

# Connecting Corresponding Identities across Communities

Reza Zafarani, Huan Liu

Department of Computer Science and Engineering  
Arizona State University  
Tempe, AZ 85281  
{reza, huanliu}@asu.edu

## Abstract

One of the most interesting challenges in the area of social computing and social media analysis is the so-called community analysis. Here, a community refers to a particular social media website. The current research in this area seeks to understand the behavior of communities by means of different techniques such as link analysis and opinion mining; however, in most cases, if not all, these analysis techniques are restricted to a single community or more specifically *domain*. In this type of analysis, the inter-connections across different communities are neglected. As an intuitive result, these methods miss the opportunity of data analysis across multiple communities, which in turn can reveal many latent behavioral patterns shared among users while interacting in the cyberspace. A well known barrier in cross-community analysis is the disconnectedness of these websites. This is due to the unrevealing structure of the web and the fact that most websites preserve the anonymity of users by allowing them to select usernames instead of their real identities. In fact, given a mapping between usernames and their real identities, it's possible to analyze these social media websites as one enormous community. In this paper, our aim is to provide evidence on the existence of this mapping. Our studies have shown that simple, yet effective approaches, which leverage social media's collective patterns can be utilized to find such mapping. The employed methods were able to successfully reveal this mapping with a 66% accuracy. Moreover, different intelligent methods have been proposed to boost this performance and to make it as robust as possible. This identification process, as we have mentioned before, alleviates the burden of disconnectedness among various communities.

## Introduction

Community analysis has been an interesting problem in the recent developments of Data Mining and Social Media Analysis (Wasserman and Faust 1994). Here, a *community* refers to a specific social media website (e.g. *StumbleUpon*). This problem has a direct application in many research areas such as correlation and influence analysis (Anagnostopoulos, Kumar, and Mahdian 2008), or the group interaction analysis (Agarwal et al. 2008; Brinkmeier, Werner, and Recknagel 2007). As an example, consider the group interaction analysis in social media. This problem can be informally stated

as follows: given a social group within a social-media community, what are the most similar groups to it?, and how can we leverage relationships among users to predict group similarities? The problem has been extensively analyzed by different researchers in the area of social computing (Agarwal et al. 2008; Brinkmeier, Werner, and Recknagel 2007).

It is worth mentioning that the current research seeks to analyze communities by means of different techniques such as Link Analysis and Opinion Mining (Flake, Lawrence, and Giles 2000; Hu and Liu 2004). However, in most cases, if not all, all the analyses are restricted to a single community. For example, in single-community group interaction analysis, the user groups are analyzed within the context of a specific community and the inter-connections between different communities and user-relationships across these communities are neglected. As an intuitive result, these methods miss the opportunity of data analysis across multiple communities, which in turn can reveal many latent behavioral patterns shared among users while interacting in the cyberspace.

A major problem when dealing with any kind of cross-community analysis is the disconnectedness of these communities. A major missing element is the connectivity among users in different communities, which is an essential factor in any link analysis algorithm. This is due to the unrevealing structure of the web and the fact that most communities preserve the anonymity of users by allowing them to select usernames instead of their real identities and the fact that different websites employ different usernames and authentication systems. Furthermore, communities rarely share Single-Sign-On procedures, where users can logon to different communities using a single username (e.g. as in *Orkut* and *YouTube*). Nevertheless, if there exists a mapping between usernames across different communities and the real identities behind them, then connecting communities across the web becomes a straightforward task. Can we find this mapping? In this paper our aim is to provide evidence on the existence of this mapping, and on a broader level, demonstrate a step-by-step procedure, by employing which corresponding identities across communities can be discovered.

As an example, consider the situation depicted in Figure 1. A specific user is active in both communities *A* and *B* (shaded circles). As long as the one-one mapping (shown as arrows) between his/her usernames in these communities is

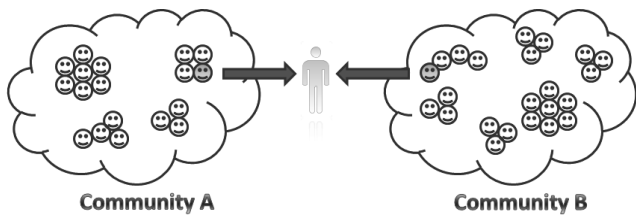


Figure 1: A Typical Inter-Community Connection Scenario

discovered, then the existing connection between these two disjoint communities can be revealed.

This paper is organized as follows. First, we reveal the motivation behind this work and review the impact of such a study in practice. Then, we formally present the corresponding identity elicitation problem. The section that follows present the empirical observations that we had regarding the behavior of users on the web. Our proposed method for identifying corresponding identities is discussed next, followed by our experimental results and our conclusions.

### How Can Corresponding Users Be Useful?

The first question that arises here is on whether there is any need for finding the corresponding users across different communities. Here, we aim to answer this question in detail and we provide evidence on the requirement of this kind of research and the impact it has on related areas.

There are different perspectives through which we can analyze the influence of this kind of research. The following are some of the major areas of impact, when multiple communities can be analyzed:

1. **Influence Analysis:** consider the situation where influence analysis is not only measured within a community, but also across different communities. In real world scenarios, influentials can influence individuals across different communities. In this case, it is interesting to know how an influential figure within one community, perform in terms of his/her influence in other communities. Moreover, do influentials move from one community to another?, and if they do, how does their influence evolve as this transition occurs? In fact, given a universal definition of influence (Anagnostopoulos, Kumar, and Mahdian 2008), we can analyze the fluctuations of influence across different communities. For example, we may deduce that if an individual is influential in community  $A$ , he/she would be also considered influential in community  $B$ , as long as communities  $A$  and  $B$  are semantically correlated.
2. **Multiple Community Group Interaction:** another interesting research area that can benefit from this research is the cross-community group interaction analysis. By locating users across different communities, multiple-community group interactions can be viewed as an instance of single-community group interactions. Hence, methods proposed for single community group interaction analysis (Agarwal et al. 2008) can be utilized to model this problem.

3. **Analyzing Network Dynamics:** the observation of network dynamics (e.g. any low-level network property such as the diameter of the network, or metrics such as the user centrality at a higher level) under the situation where multiple communities exist is another challenging problem. For example, it is phenomenologically appealing to find out how adding/removing a new community changes the overall dynamics of a network. As another example, it seems interesting to find an answer to the question of how close are the dynamics of a single community network in comparison to that of a multi-community network?

It is worth mentioning, that in order to conduct research in any of these areas, we first require the construction of a network from multiple communities (see Figure 1) for which all (or most) inter-connections between these communities have been revealed, i.e. user-mapping is found. However, as we will see, finding these connections can be tedious and is quite involved. So, before tackling this problem, we will first overview and formally define the problem of finding this mapping.

### Cross-Community Corresponding Identity Elicitation Problem

There are many properties of web communities that can be employed in order to elicit the connection between these communities. As we have mentioned before, usernames are one of the gadgets using which these connections can be revealed. However, there are other connection-revealing cues too. As another example, E-mail addresses can also play an important role in discovering inter-community connections. The uniqueness property of E-mail addresses allow them to act as a universal identifier of individuals across different communities; However, in this paper, we aim to find these connection by employing usernames. Hence, here we aim to formally state this problem when usernames are used a community-linkage tool.

Let  $u$  represent an active individual in the cyberspace. Let  $C$  represent the set of all communities and  $c_i \in C$  represent a single community. Let  $C_u \subset C$  denote the set of all communities user  $u$  is active in, i.e. has a username. We denote the set of all active users in community  $c_i$  as  $\Upsilon_{c_i}$ . Let  $U(u, c_i)$ ,  $c_i \in C_u$  represent the username user  $u$  employs in community  $c_i$  and let  $U^{-1}$  represent the inverse function (*username*  $\rightarrow$  *user*) such that  $U^{-1}(U(u, c_i), c_i) = u$ . Furthermore, a *username-username* pair  $\langle \gamma, \delta \rangle$  for some user  $u$  and communities  $c_i$  and  $c_j$  is defined as the following:

$$\langle \gamma, \delta \rangle: U(u, c_i) = \alpha, U(u, c_j) = \delta, u \in \Upsilon_{c_i}, u \in \Upsilon_{c_j}$$

whereas, a *username-community* pair  $\langle \alpha, \beta \rangle$  for some user  $u$  is defined as the following:

$$\langle \alpha, \beta \rangle: U(u, \beta) = \alpha, u \in \Upsilon_{\beta}$$

Moreover, a *username-set* for user  $u$  ( $\Lambda_u$ ) is defined as the following:

$$\Lambda_u = \{U(u, c_i) | c_i \in C_u\}$$

Similarly, a *community-username-set* for community  $c_i$  ( $\Sigma_{c_i}$ ) is defined as the following:

$$\Sigma_{c_i} = \{U(u, c_i) | u \in \Upsilon_{c_i}\}$$

Then, cross-community corresponding username elicitation can be formally stated as follows:

**Definition 0.1 Cross-Community Corresponding Username Elicitation:** *given a username-community pair  $\langle u_1, c_1 \rangle$ , called base-username and base-community, and a community  $c_2$  (target community), a solution to the cross-community corresponding username elicitation problem is a username  $u_2 \in \Sigma_{c_2}$ , called the target-username, such that  $U^{-1}(u_1, c_1) = U^{-1}(u_2, c_2)$ .*

In order to provide a sufficiently accurate solution to this problem, we first devised some hypotheses on the relationship between usernames selected by a single person in different communities, and also on some of the web phenomena regarding usernames and communities. All these hypotheses are evaluated based on the empirical experiments. The results from these evaluational experiments, as we will see, tend to be useful in devising our proposed method for corresponding-username extraction.

## Empirical Observations

We have devised 7 hypotheses, each of which, if required, is formally defined and then empirically validated. The observations gathered while evaluating these hypotheses are used later on to help us construct our proposed method for extracting corresponding identities in other communities. Before we provide any of our hypotheses and the empirical observations we had while evaluating them, we provide a brief overview of our data collection for these experiments.

### Data Collection

In order to evaluate our hypotheses we required a sufficiently large dataset from which labeled data could be acquired. For this purpose we have used the *BlogCatalog*<sup>1</sup> web community and developed a data fetching engine for this website. BlogCatalog is a comprehensive directory of blogs which not only provides useful informations about various weblogs, but also comprises different facilities for users to interact within its community. What is more interesting about BlogCatalog is that users in BlogCatalog are provided with a feature called “My Communities”. This feature enables users to list their usernames in other communities. Our engine has gained advantage of this feature of BlogCatalog and has collected a large set of usernames in this community, along with their corresponding usernames in other web communities. Overall, 38093 *username-username* pairs were gathered. Each pair consisted of the username in the BlogCatalog community and the corresponding username in another community. Besides BlogCatalog, the dataset contains usernames from 36 different communities. From this dataset, the other datasets required for all our experiments were generated. We have also further analyzed the properties of this

<sup>1</sup><http://www.blogcatalog.com/>

MySpace	<a href="http://www.myspace.com/test">http://www.myspace.com/test</a>
Flickr	<a href="http://www.flickr.com/photos/test">http://www.flickr.com/photos/test</a>
Reddit	<a href="http://www.reddit.com/user/test">http://www.reddit.com/user/test</a>
YouTube	<a href="http://www.youtube.com/test">http://www.youtube.com/test</a>
Del.icio.us	<a href="http://del.icio.us/test">http://del.icio.us/test</a>

**Table 1:** Profile URLs for Popular Social Networking Webs

dataset comprehensively and the results are provided in our experimental results. Next, we provide a set of hypotheses we devised and the validation results we obtained for them using this dataset.

## Hypotheses

Before we delve into these hypotheses, we are further required to formally define some of the notations. Let  $Domain(c_i)$  denote the Registered Domain Name of community  $c_i$ . Furthermore, for any Registered Domain Name  $d_i$  and for any URL  $URL_i$ ,  $URL_i \in d_i$  denotes that  $URL_i$  in on domain  $d_i$ . Finally, the *URL-set* of community  $c_i$  ( $\Phi_{c_i}$ ) is defined as follows:

$$\Phi_{c_i} = \{URL_i | URL_i \in Domain(c_i)\}$$

- $\mathcal{H}_1$ : **for any username  $u_i$  and community  $c_j$  s.t.  $u_i \in \Sigma_{c_j}$ , there exists a non-empty set  $S \in \Phi_{c_j}$ , for which the following holds true:**  $\forall url \in S$ ,  $u_i$  is a sub-string of  $url$ . Informally speaking, this hypothesis states that for most communities and for all usernames residing on them, there exists URLs on the community website which contain the username. These URL are most commonly pointing to the profile/homepage of the users on that community. As an example, consider how the profile page URLs of a fictional user *test* can be reached on some of the most popular social networking websites in Table 1.

In order to empirically prove this phenomenon, we have analyzed more than 36 online community websites and surprisingly in all 36 there exists URLs which contain the username, i.e. 100% accuracy. As previously mentioned, in most cases, these URLs are the username’s profile location on the community website.

- $\mathcal{H}_2$ : **given a community  $c_i$ , it’s highly probable to identify  $Domain(c_i)$  using web search engines.** In order to approximate the validity of this hypothesis, we again used all 36 communities available in our dataset. For each community, a Google search was performed with  $c_i$  as the query, e.g. Flickr. It was found that in all cases, the first retrieved URL was the community’s Registered Domain Name ( $Domain(c_i)$ ) and therefore the 100% accuracy was achieved easily. This high accuracy is due to the high PageRank (Brin and Page 1998) values of these popular websites.
- $\mathcal{H}_3$ : **for any username  $u_i$  and community  $c_j$  s.t.  $u_i \in \Sigma_{c_j}$ , it’s highly probable to discover, using web search engines, a non-empty set  $S \in \Phi_{c_j}$ , for which the following holds true:**  $\forall url \in S$ ,  $u_i$  is a sub-string of  $url$ . It has been empirically proven in the first hypothesis that if a user is active on some community, then there exists

URLs containing his/her username on the community’s domain. Given this fact, this hypothesis tends to prove that these URLs can be easily found on the web using web search engines. Note that if all the existing communities on the web were known, then we would have been able to simply use the pattern through which the user profile’s URL is generated on that specific community (see Table 1) and then, check if this generated URL existed on the community website (e.g. no HTTP 404 error is encountered); However, a more realistic scenario is the case where we do not know anything about the user-profiles’s URL pattern and we are only provided with the community name. In this scenario, the first challenge is to find the community’s Registered Domain Name (e.g. myspace.com) and then, find the URLs, such as the user’s profile, which contain the username (e.g. myspace.com/**u** for user **u**). As previously discussed, given the community name, the community’s Registered Domain Name can be found quite easily. Moreover, based on our first hypothesis, we have shown that the username exists in a non-empty set of URLs residing on the community’s domain name in all cases. Hence, the task of finding this non-empty set of URLs is reduced to the task finding URLs that not only reside on the community’s domain, but they also contain the username in them. This process can be easily performed using the `inurl` (Searches within URLs) and `site` (Searches within the webpages residing on some specific Registered Domain Name) features of the Google search engine<sup>2</sup>. Another view at this hypothesis is that it analyzes the likelihood of the set of URLs containing username (e.g. user’s profile) being indexed by a search engine (Google in our case). We analyzed more than 45565 username-community pairs  $\langle \alpha, \beta \rangle$  for this experiment. A search on Google with “`inurl: $\alpha$  site:Domain( $\beta$ )`” as the query was performed. Our experiments proved that in nearly 81% of the situations, at least one URL is retrieved satisfying our conditions.

- $\mathcal{H}_4$ : **for any user  $u$ , if  $|\Lambda_u| > 1$ , then for any two usernames  $u_1$  and  $u_2$  in  $\Lambda_u$ , there is a high chance of co-occurrence of these two in search engine results.** To evaluate this hypothesis, we generated 41241 *username-username* pairs  $\langle u_1, u_2 \rangle$ , i.e. both  $u_1$  and  $u_2$  belonged to the same person’s username-set. We found using Google search that usernames co-occur in nearly 68% of the situations. Since this hypothesis holds with a reasonable accuracy, therefore we can perform a web search using one of the usernames and then perform keyword extraction on the retrieved webpages to discover the other usernames; however, though sufficiently accurate, in some cases, the retrieved URLs are many and as a direct result, keyword extraction can be quite tedious. So, we proposed another hypothesis, which deals with a somewhat more restricted version of the current hypothesis, yet can be quite useful.
- $\mathcal{H}_5$ : **for two username-community pairs,  $\langle u_1, c_1 \rangle$  and  $\langle u_2, c_2 \rangle$ , it is sufficiently likely for  $u_1$  to exist on webpages retrieved using popular search engines**

<sup>2</sup>Other search engines have similar features.

**whose URLs are a member of a non-empty set  $S \in \Phi_{c_2}$  and for which the following holds true:**  $\forall url \in S$ ,  $u_2$  is a sub-string of  $url$ . This hypothesis analyzes the chance of a username of a person occurring on the webpages whose URL contain the other username (e.g. user’s profile). Again, to evaluate this hypothesis, we generated 41241 *username-username* pairs  $\langle u_1, u_2 \rangle$ , i.e. both  $u_1$  and  $u_2$  belonged to the same person’s username-set. For each pair, two separate queries were sent to Google (first username occurring on URLs containing the second username, and vice versa). These queries were in the following format: “`inurl: $u_1$   $u_2$` ” and “`inurl: $u_2$   $u_1$` ”. We found that this hypothesis holds in nearly 38% of the situations. Likewise our previous hypothesis, and based on the results of this hypothesis, we can perform a web search using one of the usernames and then perform keyword extraction on the URLs of the webpages retrieved to discover the other usernames.

- $\mathcal{H}_6$ : **for any user  $u$ , it’s highly probable to have  $|\Lambda_u| = 1$ .** This hypothesis states that people tend to use the same username in different communities. If this holds, then the only requirement for extracting corresponding usernames in different communities is to find a single username of an individual. In order to approximate the validity of this hypothesis, we gathered 101179 *username-username* pairs  $\langle u_1, u_2 \rangle$ , i.e. both  $u_1$  and  $u_2$  belonged to the same person’s username-set. It turns out that users have selected the same username in 59% of the situations. Moreover, a 6% of *username-username* pairs are pairs for which one of the usernames is created using the other one by adding a suffix, and another 1% are the ones that are created by adding a prefix. Finally, even if the usernames are not equal or created using a prefix or suffix, there is 2% chance that they have a small Levenstein distance (Gusfield 1997), also known as Edit distance, from each other (e.g.  $\langle \text{BobThomas}, \text{Bob1Thomas} \rangle$ ). So, if we can find the popular prefixes/suffixes, then an accuracy of around 66% is expected.
- $\mathcal{H}_7$ : **for any user  $u$ , it’s highly probable to have  $|\Lambda_u| < |C_u|$ .** This is more general than the previous hypothesis and what it states is that people tend to use one of their many usernames in different communities. if this hypothesis holds, then the requirement for extracting corresponding usernames across multiple communities is to find different usernames of a person and try each username on the community’s website (e.g. check if the profile exists). In order to approximate the likelihood of this hypothesis, we evaluated this hypothesis over 36214 usernames. It turns out that users have selected the same username as one of their many usernames in 77% of the situations. Moreover, a 5% of the usernames are created by adding suffixes to one of their other usernames, and another 1% are the ones that are created by adding a prefix. So, again if we can find all the other username and the popular prefixes/suffixes, then theoretically an accuracy of around 83% is expected.

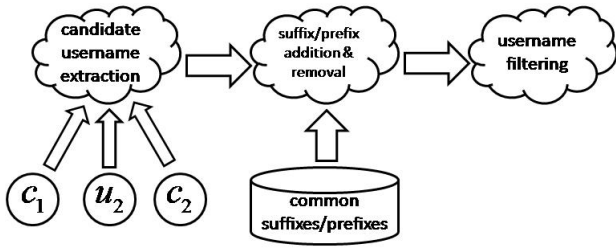


Figure 2: Corresponding Username Extraction

## An Approach to Cross-Community Corresponding Username Elicitation

In this section, we overview our proposed method which identifies corresponding usernames across communities. The procedure is depicted in Figure 2. As shown in this figure, the input to this process is the base username  $u_1$ , base community  $c_1$ , and the target community  $c_2$ . The procedure starts with finding a set of keywords, for which it believes can be candidates for the corresponding usernames in the target community. Then, in addition to keeping the original keywords, this set is expanded by adding/removing common prefixes and suffixes to its members. Note that since we have found out that *any* of the both usernames can be created by adding prefixes and suffixes ( $\mathcal{H}_6$  and  $\mathcal{H}_7$ ) to the other, hence we also remove prefixes and suffixes from these members in case they contains any of them. Finally, the members of this set are checked with the target community in order to filter out keywords which do not represent usernames in the target community.

### Candidate Usernames Extraction

One of the main hypotheses that we discussed previously ( $\mathcal{H}_5$ ) is that *usernames appear in the URLs of the profile webpages of each other*. We use this principle to extract our username sets for each username. Given a user name, based on hypothesis  $\mathcal{H}_5$ , we know that usernames co-occur in each other’s profiles; therefore, we search our base-username on Google hoping for it to be found on the user’s target-community profile or some other profiles of the same person. Since, the usernames must occur in the URL ( $\mathcal{H}_1$ ), we extract keywords from all the retrieved URLs. These keywords are preprocessed and the remaining keywords are assumed to be candidate usernames. The preprocess procedure removes common words such as the protocol names, famous sub domains, index files, extensions, etc.

### Prefix/Suffix Form Generation

As mentioned previously (hypotheses  $\mathcal{H}_6$  and  $\mathcal{H}_7$ ), after analyzing the corresponding username-username pairs, we found that users tend to create new usernames by adding prefixes or suffixes to their other usernames. Based on our data, we gathered all the prefixes and suffixes employed by the users in two separate sets. We then sorted these sets based on their frequency and the frequent prefixes and suffixes were selected. A prefix or suffix is considered frequent, if its frequency is statistically significant. In our experiments

a frequency more than  $2.5\sigma$  far from the mean frequency is considered significant, where  $\sigma$  is the standard deviation of frequencies. A sample subset of these suffixes and prefixes is provided in Table 2.

<b>Prefixes</b>	the, i, b, iam, my, free, happy, dr, x, mister, coach sexy, micro, thereal, ima, your, blogger
<b>Suffixes</b>	1, 2, s, dotcom, b, blog, 7, 07, 77, 13, a, z, art, 66, 0, 50, 18, 08, com, e, art

Table 2: A Subset of Top Prefixes / Suffixes

The set of candidate usernames is further expanded using these prefixes and suffixes in order to generate the final set of usernames. It is also worth mentioning that by using some Google search engine features (e.g. using the \* operator and searching `username*` on Google as the query) the prefix/suffix list can be further expanded.

### Username Filtering

Finally, now that we have the candidate usernames ready, in order to filter out usernames, we check for the existence of these usernames on the URLs that reside on the target community domain. Note that we are already sure ( $\mathcal{H}_1$ ) that there must exist URLs which contain these username. For each candidate username  $u$ , this procedure is performed by a web-search on Google with “`u site:Domain( $c_2$ )`”, where  $c_2$  is the target community. If the quantity of returned results were more than 0, then the username is considered valid. The accuracy of this procedure can be further improved by employing profile patterns (see Table 1) and hand-tuning.

## Experimental Results

The discussion of our experimental results is divided into two sub-sections. The first subsection reviews our in-depth analysis of the dataset we gathered and provides some interesting insight on the behavior of the community membership behavior of the users in the cyberspace. The second part however deals with our evaluation results of the proposed method as well as a comparative analysis of our method with a baseline technique.

### Dataset Analysis

As previously discussed, our dataset is gathered from the BlogCatalog community website. This dataset covers 38093 users in the BlogCatalog community. Furthermore, for each user, the communities a user belongs to and his/her username in that community is gathered. Overall, usernames from 36 web communities are found. To make sure the number of usernames where sufficient enough in each community, the first property analyzed within this dataset was the number of usernames gathered from each community. We kept every community whose usernames were at least  $400^3$  and removed the rest. We also removed the communities whose usernames where not simple strings. For

<sup>3</sup>This cut-off is experimentally decided based on the mean and standard deviation of the number of users.

	AIM	BUMPzee	del.icio.us	Digg	Flickr	Furl	Last.fm	Multiply	MyBlogLog	MySpace	Reddit	StumbleUpon	Technorati	Twitter	YouTube
AIM	1	0.17	0.54	0.6	0.39	0.12	0.24	0.1	0.61	0.58	0.16	0.58	0.69	0.54	0.52
BUMPzee	0.18	1	0.68	0.76	0.41	0.15	0.21	0.1	0.88	0.46	0.21	0.78	0.91	0.59	0.5
del.icio.us	0.14	0.17	1	0.73	0.37	0.13	0.16	0.09	0.71	0.39	0.2	0.72	0.86	0.56	0.45
Digg	0.14	0.17	0.65	1	0.31	0.11	0.14	0.08	0.69	0.37	0.19	0.73	0.83	0.55	0.43
Flickr	0.15	0.15	0.55	0.52	1	0.08	0.22	0.09	0.65	0.46	0.12	0.58	0.75	0.56	0.52
Furl	0.21	0.26	0.91	0.89	0.37	1	0.21	0.31	0.82	0.44	0.66	0.83	0.92	0.7	0.6
Last.fm	0.24	0.2	0.62	0.6	0.56	0.12	1	0.11	0.69	0.51	0.16	0.63	0.77	0.69	0.62
Multiply	0.17	0.16	0.57	0.56	0.39	0.27	0.18	1	0.8	0.4	0.28	0.56	0.77	0.58	0.64
MyBlogLog	0.12	0.16	0.51	0.56	0.32	0.08	0.13	0.09	1	0.35	0.13	0.59	0.81	0.5	0.4
MySpace	0.19	0.15	0.49	0.53	0.4	0.08	0.17	0.08	0.62	1	0.13	0.56	0.7	0.52	0.5
Reddit	0.18	0.22	0.83	0.89	0.33	0.39	0.18	0.18	0.74	0.43	1	0.85	0.89	0.71	0.53
StumbleUpon	0.13	0.16	0.59	0.68	0.32	0.1	0.14	0.07	0.67	0.36	0.17	1	0.82	0.55	0.4
Technorati	0.11	0.13	0.51	0.55	0.3	0.08	0.12	0.07	0.66	0.33	0.13	0.58	1	0.46	0.39
Twitter	0.14	0.14	0.54	0.6	0.36	0.1	0.18	0.09	0.67	0.39	0.17	0.65	0.75	1	0.44
YouTube	0.16	0.15	0.52	0.56	0.41	0.1	0.19	0.12	0.66	0.46	0.15	0.56	0.76	0.53	1

Table 3: Membership Rule Confidence

instance, we removed FaceBook since it has integer usernames and linkedIn which has some extra prefixes for usernames. The remaining 15 communities sorted based on the number of usernames were the following:  $\{Technorati, MyBlogLog, StumbleUpon, Digg, Twitter, Del.icio.us, YouTube, MySpace, Flickr, AIM, Last.fm, BUMPzee, Reddit, Multiply, Furl\}$ .

Next, we analyzed the dependency among user memberships in different communities. We first evaluated the confidence (Han and Kamber 2001) of the following rule:  $c_i \in C_u \Rightarrow c_j \in C_u$  for all users  $u$  and for all communities  $c_i$  and  $c_j$  who are members of the the top communities we just mentioned. Simply speaking, this measures the likelihood of a user being a member of community  $c_j$ , when we already know he is active in community  $c_i$ . Table 3 provides the detailed results. As shown in this table, for instance, it's highly probable to have a username in *Technorati*, if you have usernames in other communities. Since the evaluated rule was too naive, we also analyzed the memberships using Apriori<sup>4</sup> Sequential Rule Mining algorithm (Agrawal and Srikant 1995) and found out some interesting results. Table 4 provides the top 10 rules found in our dataset.

$Digg \in C_u, \Rightarrow StumbleUpon \in C_u, Del.icio.us \in C_u$
$StumbleUpon \in C_u, Del.icio.us \in C_u \Rightarrow Digg \in C_u$
$StumbleUpon \in C_u, Digg \in C_u \Rightarrow Del.icio.us \in C_u$
$Del.icio.us \in C_u, \Rightarrow StumbleUpon \in C_u, Digg \in C_u$
$MyBlogLog \in C_u, Digg \in C_u \Rightarrow Del.icio.us \in C_u$
$Del.icio.us \in C_u \Rightarrow MyBlogLog \in C_u, Digg \in C_u$
$Digg \in C_u \Rightarrow MyBlogLog \in C_u, Del.icio.us \in C_u$
$MyBlogLog \in C_u, Del.icio.us \in C_u \Rightarrow Digg \in C_u$
$StumbleUpon \in C_u \Rightarrow Twitter \in C_u, Digg \in C_u$
$Twitter \in C_u, Digg \in C_u \Rightarrow StumbleUpon \in C_u$

Table 4: Community Membership Rules

Next, we will overview our results regarding the evaluation of our proposed technique.

<sup>4</sup>As implemented in the WEKA (Witten and Frank 2005) software package. *Lift* was used instead of confidence as the sequential mining metric. Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support.

Community	Accuracy	Correctly Classified	# Users
AIM	0.36	290	803
BUMPzee	0.78	591	757
Del.icio.us	0.68	2082	3044
Digg	0.67	2315	3445
Flickr	0.61	1127	1860
Furl	0.74	324	435
Last.fm	0.62	495	800
Multiply	0.55	266	485
MyBlogLog	0.66	2788	4216
MySpace	0.43	983	2289
Reddit	0.76	564	740
StumbleUpon	0.67	2470	3684
Technorati	0.72	3679	5134
Twitter	0.67	2113	3158
YouTube	0.54	1386	2576

Table 5: Corresponding Target Username Identification Accuracy Using Proposed Method

## Evaluation Results

In order to analyze the competitiveness of the designed method, we performed a complete analysis on different communities. As mentioned before, 15 different well known communities were selected. For each community, a set of *username-username* pairs were selected, for which the base username was in the BlogCatalog community and the target one was in the community. The proposed method was employed in order to extract the set of possible usernames in the target community. The inclusion of the target username in this set is checked and the overall accuracy was recorded. Table 5 presents the detailed accuracy results for each of these communities. The results show that if the base username is from the BlogCatalog community, on average our method has a 63% accuracy, and in the best case, can be up to 78% accurate.

Note that in order to interpret these results more accurately, we have also compared our method with a conservative method which assumes that the target username is equal to the base username. As analyzed before, in hypothesis  $\mathcal{H}_6$ , this should provide reasonably accurate predictions. The detailed comparison between these two methods over all communities is provided in Table 6. As you can see in these tables, both methods perform quite well. Moreover, it seems that there is not that much correlation between the

	AIM	BUMPzee	Del.icio.us	Digg	Flickr	Furl	Last.fm	Multiply	MyBlogLog	MySpace	Reddit	StumbleUpon	Technorati	Twitter	YouTube
AIM	1	0.61	0.58	0.6	0.34	0.7	0.58	0.64	0.56	0.38	0.69	0.64	0.5	0.54	0.56
BUMPzee	0.53	1	0.76	0.74	0.6	0.7	<b>0.88</b>	0.62	0.68	0.52	0.84	0.67	0.72	0.76	0.58
Del.icio.us	0.6	0.84	1	0.68	<b>0.66</b>	0.84	0.76	0.62	<b>0.73</b>	0.47	0.9	0.72	0.78	0.76	0.58
Digg	0.62	0.72	0.7	1	0.57	0.78	0.82	0.54	0.63	0.4	0.84	0.62	0.68	0.64	0.54
Flickr	0.4	0.71	0.66	0.64	1	0.66	0.71	0.45	0.51	<b>0.58</b>	0.56	0.63	0.59	0.65	0.6
Furl	<b>0.7</b>	0.78	0.78	0.76	0.63	1	<b>0.88</b>	0.74	<b>0.73</b>	0.45	<b>0.92</b>	0.78	<b>0.82</b>	0.76	0.6
Last.fm	0.48	<b>0.88</b>	0.74	0.78	0.6	0.82	1	0.64	0.64	0.53	0.72	0.72	0.72	0.64	0.54
Multiply	0.6	0.76	0.66	0.58	0.45	0.72	0.66	1	0.52	0.43	0.8	0.58	0.48	0.53	0.42
MyBlogLog	0.56	0.72	0.71	0.67	0.47	0.63	0.66	0.46	1	0.35	0.71	0.6	0.67	0.67	0.47
MySpace	0.4	0.65	0.57	0.56	0.54	0.61	0.57	0.49	0.56	1	0.57	0.52	0.53	0.53	0.58
Reddit	0.63	0.86	<b>0.84</b>	<b>0.8</b>	0.54	<b>0.86</b>	0.68	<b>0.78</b>	0.67	0.43	1	<b>0.8</b>	0.76	<b>0.77</b>	<b>0.62</b>
StumbleUpon	0.66	0.67	0.74	0.68	0.5	0.78	0.68	0.6	0.62	0.38	0.86	1	0.66	0.6	0.58
Technorati	0.46	0.76	0.74	0.66	0.5	0.8	0.72	0.48	0.65	0.4	0.78	0.64	1	0.66	0.58
Twitter	0.56	0.8	0.64	0.64	0.53	0.68	0.7	0.53	0.65	0.33	0.81	0.58	0.62	1	0.52
YouTube	0.6	0.76	0.58	0.6	0.58	0.6	0.68	0.52	0.55	0.56	0.67	0.6	0.68	0.62	1

Table 7: Corresponding Target Username Identification Accuracy Using Proposed Method

	AIM	BUMPzee	Del.icio.us	Digg	Flickr	Furl	Last.fm	Multiply	MyBlogLog	MySpace	Reddit	StumbleUpon	Technorati	Twitter	YouTube
AIM	1	0.45	0.48	0.48	0.28	0.62	0.4	0.54	0.38	0.32	0.57	0.52	0.38	0.44	0.52
BUMPzee	0.45	1	0.66	0.68	0.49	0.6	0.78	0.58	0.58	0.44	0.73	0.61	0.64	0.65	0.5
Del.icio.us	0.48	0.66	1	0.6	0.52	0.7	0.64	0.5	0.61	0.41	0.82	0.68	0.7	0.6	0.5
Digg	0.48	0.68	0.6	1	0.45	0.7	0.72	0.48	0.57	0.31	0.76	0.6	0.54	0.6	0.48
Flickr	0.28	0.49	0.52	0.45	1	0.53	0.47	0.36	0.4	0.42	0.46	0.41	0.41	0.42	0.49
Furl	0.62	0.6	0.7	0.7	0.53	1	0.69	0.68	0.57	0.41	0.84	0.72	0.76	0.64	0.5
Last.fm	0.4	0.78	0.64	0.72	0.47	0.69	1	0.54	0.52	0.39	0.62	0.56	0.64	0.52	0.42
Multiply	0.54	0.58	0.5	0.48	0.36	0.68	0.54	1	0.38	0.37	0.73	0.48	0.42	0.45	0.36
MyBlogLog	0.38	0.58	0.61	0.57	0.4	0.57	0.52	0.38	1	0.19	0.58	0.48	0.55	0.51	0.35
MySpace	0.32	0.44	0.41	0.31	0.42	0.41	0.39	0.37	0.19	1	0.43	0.27	0.32	0.27	0.44
Reddit	0.57	0.73	0.82	0.76	0.46	0.84	0.62	0.73	0.58	0.43	1	0.73	0.71	0.71	0.48
StumbleUpon	0.52	0.61	0.68	0.6	0.41	0.72	0.56	0.48	0.48	0.27	0.73	1	0.58	0.5	0.5
Technorati	0.38	0.64	0.7	0.54	0.41	0.76	0.64	0.42	0.55	0.32	0.71	0.58	1	0.5	0.54
Twitter	0.44	0.65	0.6	0.6	0.42	0.64	0.52	0.45	0.51	0.27	0.71	0.5	0.5	1	0.42
YouTube	0.52	0.5	0.5	0.48	0.49	0.5	0.42	0.36	0.35	0.44	0.48	0.5	0.54	0.42	1

Table 8: Corresponding Target Username Identification Accuracy Using Conservative Method

Community	Accuracy	Correctly Classified	# Users
AIM	0.32	253	803
BUMPzee	0.74	558	757
Del.icio.us	0.62	1892	3044
Digg	0.61	2118	3445
Flickr	0.49	912	1860
Furl	0.69	298	435
Last.fm	0.52	418	800
Multiply	0.52	250	485
MyBlogLog	0.6	2535	4216
MySpace	0.35	811	2289
Reddit	0.7	516	740
StumbleUpon	0.61	2245	3684
Technorati	0.65	3355	5134
Twitter	0.59	1852	3158
YouTube	0.45	1161	2576

Table 6: Corresponding Target Username Identification Accuracy Using Conservative Method

usernames that are chosen for instant messaging communities (*AIM*) and the other usernames. Another interesting phenomenon is that users in the *MySpace* community tend to have little or no correlation with the others. This might have been a direct result of the number of users *MySpace* community has and the less likelihood a chosen username has not been already taken.

As already mentioned, in all these experiments, the base username was selected from the *BlogCatalog* community. Therefore, we decided to perform the same experiment with the base usernames from different communities. This allows us to analyze the fluctuations in the accuracy acquired

depending on the base community. Tables 7 and 8 present the detailed accuracy results for both the proposed and conservative methods. On average, our method has predicted the correct target-username in more than 66% of the cases and is up to 92% accurate in the best case scenario. Moreover, our conservative method is also capable of predicting the target username with 56% accuracy. Note that as highlighted in Table 7 certain communities have the tendency to be more useful in predicting the final user, i.e. *Furl*, *Reddit*, *Del.icio.us*, etc.

Finally, there is a need to compare our proposed methods to other methods as well. However, to best of our knowledge, no such study has ever been carried out and this work is the first of its kind in this area. Therefore, we first devised a baseline method for this comparison. For the baseline method, we assume that for predicting a target username the method selects a username from the set of all other usernames ( $\Lambda_u$ ) randomly, i.e. uniform distribution. The accuracy values for this method with regards to predicting the target username for different target communities are presented in Table 9.

we also compared this baseline value to the accuracies of both our proposed and conservative method in Table 10. In this table, in each cell, the first value is the accuracy improvement using the conservative method and the second value is the improvement using the proposed method.

## Conclusion and Future Work

In this paper, we have empirically studied the possibility of identifying corresponding identities across various communities on the web. Our linkage gadget has been usernames

AIM	BUMPzee	Del.icio.us	Digg	Flickr	Furl	Last.fm	Multiply	MyBlogLog	MySpace	Reddit	StumbleUpon	Technorati	Twitter	YouTube
0.23	0.16	0.19	0.2	0.23	0.12	0.15	0.23	0.24	0.24	0.14	0.21	0.26	0.22	0.23

Table 9: Baseline Method Performance

	AIM	BUMPzee	Del.icio.us	Digg	Flickr	Furl	Last.fm	Multiply	MyBlogLog	MySpace	Reddit	StumbleUpon	Technorati	Twitter	YouTube
AIM	.77/.77	.29/.45	.3/.4	.28/.4	.07/.13	.49/.57	.22/.4	.33/.43	.14/.32	.09/.15	.44/.56	.31/.43	.12/.24	.22/.32	.29/.33
BUMPzee	.22/.3	.84/.84	.48/.58	.48/.54	.28/.39	.47/.57	.6/.7	.37/.41	.34/.44	.21/.29	.6/.71	.4/.46	.38/.46	.43/.54	.27/.35
Del.icio.us	.25/.37	.5/.68	.82/.82	.4/.48	.31/.45	.57/.71	.46/.58	.29/.41	.37/.49	.18/.24	.69/.77	.47/.51	.44/.52	.38/.54	.27/.35
Digg	.25/.39	.52/.56	.42/.52	.8/.8	.24/.36	.57/.65	.54/.64	.27/.33	.33/.39	.08/.17	.63/.71	.39/.41	.28/.42	.38/.42	.25/.31
Flickr	.05/.17	.33/.55	.34/.48	.25/.44	.79/.79	.4/.53	.29/.53	.15/.24	.16/.27	.19/.35	.33/.43	.2/.42	.15/.33	.2/.43	.26/.37
Furl	.39/.47	.44/.62	.52/.6	.5/.56	.32/.42	.87/.87	.51/.7	.47/.53	.33/.49	.18/.22	.71/.79	.51/.57	.5/.56	.42/.54	.27/.37
Last.fm	.17/.25	.62/.72	.46/.56	.52/.58	.26/.39	.56/.69	.82/.82	.33/.43	.28/.4	.16/.3	.49/.59	.35/.43	.38/.46	.3/.42	.19/.31
Multiply	.31/.37	.42/.6	.32/.48	.28/.38	.15/.24	.55/.59	.36/.48	.79/.79	.14/.28	.14/.2	.6/.67	.27/.37	.16/.22	.23/.31	.13/.19
MyBlogLog	.15/.33	.42/.56	.43/.53	.37/.47	.19/.26	.44/.5	.34/.48	.17/.25	.76/.76	-.04/.12	.45/.58	.27/.39	.29/.41	.29/.45	.12/.24
MySpace	.09/.17	.28/.49	.23/.39	.11/.36	.21/.33	.28/.48	.21/.39	.16/.28	-.05/.32	.77/.77	.3/.44	.06/.31	.06/.27	.05/.31	.21/.35
Reddit	.34/.4	.57/.7	.64/.66	.56/.6	.25/.33	.71/.73	.44/.5	.52/.57	.34/.43	.2/.2	.87/.87	.52/.59	.45/.5	.49/.55	.25/.4
StumbleUpon	.29/.43	.45/.51	.5/.56	.4/.48	.2/.29	.59/.65	.38/.5	.27/.39	.24/.38	.04/.14	.6/.73	.79/.79	.32/.4	.28/.38	.27/.35
Technorati	.15/.23	.48/.6	.52/.56	.34/.46	.2/.29	.63/.67	.46/.54	.21/.27	.31/.41	.09/.17	.58/.65	.37/.43	.74/.74	.28/.44	.31/.35
Twitter	.21/.33	.49/.64	.42/.46	.4/.44	.21/.32	.51/.55	.34/.52	.24/.32	.27/.41	.04/.1	.58/.68	.29/.37	.24/.36	.78/.78	.19/.29
YouTube	.29/.37	.34/.6	.32/.4	.28/.4	.28/.37	.37/.47	.24/.5	.15/.31	.11/.31	.21/.33	.35/.54	.29/.39	.28/.42	.2/.4	.77/.77

Table 10: Improvement over Baseline Method

and several hypotheses were proposed with regard to the possibility of use of usernames as a linkage gadget and were evaluated based on the data collected from BlogCatalog. Based on these evaluations, it turns out that usernames can be used quite successfully to identify corresponding usernames in various communities. We have also proposed a method to identify corresponding usernames in various communities. The method has been successfully evaluated over 15 different communities and thousands of usernames with the average accuracy of around 66%.

For the future, we aim to deal with the many challenges that we faced during the course of this research. For instance, when dealing with usernames, there are many cases where same usernames does not necessarily guarantee the same identity. For instance, while a username such as *pat&mat1989prague* might represent the same identity, but common usernames such as *john.smith* can be employed by different identities in various communities and do not necessarily represent a unique individual. A simple solution for verifying the identity-uniqueness property for a given username is to check the probability that username can be generated at random or whether it's a dictionary word or not.

Finally, we focused on the demonstration that (1) more often than not, there exists a corresponding-username mapping, and (2) one can devise effective techniques to identify corresponding users across communities. It is evident that the accuracy of these identification methods can be further improved.

## References

[Agarwal et al. 2008] Agarwal, N.; Liu, H.; Salerno, J.; and Sundarajan, S. 2008. Understanding Group Interaction in Blogosphere: A Case Study. In *Proceedings of the Second International Conference on Computational Cultural Dynamics*, 9–17. AAAI.

[Agrawal and Srikant 1995] Agrawal, R., and Srikant, R. 1995. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, 3–14.

[Anagnostopoulos, Kumar, and Mahdian 2008] Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 7–15. New York, NY, USA: ACM.

[Brin and Page 1998] Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1-7):107–117.

[Brinkmeier, Werner, and Recknagel 2007] Brinkmeier, M.; Werner, J.; and Recknagel, S. 2007. Communities in graphs and hypergraphs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 869–872. ACM New York, NY, USA.

[Flake, Lawrence, and Giles 2000] Flake, G.; Lawrence, S.; and Giles, C. 2000. Efficient identification of Web communities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 150–160. ACM New York, NY, USA.

[Gusfield 1997] Gusfield, D. 1997. Algorithms on Stings, Trees, and Sequences: Computer Science and Computational Biology. *ACM SIGACT News* 28(4):41–60.

[Han and Kamber 2001] Han, J., and Kamber, M. 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

[Hu and Liu 2004] Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM New York, NY, USA.

[Wasserman and Faust 1994] Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

[Witten and Frank 2005] Witten, I., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.