

# Who's Who: Linking User's Multiple Identities on Online Social Media

Paridhi Jain  
IIIT-Delhi, India  
paridhij@iiitd.ac.in

Ponnurangam Kumaraguru  
IIIT-Delhi, India  
pk@iiitd.ac.in

Anupam Joshi  
UMBC, United States  
joshi@cs.umbc.edu

## 1. ABSTRACT

On online social media, users join new online social networks (OSNs) to exploit variety of services while maintaining their old identities on other OSNs. A user maintains an identity on each OSN mentioning metadata (e.g. profile information) about her. Heterogeneity of metadata shared by user across OSNs leads to a problem of finding if two online identities on multiple OSNs belong to the same user or different users. In this work, we attempt to understand that to what extent can we link multiple online identities of a user or disambiguate identities of different users, using an easily accessible and public attribute – username. The solution to the problem has multiple applications. In privacy domain, the problem finds its application in understanding the quantity and quality of the user’s information leakages via either aggregation of user’s information from OSNs or differences in privacy policies of multiple social networks. In system building domain, the solution can help in building recommendation feature for social aggregation sites. In security domain, the solution can help in linking malicious user accounts present on multiple OSNs.

## 1.1 Methodology

We collected usernames of 1,193 users on different social networks and created two datasets by different methods. In dataset 1, no two users shared the same name and hence their usernames were distinct and easily separable. However in real world, disambiguation of two users with similar names was a challenge. Therefore in dataset 2, there exists users who shared the same first name and had similar usernames. The existence of similar usernames belonging to different users challenged the techniques we proposed to link identities of a user and disambiguate identities of different users.

We proposed a set of string based features to capture the possible similarities a user’s two usernames had, in order to predict if two usernames belong to the same user. Some of the strong features were – n-gram coefficient, Jaccard coefficient, Affine gap and Smith-Waterman distance.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*The 18th International Conference on Management of Data (COMAD),  
14th-16th Dec, 2012 at Pune, India.*

Copyright ©2012 Computer Society of India (CSI).

## 1.2 Analysis and Results

We analyzed 1,193 users and found that 359 users (30%) used same username, and 327 (27.4%) users had twisted versions of the user’s most used username on every OSN. Rest of the users had atleast one different username on at least one OSN. We observed that a more than half the users (30% + 27.4%) had same or similar users, with possible reasons as – the username they wished was already taken or they modified their username on the basis of social network nature. This motivated the string based feature set and the techniques we discuss to link users.

### 1.2.1 Classification

We extracted a set of string based features for a username pair (437,836 pairs for dataset 1 and 4,384 pairs for dataset 2) either belonging to same user or different users. We performed random sub-sampling validation 10 times with 50% training and 50% testing dataset. We used SVM with RBF kernel to classify the username pair if it belonged to the same user or two different users. SVM with the training accuracy of 93.7% for dataset 1 and 85% for dataset 2, yielded a classification accuracy of 99.85% on dataset 1 while 75.37% on dataset 2.

The classification accuracy (99.85%) is higher than the state-of-the-art accuracy (71%) by Perito et. al [1] which experiments with 10,000 username pairs of the users where no two users have same names (similar to dataset 1). The higher accuracy shows that string based features are efficient in predicting similarities of two usernames of a user. However, the classifier makes errors when different users with similar usernames are marked as the usernames belong to the same user (accuracy - 75.37%). To distinguish between users with similar name, we need to incorporate other attributes e.g. profile and network attributes.

## 1.3 Conclusion

In conclusion, we observe that majority (57.4%) of users use same or similar usernames across multiple online networks. Therefore we argue that username can be used as a unique identifier to link user identities across OSNs. With string based features of a username pair, accuracy of correct prediction can be improved from 71% to 99.85%.

## 2. REFERENCES

- [1] D. Perito, C. Castelluccia, M. A. Kâafar, and P. Manils, “How unique and traceable are usernames?” in *PETS*, 2011.