Matching User Profiles Across Social Networks

Nacéra Bennacer[1], Coriane Nana Jipmo[1],
Antonio Penta[2], and Gianluca Quercini[1]

[1] Supélec E3S,
3, rue Joliot-Curie, 91190 Gif-sur-Yvette (France)
{nacera.bennacer,coriane.nanajipmo,gianluca.quercini}@supelec.fr
[2] Università di Torino
Corso Svizzera 185, Torino (Italy)
penta@di.unito.it

Abstract. *Social Networking Sites*, such as Facebook and LinkedIn, are clear examples of the impact that the Web 2.0 has on people around the world, because they target an aspect of life that is extremely important to anyone: social relationships. The key to building a social network is the ability of finding people that we know in real life, which, in turn, requires those people to make publicly available some personal information, such as their names, family names, locations and birth dates, just to name a few. However, it is not uncommon that individuals create multiple profiles in several social networks, each containing partially overlapping sets of personal information. Matching those different profiles allows to create a global profile that gives a holistic view of the information of an individual. In this paper, we present an algorithm that uses the network topology and the publicly available personal information to iteratively match profiles across n social networks, based on those individuals who disclose the links to their multiple profiles. The evaluation results, obtained on a real dataset composed of around 2 million profiles, show that our algorithm achieves a high accuracy.

1 Introduction

A social network is a set of individuals and their relationships. In a broader sense, the term social network also refers to a website, such as Facebook and LinkedIn, which enables individuals to create a personal page, or *profile*, and to stay in contact with their acquaintances. The key to building a social network is the ability of finding people that we know in real life, which in turn requires those people to make publicly available on their profiles some personal information, such as their names, family names, locations and birth dates, just to name a few. Several surveys showed that Social Networking Services (SNSs) users tend to share many of their personal data, including sensitive information, such as home addresses and phone numbers [1–3].

However, it is not uncommon that an individual creates multiple profiles in different SNSs, each disclosing sets of personal information that are unlikely to be identical, though they might overlap. Indeed, profile information might not be

M. Jarke et al. (Eds.): CAiSE 2014, LNCS 8484, pp. 424–438, 2014.

updated regularly and are not necessarily created at the same time. Moreover, the differences between two profiles of an individual might reflect the fact that they are created in SNSs that target different aspects of the individual life. For instance, information on the career of individuals are more likely to be found on their LinkedIn profiles than Facebook's, as LinkedIn is mainly used for professional networking. As a result, finding a person based on a limited knowledge of her personal information might require several manual searches across social networks, which is obviously annoying and time-consuming. It would be useful to create a global profile that provides a holistic view of the personal information of an individual by automatically integrating all her profiles. This calls for efficient methods for automatically determining the profiles that an individual owns across different SNSs, which is the focus of our paper. We say that two profiles *match* if they are owned by the same individual.

In this paper, we present an algorithm that matches profiles across n distinct social networks by using the network topology and the personal information that are publicly available in the profiles. The algorithm first selects a candidate set of profile pairs that are likely to match; this selection exploits the fact that some links between profiles referring to the same individual already exist, as explicitly disclosed in the profiles themselves. Next, the algorithm applies a set of rules that compare the values of the *profile attributes* (such as names, family names, usernames) to determine the pairs that match. Finally, the algorithm uses the newly found matches to retrieve more candidates and further determine other matches in an iterative way. The key contributions of our paper are the following:

- We define sets of rules that use the values of a limited set of attributes to determine whether two profiles match. Unlike the existing rule-based approaches (i) we consider that all attributes are equally important, which relieves us from assigning each attribute an empirical and, inevitably, arbitrary weight and (ii) we study the combined contribution of the different attributes, when used in the same rule.
- Our algorithm matches new profiles in a iterative way, which means that the new found matches are used to discover new matches. Moreover, the discovered matches are propagated by transitive closure across all considered social networks. To the best of our knowledge, no existing method is iterative in this sense.
- We evaluate our algorithm on four real social networks, namely Flickr, LiveJournal, Twitter and YouTube, which combined form a graph composed of around 2 million nodes and more than 17 million links. On this dataset, our algorithm achieves a precision of 94%. No existing approach is evaluated on such a big dataset.

The remainder of the paper is organized as follows. We survey the research work that is related to ours in Section 2 and we introduce basic concepts and notation in Section 3. Section 4 is the central part of the paper, in which we detail our algorithm, which is then thoroughly evaluated in Section 5. Section 6 concludes the presentation.

2 Related Work

Numerous solutions have been proposed to the problem that we study in this paper. Interestingly, two of them focus only on the username of an individual as a way to match different profiles, based on the observation that individuals tend to use the same or a similar username across distinct social networks [4, 5]. Although in our evaluation we confirm this observation, we also consider other attributes, in order to match profiles of individuals who choose to use unrelated usernames.

The use of the attributes to match profiles across distinct social networks has been largely investigated [6–12]. Two approaches describe each pair of profiles as a vector of scores, which represent the similarity between the values of the attributes, and use machine learning techniques to determine whether they match [9, 10]. While the results are promising, both approaches need a training set, which is not easy to determine. In fact, a careful analysis of the available data is necessary to create a training set that is representative of all possible situations where profile pairs match or not. Moreover, a model trained on a given pair of social networks might not be generalizable to other networks, which implies that a training set should be created for each network pair. Some social networks allow the exportation of profiles that are described with the Friend of a Friend ontology (FOAF); the advantage is that standard Semantic Web techniques, such as OWL reasoning, can be used to match profiles [8, 12]. However, these techniques are applied to a limited set of attributes, and in particular to those, such as the email, that are likely to identify uniquely an individual. Similarly to us, Carmagnola et al. determine the profile attributes that are more likely to identify uniquely an individual, by assigning them an *importance factor* [6]. The importance factor is used to weight the similarity score that is computed between two profiles that have similar attributes. Our approach goes a step further and uses the pairs of profiles that are found to match to iteratively discover new matches. Moreover, our evaluation is based on a real large social internetwork, while theirs uses different closed user-adaptive systems. The key difference is that in Web social networks often individuals are reluctant to disclose their real identities, while in closed user-adaptive systems they feel that their privacy is less threatened; as a result, data in social networks are likely to be erroneous and messy, which constitutes a real challenge. Some researchers also propose the computation of semantic similarity between profile attributes [7, 11]. Although these approaches are original, they provide little (50 user profiles [11]) or no evaluation.

Some authors proposed to go beyond the profile attributes and investigated the possibility of using the network properties [13–16]. The approach proposed by Buccafurri et al. considers that two nodes are similar, and therefore likely to refer to the same individual, if they have similar usernames and the nodes to which they are connected are recursively similar [14]. This approach presents two major drawbacks. First, profiles associated with dissimilar nicknames are ignored and discarded with no further analysis, although they might very well refer to the same person; second, the discovered associations between profiles are not used to

re-iterate the algorithm and discover new associations. Our approach overcomes these two limitations. Besides considering the network structure, Jain et al. also propose to use of the content that an individual publishes in the form of short texts [15]. This approach has the merit of exploring the use of the content and the shared connections to match profiles. However, the experiments reveal that this information is not very effective alone, as only 4 out of 543 profiles are matched correctly. We found an elegant approach that combines profile attributes and network by using conditional random fields [13]. The key advantage is that it is robust to the absence of profile and/or network information and therefore can also be applied to cases where no profile information is available except the network, although with a significant drop in recall. The disadvantage is that the proposed model needs training data, which, as recalled before, might not be easy to find. Finally, Narayanan et al. consider the case of anonymized networks where little or no profile attributes are available and only the network structure can be exploited [16]. They propose a method that first selects a small set of seed profiles in both networks that are highly likely to belong to the same individual. Then, new matchings are propagated iteratively by using the seed. This is similar in spirit to our approach. However, since they only use the network structure the accuracy of their approach is quite low compared to ours.

Finally, social aggregators, such as *FriendFeed* [17] or *Plaxo* [18], provide a platform for people to manage their own profiles but they make no attempt at automatically discovering profiles linked to an individual across social networks. *Spokeo* [19] seems to be quite accurate in finding personal information from different sources (not necessarily social networks), but it shows its limits when it comes to aggregating them. To the best of our knowledge, there is no existing tool that is able to automatically match profiles across social networks.

3 Background

We define a *social internetwork* as a collection of n distinct social networks and we model it as a directed graph. Its nodes correspond to the profiles of the individuals or, with an abuse of language, to the individuals themselves. A *profile* consists of a set of attributes (such as username, name, email address), which are usually described in a Web page created by an individual, and a *uri*, identifying that page on the Web. A link in a social internetwork connects either two profiles within the same social network, in which case we call it a *friendship link*, or two profiles that refer to the same individual in two different social networks, and we call it a *cross-link*.

Formally, a social internetwork with n social networks is a directed labelled graph defined as follows:

$$\mathcal{G} = < \bigcup_{i=1}^{n} V_i \ , \bigcup_{i=1}^{n} E_i \ , \bigcup_{1,i \neq j}^{n} E_{i,j} >$$

where:

- V_i is the node set of the social network i. Since the social networks are distinct, $V_i \cap V_j = \emptyset, \forall i, j,\ i \neq j$. Each node v_i is the profile of an individual in the social network i. A is the set of the attributes defined in a profile, while $P_a(v_i)$ denotes the value(s) of the attribute $a \in A$ in the profile v_i.
- E_i is the set of friendship links, which are identified by the label $friend$. Each link $(v_i^1, friend, v_i^2)$ represents a friendship link from the individual v_i^1 to the individual v_i^2 within the social network i.
- $E_{i,j}$ is the set of cross-links, which are identified by the label me. Each link (v_i, me, v_j) represents a cross-link between two profiles v_i and v_j owned by the same individual in the social networks i and j, $i \neq j$. By definition, this type of link is symmetrical and transitive. For instance, *Bob* might indicate in his Flickr profile, represented by the node v_f, the uri of the page of his profile LiveJournal, represented by the node v_l, and in this page he declares the uri of the page of his profile Twitter, represented by the node v_t. In this case, $E_{f,l} = \{(v_f, me, v_l), (v_l, me, v_f)\}$, $E_{t,l} = \{(v_t, me, v_l), (v_l, me, v_t)\}$ and $E_{t,f} = \{(v_t, me, v_f), (v_f, me, v_t)\}$.

The problem of matching the profiles that are owned by the same individual across social networks is the problem of discovering the missing cross-links in a social internetwork and is formalized as follows:

Input: $\mathcal{G} = < \bigcup_{i=1}^{n} V_i\ ,\ \bigcup_{i=1}^{n} E_i\ ,\ \bigcup_{1,i\neq j}^{n} E_{i,j} >$

Output: $\mathcal{G}' = < \bigcup_{i=1}^{n} V_i\ ,\ \bigcup_{i=1}^{n} E_i\ ,\ \bigcup_{1,i\neq j}^{n} E_{i,j}\ \bigcup_{1,i\neq j}^{n} D_{i,j} >$ with

$D_{i,j} = \{(v_i, me, v_j) | v_i \in V_i, v_j \in V_j, 1 \leq i \neq j \leq n, (v_i, me, v_j) \notin E_{i,j}\}$

$D_{i,j}$ is the set of the discovered cross-links between the social networks i and j.

4 Our Approach

A first intuitive solution to our problem is to compare each pair of profiles (v_i, v_j), which are not connected via a cross-link, for each pair of networks (i, j), $1 \leq i \neq j \leq n$. However, this amounts to analyse $\sum_{i,j} |V_i| \times |V_j| - |E_{i,j}|$ pairs of nodes, which is not feasible when social networks are large, as it is usual the case. Based on this observation, our matching approach goes through two steps:

- **Candidate selection.** A subset of profile pairs is selected which are likely to represent the same individual and are therefore candidate profiles for the matching approach. The candidates are identified based solely on the topology of the graph.
- **Cross-links determination.** The pairs of profiles that are deemed to correspond to the same individual are identified among the selected candidates. The determination is based on a set of rules which compare the attribute values of the candidate pairs.

The two steps are iterated until no new cross-links can be determined. The remainder of this section describes both steps in greater detail.

4.1 Candidate Selection

The selection of the candidates is based on the observation that a small percentage of individuals own multiple accounts, but they tend to be connected with friends who also have multiple profiles; moreover, when two friends have both multiple profiles, they are frequently friends in multiple networks [8]. Therefore, we consider as candidates the friends of the same individual across different social networks. More specifically, if $\exists(v_i, me, v_j) \in E_{i,j},\ v'_i \in friend(v_i), v'_j \in friend(v_j)$ then (v'_i, v'_j) is a candidate, where :

$$friend(v_i) = \{v'_i | (v'_i, friend, v_i) \vee (v_i, friend, v'_i) \in E_i\}$$

represents the set of the profiles of the friends of an individual v_i. The set of candidates, denoted $C_{i,j}$, for the social networks i and j is formally defined as follows:

$$\{(v'_i, v'_j) | \exists v'_i \in friend(v_i) \wedge v'_j \in friend(v_j) \wedge (v_j, me, v_i) \in E_{i,j} \wedge (v'_i, me, v'_j) \notin E_{i,j}\}$$

4.2 Cross-links Determination

Once the candidate set $C_{i,j}$ is created for a social network pair (i, j), we need a method to determine if a pair of candidate profiles (v_i, v_j) represents the same individual. More precisely, we need to determine the set of cross-links $D_{i,j},\ \forall i, j,\ i \neq j$. We introduce the attributes that we use in our approach and then we detail the rules that allow the determination of the new cross-links, as well as the algorithm that we defined.

The Attributes. In all major social networks the values of some attributes are publicly accessible as per default privacy policy and/or left accessible by the individuals. It is therefore natural to analyse these data to establish new cross-links between i and j.

Based on the observations by Krishnamurthy et al., who identified a set of attributes that are generally publicly available in 12 of the most important social networks [20], we focus our attention on the following: *username*, *name* (which includes first name and last name), *email*, and *links* to other Web pages.

Username. Denoted as u, the username is always publicly accessible, as it is the only way to uniquely identify an individual within a social network, and is generally a part of the URL of the web page that hosts the profile. Studies have shown that individuals tend to use the same username, or a similar one, when registering different profiles [4, 5]. In order to determine the similarity of two usernames, we chose the *Levenshtein* distance d_{lev}, which is the minimum number of single character edits (insertion, deletion and substitution), as several studies have revealed that is quite effective in capturing the variations in the usernames chosen by the individuals [4, 5, 14]. The similarity of two usernames

u_1 and u_2 is computed as $1 - \frac{d_{lev}(u_1,u_2)}{max(l(u_1),l(u_2))}$, where $l(u_i)$ is number of characters of $u_i, i = 1, 2$. The Levenshtein distance between the username *cospics* of the Flickr profile at `www.flickr.com/photos/cospics` and the username *cos* of the LiveJournal profile at `www.livejournal.com/users/cos/profile` is 4, because we need to suppress the substring "pics", composed of four characters, to obtain the second username from the first. As a result, their similarity is 0.43. To determine whether two usernames are similar or not, we empirically define a threshold θ_u, whose value is discussed in Section 5.

Name. Denoted as n, the first and family name are also present in most of the networks we came across, but their values cannot be trusted as much as the usernames. Indeed, in some social networks, such as LiveJournal, the profile of a person is almost entirely public and consequently individuals do not feel confident in revealing their real names. Moreover, names are often ambiguous, and do not generally identify uniquely an individual. As a result, we do not expect the name of an individual to reveal many profile matches, if not in combination with other attributes. The similarity of two names n_1 and n_2 is computed with the *Jaccard* similarity measure as $\frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}$, where N_1 and N_2 are the sets of the words that compose n_1 and n_2 respectively. For example, if n_1 is "Barack Obama" and n_2 is "Barack Hussein Obama", then $N_1 = \{Barack, Obama\}$, $N_2 = \{Barack, Hussein, Obama\}$ and their similarity is $\frac{2}{3}$. The reason why we select the Jaccard measure instead of the Levenshtein distance is that generally social networks do not force their individuals to specify their first names before the last names. Moreover, some individuals might specify their middle names in a profile, while omitting them in another. Therefore, a comparison between "Barack Obama" and "Obama Barack" would give a Levenshtein distance of 10, although the two strings are equivalent, while Jaccard gives score 1. Similarly to the username, we define a threshold θ_n to determine whether two names are similar.

Email. Denoted as e, *email* is a multi-valued attribute whose values correspond to the different email addresses disclosed by an individual. The email address is a very sensitive attribute, because it could identify uniquely a person. If two profiles are associated with the same email address, there are high chances that the two profiles refer to the same individual. It is certainly possible that two individuals share the same email address, as in the case of people that work within the same organization. But these are particular cases, and in general email addresses can be trusted. The only problem is that only a small percentage of people grant public access to their email addresses. In order to compare the values of the attribute e of two profiles, we need to determine whether one of the email addresses of a profile is identical to one of the email address of the other profile. In this case, the similarity score is 1, otherwise it is 0.

Links to other Web pages. This is a multi-valued attribute whose different values correspond to different URLs of Web pages. We distinguish between two types of links, those to Web pages that describe online profiles in social networks (denoted

as s), and those to other Web pages (denoted as w). We aim at investigating the contribution of the two attributes separately. Indeed, the former are likely to indicate the links to the different social network profiles that an individual owns, which might indicate a cross-link in our graph. On the other side, the links to other Web pages are links to resources that an individual wants to share and does not necessarily identify the individual. The similarity score for s (respectively, w) for a profile pair is 1 if one of the values of s (resp., w) of one profile is identical to one of the values of s (resp., w) of the other profile. In this paper, we limit ourselves to determine whether the values of the attribute s or w for two profiles have at least a URL in common without analysing the content of these pages.

Another attribute that is worth considering is the location, whose values are often publicly accessible in different social networks. However, the location name poses some challenges, such as their ambiguity, which fall out of the scope of this paper. For this reason, we leave the use of this attribute for future work.

The Rules. In order to determine whether two profiles v_i et v_j refer to the same individual, we defined a set of rules based on the attributes introduced above. Each rule considers the contribution of one or several attributes. We assume that the higher the number of attributes that match, based on the defined similarity measures, the higher the probability for two profiles to refer to the same individual. We therefore define the *order* k of a rule as the number of attributes that the rule uses. The rule with the highest confidence is the one that uses all the attributes ($k = |A|$). The rules with the lowest confidence are those that use just one attribute ($k = 1$).

Let $match(P_a(v_i), P_a(v_j))$ be the predicate which is true when the values of the attribute a match for the profiles v_i and v_j, based on the similarity measure defined for the attribute a. A rule with the order k, or $k-$rule, $\mathcal{R}^k$ is defined as follows:

$$\mathcal{R}^k(v_i, v_j) = \begin{cases} \bigwedge_{a \in A} match(P_a(v_i), P_a(v_j)) & \text{if } k = |A| \\ \bigvee_{B \in [A]^k} \bigwedge_{a \in B} match(P_a(v_i), P_a(v_j)) & \text{if } 1 \leq k < |A| \end{cases}$$

where $[A]^k$ is the set of all subsets of A with k elements.

Therefore, if $\bigvee_{1 \leq k \leq |A|} \mathcal{R}^k(v_i, v_j)$ is true, then v_i and v_j are considered to refer to the same individual. If for one pair of candidate profiles (v_i, v_j) at least one rule $\mathcal{R}^k(v_i, v_j)$ is *true*, then no rule with order $l < k$ is applied. In the worst case, for (v_i, v_j) no rule is true, in which case the two profiles are considered to refer to two distinct individuals. When there is a rule $\mathcal{R}^k(v_i, v_j)$ which is *true*, the pair (v_i, v_j) is added to $D_{i,j}$, meaning that a new cross-link is discovered.

4.3 The Algorithm

We here detail the procedure that adds missing cross-links to a social internetwork $\mathcal{G}$ with n distinct social networks (cf. Algorithm 1). For each pair of social networks i and j in $\mathcal{G}$, the set $C_{i,j}$ of candidate profile pairs is computed

(Line 3), as described in Section 4.1. Next, the $k-$rules are applied to each candidate (v_i, v_j) by decreasing order, starting with $k = |A|$, until either one is true or none applies (Lines 5 through 7). If one rule is verified, a cross-link is added between v_i and v_j (Line 8).

Algorithm 1. The algorithm to match profiles across n social networks

Data: $\mathcal{G} =< \bigcup_{i=1}^{n} V_i \ , \bigcup_{i=1}^{n} E_i \ , \bigcup_{1,i\neq j}^{n} E_{i,j} >$

Result: $\mathcal{G}' =< \bigcup_{i=1}^{n} V_i \ , \bigcup_{i=1}^{n} E_i \ , \bigcup_{1,i\neq j}^{n} E_{i,j} \ \bigcup_{1,i\neq j}^{n} D_{i,j} >$

1 **foreach** *social network pair* (i, j) **do** $D_{i,j} \leftarrow \emptyset$;
2 **foreach** *social network pair* (i, j) **do**
3 $C_{i,j} \leftarrow candidateSelection(\mathcal{G}, i, j)$ $newCl \leftarrow false$;
4 **while** $C_{i,j} \neq \emptyset$ **do**
5 **foreach** $(v_i, v_j) \in C_{i,j}$ **do**
6 $k \leftarrow |A|$;
7 **while** $\neg\mathcal{R}^k(v_i, v_j) \wedge k \geq 1$ **do** $k \leftarrow k - 1$;
8 **if** $k > 0$ **then** $D_{i,j} \leftarrow D_{i,j} \cup (v_i, v_j)$; $newCl \leftarrow true$;
9 **if** $newCl$ **then** $C_{i,j} \leftarrow candidateSelection(\mathcal{G}', i, j)$;
10 **else** $C_{i,j} \leftarrow \emptyset$;
11 $transitiveClosure(\mathcal{G}')$;

Once all pairs of candidate profiles are processed, the newly discovered cross-links are used to get new candidates (Line 9), on which the algorithm applies again the rules, and this is iterated until no more candidates can be found ($C_{i,j} = \emptyset$). Finally, before considering the next pair of social networks, the discovered cross-links are propagated over the n social networks by transitive closure (Line 11).

The example depicted in Figure 1 represents a social internetwork composed of four social networks. The arrows represent the friendship links, while the black dashed lines are the existing cross-links, that the algorithm uses to find the candidates. The algorithm starts from the pair of profiles of Lisa in Flickr and LiveJournal. The candidate set is the result of the Cartesian product between the set of the friends of Lisa in Flickr (Bob, Mark and Alice) and the set of friends of Lisa in LiveJournal (Alice and Ben). After applying the rules to the candidate set, the algorithm unveils a cross-link between the profiles of Alice in Flickr and LiveJournal, which is represented as a dotted line. By transitive closure, three more cross-links are found, represented as dash-dotted lines.

Complexity. Let (i, j) be a pair of social networks.
The selection of candidates (Line 3, for the first iteration, Line 9, for the others) costs $T_S = \sum_{(v_i,v_j)\in D_{i,j}} |friends(v_i)| \times |friends(v_j)|$ ($D_{i,j} = E_{i,j}$ for the first iteration). We observe that in our dataset, the number of cross-links $|D_{i,j}| << min(|V_i|, |V_j|)$; moreover, each node has 74 friends on average (the degree of a node ranges from 1 to 18305). As a result, the cost of selecting the candidates

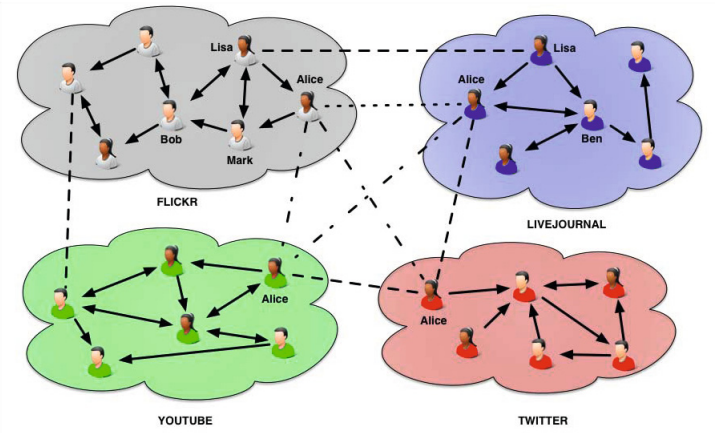

Fig. 1. Description of the algorithm on a small social internetwork. Continuous lines are friendship links, dashed lines are the existing cross-links. Cross-links established after applying the rules are dotted. Cross-links established by transitive closure are dash-dotted.

(and the number of candidates $|C_{i,j}|$ itself) is much smaller than $|V_i| \times |V_j|$, which would be the cost of considering all possible pairs of nodes. The cost of applying all rules on one candidate (Line 7) is $\sum_{k=1}^{|A|} \binom{|A|}{k} \times k \times \alpha$, α being the cost of the comparison of the values of one attribute. Since $k \leq |A|$, this cost is bounded by $r = \alpha \times |A| \times (2^{|A|} - 1)$. The cost of applying all rules to all candidates (Lines 5-7) is $T_R = r \times O(|C_{i,j}|)$. Therefore, the cost of Lines 4-9 is $p \times (T_S + T_R)$, where p is the number of times Lines 4 - 9 are repeated. Finally, the transitive closure (Line 11) cost is $O(|V| \times |E_c|)$, where $|V|$ (respectively, $|E_c|$) is the number of nodes (respectively, cross-links) in the social internetwork. Since $|E_C| << |V|$, this amounts to $O(|V|)$.

5 Evaluation Results

In order to evaluate our approach, we considered the dataset used by Buccafurri et al. in their experiments [14]. The original dataset includes a social internetwork with four social networks, namely LiveJournal, Flickr Twitter and YouTube [1]. The graph is composed of 93,169 nodes, 145,580 friendship links and 462 cross-links. We note that the number of cross-links declared by Buccafurri et al. [14] is 745, but this also includes duplicate links, which we removed.

After a careful analysis of the data, we found that many *friend* links were missing between a large number of nodes, probably because they were added after the internetwork was crawled. Moreover, the only available profile attribute is

[1] `http://www.ursino.unirc.it/pkdd-12.html`

Table 1. Statistics on the social internetwork used in our evaluation

Network	Nodes	**Links**		
		friend	*me*	Total
Flickr	1,814,405	15,415,083	189	154,152,72
LiveJournal	211,045	2,093,737	161	2,093,898
Twitter	8,842	19,008	312	19,320
YouTube	1,210	1,367	286	1,653
Total	**2,035,502**	**17,529,195**	**474**	**17,529,669**

Table 2. Cross-links between all pairs of social networks

Network	Flickr	LiveJournal	Twitter	YouTube
Flickr	–	148	29	12
LiveJournal	148	–	11	2
Twitter	29	11	–	272
YouTube	12	2	272	–

the username. For this reason, we updated the internetwork by obtaining the missing information using the API of the four SNSs under evaluation. While we were at that, we also enriched the graphs by adding new nodes that are linked via a *friend* link to the existing nodes. As a result, we obtained a much larger internetwork, whose properties are shown in Table 1. In total, we have more than 2 million nodes, more than 17 million links and 474 cross-links after transitive closure. We note that the cross-links are sparse, compared to the friendship links. The number of cross-links between each pair of networks is detailed in Table 2.

In the implementation of our approach, the social internetwork is stored in a Neo4j database[2], which is particularly indicated to handle large graphs.

5.1 Evaluation of the Rules

We observe that the accuracy of our algorithm might degrade if cross-links are established between pairs of profiles that do not match. In fact, the algorithm determines at each iteration the missing cross-links based on those discovered at the previous iterations. Therefore, the set of rules that we described in Section 4.2 need to be highly effective in determining whether any two profiles match.

In this section, we describe a first evaluation that aims at identifying the attributes that are the most relevant, as well as tuning the thresholds of the approximate similarity measures that we defined in Section 4.2 to compare the profile attributes. To this extent, we consider Flickr and LiveJournal, the two largest networks in our dataset. We arbitrarily set to 0.7 the thresholds θ_u and θ_n for the similarity measures that compare the usernames and names respectively, and we run only the first iteration of our algorithm.

[2] `www.neo4j.org/`

Table 3. Evaluation of the rules on Flickr and LiveJournal $\theta_u = 0.7$ and $\theta_n = 0.7$

Rule	Attributes	ratio of $\lvert\mathcal{M}\rvert$%	$\frac{\lvert C\rvert}{\lvert\mathcal{M}\rvert}$ %	$\frac{\lvert W\rvert}{\lvert\mathcal{M}\rvert}$%	$\frac{\lvert U\rvert}{\lvert\mathcal{M}\rvert}$%
$\mathcal{R}^1$	$\{u\}$	83.09	60.39	30.19	9.42
	$\{n\}$	3.21	42.72	54.55	2.73
	$\{s\}$	2.92	100.00	0.00	0.00
	$\{w\}$	2.19	96.00	2.00	2.00
	Total	91.41	**61.85**	**29.42**	**8.73**
$\mathcal{R}^2$	$\{u, s\}$	3.21	100	0	0
	$\{u, w\}$	2.07	99.09	0	0.91
	$\{u, n\}$	1.93	92.42	1.52	6.06
	$\{n, s\}$	0.26	100	0	0
	$\{n, w\}$	0.18	100	0	0
	Total	7.65	**97.71**	**0.38**	**1.91**
$\mathcal{R}^3$	$\{u, n, s\}$	0.55	100	0	0
	$\{u, w, n\}$	0.32	100	0	0
	$\{u, w, s\}$	0.03	100	0	0
	Total	0.91	**100**	**0**	**0**
$\mathcal{R}^4$	$\{u, n, w, s\}$	0.03	100	0	0
	Total	0.03	**100**	**0**	**0**
	Grand total	100	**64.95**	**26.93**	**8.12**

Table 3 shows the results for each rule $\mathcal{R}^k$ that is verified by at least one candidate. The second column shows the set of attributes used by each rule; we recall from Section 4.2 that u, n, s and w refer to the attributes username, name, link to a social network profile and link to a web page respectively. We note that the attribute e (email) does not contribute to any rule, which is due to the fact that the value for this attribute is almost never disclosed in both profiles of the candidate pairs; this is why the rule $\mathcal{R}^5$ does not appear in the table. In total, the algorithm retrieves 16,000 candidates and determines a set $\mathcal{M}$ of 3,424 cross-links. As shown in the third column, 91.41% verify a 1−rule, 7.65% verify a 2−rule, 0.91% verify a 3−rule and 0.03% verify a 4−rule. The results clearly show that only a small percentage of profile pairs verify a k−rule, with $k \geq 2$, and the vast majority verifies a 1−rule, which indicates that in the selected networks the information disclosed by the individuals have little overlapping. We note also that the attribute username is present in a large number of rules verified by the profile candidates.

In order to evaluate the accuracy of the rules, we determined a ground truth by tagging each cross-link $(v_f, me, v_l) \in \mathcal{M}$ as either *correct*, if v_f and v_l match, or *incorrect*, if they do not match, or *undetermined*, if no decision can be taken. To this extent, we split set $\mathcal{M}$ into four equal-size independent subsets, one for each author of this paper, who had to assign the proper tag to each cross-link, based on a visual inspection of the profile web pages of the individuals concerned. Most of the time the information on the profile web pages were enough to determine

whether two profiles referred to the same individual; however, in some cases the information are so scarce that no conclusive evidence as to whether the two profiles match can be found. In order to avoid errors in the ground truth, which would inevitably invalidate the results of our evaluation, we introduced the tag *undetermined*, which we assigned to all cross-links that we could not determine with certainty either as *correct* or *incorrect*. As a result, we determined three subsets of $\mathcal{M}$: (i) C, the set of the cross-links tagged as *correct*; (ii) W, the set of the cross-links tagged as *incorrect* and (iii) U, the set of the cross-links tagged as *undetermined*. The *precision*, computed as $P = \frac{|C|}{|\mathcal{M}|}$, is reported in the fourth column while the *error rate* is shown in the fifth column. It took approximately 300 hours in total to tag all the cross-links in $\mathcal{M}$. Since this work was split among 4 people, it took 10 days to have the ground truth available. We note that while this is acceptable for the preliminaries results that we discuss in this paper, we are aware that it is not feasible for the larger scale experiments that we are organizing. We will involve more evaluators and we will make sure that each cross-link is tagged by more than one person; the agreement among the evaluators on the assigned tags, which can be computed by using the Pearson correlation coefficient, will be a solid evidence that the ground truth is error-free, or, at least, contains a negligible amount of errors.

As for the 1−rules, those that show the highest error rate (30.19% and 54.55% respectively) are the ones that use u and n. On the other hand, the 1−rules that have the highest precision (100% and 96% respectively) are the ones that use s and w, which confirms our intuition that the attribute link to other profiles is highly relevant. We also note that the rules that combine at least two attributes, including those that use n and u, achieve a precision between 92% et 100%. This confirms our hypothesis that the more the attributes that match, the higher the probability that two profiles refer to the same individual.

We further studied the two 1−rules that achieve a high error rate and we raised the value of thresholds θ_u and θ_n. While the precision significantly improves for the 1−rule that uses u, no significant change is observed for the 1−rule that uses n. This undeniably shows that a even high similarity of two names is not alone a good indicator that two profiles refer to the same individual. As a matter of fact, n is not only ambiguous, but also sensitive, which implies that very often an individual omits it or provides a fake name in order not to reveal her identity. This is evident by just looking at any two profiles belonging to the same individual in our dataset. Most of those that we came across disclose names that are partially or completely different across the two profiles.

5.2 Evaluation of the Algorithm and Comparison

Based on the discussion in the previous section, we discarded the 1−rule that uses the attribute name and we set to 0.9 the value of θ_u. We run our algorithm on the four social networks of the dataset, namely Flickr, LiveJournal, Twitter and YouTube. The algorithm terminated after four iterations and discovered 2,788 new cross-links: 1,053 after the first iteration, 1,005 after the second, 654

after the third and 76 after the fourth. The precision is 94% with 2% of error rate and 4% of undetermined.

We first compare our algorithm against the one proposed by Buccafurri et al. [14], as we built our dataset on top of theirs. As explained above, we considerably enriched their dataset by adding new nodes and friendship links (but no new cross-link); as a result, we evaluated our algorithm on a much larger social internetwork. Their evaluation consists in selecting 160 cross-links that are given as input to their algorithm, which discovers 22 new cross-links across the four social networks with a precision of 85%. Thus, their algorithm discovers a considerably lower number of cross-links than ours, which is likely to be due to the iterative nature of our algorithm.

Finally, we compare our algorithm against the approach proposed by Malhotra et al. [9], which uses machine learning techniques. Similarly to our algorithm, they consider multiple attributes, such as username and name, but they ignore the network topology. They train four classifiers to determine whether two profiles match. Their model, applied to a real world scenario, which includes two social networks, namely LinkedIn and Twitter, achieves an accuracy of 64% [9]. No information is given on the number of discovered cross-links, nor the number of considered profile pairs.

6 Concluding Remarks

In this paper, we presented an algorithm to match profiles of individuals across several social networks by using the network topology and the personal information that are publicly available in the profiles. We thoroughly evaluated the algorithm on a large dataset of four real social networks, which constitutes a real challenge, because data are likely to be erroneous and messy. The evaluation showed the robustness of our algorithm, as it achieves a high precision (94%). We also presented a comparison against two existing approaches and discussed the results. We note that our algorithm relies on the attributes whose values are publicly available on the profiles of the individuals. It would be interesting to further explore the use of the network topology to generalize the algorithm to networks where the attribute values are anonymized. Moreover, we are currently fetching data from other social networks, to evaluate our algorithm on a mix of heterogeneous kinds of networks. Finally, we are migrating our dataset to the newest version of Neo4j and optimizing the code of the algorithm to fully take advantage of the new features of Neo4j. The time performance of the optimized code will be thoroughly assessed.

References

1. Gross, R., Acquisti, A.: Information Revelation and Privacy in Online Social Networks. In: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, WPES 2005, pp. 71–80. ACM, New York (2005)

2. Little, L., Briggs, P., Coventry, L.: Who Knows about Me?: An Analysis of Age-related Disclosure Preferences. In: Proceedings of the 25th BCS Conference on Human-Computer Interaction, BCS-HCI 2011, pp. 84–87. British Computer Society, Swinton (2011)
3. Stutzman, F.: An Evaluation of Identity-Sharing Behavior in Social Network Communities. iDMAa Journal 3(1) (2006)
4. Perito, D., Castelluccia, C., Kaafar, M.A., Manils, P.: How Unique and Traceable Are Usernames? In: Fischer-Hübner, S., Hopper, N. (eds.) PETS 2011. LNCS, vol. 6794, pp. 1–17. Springer, Heidelberg (2011)
5. Zafarani, R., Liu, H.: Connecting Corresponding Identities across Communities. In: Third International AAAI Conference on Weblogs and Social Media (2009)
6. Carmagnola, F., Cena, F.: User Identification for Cross-system Personalisation. Inf. Sci. 179, 16–32 (2009)
7. Cortis, K., Scerri, S., Rivera, I., Handschuh, S.: Discovering Semantic Equivalence of People Behind Online Profiles. In: Proceedings of the Resource Discovery (RED) Workshop. ESWC (2012)
8. Golbeck, J., Rothstein, M.: Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. In: AAAI, vol. 8, pp. 1138–1143 (2008)
9. Malhotra, A., Totti, L., Meira, W., Kumaraguru, P., Almeida, V.: Studying User Footprints in Different Online Social Networks. In: International Workshop on Cybersecurity of Online Social Network, ACM ASONAM 2012 (2012)
10. Motoyama, M., Varghese, G.: I Seek You: Searching and Matching Individuals in Social Networks. In: Proceedings of the Eleventh International Workshop on Web Information and Data Management, pp. 67–75. ACM (2009)
11. Raad, E., Chbeir, R., Dipanda, A.: User Profile Matching in Social Networks. In: 2010 13th International Conference on Network-Based Information Systems (NBiS), pp. 297–304. IEEE (2010)
12. Rowe, M.: Interlinking Distributed Social Graphs. In: Linked Data on the Web Workshop, WWW (2009)
13. Bartunov, S., Korshunov, A., Park, S., Ryu, W., Lee, H.: Joint Link-attribute User Identity Resolution in Online Social Networks. In: SNA-KDD Workshop (2012)
14. Buccafurri, F., Lax, G., Nocera, A., Ursino, D.: Discovering Links among Social Networks. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part II. LNCS, vol. 7524, pp. 467–482. Springer, Heidelberg (2012)
15. Jain, P., Kumaraguru, P., Joshi, A.: @i Seek 'fb.me': Identifying Users Across Multiple Online Social Networks. In: WWW (Companion Volume), pp. 1259–1268 (2013)
16. Narayanan, A., Shmatikov, V.: De-anonymizing Social Networks. In: 30th IEEE Symposium on Security and Privacy, pp. 173–187. IEEE (2009)
17. FriendFeed, `http://friendfeed.com`
18. Plaxo, `http://www.plaxo.com`
19. Spokeo, `http://www.spokeo.com`
20. Krishnamurthy, B., Wills, C.E.: On the Leakage of Personally Identifiable Information via Online Social Networks. In: Proceedings of the 2nd ACM Workshop on Online Social Networks, pp. 7–12. ACM (2009)