# CROSS-SYSTEMS IDENTIFICATION OF USERS IN THE SOCIAL WEB

Francesca Carmagnola
*Department of Computer Science, University of Turin*
*Corso Svizzera, 185 - 10149 Torino (Italy)*

Francesco Osborne
*Department of Computer Science, University of Turin*
*Corso Svizzera, 185 - 10149 Torino (Italy)*

Ilaria Torre
*Department of Computer Science, University of Turin*
*Corso Svizzera, 185 - 10149 Torino (Italy)*

**ABSTRACT**

In the *Social Web*, users typically interact with different S*ocial network systems* and can have different accounts and different identities. Identifying the users across the Web, independently of the protocols supported by each social system and independently of the authentication data supplied by the user, is, so far, a challenge. This paper presents an approach to perform uniquely user identification by using the public user data distributed across the systems in the Social Web.

**KEYWORDS**
Cross-systems identification, user profile data interoperability, Social Web

## 1. INTRODUCTION AND RELATED WORKS

In the so called *Social Web*, people interact one with each other throughout the World Wide Web, sharing knowledge and interests. There are different ways to socialize, collaborate and share interests on the Social Web, supported by different *social systems*. These systems allow to create relations among users. The notion *social networks* specifically identifies connections of people interacting and creating relationships (Scott, 2000) and *social network systems* identify the specific systems that allow to manage and support these relationships. They offer people tools to create personal pages where they describe themselves, providing information about who they are (e.g. age, gender, region, education, interests etc.) and who they know (friends). Examples of social network systems are Myspace (myspace.com) and Facebook (facebook.com).

Other social systems, like YouTube (youtube.com) and Flickr (flickr.com), allow users to share multimedia objects and annotate them (comment, rate, tag, etc.). Instead, Delicious (del.icio.us), Digg (dig.com) and Magnolia (ma.gnolia.com), are examples of websites that allow users to share bookmarks and tag them and Twitter is an example of social network and microblogging website that allows people to post brief messages (about what they are doing, their mood, what happens around them) on their personal pages. There are also tools, such as Friend Connect[1] that allow websites' owners to integrate social features onto their website, creating a community.

A relevant feature is that all these social systems collect data about users. Users have an account and, in different ways, provide data about themselves. A vast literature describes the phenomenon of massive participation of users to content production and the users' personal and social satisfaction in doing this (see for example Marlow et al., 2006). In this Social Web, users typically interact with different systems and can have different accounts and different identities.

[1] http://google.com/friendconnect

Equal

Currently, several initiatives, like OpenID (opened.net), OpenSocial (opensocial.org), Connect[2] and MySpaceID,[3] previously called Data Availability, try to face the issue of user identification and profile portability across systems. More in detail, OpenID is an initiative to provide a single digital identity across the Internet. OpenSocial is a set of APIs, developed by Google along with MySpace and other social networks for the portability of the user personal profile, letting the user authorize access to her data stored in social networks. It is important to underline that interoperability between social networks requires the applications to implement the OpenSocial APIs. Similar initiatives, specifically carried out by Facebook and MySpace, are Connect and MySpaceID, which allow users to share their data on other social systems and require, as well as OpenSocial, that the participating social system supports the proper protocol. Identifying the users across the Web, independently of the protocols supported by each social system and independently of the authentication data supplied by the user, is, so far, a challenge.

This paper presents an approach to perform uniquely user identification by using the public user data distributed across the social systems. We consider the user identity as a collection of attributes. For identifying users across systems we perform a semi-combinatorial weighed match between all the matching attributes collected by the distributed user profiles. For the nickname attribute, we also calculate a specificity function that computes its rareness. Profiles whose match is over a specified threshold are considered belonging to the same user.

What can this identification be useful for? Identifying users across systems can have several uses. Consider John who likes the bookmarks posted by another user, or John who is interested in blog posts, photos or messages of another user. The only public data about this user are, for instance, her username and her country. It is likely that John would be interested in having more information about this user. For example, he could be interested in knowing her age and gender, her interests, but also her home page, her account on other social systems, and so on. Our approach to crawl other social systems comparing profiles' attributes (in this case username and country), can be used to identify this user on other systems and collect the public data they expose. A related work regarding the aggregation of user profiles sparse over the Web is sketched in Ghosh and Dekhil (2009). Another important use of our approach for cross-systems identification is for the owner of the profiles, who can use this service to monitor and control her public data distributed over the Web. The issue of control of information in user profiles in social networks is analyzed, for example, in Blanco et al (2008).

Notice, moreover, that our approach can also be used as an integration to other approaches like OpensSocial, seen above, based on proprietary protocols. The limit is that, using our approach, we can only collect public data, but the advantage is that it does not require the user to provide her credentials to the other websites.

An approach that, similarly to ours, tries to identify users over other social systems automatically is based on the Google Social Graph APIs. Given an URL about the user, these APIs find public connections of this user in other web services. The current indexing is limited to pages that use XHTML Friends Network (XFN[4]) and Friend of a Friend (FOAF[5]) and some other declared connections. This approach is used, for example, in Szomszor et al (2008). Other applications for user identification start from personal data to find information about a user (see for example the projects FindMeOn[6], profilelinker[7]). However, on the Web, users are often identified only by a few variable number of attributes (e.g., in Delicious nearly the 80 % of the users have provided only their nickname). Our project tries to address this issue: identifying users, given a variable number of attributes available.

The paper is structured as follows: the next section describes the approach and the specific algorithms we defined for the user identification, Section 3 provides and discusses some statistics obtained by running the algorithm over two popular systems and Section 4 concludes the paper.

---

[2] http://developers.facebook.com/connect.php
[3] http://wiki.developer.myspace.com/index.php?title=Category:MySpaceID
[4] http://gmpg.org/xfn/
[5] http://www.foaf-project.org/
[6] http://www.findmeon.org/
[7] http://www.profilelinker.com

## 2. AN APPROACH FOR CROSS-SYSTEMS USER IDENTIFICATION

The notion of *identity* is extremely important to disambiguate individuals. In the object-oriented programming, "identity" is defined as the set of properties of an object that allows it to be distinguished from the others. Referring to this definition, we consider the user identity as a collection of properties that uniquely represents a user. Among the identity properties, the persistent features (e.g. full name, gender, age) are the most relevant when disambiguating one individual from another (Windley, 2005).

**User Identification Algorithm.** To identify users across systems, we perform a semi-combinatorial weighed match between the available persistent and non-persistent identity properties included in the individual public profiles maintained by the different social systems the user interacts with. To collect such identity properties, we crawl the user profiles in the social systems, using the system's public APIs, when available, or, otherwise, parsing the public profiles web pages. Comparing persistent information provides strong indicators for assessing the match between two profiles. In particular, we distinguish properties that are strong indicators when a match occurs between them (positive match) and properties that are strong indicators when they do not match (negative match). In the first case, they increase the probability that the profiles belong to the same user, while in the second case, they work as negative triggers, that drastically reduce the probability of match. For example, the mere co-occurrence of the same gender in two profiles is not very significant, since the probability for each value is 0.5. Conversely, if two profiles have different gender, this information is a trigger to exclude a match between such profiles. Considering non-persistent properties, an interesting difference from persistent information is that, in case of positive match, they can be strong indicators (e.g. two profiles with the same url of the user blog) or weak indicators (e.g. two profiles sharing the same country), but they are never strong indicators in case of negative match, since different values may simply reflect a not suitable update of the profile. In this last case this is a weak negative indicator of matching identities.

Our algorithm exploits this mechanism, introducing strong and weak weights (ranging [0,0.6]) for positive and negative matches of properties. To this aim, for each property, we have assessed a set of markers, as shown in Table 1.

Table 1. Sample of properties and of their weights in case of positive or negative match

| *ID* | *Property to be compared* | *Weight for Positive match (Wp)* | | | *Weight for Negative match (Wn)* |
|---|---|---|---|---|---|
| | | Same | Similar | Included | |
| 1 | Full name | 0.4 | 0.2 | - | 0.5 |
| 2 | Age | 0.2 | 0.05 | - | 0.4 |
| 3 | Gender | 0.05 | - | - | 0.5 |
| 4 | Province | 0.1 | 0.05 | 0.05 | 0.1 |
| n | …. | …. | …. | …. | …. |

Notice that a positive match can be obtained when values are exactly the same, are similar or one is included into the other (for example, a province included into a country). These options can be applied just to some properties. When they cannot be applied, Table 1 displays a "-" symbol.

The identification score is assessed by measuring the difference between the weighted values of the positive matches and the weighted values of the negative ones. This approach allows to manage different combinations of properties without requiring the development of ad hoc formulas. We call this identification score *IdScore(x,y)*, with x and y being two user profiles in two different social systems, and we define it as:

$$IdScore(x,y) = \alpha \sum_{i=1}^{n} Wp(i) - \beta \sum_{i=1}^{n} Wn(i) \quad (1)$$

where $\sum Wp(i)$ indicates the addition of the weights of the positive matches for each property (i) in Table 1, and $\sum Wn(i)$ indicates the sum of the weights of the negative matches.
If a property in a profile is not available, its weight is 0, that is, it does not affect the computation.

$\alpha$ and $\beta$ represent two constants, they can be used to weight more or less positive and negative matches, depending, for example on the kind of social systems used in the comparison. Currently they are set to 1.

**Specificity Algorithm**. As said in the Introduction, in the Social Web the nickname is a fundamental identification property and, in many cases, the only one available (e.g. 29.16% of Skype users and 82,3% of Delicious users have empty profiles). Therefore, analyzing the nickname is a very relevant component to identify users across systems. To this aim, when there is a perfect match between the nickname in different systems, we estimate *IdScore*(x,y) by taking into account also the *specificity* of the nickname, that is how uncommon and rare it is. We define the specificity score (*SpScore*) as a function of the lengths of the nickname (long nicknames are probably more specific than shorter ones) and of a score (*characterScore*) indicating its rareness in a given sample population. To estimate the characterScore, we have defined ten categories of nicknames, considering the different possible combinations of letters, numbers and alphanumeric characters, and we have assigned to each category a weight which is inversely related to the number of nicknames of that type contained in a given sample.

For example, considering 20.000 random MySpace nicknames, 63.9% of them belong to the category *alpha_min* (nickname composed only of small letters, like "bill"), while only 0.03% of them belong to the category *alpha_min_numeric2* (nickname composed of a numeric series followed by small letters, like "22bill"). Thus, this second type of nickname, among the MySpace population, will get a higher characterScore.

**New user identification algorithm.** Therefore, we modify function (1) defined above combining it with the nickname SpScore, using the Bernoulli addictive formula.

$$IdScore(x,y) = \left(\alpha \sum_{i=1}^{n} Wp(i) - \beta \sum_{i=1}^{n} Wn(i)\right) + \left(\gamma\, SpScore\right) - \left(\left(\alpha \sum_{i=1}^{n} Wp(i) - \beta \sum_{i=1}^{n} Wn(i)\right)\left(\gamma\, SpScore\right)\right) \quad (2)$$

An example of estimation of *IdScore(x,y)* based on this new algorithm is reported in Table 2.

Table 2. Estimation of *IdScore*(x,y) given two examples of nicknames perfectly matching between MySpace and Skype but having different SpScores and different types and numbers of matching identity properties (N.m.p*)

| *Nickname* | *SpScore* | *N.m.p.** | *∑Wp(i)* | *∑Wn(i)* | *IdScore(x,y)* |
|---|---|---|---|---|---|
| anafpres | 0.4 | 4 (gender, city, country, full name) | 0.65 | 0 | 0.8 |
| adamuk98 | 0.64 | 2 (city, country) | 0.2 | 0 | 0.71 |

* Number of matching properties (same, similar or included match) other than nickname

When the identification score *IdScore(x,y)* exceeds a specified threshold, the profiles x and y are considered as belonging to the same user. The estimation of the threshold, as well as a preliminary evaluation of the algorithm, will be discussed in the following section.

# 3. RESULTS AND DISCUSSION

To evaluate our approach, we crawled two popular systems with public profiles, MySpace and Skype. The first step was to use them to test the feasibility and validity of the approach, getting some statistics about the profiles of their users. This analysis aimed at estimating the percentage of user profiles' properties with public data. Table 3 shows the results on a random sample of 5000 MySpace profiles and 5000 Skype profiles.

Table 3. Percentage distribution of user profiles' public properties

| *System* | *Full name* | *Gender* | *Age* | *City* | *Province* | *Country* | ……. | *Empty profile* |
|---|---|---|---|---|---|---|---|---|
| MySpace | 8.3 | 91.5 | 100 | 38.9 | 74.1 | 95 | ……. | 0 |
| Skype | 52.4 | 2.1 | 1.3 | 58.4 | 9.9 | 69.5 | ……. | 29.2 |

These results seem promising for the definition of our approach, showing an high percentage of persistent (see Sec. 2) properties filled in and made public. In particular, see the high percentage of the strong negative indicators *age* and *gender* (for MySpace) and that one of the strong positive indicator *full name* (for Skype).

The second step was to evaluate the identification algorithm and set a reasonable threshold for the identification score, over which, a user can be identified. We used the dataset obtained by crawling the systems mentioned above. We ran the identification algorithm and compared the previous 5000 MySpace profiles with all the available Skype profiles. For each pair of profiles (one from MySpace and one from Skype), the algorithm produces an identification score (*IdScore*) in the range 0 – 1. Figure 1 shows the distribution of the identified users (x-axis) in correspondence of different identification scores (y-axis). To set the threshold over which two profiles can be identified as belonging to the same user, we performed a qualitative analysis on all the pairs of profiles with identification score higher than 0.4 Three subjects evaluated each pair of profiles, assigning a score to the match: 0=different users, 0.5=probably same user, 1=same user. This analysis led to fix to 0.7 the threshold for user identification. As an example of match with identification score higher than 0.7, see Table 2 in the previous page.

Figure 1. Users identified on MySpace and Skype

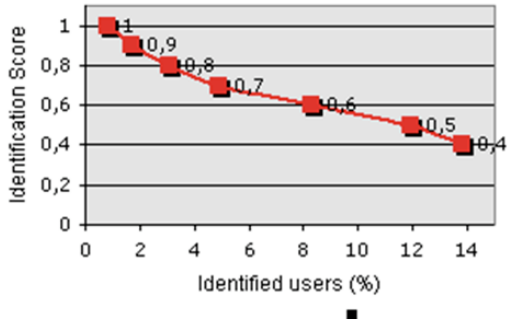
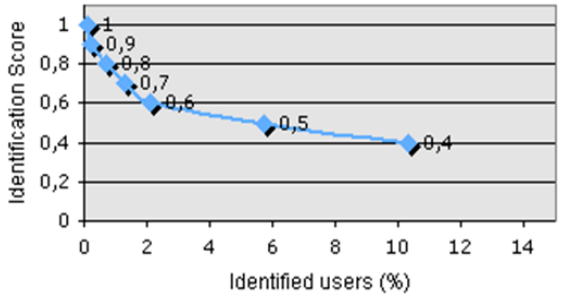



Figure 2. Users identified on MySpace and Delicious




At a first glance, results seem quite negative, since just 4,87 % of the MySpace profiles match with a profile on Skype with an identification score higher than 0.7. However, some considerations will show that this results are biased by the fact of using just two systems, and in particular systems with a different distribution of public properties.

1) *Different distribution of user profiles' public properties on MySpace and Skype.* Discussing Table 3, we emphasized the high percentage of strong indicators in both the profiles. However, we can also notice that these properties are not filled in with similar distributions on the two systems. As the algorithm works comparing pairs of corresponding properties, it is clear that the probability of uniquely identify users is low if just these two specific systems are used for the identification. An extreme case is using a system, as Delicious showed in Figure 2, where almost only the nickname is available in the profiles. In this case, the identification is nearly only based on the *Specificity Score* of the nicknames.

2) *IdScore threshold* is currently set to 0.7, but this value is highly dependent on the weights we assigned to the profile's properties and to the types of nicknames, discussed in the previous section, which still require to be tuned.

3) *Low probability of a user to have an account, and thus a profile, on both MySpace and Skype*. This is an intrinsic consequence of crawling only two systems on a total of over 150 social systems, counted just considering social networks[8]. Consider, in addition, that MySpace accounts represent less than 13% of total accounts. This data are consistent with the percentage of users we identified by using MySpace and Skype. We expect more significant results by crawling more social systems. Currently we are working on crawling Twitter and Facebook.

[8] http://en.wikipedia.org/wiki/List_of_social_networking_websites

## 4. CONCLUSION

To conclude, we summarize our approach as a way to uniquely identify users on the Web, but especially on the Social Web, by comparing pairs of properties of the user's profiles and by considering the *specificity* of the nickname accounts. The final objective is to be able to reunify the public information about a user distributed on different social systems. This can be useful to other users, which can get information about "anonymous" nicknames, but also to the user herself, who can use the application for automatic identification, profile portability and as a monitor of all her current public data on the Web, independently of specific protocols and platforms, as explained in the Introduction.

This is an ongoing project. Besides extending the analysis to more social systems, we are also planning a deeper evaluation to tune the value of the score threshold, by contacting users identified with score over 0.7 to have a confirmation of the identification. The main goal is to avoid false positives that may entail to unify data from two different users.

## REFERENCES


Blanco, D. et al, 2008, Social Identity Management in Social Networks. *Proceedings of the International Symposium on Distributed Computing and Artificial Intelligence 2009 (DCAI 2008)*. Advances in Soft Computing, Vol. 50, Springer Berlin / Heidelberg, pp. 62-70.

Ghosh, R. and Dekhil, M., 2009, Discovering user profiles. *Proceedings of the 18th international Conference on World Wide Web (WWW '09)*. Madrid, Spain, April 20-24, pp. 1233-1234.

Marlow, C. et al, 2006, HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the seventeenth conference on Hypertext and hypermedia (HT '06)*. New York, NY, USA, pp. 31-40.

Scott, J. P., 2000. *Social Network Analysis: A Handbook*. SAGE Publications, London.

Szomszor, M. et al, 2008, Semantic modelling of user interests based on cross-folksonomy analysis. *Proceedings of the 7th International Semantic Web Conference ISWC*, Karlsruhe, Germany, vol. 5318 of Lecture Notes in Computer Science, Springer, pp. 632–648.

Windley, P., 2005. *Digital Identity*. O'Reilly Media, Inc., Sebastopol, CA.