many users would like to form a new display name by abbreviating their display names currently used. From the above analysis, we easily reach a conclusion that the metric Sim_{lcseq} is very helpful to determine whether two display names belong to the same individual or not.



(a) FB-TW (b) FS-TW (c) FB-FS

Fig.12. Distribution of Sim_{lcseq} on three Datasets

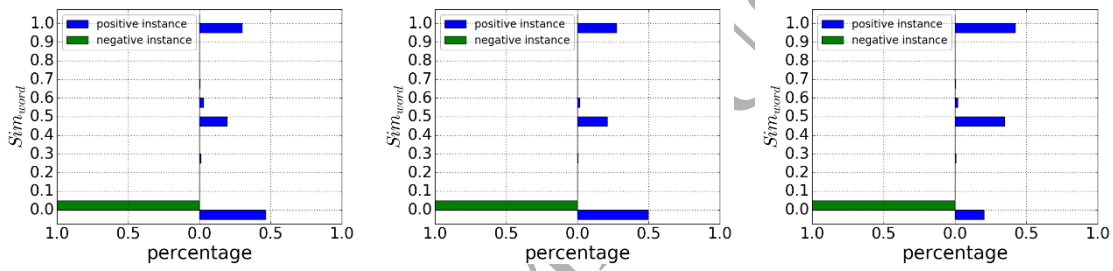
No. of common words /No. of short name words The difference between name and ordinary string is that name can be divided into first name, last name or even

middle name. Assume two display names are $name_1$ and $name_2$, respectively. Each name contains several words. We consider the number of the common words between two names, and is expressed by Eq.(5).

$$Sim_{word} = \frac{commonword(name_1, name_2)}{\min(word(name_1), word(name_2))} \quad (5)$$

where *commonword*($name_1$, $name_2$) counts the number of the common words between $name_1$and $name_2$; *word*(*name*) count the number of words contained in *name*.

We illustrate the results in Fig.13. The values of all negative instances are 0, that is, there is no common word in the negative instances. However, nearly 50% of positive instances also have no common word between two display names, but it still has 30%, 27%, and 42% of positive instances with value 1.0 on three datasets, respectively. It should be mentioned that we remove these positive instances with two same names from three datasets. The Sim_{word} values of all these instances are 1.0. Besides, there are more than 20% of positive instances with value 0.5. The cases arise mainly because individual omitted first name or last name.



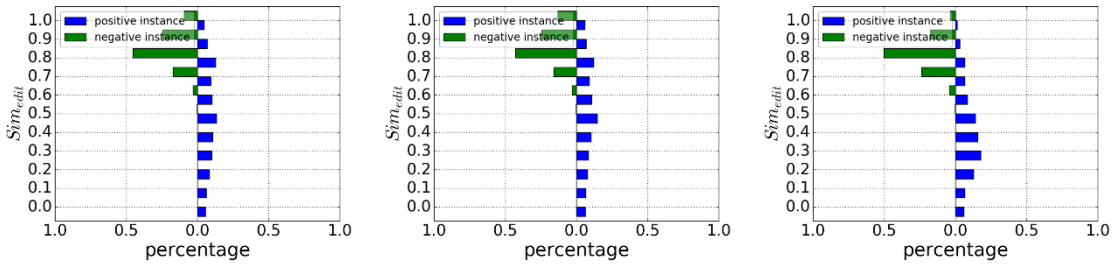
(a) FB-TW (b) FS-TW (c) FB-FS

Fig.13. Distribution of Sim_{word} on three Datasets

Edit Distance/Longest Length: The edit distance [27] reflects the difference between two strings by counting the minimum number of operations required to transform one string to the other. It is a commonly used metric to evaluate the difference of two strings.

The edit distance of two names relates to the name length. In our previous works [19, 20], we introduce the name length to this metric and express it by Eq.(6). The smaller the value is, the larger the similarity is.

$$Sim_{edit} = \frac{edit(name_1, name_2)}{\max(len(name_1), len(name_2))} \quad (6)$$



(a) FB-TW (b) FS-TW (c) FB-FS

Fig.14. Distribution of Sim_{edit} on three Datasets

We also show its distribution in Fig.14. The values of all negative instances are larger than 0.5. Conversely, the values of most positive instances is smaller than 0.5 with percentages 54.36%, 53.90%, and 72.68% on three datasets, respectively. That is, if the edit distance of two display names is less than half of the longest name length, these two display names belong to the same individual with high probability.

Max of Best Match: Normally, a user's display name always consists of several parts, such as first name, middle name, last name, or other title. However, not everyone writes all parts, some omit the middle name, some omit last name, or even reverse the first name and last name. In this situation, if we just compare the name as a whole, it will neglect the name's identical part. Therefore, we consider the max of best part match based on the longest common substring.

Suppose s_1 and s_2 are two strings. The similarity of s_1 and s_2 is expressed by Eq.(7).

$$Sim_{str} = \frac{len(lcs(s_1, s_2))}{(len(s_1) + len(s_2))/2} \tag{7}$$

Suppose $name_1$ and $name_2$ are two display names. The detailed implementation steps of Max of Best Match of $name_1$ and $name_2$ are shown as follows.

Step 1: Segment the two names into words, respectively, and get two name arrays $\boldsymbol{Arr}_1$ and $\boldsymbol{Arr}_2$;

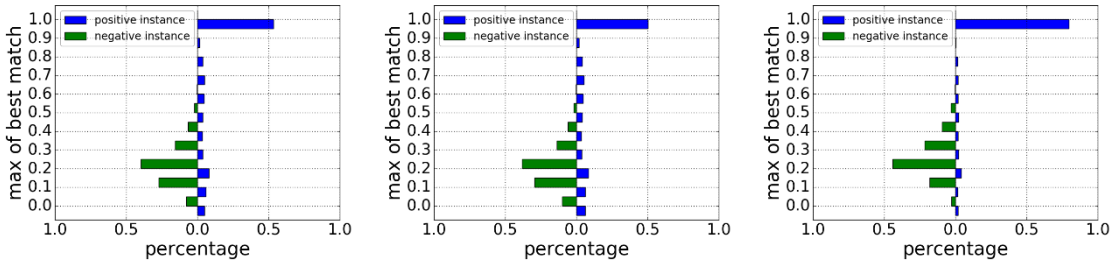
Step 2: Calculate similarity of each word in $\boldsymbol{Arr}_1$ with word in $\boldsymbol{Arr}_2$ based on Eq.(7), and get a similarity matrix $\boldsymbol{A}$;

Step 3: Find the largest value in matrix $\boldsymbol{A}$, and this value is the max of best match.

For example, if an individual's name in Facebook is 'David J. Whelan', and in Twitter called 'Dave Whelan', we first segment them and get two arrays, ['David', 'J.', 'Whelan'] and ['Dave', 'Whelan']. Then we calculate the similarity by Eq.(7), and get the similarity matrix as shown in Table3.

Table 3. Overview of similarity matrix

	David	J.	Whelan
Dave	0.667	0.0	0.2
Whelan	0.182	0.0	1.0



(a) FB-TW (b) FS-TW (c) FB-FS

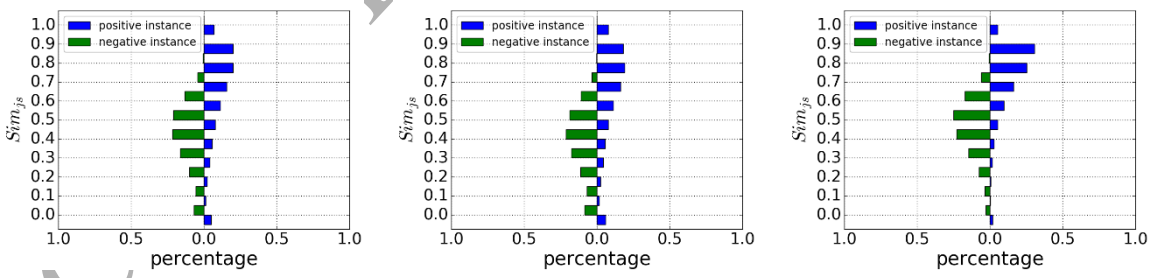
Fig.15 Distribution of max of best match on three Datasets

Find the largest value 1.0, and the maximum value 1.0 is the metric we want, max of best match. Fig.15 shows our measurement results on max of best match.

From Fig.15, we can see that most of the metric values on the positive instances are 1.0, with percentage of 53.43%, 50.37%, 80% on FB-TW, FS-TW, FB-FS respectively. That is, more than half of the users always use the same name part across social network sites. Besides, the percentage of FB-FS datasets with metric value 1.0 is higher than other two datasets. This mainly because most users always select the display names similar to their real names on Facebook and Foursquare. There are also about 20% of positive instances whose values are in [0.5, 0.9] on FB-TW and FS-TW. These users do not use the completely same name part but make some changes on their first names, or last names, or middle names, when they select the display names for the different social networks. While on the negative instance, most metric values are below 0.5. There is a great difference between the distribution of positive instances and negative instances, which is helpful to improve user identifiability on social network.

4.3 Letter Distribution Similarity

The letter distribution presents the occurrence probability of each letter in a display name. Two identical display names have the same distribution of letters. However, for two similar display names, their letter distribution is also similar. For example, name “gate man” and name “man gate” have same letter distribution. The quantity of letters is so large that we cannot consider all language letters. For simplicity, we only consider twenty-six English letters. We measure these letter distributions based on Jensen-Shannon distance, Cosine similarity and Jaccard similarity, respectively.



(a) FB-TW (b) FS-TW (c) FB-FS

Fig.16 Distribution of Sim_{jc} on three Datasets

Jensen-Shannon Similarity Jensen-Shannon distance [11], which is an improvement on Kullback–Leibler distance [17], is used to calculate the difference between two probability distributions. Its value ranges from 0 to 1. We use JS distance to measure letter distribution difference between two display names. The smaller the distance is, the greater the similarity between two distributions is. Assume P and Q are the letter distributions of display name $name_1$ and $name_2$, respectively. The JS similarity of name $name_1$ and $name_2$ is expressed by Eq.(8).

$$Sim_{JS} = 1 - \frac{1}{2}(KL(P||M) + KL(Q||M))$$
$$\mathrm{M} = \frac{1}{2}(\mathrm{P}+\mathrm{Q}), \quad \mathrm{KL}(\mathrm{P}||\mathrm{Q}) = \sum_{i=1}^{|P|} P_i \cdot \log\frac{P_i}{Q_i} \tag{8}$$

where p_i is the occurrence probability of the i^{th} character.

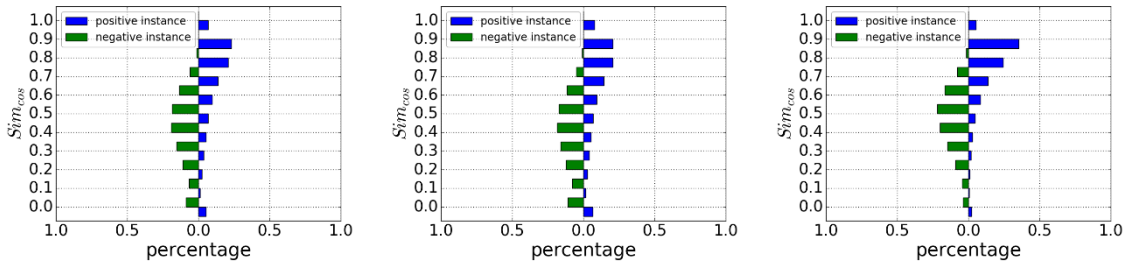
To avoid the situation that the logarithm does not make sense, we use a very low value e to smooth for the letters whose probability is zero. In this paper, we set the e to 2.2204460492503131e^{-16}. The measurement results are illustrated in Fig.16.

For the negative instances, the Sim_{JS} values are concentrated on the lower interval, and show a trend of rising first and then decreasing in [0.0,0.7], while for the positive instances, their Sim_{JS} values focus on the higher interval, and show an increasing trend in [0.1,0.8]. On FB-TW and FS-TW's positive instances, there are 5%, 6% of the instances with values less than 0.1. Sim_{JS} values less than 0.1, indicate that the letters' distributions of two display names are almost completely different. This is mainly due to the fact that the display names a user selected for different social networks are in different languages. For example, a user whose display name is 'デーブミナナスス一一' on the Twitter, has a display name 'Dave Mianowski' on the Facebook. Although these two display names have same meaning, the letters in the two display names are completely different. In data preprocessing, we first translate names into the same language using the machine translation software. From Fig.16, we also easily find that the percentage of the positive instances with values larger than 0.8 is about 45%, while the percentage of the negative instances is less than 2%. Obviously, two display names with Sim_{JS} value larger than 0.8 belong to the same user with high probability.

Cosine Similarity The cosine distance is mainly used to measure the similarity between two vectors. After calculating the frequency of each letter in the display name, we get two vectors and then we express their similarity by Eq.(9). The larger the value is, the more similar two vectors are.

$$Sim_{cos} = 1 - \cos(\mathrm{P}||Q)$$
$$\cos(\mathrm{P}||Q) = \frac{\sum_{i=1}^{n}(P_i \times Q_i)}{\sqrt{\sum_{i=1}^{n}(P_i)^2} \times \sqrt{\sum_{i=1}^{n}(Q_i)^2}} \tag{9}$$

where P is the frequency vector of letters in $name_1$, and Q is the frequency vector of letters in $name_2$. P_i, Q_i is the frequency of i^{th} letter in $name_1$ and $name_2$ respectively. For clearly explaining the frequency vector of letters in a display name, we take the name 'mangate' for instance. its frequency vector is [2,0,0,0,1,0,1,0,0, 0,0,0,1,1,0,0 ,0,0,0,1,0,0,0,0,0,0].



(a) FB-TW (b) FS-TW (c) FB-FS

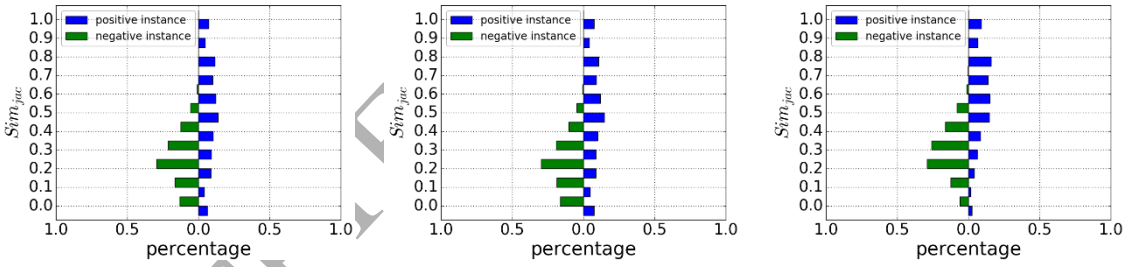
Fig.17 Distribution of Sim_{cos} on three Datasets

Fig.17 shows the measurement results. From Fig.17, we find that the cosine similarity distribution is very similar with the Jensen-Shannon similarity distribution. The percentages of the positive instances with values larger than 0.8 are 51% on FB-TW, 50% on FS-TW and 65% on FB-FS, respectively, while the percentage of negative datasets is less than 2%.

Jaccard similarity Jaccard similarity [28] is used to compare the similarity of two sets. It is the ratio of the size of the intersection to the size of union of two sets. We consider the letters in a display name as a set and calculate the Jaccard similarity by Eq.(10). Fig.18 illustrates the results of Jaccard Similarity.

$$Sim_{jac} = \frac{len(set(name1) \cap set(name2))}{len(set(name1) \cup set(name2))} \tag{10}$$

where *set*(*name*) is the set of letters in the name.



(a) FB-TW (b) FS-TW (c) FB-FS

Fig.18 Distribution of Sim_{jac} on three Datasets

The Jaccard similarity values of negative instances mainly concentrate on the lower interval and are less than 0.5. For the positive datasets, the value distributions are more uniform. It should be noticed that we remove the positive instances with two same display names, but the values of the positive instances are still much larger than the values of the negative instances on average.

5. Evolution analysis

In the above analysis, we only consider a single snapshot of the social network, neglecting an important aspect of these social networks, viz: their evolution over time. The social network continuously evolves in response to many factors, such as the underlying social dynamics. Will these factors affect his behavior when a user selects

the display names for different social networks? The redundant information between two display names is consistent over time? In this subsection, we focus on the evolution analysis on the above attributions.

In Foursquare, the user ID is assigned in an ascending order. That is, the larger the user ID is, the later the registration time of this user account is. To obtain multiple snapshot of Foursquare, we divide the total ID into nine chunks, and crawl a part of IDs on each chunk. The Foursquare ID ranges of nine chunks are [0-100,000], [10,000,000-10,150,000], [20,000,000-20,150,000], [30,000,000-30,150,000], [40,000,000-40,150,000], [50,000,000-50,150,000], [60,000,000-60,150,000], [70,000,000-70,150,000], [80,000,000-80,150,000], respectively. After repeat the data collection described in section 2.1, we obtain 3 datasets based on each chunk, and totally 27 datasets on all chunks. For the sake of convenience, the nine datasets across Facebook and Twitter are denoted by $FB\text{-}TW_i$ (i=0,1,…,8). Similarly, we have datasets $FB\text{-}FS_i$ (i=0,1,…,8) and $FS\text{-}TW_i$ (i=0,1,…,8).

We first calculate the percentage of the positive instances with the duplicate display names on 27 datasets, respectively. The results are shown in Table 5.

From table 5, we can find, the size of the datasets we obtained for each chunk is continually decreasing. As the time goes on, fewer and fewer Foursquare users simultaneously reveal their Facebook and Twitter accounts. The percentages of the duplicate display names also decrease gradually. With the development of Internet, increasing people pay attention to privacy protection and make their profile only open to their friends, not everyone on the Internet.

Table 5. the ratio of the duplicate display names

	FB-TW		FS-TW		FB-FS	
	size	%	size	%	size	%
0	26062	63.86%	37470	60.81%	36358	75.59%
1	10426	39.64%	14711	41.17%	52277	61.01%
2	6555	37.80%	10196	39.71%	31598	62.23%
3	6733	38.13%	9650	33.10%	35532	49.32%
4	4158	38.91%	7694	37.21%	21040	63.72%
5	4516	38.22%	7056	38.00%	29757	64.90%
6	4375	37.26%	7305	36.56%	33096	65.24%
7	3092	38.62%	5069	36.93%	27732	66.94%
8	1909	38.66%	3164	36.09%	21090	67.87%

5.1 Evolution analysis on name length

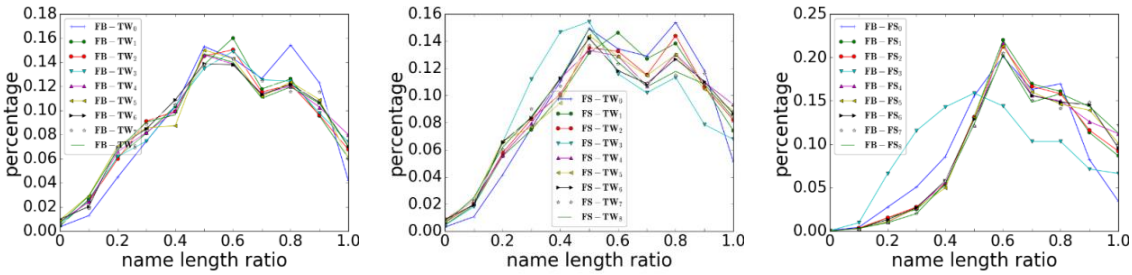


Fig.19. evolution analysis on length similarity

Table 6. the percentage of user with only first name (%)

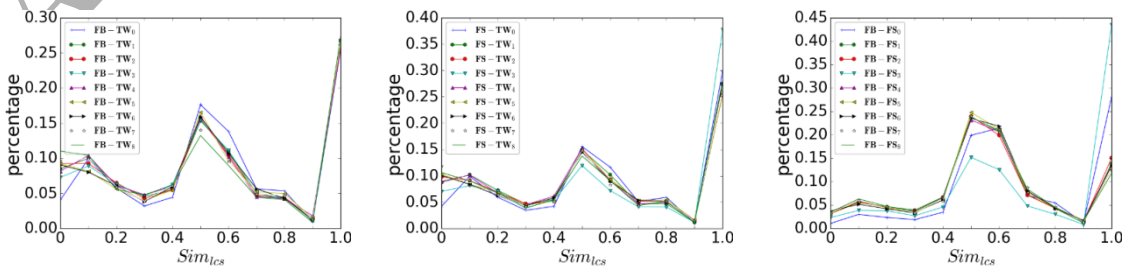
Dataset		0	1	2	3	4	5	6	7	8
FB-FS	FS	14.75	4.31	5.11	**39.24**	4.32	3.62	5.08	3.17	3.38
	FB	0.10	1.29	0.99	0.52	1.39	1.18	1.07	1.30	1.39
FS-TW	FS	8.73	5.18	5.65	**30.11**	5.24	5.35	6.73	4.94	5.74
	TW	49.14	44.98	44.62	36.74	45.33	45.78	48.32	49.42	51.14
FB-TW	FB	0.03	0.64	0.56	0.34	0.75	0.79	0.66	0.37	1.11
	TW	51.27	43.33	42.56	41.00	43.46	44.70	46.56	45.94	49.02

The evolution analysis results are illustrated in Fig.19. The curves on different datasets are very close except on FS-TW_3 and FB-FS_3. We make a further analysis on FS-TW_3 and FB-FS_3. It is mainly due to the Foursquare changes its privacy policy on Jan. 28th, 2013 [3], when the number of its registered users reaches 30 million. The main change is that the users would see the complete first and last names on the profile page. Before that time, the Foursquare sometimes shows the user's full name and sometimes shows his first name and the initial of last name ("John Smith" vs. "John S.").

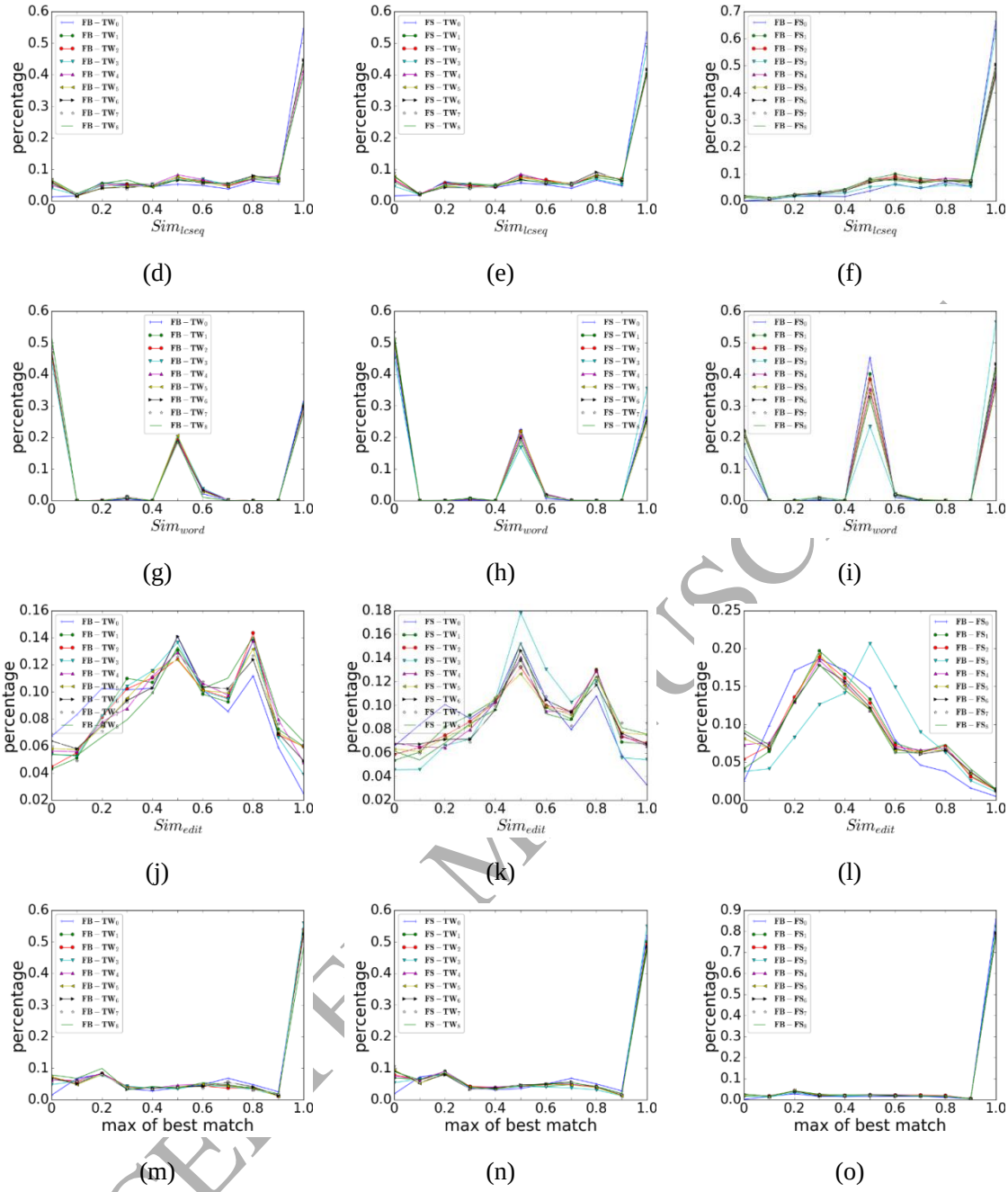
This change has a great impact on the user's display name. In order to avoid long full names displayed, the user only presents his first name. Table 6 shows the percentage of only presenting his first name in display name on FB-FS, FS-TW and FB-TW. We find that the users corresponding to chunk3 have higher percentage of user with only first name on Foursquare.

5.2 Evolution Analysis on Character Similarity and Letter Distribution

We also make the evolution analysis on the character similarity and letter distributions. These attributes are consistent on most datasets except on the datasets containing the Foursquare users corresponding to chunk3, which caused by the rules change on Foursquare. That is, these attributes remain unchanged over time.



(a) (b) (c)



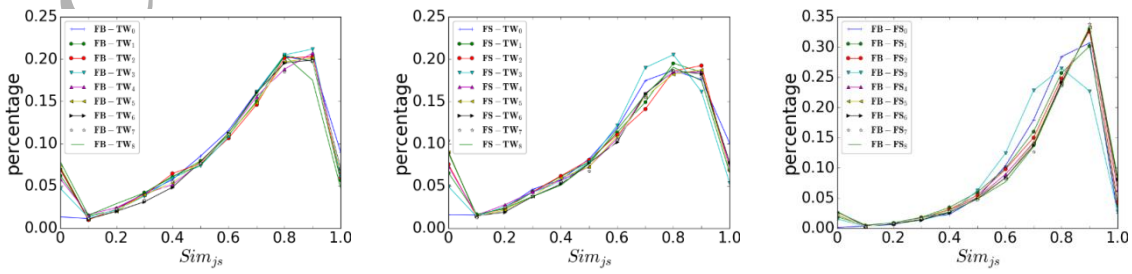
(d) (e) (f)

(g) (h) (i)

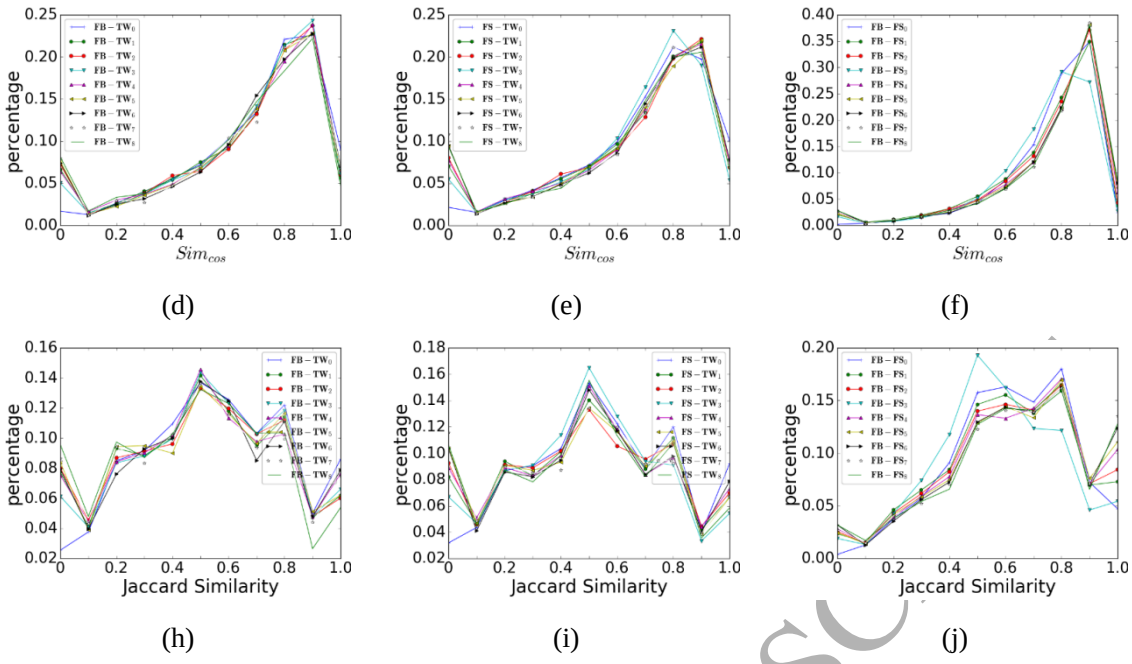
(j) (k) (l)

(m) (n) (o)

Fig.20. evolution analysis on character similarity. (a-c): Sim_{lcs} similarity, (d-f): Sim_{lcseq} similarity, (g-i): Sim_{word} similarity, (j-l): Sim_{edit} similarity, (m-o):max of best match



(a) (b) (c)



(d) (e) (f)

(h) (i) (j)

Fig.21. evolution analysis on character distribution. (a-c): Sim_{js} similarity, (d-f): Sim_{cos} similarity, (g-i): Sim_{jac} similarity

6. Discovery

Through above analysis, we conclude:

(1) More than 45% of users tend to use the same display name in different OSNs. This is mainly because the users have limited memory and the needs of maintaining their personal image and reputation on different social network sites[20]. The information contained in the display name can make people associate with this person or related product. It is particularly important for stars or people in the marketing.

(2) The display names of an individual selected for the different social networks have no fixed lengths. Except the duplicate names, which has same length, the length of the display names always vary greatly. For two display names, even if generated by two individuals, the length difference may be the same as an individual generated.

(3) For the positive instances, the character similarity is striking, although two names are not exactly same. Specifically as follows:

- For more than 64% of the positive instances, the length of the longest common substring is more than half of the shorter name length. Moreover, there are 20% of the users whose one name is fully contained in the other name;
- there are about 76% of users whose length of longest common subsequence is more than half of his shorter name length;
- 27% or more of users have the same surname or last name;
- There are 53% of positive instances whose edit distance is less than half of his longer name length;
- As for the best match of the positive instances, the values of more than 50% of users are 1.0.

A user usually selects different names in different OSNs for the purpose of

privacy. However, if the display names of a user are completely irrelevant, he can hardly remember them clearly. Actually, most individuals just change part of their real names, and retain some of the basic information. This information tends to make the display names of a user having high character similarity.

(4) The letter distributions of the positive instance are very similar.

- The Jensen-Shannon similarity of more than 45% of positive instances are larger than 0.8;
- The positive instances whose Cosine distance are over 0.8 also account more than 50%;
- The Jaccard similarities of more than 47% of positive instances are over 0.5. On the contrary, the corresponding percentage of the negative instances is only 2%.

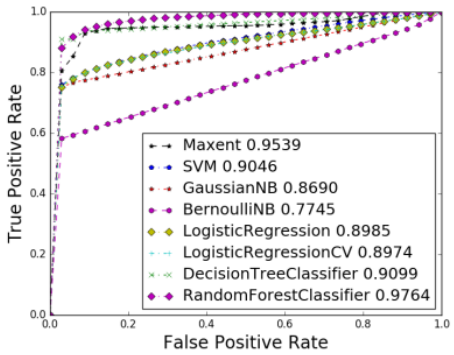
The alphabet distribution reflects the user preference for specific letter. Some letter can also reflect a user's country or region to a certain extent. For example, Zafarani et al. [34] mentioned that the frequency of using letter "*x*" in Chinese name is higher than other countries or regions. Therefore, the closer the letter distribution of two display names is, the more likely the two names belong to the same user.

(5) The evolutionary analysis results show that the above attributes remain unchanged over time.

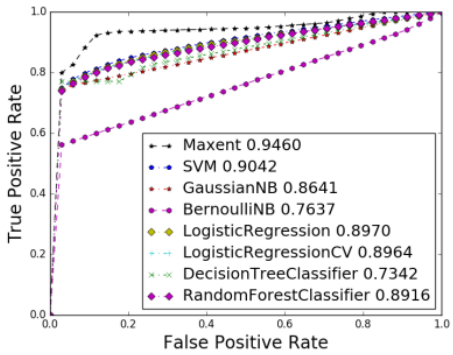
(6) The similarity of two display names from Facebook and Foursquare is generally more striking. This is mainly due to the user tend to choose his display name closer to his real name on these two social networks.

7. Application

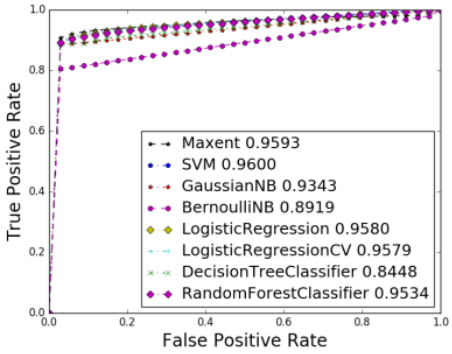
We apply the measurement results to the user identification across social networks. The user identification is a fundamental problem of information fusion. Assume two given display names from two different social networks, the user identification problem is to determine whether these two display names belong to the same individual or not.



(a) FB-TW



(b) FS-TW



(c) FB-FS

Fig.22. user identification performance based on display names

Based on the datasets and features described in the previous section, several supervised machine learning models are introduced for user identification. We use eight classifiers including Gaussian Naïve Bayes (GaussianNB), Bernoulli Naïve Bayes (BernoulliNB), Logistic Regression, Logistic Regression with builtin Cross-Validation (LogisticRegressionCV), Support Vector Machine (SVM), Decision Tree, Random Forest, and Maximum Entropy Model to train the identification model, respectively. The first seven classifiers could be achieved through the implementation provided by scikit-learn [5], and the Maximum Entropy Model is achieved by Zhang's maxent toolkit [7]. All parameters of these classifiers are default. For each classifier and dataset, we perform 10 runs, and then report the average of the results. The identification results on three datasets are illustrated in Fig.22, and the corresponding AUC value of every classifier is also list in the legend.

The identification results show these classifiers could achieve good precision on three datasets, especially SVM and Maxent with all AUC values more than 0.9 on three datasets. This indicates that these suitable features we measured and analyzed above are capable to identify user across OSN sites effectively.

8. Conclusion

A display name is a name that an individual chooses shown to other avatars on an OSN site. By comparing the display names from the same users and the different users, we know that the character similarity and the letter distribution similarity of the positive instances are very high. The results of our measurements demonstrate that the same individual on different OSNs tends to use the same display names or similar display names. We final apply the measurement results to identify a user across social networks and the results proved that the presented attributes are very helpful for identifying whether accounts belong to the same individual or not based on their display names.

Acknowledgements This research is supported in part by Shaanxi Provincial Natural Science Foundation Research, China under grant No. 2014JM2-6104.

Reference

[1] Datatang Names Corpus, 2016. <http://www.datatang.com/data/12061>

[2] Foursquare. about, 2016. <https://Foursquare.com/about>

[3] Foursquare changes privacy policy but it shouldn't cause a ruckus, 2016. <http://www.phonearena.com/news/Foursquare-changes-privacy-policy-but-it-shouldnt-cause-a-ruckus_id38169>

[4] Most famous social network sites worldwide as of April 2017, 2017. <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

[5] Scikit-learn-Machine Learning in Python, <http://scikit-learn.org/stable/>.

[6] Twitter Name Generator, < http://Twitternamegenerator.com>

[7] Le Zhang. Maximum Entropy Modeling Toolkit for Python and C++, <http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html>.

[8] Francesco Buccafurri, Gianluca Lax, Antonino Nocera, Domenico Ursino. Discovering Links among Social Networks, in: Proceedings of 2012 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2012, pp. 467-482.

[9] Yang Chen, Chenfan Zhuang, Qiang Cao, Pan Hui. Understanding Cross-site Linking in Online Social Networks, in: Proceedings of the 8th Workshop on Social Network Mining and Analysis, 2014, Article No. 6.

[10] William W. Cohen, Pradeep Ravikumar, Stephen E. Fienberg. A comparison of string metrics for matching names and records, in: Kdd workshop on data cleaning and object consolidation. 2003, 3: 73-78.

[11] B. Fuglede, F. Topsoe. Jensen-Shannon divergence and Hilbert space embedding, IEEE International Symposium on Information Theory. 2004: 31-31.

[12] Oana Goga, Howard Lei, Sree Hari Krishnan, Gerald Friedland, Robin Sommer, Renata Teixeira. Exploiting innocuous activity for correlating users across sites, in: Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 447-458.

[13] James W. Hunt, Thomas G. Szymanski. A fast algorithm for computing longest common subsequences. Communications of the ACM, 20(5): 350-353, 1977.

[14] Faiza Hussain, Usman Qamar. Identification and Correction of Misspelled Drugs' Names in Electronic Medical Records, in: Proceedings of 18th International Conference on Enterprise Information Systems, 2016, pp. 333-338.

[15] Tereza Iofciu, Peter Fankhauser, Fabian Abel, Kerstin Bischoff. Identifying Users Across Social Tagging Systems, in: Proceedings of the Fifth International Conference on Weblogs and Social Media, 2011, pp. 522-525

[16] Paridhi Jain, Ponnurangam Kumaraguru, Anupam Joshi. @ i seek'fb. me': identifying users across multiple online social networks, in: Proceedings of the 22nd international conference on World Wide Web companion, 2013, pp. 1259-1268.

[17] James M. Joyce. Kullback-leibler divergence, International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, 2011: 720-722.

[18] Kunho Kim, Madian Khabsa, C. Lee Giles. Random Forest DBSCAN for USPTO Inventor Name Disambiguation. arXiv preprint arXiv:1602.01792, 2016.

[19] Yongjun LI, You Peng, Wenli JI, Zhen Zhang, Quanqing Xu. User Identification based on Display Names across Online Social Networks. IEEE Access. 5: 17342-17353, 2017.

[20] Yongjun Li, You Peng, Zhen Zhang, Quanqing Xu, and Hongzhi Yin. Understanding the User Display Names across Social Networks. In: Proceedings of the 26th International Conference on

World Wide Web Companion, 2017, pp. 1319-1326.
[21] Yongjun LI, Zhen Zhang, You Peng, Hongzhi Yin, Quanqing Xu. Matching User Accounts based on User Generated Content across Social Networks. Future Generation Computer Systems. 83: 104-115, 2018.
[22] Dong Liu, Quanyuan Wu, Weihong Han, Bin Zhou. User Identification across Multiple Websites Based on Username Features. Chinese Journal of Computers. 38(10): 2028-2040, 2015
[23] Jing Liu, Fan Zhang, Xinying Song, Young-In Song, Chin-Yew Lin, Hsiao-Wuen Hon. What's in a name?: an unsupervised approach to link users across communities, in: Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp.495-504.
[24] Anshu Malhotra, Luam Totti, Wagner Meira Jr, Ponnurangam Kumaraguru, Virgiio Almeida. Studying User Footprints in Different Online Social Networks. in: Proceedings of Advances in Social Networks Analysis and Mining, 2012, pp. 1065-1070.
[25] Wataru Matsubara, Shunsuke Inenaga, Akira Ishino, Ayumi Shinohara, Tomoyuki Nakamura, Kazuo Hashimoto. Efficient algorithms to compute compressed longest common substrings and compressed palindromes, Theoretical Computer Science, 410(8):900-913, 2009.
[26] Marti Motoyama, George Varghese. I seek you: searching and matching individuals in social networks, in: Proceedings of the eleventh international workshop on Web information and data management, 2009, pp. 67-75.
[27] Gonzalo Navarro. A guided tour to approximate string matching, ACM computing surveys, 33(1): 31-88, 2001.
[28] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, Supachanun Wanapu. Using of Jaccard coefficient for keywords similarity, in: Proceedings of the International MultiConference of Engineers and Computer Scientists, 2013, pp. 380-384
[29] Raphael Ottoni, Diego Las Casas, Joao Paulo Pesce, Wagner Meira Jr, Christo Wilson, Alan Mislove, Virgilio Almeida. Of Pins and Tweets: Investigating How Users Behave Across Image- and Text-Based Social Networks, in: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014, pp. 386-395.
[30] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, Pere Manils. How unique and traceable are usernames? in: Proceedings of International Symposium on Privacy Enhancing Technologies, 2011, pp. 1-17.
[31] Elie Raad, Richard Chbeir, Albert Dipanda. User profile matching in social networks, in: Proceedings of 13th International Conference on Network-Based Information Systems, 2010, pp. 297-304.
[32] Jan Vosecky, Dan Hong, Vincent Y. Shen. User identification across multiple social networks, in: Proceedings of the 1st International Conference on Networked Digital Technologies, 2009, pp. 360-365.
[33] Pinghui Wang, Wenbo He, Junzhou Zhao. A Tale of Three Social Networks: User Activity Comparisons across Facebook, Twitter, and Foursquare, IEEE Internet Computing, 18(2): 10-15, 2014.
[34] Reza Zafarani, Huan Liu. Connecting users across social media sites: a behavioral-modeling approach, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 41-49.